An Introduction to the Science of Statistics:
From Theory to Implementation

Preliminary Edition

©Joseph C. Watkins

Contents

I Organizing and Producing Data **1**

1 Displaying Data **3**
1.1 Types of Data . . . 3
1.2 Categorical Data . . . 4
1.2.1 Pie Chart . . . 4
1.2.2 Bar Charts . . . 7
1.3 Two-way Tables . . . 7
1.4 Histograms and the Empirical Cumulative Distribution Function . . . 10
1.5 Scatterplots . . . 13
1.6 Time Plots . . . 15
1.7 Answers to Selected Exercises . . . 18

2 Describing Distributions with Numbers **21**
2.1 Measuring Center . . . 21
2.1.1 Medians . . . 21
2.1.2 Means . . . 21
2.2 Measuring Spread . . . 24
2.2.1 Five Number Summary . . . 24
2.2.2 Sample Variance and Standard Deviation . . . 25
2.3 Quantiles and Standardized Variables . . . 27
2.4 Quantile-Quantile Plots . . . 28
2.5 Answers to Selected Exercises . . . 30

3 Correlation and Regression **33**
3.1 Covariance and Correlation . . . 33
3.2 Linear Regression . . . 37
3.2.1 Transformed Variables . . . 45
3.3 Extensions . . . 50
3.3.1 Nonlinear Regression . . . 51
3.3.2 Multiple Linear Regression . . . 51
3.4 Answers to Selected Exercises . . . 56

4 Producing Data **63**
4.1 Preliminary Steps . . . 63
4.2 Professional Ethics . . . 64
4.3 Formal Statistical Procedures . . . 64
4.3.1 Observational Studies . . . 64
4.3.2 Randomized Controlled Experiments . . . 65
4.3.3 Natural experiments . . . 68
4.4 Case Studies . . . 69

4.4.1 Observational Studies . . . 69
4.4.2 Experiments . . . 71

II Probability **77**

5 The Basics of Probability **79**
5.1 Introduction . . . 79
5.2 Equally Likely Outcomes and the Axioms of Probability . . . 80
5.3 Consequences of the Axioms . . . 82
5.4 Counting . . . 84
5.4.1 Fundamental Principle of Counting . . . 84
5.4.2 Permutations . . . 85
5.4.3 Combinations . . . 86
5.5 Answers to Selected Exercises . . . 89
5.6 Set Theory - Probability Theory Dictionary . . . 94

6 Conditional Probability and Independence **95**
6.1 Restricting the Sample Space - Conditional Probability . . . 95
6.2 The Multiplication Principle . . . 96
6.3 The Law of Total Probability . . . 97
6.4 Bayes formula . . . 98
6.5 Independence . . . 102
6.6 Answers to Selected Exercises . . . 104

7 Random Variables and Distribution Functions **109**
7.1 Introduction . . . 109
7.2 Distribution Functions . . . 110
7.3 Properties of the Distribution Function . . . 112
7.4 Mass Functions . . . 114
7.5 Density Functions . . . 116
7.6 Mixtures . . . 118
7.7 Joint Distributions . . . 119
7.7.1 Independent Random Variables . . . 119
7.8 Simulating Random Variables . . . 120
7.8.1 Discrete Random Variables and the `sample` Command . . . 120
7.8.2 Continuous Random Variables and the Probability Transform . . . 122
7.9 Answers to Selected Exercises . . . 123

8 The Expected Value **131**
8.1 Definition and Properties . . . 131
8.2 Discrete Random Variables . . . 133
8.3 Bernoulli Trials . . . 134
8.4 Continuous Random Variables . . . 136
8.5 Quantile Plots and Probability Plots . . . 139
8.6 Summary . . . 140
8.7 Names for $Eg(X)$. . . . 141
8.8 Independence . . . 143
8.9 Covariance and Correlation . . . 143
8.9.1 Equivalent Conditions for Independence . . . 144
8.10 Answers to Selected Exercises . . . 144

9 Examples of Mass Functions and Densities **149**
9.1 Examples of Discrete Random Variables . . . 149
9.2 Examples of Continuous Random Variables . . . 155
9.3 R Commands . . . 160
9.4 Summary of Properties of Random Variables . . . 160
9.4.1 Discrete Random Variables . . . 161
9.4.2 Continuous Random Variables . . . 162
9.5 Answers to Selected Exercises . . . 163

10 The Law of Large Numbers **169**
10.1 Introduction . . . 169
10.2 Monte Carlo Integration . . . 172
10.3 Importance Sampling . . . 176
10.4 Answers to Selected Exercises . . . 178

11 The Central Limit Theorem **183**
11.1 Introduction . . . 183
11.2 The Classical Central Limit Theorem . . . 184
11.3 Propagation of Error . . . 189
11.4 Delta Method . . . 192
11.5 Summary of Normal Approximations . . . 195
11.5.1 Sample Sum . . . 195
11.5.2 Sample Mean . . . 196
11.5.3 Sample Proportion . . . 196
11.5.4 Delta Method . . . 196
11.6 Answers to Selected Exercises . . . 197

III Estimation 203

12 Overview of Estimation **205**
12.1 Introduction . . . 205
12.2 Classical Statistics . . . 207
12.3 Bayesian Statistics . . . 208
12.4 Answers to Selected Exercises . . . 215

13 Method of Moments **219**
13.1 Introduction . . . 219
13.2 The Procedure . . . 220
13.3 Examples . . . 220
13.4 Answers to Selected Exercises . . . 227

14 Unbiased Estimation **229**
14.1 Introduction . . . 229
14.2 Computing Bias . . . 230
14.3 Compensating for Bias . . . 233
14.4 Consistency . . . 237
14.5 Cramér-Rao Bound . . . 238
14.6 A Note on Efficient Estimators . . . 241
14.7 Answers to Selected Exercises . . . 242

15 Maximum Likelihood Estimation **247**
15.1 Introduction . . . 247
15.2 Examples . . . 249
15.3 Summary of Estimators . . . 255
15.4 Asymptotic Properties . . . 255
15.5 Multidimensional Estimation . . . 256
15.6 Comparison of Estimation Procedures . . . 259
15.7 Choice of Estimators . . . 259
15.8 Technical Aspects . . . 260
15.9 Answers to Selected Exercises . . . 260

16 Interval Estimation **265**
16.1 Classical Statistics . . . 265
16.1.1 Means . . . 266
16.1.2 Linear Regression . . . 272
16.1.3 Sample Proportions . . . 272
16.1.4 Summary of Standard Confidence Intervals . . . 273
16.1.5 Interpretation of the Confidence Interval . . . 274
16.1.6 Extensions on the Use of Confidence Intervals . . . 276
16.2 The Bootstrap . . . 278
16.3 Bayesian Statistics . . . 279
16.4 Answers to Selected Exercises . . . 281

IV Hypothesis Testing **285**

17 Simple Hypotheses **287**
17.1 Overview and Terminology . . . 287
17.2 The Neyman-Pearson Lemma . . . 288
17.2.1 The Receiver operating Characteristic . . . 290
17.3 Examples . . . 291
17.4 Summary . . . 297
17.5 Proof of the Neyman-Pearson Lemma . . . 298
17.6 An Brief Introduction to the Bayesian Approach . . . 300
17.7 Answers to Selected Exercises . . . 302

18 Composite Hypotheses **307**
18.1 Partitioning the Parameter Space . . . 307
18.2 The Power Function . . . 307
18.3 The p-value . . . 315
18.4 Multiple Hypothesis Testing . . . 317
18.4.1 Familywise Error Rate . . . 317
18.4.2 False Discovery Rate . . . 318
18.5 Answers to Selected Exercises . . . 320

19 Extensions on the Likelihood Ratio **323**
19.1 One-Sided Tests . . . 323
19.2 Likelihood Ratio Tests . . . 327
19.3 Chi-square Tests . . . 330
19.4 Answers to Selected Exercises . . . 333

20 t Procedures **343**
20.1 Guidelines for Using the t Procedures . . . 343
20.2 One Sample t Tests . . . 344
20.3 Correspondence between Two-Sided Tests and Confidence Intervals . . . 347
20.4 Matched Pairs Procedures . . . 348
20.5 Two Sample Procedures . . . 350
20.6 Summary of Tests of Significance . . . 355
20.6.1 General Guidelines . . . 355
20.6.2 Test for Population Proportions . . . 356
20.6.3 Test for Population Means . . . 356
20.7 A Note on the Delta Method . . . 357
20.8 The t Test as a Likelihood Ratio Test . . . 357
20.9 Non-parametric alternatives . . . 359
20.9.1 Permutation Test . . . 359
20.9.2 Mann-Whitney or Wilcoxon Rank Sum Test . . . 360
20.9.3 Wilcoxon Signed-Rank Test . . . 361
20.10 Answers to Selected Exercises . . . 362

21 Goodness of Fit **365**
21.1 Fit of a Distribution . . . 365
21.2 Contingency tables . . . 371
21.3 Applicability and Alternatives to Chi-squared Tests . . . 375
21.4 Answer to Selected Exercise . . . 378

22 Analysis of Variance **383**
22.1 Overview . . . 383
22.2 One Way Analysis of Variance . . . 384
22.3 Contrasts . . . 388
22.4 Two Sample Procedures . . . 390
22.5 Kruskal-Wallis Rank-Sum Test . . . 393
22.6 Answer to Selected Exercises . . . 394

Appendix A: A Sample *R* Session **397**

Index **403**

Preface

Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write. – Samuel Wilkes, 1951, paraphrasing H. G. Wells from *Mankind in the Making*

The value of statistical thinking is now accepted by researchers and practitioners from a broad range of endeavors. This viewpoint has become common wisdom in a world of big data. The challenge for statistics educators is to adapt their pedagogy to accommodate the circumstances associated to the information age. This choice of pedagogy should be attuned to the quantitative capabilities and scientific background of the students as well as the intended use of their newly acquired knowledge of statistics.

Many university students, presumed to be proficient in college algebra, are taught a variety of procedures and standard tests under a well-developed pedagogy. This approach is sufficiently refined so that students have a good intuitive understanding of the underlying principles presented in the course. However, if the statistical needs presented by a given scientific question fall outside the battery of methods presented in the standard curriculum, then students are typically at a loss to adjust the procedures to accommodate the additional demand.

On the other hand, undergraduate students majoring in mathematics frequently have a course on the theory of statistics as a part of their program of study. In this case, the standard curriculum repeatedly finds itself close to the very practically minded subject that statistics is. However, the demands of the syllabus provide very little time to explore these applications with any sustained attention.

Our goal is to find a middle ground.

Despite the fact that calculus is a routine tool in the development of statistics, the benefits to students who have learned calculus are infrequently employed in the statistics curriculum. The objective of this book is to meet this need with a one semester course in statistics that moves forward in recognition of the coherent body of knowledge provided by statistical theory having an eye consistently on the application of the subject. Such a course may not be able to achieve the same degree of completeness now presented by the two more standard courses described above. However, it ought to able to achieve some important goals:

- leaving students capable of understanding what statistical thinking is and how to integrate this with scientific procedures and quantitative modeling and

- learning how to ask statistics experts productive questions, and how to implement their ideas using statistical software and other computational tools.

Inevitably, many important topics are not included in this book. In addition, I have chosen to incorporate abbreviated introductions of some more advanced topics. Such topics can be skipped in a first pass through the material. However, one value of a textbook is that it can serve as a reference in future years. The context for some parts of the exposition will become more clear as students continue their own education in statistics. In these cases, the more advanced pieces can serve as a bridge from this book to more well developed accounts. My goal is not to compose a stand alone treatise, but rather to build a foundation that allows those who have worked through this book to introduce themselves to many exciting topics both in statistics and in its areas of application.

Who Should Use this Book

The major prerequisites are comfort with calculus and a strong interest in questions that can benefit from statistical analysis. Willingness to engage in explorations utilizing statistical software is an important additional requirement. The original audience for the course associated to this book are undergraduate students minoring in mathematics. These student have typically completed a course in multivariate calculus. Many have been exposed to either linear algebra or differential equations. They enroll in this course because they want to obtain a better understanding of their own core subject. Even though we regularly rely on the mechanics of calculus and occasionally need to work with matrices, this is *not* a textbook for a mathematics course, but rather a textbook that is dedicated to a higher level of understanding of the concepts and practical applications of statistics. In this regard, it relies on a solid grasp of concepts and structures in calculus and algebra.

With the advance and adoption of the *Common Core State Standards* in mathematics, we can anticipate that primary and secondary school students will experience a broader exposure to statistics through their school years. As a consequence, we will need to develop a curriculum for teachers and future teachers so that they can take content in statistics and turn that into curriculum for their students. This book can serve as a source of that content.

In addition, those engaged both in industry and in scholarly research are experiencing a surge in the need to design more complex experiments and analyze more diverse data types. Universities and industry are responding with advanced educational opportunities to extend statistics education beyond the theory of probability and statistics, linear models and design of experiments to more modern approaches that include stochastic processes, machine learning and data mining, Bayesian statistics, and statistical computing. This book can serve as an entry point for these critical topics in statistics.

An Annotated Syllabus

The four parts of the course - organizing and collecting data, an introduction to probability, estimation procedures and hypothesis testing - are the building blocks of many statistics courses. We highlight some of the particular features in this book.

Organizing and Collecting Data

Much of this is standard and essential - organizing categorical and quantitative data, appropriately displayed as contingency tables, bar charts, histograms, boxplots, time plots, and scatterplots, and summarized using medians, quartiles, means, weighted means, trimmed means, standard deviations, correlations and regression lines. We use this as an opportunity to introduce to the statistical software package R and to add additional summaries like the empirical cumulative distribution function and the empirical survival function. One example incorporating the use of this is the comparison of the lifetimes of wildtype and transgenic mosquitoes and a discussion of the best strategy to display and summarize data if the goal is to examine the differences in these two genotypes of mosquitoes in their ability to carry and spread malaria. A bit later, we will do an integration by parts exercise to show that the mean of a non-negative continuous random variable is the area under its survival function.

Collecting data under a good design is introduced early in the text and discussion of the underlying principles of experimental design is an abiding issue throughout the text. With each new mathematical or statistical concept comes an enhanced understanding of what an experiment might uncover through a more sophisticated design than what was previously thought possible. The students are given readings on design of experiment and examples using R to create a sample under variety of protocols.

Introduction to Probability

Probability theory is the analysis of random phenomena. It is built on the axioms of probability and is explored, for example, through the introduction of random variables. The goal of probability theory is to uncover properties arising from the phenomena under study. Statistics is devoted to the analysis of data. One goal of statistical science is to

articulate as well as possible what model of random phenomena underlies the production of the data. The focus of this section of the course is to develop those probabilistic ideas that relate most directly to the needs of statistics.

Thus, we must study the axioms and basic properties of probability to the extent that the students understand conditional probability and independence. Conditional probability is necessary to develop Bayes formula which we will later use to give a taste of the Bayesian approach to statistics. Independence will be needed to describe the likelihood function in the case of an experimental design that is based on independent observations. Densities for continuous random variables and mass function for discrete random variables are necessary to write these likelihood functions explicitly. Expectation will be used to standardize a sample sum or sample mean and to perform method of moments estimates.

Random variables are developed for a variety of reasons. Some, like the binomial, negative binomial, Poisson or the gamma random variable, arise from considerations based on Bernoulli trials or exponential waiting. The hypergeometric random variable helps us understand the difference between sampling with and without replacement. The F, t and chi-square random variables will later become test statistics. Uniform random variables are the ones simulated by random number generators. Because of the central limit theorem, the normal family is the most important among the list of parametric families of random variables.

The flavor of the text returns to becoming more authentically statistical with the law of large numbers and the central limit theorem. These are largely developed using simulation explorations and first applied to simple Monte Carlo techniques and importance sampling to estimate the value of an definite integrals. One cautionary tale is an example of the failure of these simulation techniques when applied without careful analysis. If one uses, for example, Cauchy random variables in the evaluation of some quantity, then the simulated sample means can appear to be converging only to experience an abrupt and unpredictable jump. The lack of convergence of an improper integral reveals the difficulty. The central object of study is, of course, the central limit theorem. It is developed both in terms of sample sums and sample means and proportions and used in relatively standard ways to estimate probabilities. However, in this book, we can introduce the delta method which adds ideas associated to the central limit theorem to the context of propagation of error.

Estimation

In the simplest possible terms, the goal of estimation theory is to answer the question: *What is that number?* An estimate is a statistic, i. e., a function of the data. We look to two types of estimation techniques - method of moments and maximum likelihood and several criteria for an estimator using, for example, variance and bias. Several examples including mark and recapture and the distribution of fitness effects from genetic data are developed for both types of estimators. The variance of an estimator is approximated using the delta method for method of moments estimators and using Fisher information for maximum likelihood estimators. An analysis of bias is based on quadratic Taylor series approximations and the properties of expectations. Both classes of estimators are often consistent. This implies that the bias decreases towards zero with an increasing number of observations. R is routinely used in simulations to gain insight into the quality of estimators.

The point estimation techniques are followed by interval estimation and, notably, by confidence intervals. This brings us to the familiar one and two sample t-intervals for population means and one and two sample z-intervals for population proportions. In addition, we can return to the delta method and the observed Fisher information to construct confidence intervals associated respectively to method of moment estimators and and maximum likelihood estimators. We also add a brief introduction on bootstrap confidence intervals and Bayesian credible intervals in order to provide a broader introduction to strategies for parameter estimation.

Hypothesis Testing

For hypothesis testing, we first establish the central issues - null and alternative hypotheses, type I and type II errors, test statistics and critical regions, significance and power. We then present the ideas behind the use of likelihood ratio tests as best tests for a simple hypothesis. This is motivated by a game designed to explain the importance of the Neyman Pearson lemma. This approach leads us to well-known diagnostics of an experimental design, notably, the receiver operating characteristic and power curves.

Extensions of the Neyman Pearson lemma form the basis for the t test for means, the chi-square test for goodness of fit, and the F test for analysis of variance. These results follow from the application of optimization techniques from calculus, including the use of Lagrange multipliers to develop goodness of fit tests. The Bayesian approach to hypothesis testing is explored for the case of simple hypothesis using morphometric measurements, in this case a butterfly wingspan, to test whether a habitat has been invaded by a mimic species.

The desire of a powerful test is articulated in a variety of ways. In engineering terms, power is called sensitivity. We illustrate this with a radon detector. An insensitive instrument is a risky purchase. This can be either because the instrument is substandard in the detection of fluctuations or poor in the statistical basis for the algorithm used to determine a change in radon level. An insensitive detector has the undesirable property of not sounding its alarm when the radon level has indeed risen.

The course ends by looking at the logic of hypotheses testing and the results of different likelihood ratio analyses applied to a variety of experimental designs. The delta method allows us to extend the resulting test statistics to multivariate nonlinear transformations of the data. The textbook concludes with a practical view of the consequences of this analysis through case studies in a variety of disciplines including, for example, genetics, health, ecology, and bee biology. This will serve to introduce us to the well known t procedure for inference of the mean, both the likelihood-based G^2 test and the traditional chi-square test for discrete distributions and contingency tables, and the F test for one-way analysis of variance. We add short descriptions for the corresponding non-parametric procedures, namely, permutation, ranked-sum and signed-rank tests for quantitative data, and exact tests for categorical data

Exercises and Problems

One obligatory statement in the preface of a book such as this is to note the necessity of working problems. The material can only be mastered by grappling with the issues through the application to engaging and substantive questions. In this book, we address this imperative through exercises and through problems. The exercises, integrated into the textbook narrative, are of two basic types. The first is largely mathematical or computational exercises that are meant to provide or extend the derivation of a useful identity or data analysis technique. These experiences will prepare the student to perform the calculations that routinely occur in investigations that use statistical thinking. The second type form a collection of questions that are meant to affirm the understanding of a particular concept.

Problems are collected at the end of each of the four parts of the book. While the ordering of the problems generally follows the flow of the text, they are designed to be more extensive and integrative. These problems often incorporate several concepts and will call on a variety of problem solving strategies combining handwritten work with the use of statistical software. Without question, the best problems are those that the students chose from their own interests.

Acknowledgements

The concept that let to this book grew out of a conversation with the late Michael Wells, Professor of Biochemistry at the University of Arizona. He felt that if we are asking future life scientist researchers to take the time to learn calculus and differential equations, we should also provide a statistics course that adds value to their abilities to design experiments and analyze data while reinforcing both the practical and conceptual sides of calculus. As a consequence, course development received initial funding from a Howard Hughes Medical Institute grant (52005889). Christopher Bergevin, an HHMI postdoctoral fellow, provided a valuable initial collaboration.

Since that time, I have had the great fortune to be the teacher of many bright and dedicated students whose future contribution to our general well-being is beyond dispute. Their cheerfulness and inquisitiveness has been a source of inspiration for me. More practically, their questions and their persistence led to a much clearer exposition and the addition of many dozens of figures to the text. Through their end of semester projects, I have been introduced to many interesting questions that are intriguing in their own right, but also have added to the range of applications presented throughout the text. Four of these students - Beryl Jones, Clayton Mosher, Laurel Watkins de Jong, and Taylor Corcoran - have gone on to become assistants in the course. I am particularly thankful to these four for their contributions to the dynamical atmosphere that characterizes the class experience.

Part I

Organizing and Producing Data

Topic 1

Displaying Data

There are two goals when presenting data: convey your story and establish credibility. - Edward Tufte

Statistics is a mathematical science that is concerned with the collection, analysis, interpretation or explanation, and presentation of data.

The first encounters one has to data are through graphical displays and numerical summaries. The goal is to find an elegant method for this presentation that is at the same time both objective and informative - making clear with a few lines or a few numbers the salient features of the data. In this sense, data presentation is at the same time an art, a science, and an obligation to impartiality.

In the section, we will describe some of the standard presentations of data and at the same time, taking the opportunity to introduce some of the commands that the software package R provides to draw figures and compute summaries.

1.1 Types of Data

A data set provides information about a group of individuals. These individuals are, typically, representatives chosen from a **population** under study. Data on the individuals are meant, either informally or formally, to allow us to make inferences about the population. We shall later discuss how to define a population, how to choose individuals in the population and how to collect data on these individuals.

- **Individuals** are the objects described by the data.
- **Variables** are characteristics of an individual. In order to present data, we must first recognize the types of data under consideration.
  - **Categorical variables** partition the individuals into classes. Other names for categorical variables are **levels** or **factors**.
  - **Quantitative variables** are those for which arithmetic operations like addition and differences make sense.

Example 1.1 (individuals and variables)**.** *We consider two populations - the first is the nations of the world and the second is the people who live in those countries. Below is a collection of variables that might be used to study these populations.*

nations	people
population size	age
time zones	height
average rainfall	gender
life expectancy	ethnicities
mean income	annual income
literacy rate	literacy
capital city	mother's maiden name
largest river	marital status

Exercise 1.2. *Classify the variables as quantitative or categorical in the example above.*

The naming of variables and their classification as categorical or quantitative may seem like a simple, even trite, exercise. However, the first steps in designing an experiment and deciding on which individuals to include and which information to collect are vital to the success of the experiment. For example, if your goal is to measure the time for an animal (insect, bird, mammal) to complete some task under different (genetic, environmental, learning) conditions, then, you may decide to have a single quantitative variable - the time to complete the task. However, an animal in your study may not attempt the task, may not complete the task, or may perform the task. As a consequence, your data analysis will run into difficulties if you do not add a categorical variable to include these possible outcomes of an experiment.

Exercise 1.3. *Give examples of variables for the population of vertebrates, of proteins.*

1.2 Categorical Data

1.2.1 Pie Chart

A **pie chart** is a circular chart divided into sectors, illustrating relative magnitudes in frequencies or percents. In a pie chart, the area is proportional to the quantity it represents.

Example 1.4. *As the nation debates strategies for delivering health insurance, let's look at the sources of funds and the types of expenditures.*

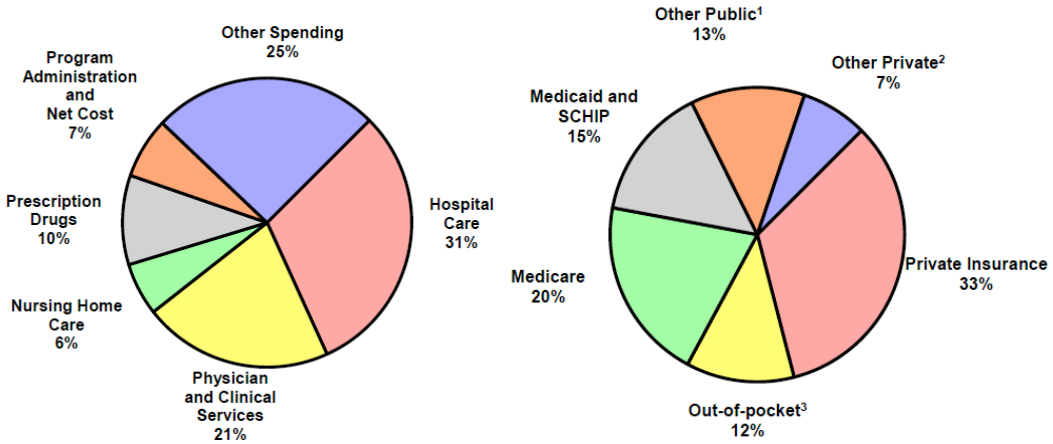
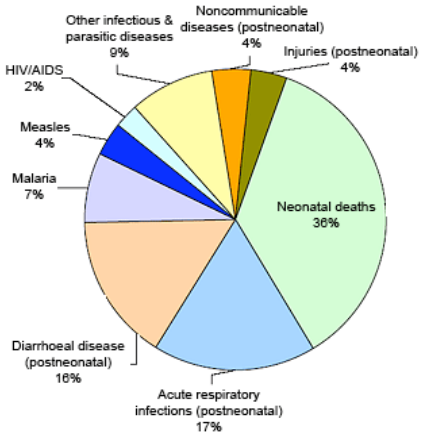


Figure 1.1: 2008 United States health care (a) expenditures (b) income sources, Source: Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group

Exercise 1.5. *How do you anticipate that this pie chart will evolve over the next decade? Which pie slices are likely to become larger? smaller? On what do you base your predictions?*

Example 1.6. *From UNICEF, we read "The proportion of children who reach their fifth birthday is one of the most fundamental indicators of a country's concern for its people. Child survival statistics are a poignant indicator of the priority given to the services that help a child to flourish: adequate supplies of nutritious food, the availability of high-quality health care and easy access to safe water and sanitation facilities, as well as the family's overall economic condition and the health and status of women in the community. "*



Major causes of death in neonates and children under five, 2004

Source: WHO The Global burden of disease:2004 update (2008)

Example 1.7. Gene Ontology (GO) *project is a bioinformatics initiative whose goal is to provide unified terminology of genes and their products. The project began in 1998 as a collaboration between three model organism databases,* Drosophila*, yeast, and mouse. The GO Consortium presently includes many databases, spanning repositories for plant, animal and microbial genomes. This project is supported by National Human Genome Research Institute. See*

`http://www.geneontology.org/`

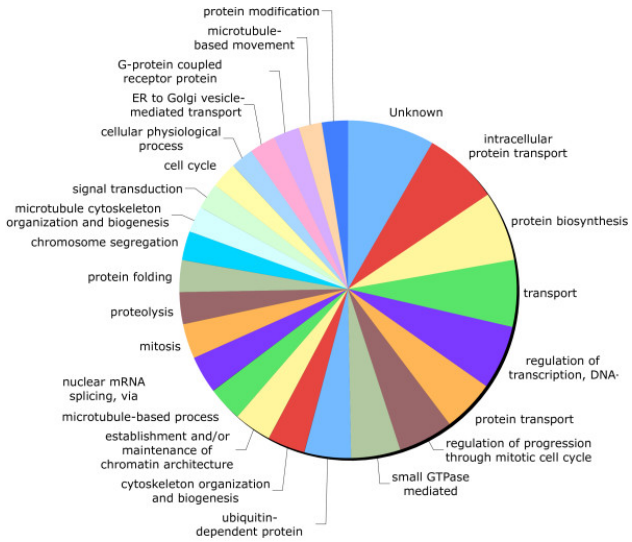


Figure 1.2: The 25 most frequent Biological Process Gene Ontology (GO) terms.

To make a simple **pie chart** in R for the proportion of AIDS cases among US males by transmission category.

```
> males<- c(58,18,16,7,1)
> pie(males)
```

This many be sufficient for your own personal use. However, if we want to use a pie chart in a presentation, we will have to provide some essential details. For a more descriptive pie chart, one has to become accustomed to learning to interact with the software to settle on a graph that is satisfactory to the situation.

- Define some colors ideal for black and white print.

  ```
  > colors <- c("white","grey70","grey90","grey50","black")
  ```

- Calculate the percentage for each category.

  ```
  > male_labels <- round(males/sum(males)*100, 1)
  ```

  The number 1 indicates rounded to one decimal place.

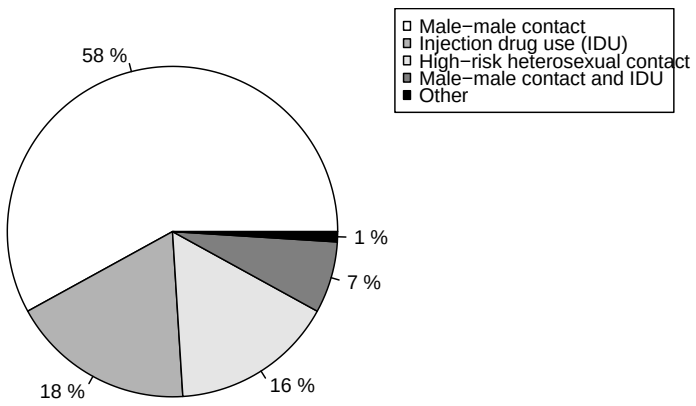
  ```
  > male_labels <- paste(male_labels, "\%", sep=" ")
  ```

  This adds a space and a percent sign.

- Create a pie chart with defined heading and custom colors and labels and create a legend.

  ```
  > pie(males, main="Proportion of AIDS Cases among Males by  Transmission Category
  + Diagnosed - USA, 2005", col=colors, labels=male_labels, cex=0.8)
  > legend("topright", c("Male-male contact","Injection drug use (IDU)",
  + "High-risk heterosexual contact","Male-male contact and IDU","Other"),
  + cex=0.8,fill=colors)
  ```

  The entry `cex=0.8` indicates that the legend has a type set that is 80% of the font size of the main title.



1.2.2 Bar Charts

Because the human eye is good at judging linear measures and poor at judging relative areas, a **bar chart** or **bar graph** is often preferable to pie charts as a way to display categorical data.

To make a simple bar graph in R,

```
> barplot(males)
```

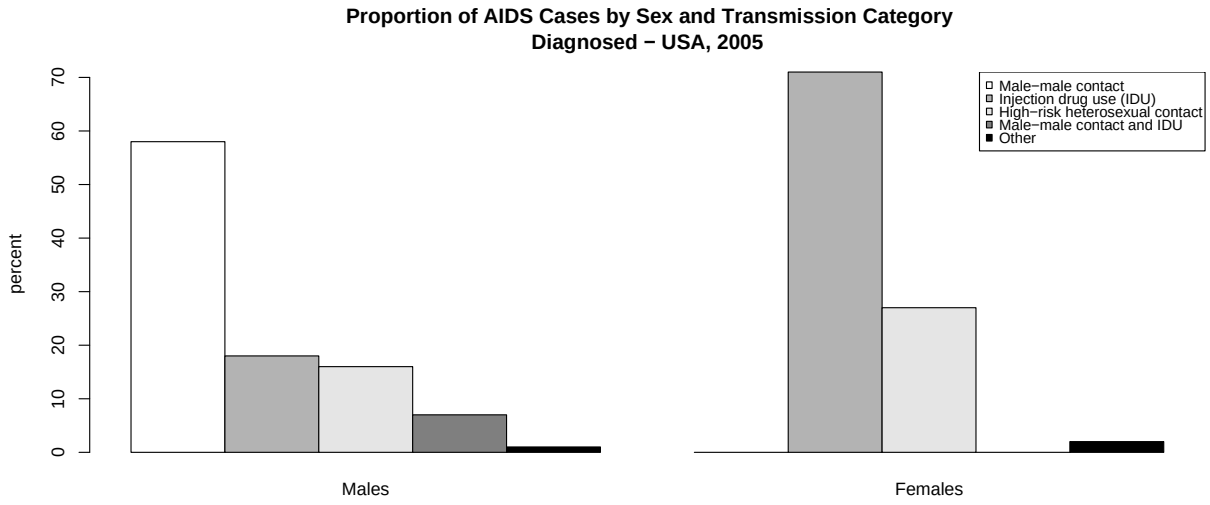
For a more descriptive bar chart with information on females:

- Enter the data for females and create a 5×2 array.

```
> females <- c(0,71,27,0,2)
> hiv<-array(c(males,females), dim=c(5,2))
```

- Generate side-by-side bar graphs and create a legend,

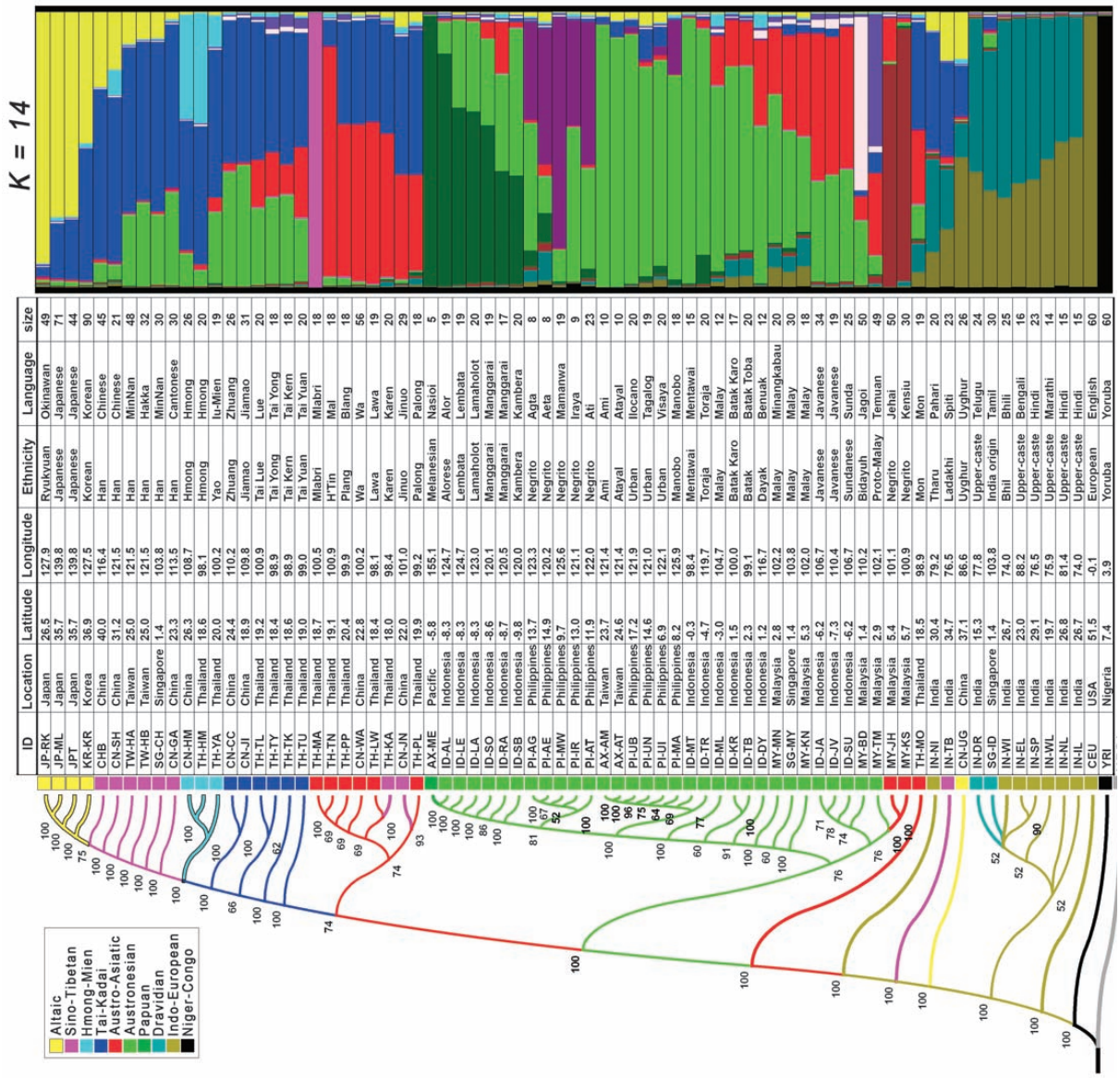
```
> barplot(hiv, main="Proportion of AIDS Cases by Sex and Transmission Category
+ Diagnosed - USA, 2005", ylab= "percent",  beside=TRUE,
+ names.arg = c("Males", "Females"),col=colors)
> legend("topright", c("Male-male contact","Injection drug use (IDU)",
+ "High-risk heterosexual contact","Male-male contact and IDU","Other"),
+ cex=0.8,fill=colors)
```



Example 1.8. *Next we examine a segmented bar plot. This shows the ancestral sources of genes for 75 populations throughout Asia. the data are based on information gathered from 50,000 genetic markers. The designations for the groups were decided by the software package STRUCTURE.*

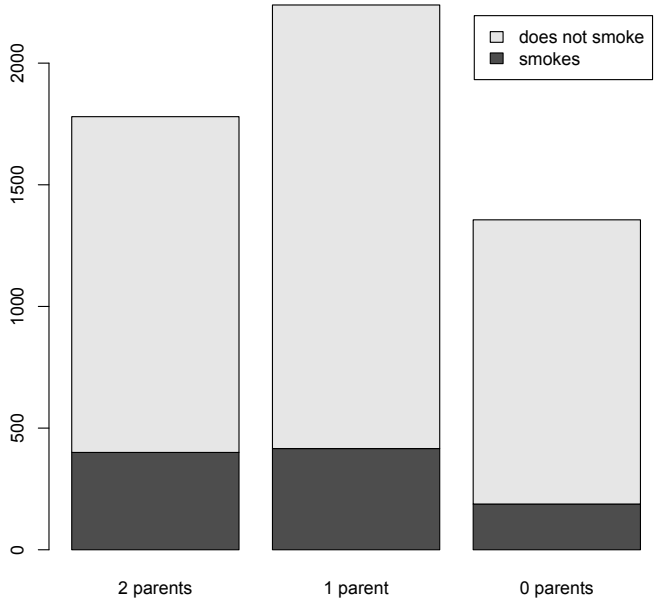
1.3 Two-way Tables

Relationships between two categorical variables can be shown through a **two-way table** (also known as a contingency table , cross tabulation table or a cross classifying table).



ID	Location	Latitude	Longitude	Ethnicity	Language	size
JP-RK	Japan	26.5	127.9	Ryukyuan	Okinawan	49
JP-ML	Japan	35.7	139.8	Japanese	Japanese	71
JPT	Japan	35.7	139.8	Japanese	Japanese	44
KR-KR	Korea	36.9	127.5	Korean	Korean	90
CHB	China	40.0	116.4	Han	Chinese	45
CN-SH	China	31.2	121.5	Han	Chinese	21
TW-HA	Taiwan	25.0	121.5	Han	MinNan	48
TW-HB	Taiwan	25.0	121.5	Han	Hakka	32
SG-CH	Singapore	1.4	103.8	Han	MinNan	30
CN-GA	China	23.3	113.5	Han	Cantonese	30
CN-HM	China	26.3	108.7	Hmong	Hmong	26
TH-HM	Thailand	18.6	98.1	Hmong	Hmong	20
TH-YA	Thailand	20.0	100.2	Yao	Iu-Mien	19
CN-CC	China	24.4	110.2	Zhuang	Zhuang	26
CN-JI	China	18.9	109.8	Jiamao	Jiamao	31
TH-TL	Thailand	19.2	100.9	Tai Lue	Lue	20
TH-TY	Thailand	18.4	98.9	Tai Yong	Tai Yong	18
TH-TK	Thailand	18.6	98.9	Tai Kern	Tai Kern	18
TH-TU	Thailand	19.0	99.0	Tai Yuan	Tai Yuan	20
TH-MA	Thailand	18.7	100.5	Mlabri	Mlabri	18
TH-TN	Thailand	19.1	100.9	H'Tin	Mal	18
TH-PP	Thailand	20.4	99.9	Pliang	Blang	18
CN-WA	China	22.8	100.2	Wa	Wa	56
TH-LW	Thailand	18.4	98.1	Lawa	Lawa	19
TH-KA	Thailand	18.0	98.4	Karen	Karen	20
CN-JN	China	22.0	101.0	Jinuo	Jinuo	29
TH-PL	Thailand	19.9	99.2	Palong	Palong	18
AX-ME	Pacific	-5.8	155.1	Melanesian	Nasioi	5
ID-AL	Indonesia	-8.3	124.7	Alorese	Alor	19
ID-LE	Indonesia	-8.3	124.7	Lembata	Lembata	19
ID-LA	Indonesia	-8.3	123.0	Lamaholot	Lamaholot	20
ID-SO	Indonesia	-8.6	120.1	Manggarai	Manggarai	19
ID-RA	Indonesia	-8.7	120.5	Manggarai	Manggarai	17
ID-SB	Indonesia	-9.8	120.0	Kambera	Kambera	20
PI-AG	Philippines	13.7	123.3	Negrito	Agta	8
PI-AE	Philippines	14.9	120.2	Negrito	Aeta	8
PI-MW	Philippines	9.7	125.6	Negrito	Mamanwa	19
PI-IR	Philippines	13.0	121.1	Negrito	Iraya	9
PI-AT	Philippines	11.9	122.0	Negrito	Ati	23
AX-AM	Taiwan	23.7	121.4	Ami	Ami	10
AX-AT	Taiwan	24.6	121.4	Atayal	Atayal	10
PI-UB	Philippines	17.2	121.9	Urban	Ilocano	20
PI-UN	Philippines	14.6	121.0	Urban	Tagalog	19
PI-UI	Philippines	6.9	122.1	Urban	Visaya	20
PI-MA	Philippines	8.2	125.9	Manobo	Manobo	18
ID-MT	Indonesia	-0.3	98.4	Mentawai	Mentawai	15
ID-TR	Indonesia	-4.7	119.7	Toraja	Toraja	20
ID-ML	Indonesia	-3.0	104.7	Malay	Malay	12
ID-KR	Indonesia	1.5	100.0	Batak Karo	Batak Karo	17
ID-TB	Indonesia	2.3	99.1	Batak	Batak Toba	20
ID-DY	Indonesia	1.2	116.7	Dayak	Benuak	12
MY-MN	Malaysia	2.8	102.2	Malay	Minangkabau	20
SG-MY	Singapore	1.4	103.8	Malay	Malay	30
MY-KN	Malaysia	5.3	102.0	Malay	Malay	18
ID-JA	Indonesia	-6.2	106.7	Javanese	Javanese	34
ID-JV	Indonesia	-7.3	110.4	Javanese	Javanese	19
ID-SU	Indonesia	-6.2	106.7	Sundanese	Sunda	25
MY-BD	Malaysia	1.4	110.2	Bidayuh	Jagoi	50
MY-TM	Malaysia	2.9	102.1	Proto-Malay	Temuan	49
MY-JH	Malaysia	5.4	101.1	Negrito	Jehai	50
MY-KS	Malaysia	5.7	100.9	Negrito	Kensiu	30
TH-MO	Thailand	18.5	98.9	Mon	Mon	19
IN-NI	India	30.4	79.2	Tharu	Pahari	20
IN-TB	India	34.7	76.5	Ladakhi	Spiti	23
CN-UG	China	37.1	86.6	Uyghur	Uyghur	26
IN-DR	India	15.3	77.8	Upper-caste	Telugu	24
SG-ID	Singapore	1.4	103.8	India origin	Tamil	30
IN-WI	India	26.7	74.0	Bhil	Bhili	25
IN-EL	India	23.0	88.2	Upper-caste	Bengali	16
IN-SP	India	29.1	76.5	Upper-caste	Hindi	23
IN-WL	India	19.7	75.9	Upper-caste	Marathi	14
IN-NL	India	26.8	81.4	Upper-caste	Hindi	15
IN-IL	India	26.7	74.0	Upper-caste	Hindi	15
CEU	USA	51.5	-0.1	European	English	60
YRI	Nigeria	7.4	3.9	Yoruba	Yoruba	60

Figure 1.3: Dispaying human genetic diversity for 75 populations in Asia. The software program STRUCTURE here infers 14 source populations, 10 of them major. The length of each segment in the bar is the estimate by STRUCTURE of the fraction of the genome in the sample that has ancestors among the given source population.



Example 1.9. *In 1964, Surgeon General Dr. Luther Leonidas Terry published a landmark report saying that smoking may be hazardous to health. This led to many influential reports on the topic, including the study of the smoking habits of 5375 high school children in Tucson in 1967. Here is a two-way table summarizing some of the results.*

	student smokes	student does not smoke	total
2 parents smoke	400	1380	1780
1 parent smokes	416	1823	2239
0 parents smoke	188	1168	1356
total	1004	4371	5375

- The **row variable** is the parents smoking habits.
- The **column variable** is the student smoking habits.
- The **cells** display the counts for each of the categories of row and column variables.

A two-way table with r rows and c columns is often called an r by c table (written $r \times c$).

The totals along each of the rows and columns give the **marginal distributions**. We can create a **segmented bar graph** as follows:

```
> smoking<-matrix(c(400,1380,416,1823,188,1168),ncol=3)
> colnames(smoking)<-c("2 parents","1 parent", "0 parents")
> rownames(smoking)<-c("smokes","does not smoke")
> smoking
              2 parents 1 parent 0 parents
smokes              400      416       188
does not smoke     1380     1823      1168
> barplot(smoking,legend=rownames(smoking))
```

Example 1.10. *Hemoglobin E is a variant of hemoglobin with a mutation in the β globin gene causing substitution of glutamic acid for lysine at position 26 of the β globin chain. HbE (E is the one letter abbreviation for glutamic acid.) is the second most common abnormal hemoglobin after sickle cell hemoglobin (HbS). HbE is common from India to Southeast Asia. The β chain of HbE is synthesized at a reduced rate compare to normal hemoglobin (HbA) as the HbE produces an alternate splicing site within an exon.*

It has been suggested that Hemoglobin E provides some protection against malaria virulence when heterozygous, but is causes anemia when homozygous. The circumstance in which the heterozygotes for the alleles under consideration have a higher adaptive value than the homozygote is called **balancing selection**.

The table below gives the counts of differing hemoglobin genotypes on two Indonesian islands.

genotype	AA	AE	EE
Flores	128	6	0
Sumba	119	78	4

Because the heterozygotes are rare on Flores, it appears malaria is less prevalent there since the heterozygote does not provide an adaptive advantage.

Exercise 1.11. *Make a segmented barchart of the data on hemoglobin genotypes. Have each bar display the distribution of genotypes on the two Indonesian islands.*

1.4 Histograms and the Empirical Cumulative Distribution Function

Histograms are a common visual representation of a quantitative variable. Histograms summarize the data using rectangles to display either frequencies or proportions as normalized frequencies. In making a histogram, we

- Divide the range of data into bins of equal width (usually, but not always).
- Count the number of observations in each class.
- Draw the histogram rectangles representing frequencies or percents by *area*.

Interpret the histogram by giving

- the overall pattern
  - the center
  - the spread
  - the shape (symmetry, skewness, peaks)
- and deviations from the pattern
  - outliers
  - gaps

The direction of the skewness is the direction of the longer of the two tails (left or right) of the distribution.

No one choice for the number of bins is considered best. One possible choice for larger data sets is Sturges' formula to choose $\lfloor 1 + \log_2 n \rfloor$ bins. ($\lfloor \cdot \rfloor$, the **floor function**, is obtained by rounding down to the next integer.)

Exercise 1.12. *The histograms in Figure 1.4 shows the distribution of lengths of a normal strain and mutant strain of* Bacillus subtilis*. Describe the distributions.*

Example 1.13. *Taking the age of the presidents of the United States at the time of their inauguration and creating its histogram in* R *is accomplished as follows.*

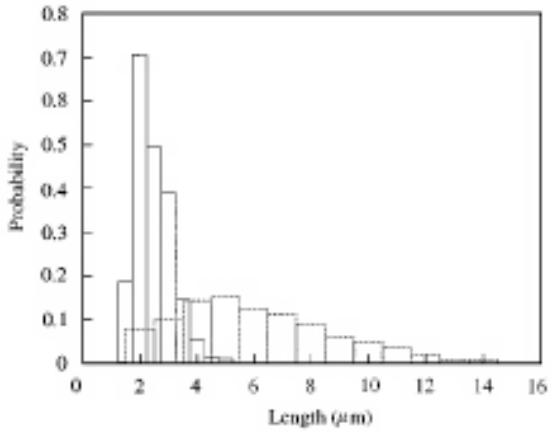
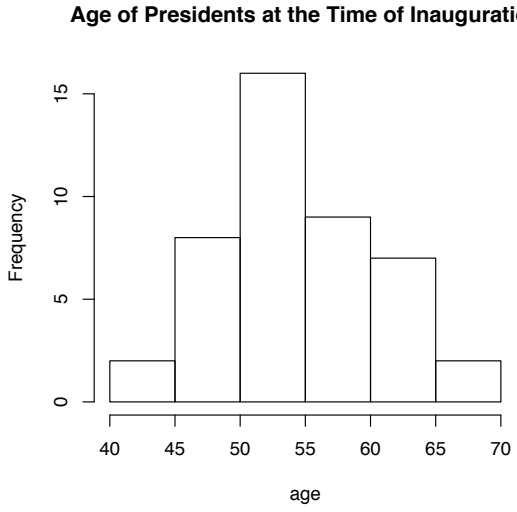
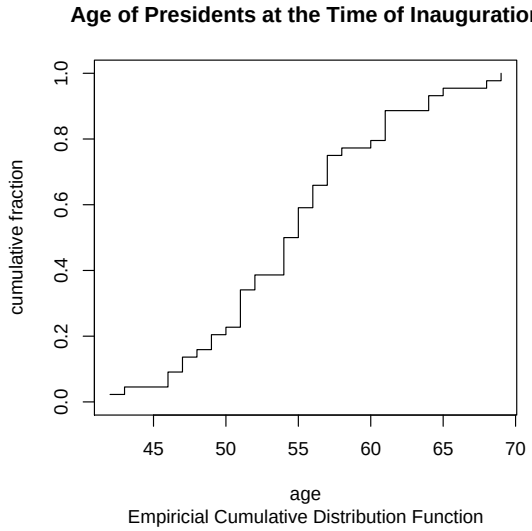


Figure 1.4: Histogram of lengths of *Bacillus subtilis*. Solid lines indicate wild type and dashed line mutant strain.

```
> age<- c(57,61,57,57,58,57,61,54,68,51,49,64,50,48,65,52,56,46,54,49,51,47,55,55,
54,42,51,56,55,51,54,51,60,61,43,55,56,61,52,69,64,46,54,47)
> hist(age, main = c("Age of Presidents at the Time of Inauguration"))
```





So the age of presidents at the time of inauguration range from the early forties to the late sixties with the frequency starting their tenure peaking in the early fifties. The histogram in generally symmetric about 55 years with spread from around 40 to 70 years.

The **empirical cumulative distribution function** $F_n(x)$ *gives, for each value* x*, the fraction of the data* **less** *than or equal to* x*. If the number of observations is* n*, then*

$$F_n(x) = \frac{1}{n}\#(\text{observations less than or equal to } x).$$

Thus, $F_n(x) = 0$ *for any value of* x *less than all of the observed values and* $F_n(x) = 1$ *for any* x *greater than all of the observed values. In between, we will see jumps that are multiples of the* $1/n$*. For example, in the empirical*

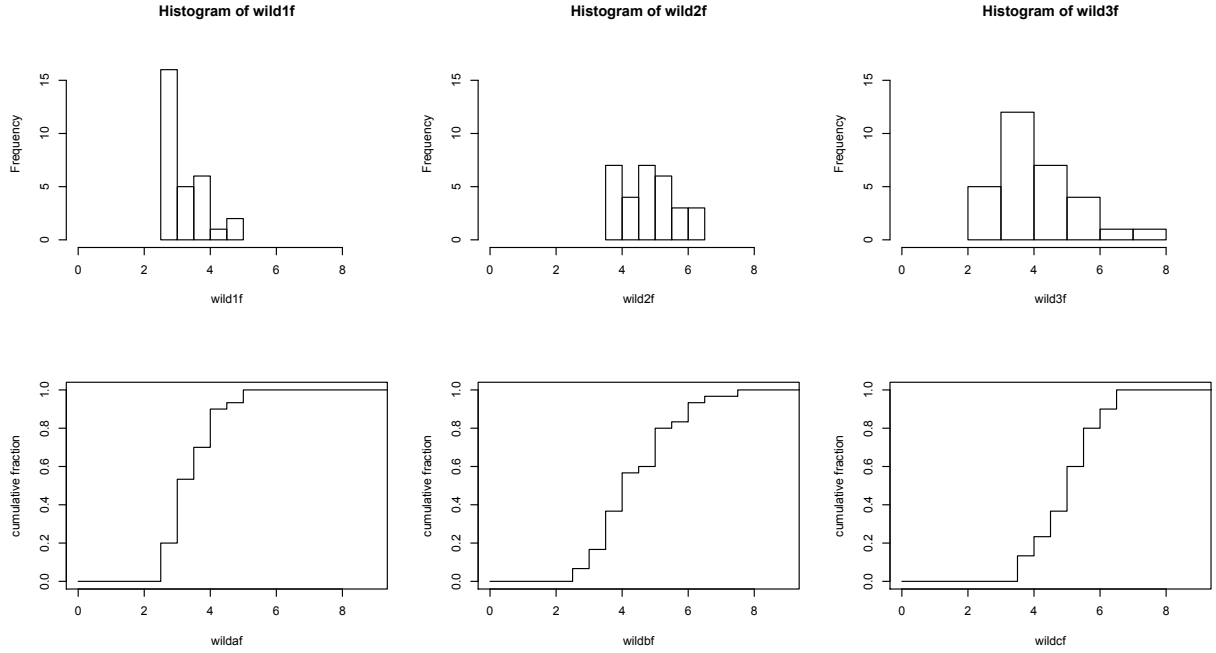
cumulative distribution function for the age of the presidents, we will see a jump of size $4/44 = 1/11$ *at* $x = 57$ *to indicate the fact that 4 of the 44 presidents were 57 at the time of their inauguration.*

In order to create a graph of the empirical cumulative distribution function, first place the observations in order from smallest to largest. For the age of presidents data, we can accomplish this in R *by writing* `sort(age)`*. Next match these up with the integral multiples of the 1 over the number of observations. In* R*, we enter* `1:length(age)/length(age)`*. Finally,* `type="s"` *to give us the steps described above.*

```
> plot(sort(age),1:length(age)/length(age),type="s",ylim=c(0,1),
main = c("Age of Presidents at the Time of Inauguration"),
sub=("Empiricial Cumulative Distribution Function"),
xlab=c("age"),ylab=c("cumulative fraction"))
```

Exercise 1.14. *Give the fraction of presidents whose age at inauguration was under 60. What is the range for the age at inauguration of the youngest fifth of the presidents?*

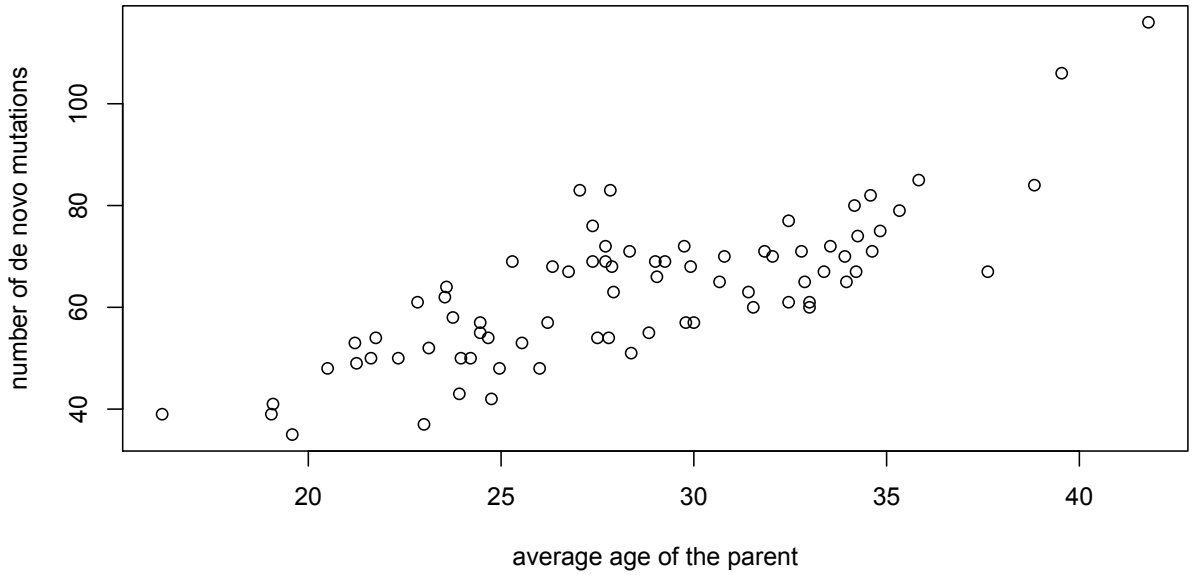
Exercise 1.15. *The histogram for data on the length of three bacterial strains is shown below. Lengths are given in microns. Below the histograms (but not necessarily directly below) are empirical cumulative distribution functions corresponding to these three histograms.*



Match the histograms to their respective empirical cumulative distribution functions.

In looking at life span data, the natural question is "What fraction of the individuals have survived a given length of time?" The **survival function** $S_n(x)$ gives, for each value x, the fraction of the data **greater** than or equal to x. If the number of observations is n, then

$$\begin{aligned} S_n(x) &= \frac{1}{n}\#(\text{observations greater than } x) = \frac{1}{n}(n - \#(\text{observations less than or equal to } x)) \\ &= 1 - \frac{1}{n}\#(\text{observations less than or equal to } x) = 1 - F_n(x) \end{aligned}$$



1.5 Scatterplots

We now consider two dimensional data. The values of the first variable $x_1, x_2, \ldots, x_n$ are assumed known and in an experiment and are often set by the experimenter. This variable is called the **explanatory, predictor, discriptor, or input variables** and in a two dimensional **scatterplot** of the data display its values on the horizontal axis. The values $y_1, y_2 \ldots, y_n$, taken from observations with input $x_1, x_2, \ldots, x_n$ are called the **response or target variable** and its values are displayed on the vertical axis. In describing a scatterplot, take into consideration

- the form, for example,
  - linear
  - curved relationships
  - clusters
- the direction,
  - a positive or negative association
- and the strength of the aspects of the scatterplot.

Example 1.16. *Genetic evolution is based on mutation. Consequently, one fundamental question in evolutionary biology is the rate of* de novo *mutations. To investigate this question in humans, Kong et al, sequenced the entire genomes of 78 Icelandic trios and recorded the age of the parents and the number of* de novo *mutations in the offspring.*

The plot shows a moderate positive linear association, children of older parent have, on average, more mutations. The number of mutations range from ~ 40 *for children of younger parents to* ~ 100 *for children of older parents. We will later learn that the father is the major source of this difference with age.*

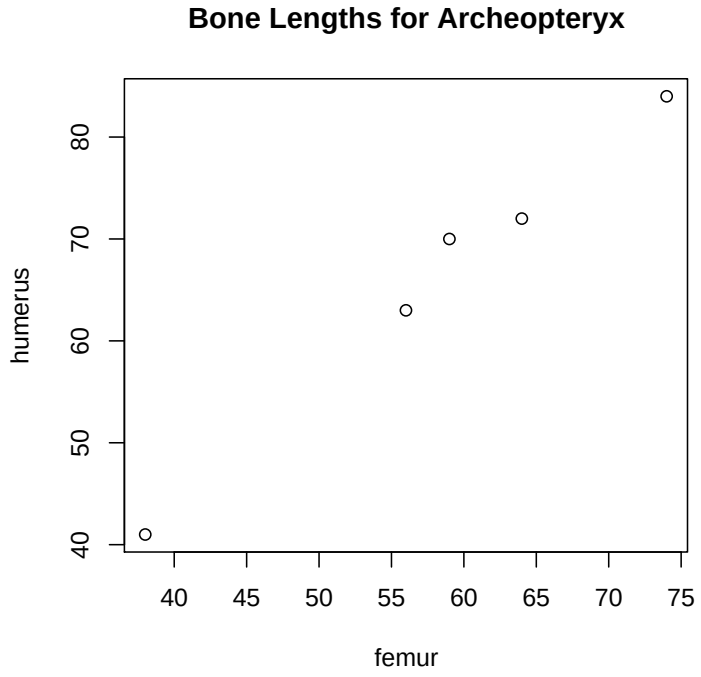
Example 1.17 (Fossils of the *Archeopteryx*)**.** *The name Archeopteryx derives from the ancient Greek meaning "ancient feather" or "ancient wing". Archeopteryx is generally accepted by palaeontologists as being the oldest known bird. Archaeopteryx lived in the Late Jurassic Period around 150 million years ago, in what is now southern Germany during a time when Europe was an archipelago of islands in a shallow warm tropical sea. The first complete specimen of Archaeopteryx was announced in 1861, only two years after Charles Darwin published* On the Origin of Species*,*

and thus became a key piece of evidence in the debate over evolution. Below are the lengths in centimeters of the femur and humerus for the five specimens of Archeopteryx *that have preserved both bones.*

femur	38	56	59	64	74
humerus	41	63	70	72	84

```
> femur<-c(38,56,59,64,74)
> humerus<-c(41,63,70,72,84)
> plot(femur, humerus,main=c("Bone Lengths for Archeopteryx"))
```

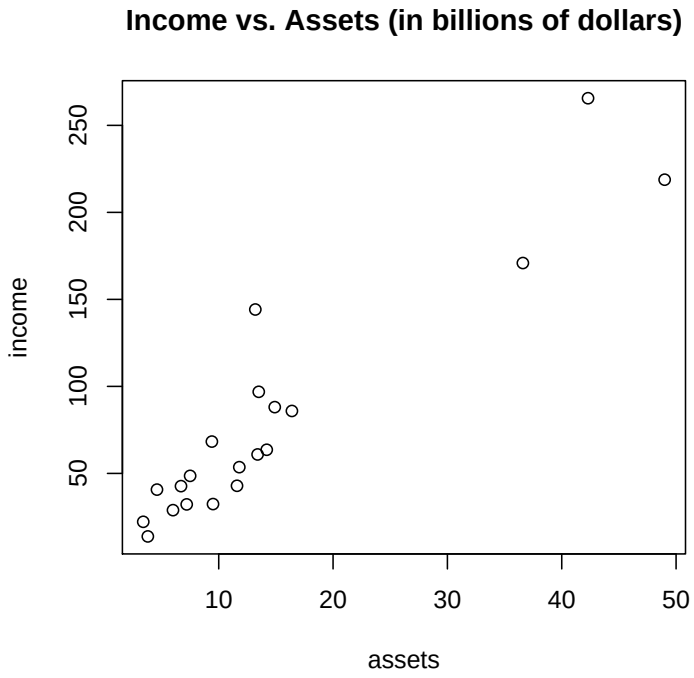
Unless we have a specific scientific question, we have no real reason for a choice of the explanatory variable.



Describe the scatterplot.

Example 1.18. *This historical data show the 20 largest banks in 1974. Values given in billions of dollars.*

Bank	1	2	3	4	5	6	7	8	9	10
Assets	49.0	42.3	36.6	16.4	14.9	14.2	13.5	13.4	13.2	11.8
Income	218.8	265.6	170.9	85.9	88.1	63.6	96.9	60.9	144.2	53.6
Bank	11	12	13	14	15	16	17	18	19	20
Assets	11.6	9.5	9.4	7.5	7.2	6.7	6.0	4.6	3.8	3.4
Income	42.9	32.4	68.3	48.6	32.2	42.7	28.9	40.7	13.8	22.2



Describe the scatterplot.

In 1972, Michele Sindona, a banker with close ties to the Mafia, along with a purportedly bogus Freemasonic lodge, and the Nixon administration purchased controlling interest in Bank 19, Long Island's Franklin National Bank. As a result of his acquisition of a controlling stake in Franklin, Sindona had a money laundering operation to aid his alleged ties to Vatican Bank and the Sicilian drug cartel. Sindona used the bank's ability to transfer funds, produce letters of credit, and trade in foreign currencies to begin building a banking empire in the United States. In mid-1974, management revealed huge losses and depositors started taking out large withdrawals, causing the bank to have to borrow over $1 billion from the Federal Reserve Bank. On 8 October 1974, the bank was declared insolvent due to mismanagement and fraud, involving losses in foreign currency speculation and poor loan policies.

What would you expect to be a feature on this scatterplot of a failing bank? Does the Franklin Bank have this feature?

1.6 Time Plots

Some data sets come with an order of events, say ordered by time.

Example 1.19. *The modern history of petroleum began in the 19th century with the refining of kerosene from crude oil. The world's first commercial oil wells were drilled in the 1850s in Poland and in Romania.The first oil well in North America was in Oil Springs, Ontario, Canada in 1858. The US petroleum industry began with Edwin Drake's drilling of a 69-foot deep oil well in 1859 on Oil Creek near Titusville, Pennsylvania for the Seneca Oil Company. The industry grew through the 1800s, driven by the demand for kerosene and oil lamps. The introduction of the internal combustion engine in the early part of the 20th century provided a demand that has largely sustained the industry to this day. Today, about 90% of vehicular fuel needs are met by oil. Petroleum also makes up 40% of total energy consumption in the United States, but is responsible for only 2% of electricity generation. Oil use increased exponentially until the world oil crises of the 1970s.*

Worldwide Oil Production

Year	Million Barrels	Year	Million Barrels	Year	Million Barrels
1880	30	1940	2150	1972	18584
1890	77	1945	2595	1974	20389
1900	149	1950	3803	1976	20188
1905	215	1955	5626	1978	21922
1910	328	1960	7674	1980	21722
1915	432	1962	8882	1982	19411
1920	689	1964	10310	1984	19837
1925	1069	1966	12016	1986	20246
1930	1412	1968	14014	1988	21338
1935	1655	1970	16690		

With the data given in two columns `oil` *and* `year`*, the time plot* `plot(year,oil,type="b")` *is given on the left side of the figure below. This uses* `type="b"` *that puts* **both** *lines and circles on the plot.*

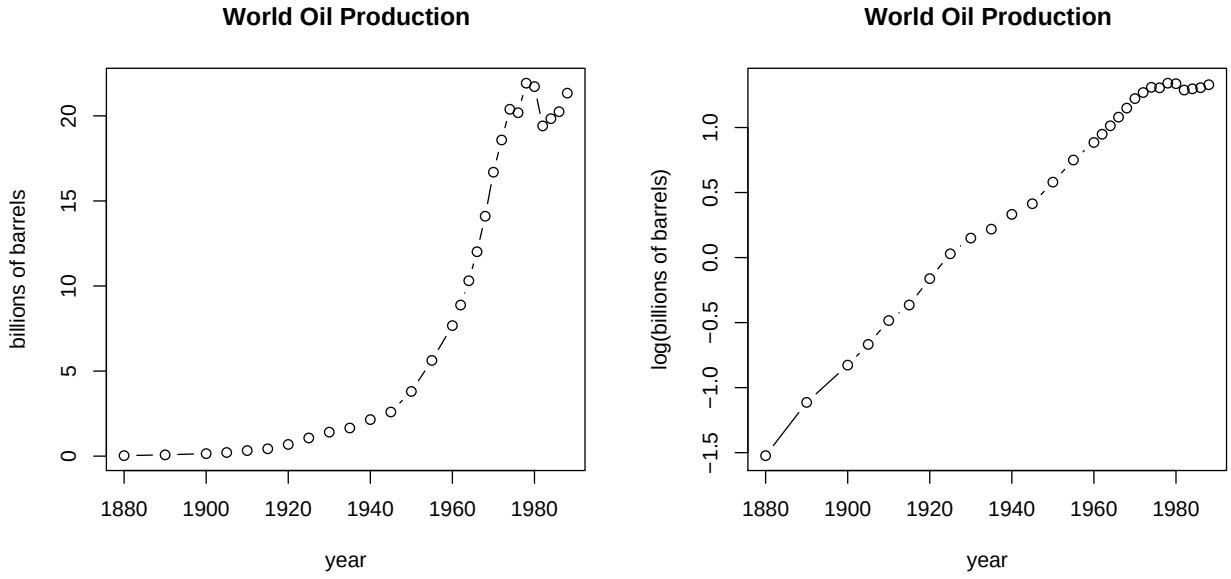


Figure 1.5: Oil production (left) and the logarithm of oil production (right) from 1880 to 1988.

Sometimes a **transformation of the data** *can reveal the structure of the time series. For example, if we wish to examine an exponential increase displayed in the oil production plot, then we can take the base 10 logarithm of the production and give its time series plot. This is shown in the plot on the right above. (In* R*, we write* `log(x)` *for the natural logarithm and* `log(x,10)` *for the base 10 logarithm.)*

Exercise 1.20. *What happened in the mid 1970s that resulted in the long term departure from exponential growth in the use of oil?*

Example 1.21. *The Intergovernmental Panel on Climate Change (IPCC) is a scientific intergovernmental body tasked with evaluating the risk of climate change caused by human activity. The panel was established in 1988 by the World Meteorological Organization and the United Nations Environment Programme, two organizations of the United Nations. The IPCC does not perform original research but rather uses three working groups who synthesize research and prepare a report. In addition, the IPCC prepares a summary report. The Fourth Assessment Report (AR4) was completed in early 2007. The fifth was released in 2014.*

Below is the first graph from the 2007 Climate Change Synthesis Report: Summary for Policymakers.

The technique used to draw the curves on the graphs is called **local regression**. *At the risk of discussing concepts that have not yet been introduced, let's describe the technique behind local regression. Typically, at each point in the data set, the goal is to draw a linear or quadratic function. The function is determined using weighted least squares, giving most weight to nearby points and less weight to points further away. The graphs above show the approximating curves. The blue regions show areas within two standard deviations of the estimate (called a confidence interval). The goal of local regression is to provide a smooth approximation to the data and a sense of the uncertainty of the data. In practice, local regression requires a large data set to work well.*

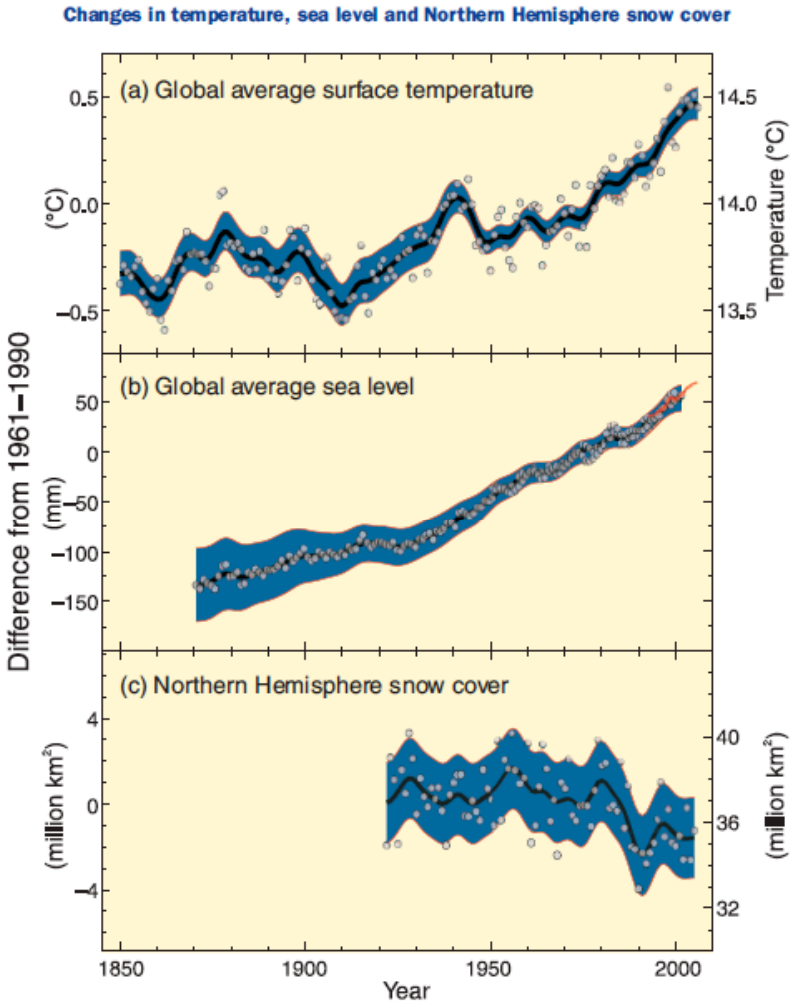
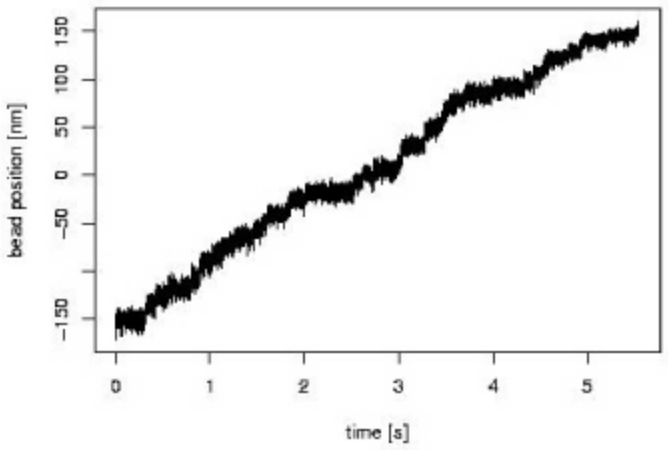
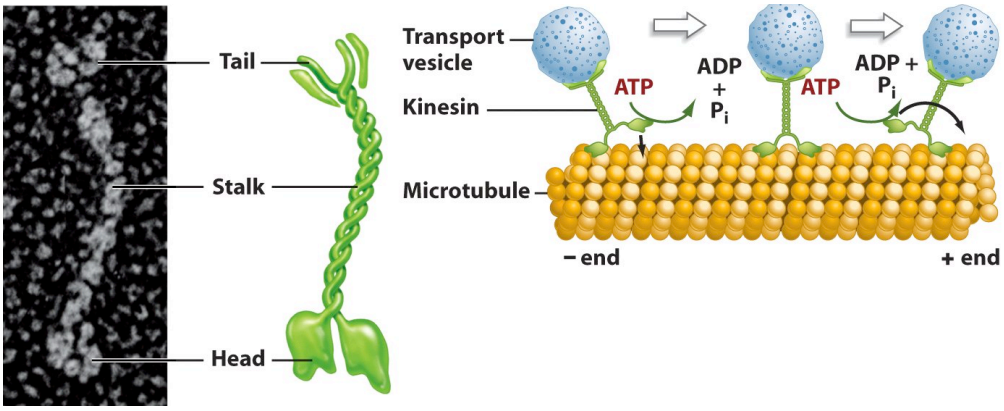


Figure SPM.1. *Observed changes in (a) global average surface temperature; (b) global average sea level from tide gauge (blue) and satellite (red) data and (c) Northern Hemisphere snow cover for March-April. All differences are relative to corresponding averages for the period 1961-1990. Smoothed curves represent decadal averaged values while circles show yearly values. The shaded areas are the uncertainty intervals estimated from a comprehensive analysis of known uncertainties (a and b) and from the time series (c). [Figure 1.1]*

Example 1.22. *The next figure give a time series plot of a single molecule experiment showing the movement of kinesin along a microtubule. In this case the kinesin has at its foot a glass bead and its heads are attached to a microtubule. The position of the glass bead is determined by using a laser beam and the optical properties of the bead to locate the bead and provide a force on the kinesin molecule. In this time plot, the load on the microtubule has a force of 3.5 pN and the concentration of ATP is 100μM. What is the source of fluctuations in this time series plot of bead position? How would you expect this time plot to change with changes in ATP concentration and with changes in force?*



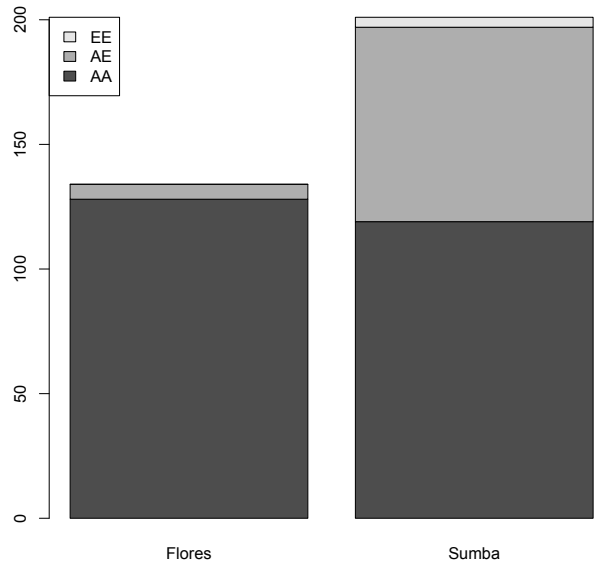


1.7 Answers to Selected Exercises

1.11. Here are the R commands:

```
> genotypes<-matrix(c(128,6,0,119,78,4),ncol=2)
> colnames(genotypes)<-c("Flores","Sumba")
> rownames(genotypes)<-c("AA","AE","EE")
> genotypes
   Flores Sumba
AA    128   119
AE      6    78
EE      0     4
> barplot(genotypes,legend=rownames(genotypes),args.legend=list(x="topleft"))
```

The legend was moved to the left side to avoid crowding with the taller bar for the data on Sumba.



1.12. The lengths of the normal strain has its center at 2.5 microns and range from 1.5 to 5 microns. It is somewhat skewed right with no outliers. The mutant strain has its center at 5 or 6 microns. Its range is from 2 to 14 microns and it is slightly skewed right. It has not outliers.

1.14. Look at the graph to the point above the value 60 years. Look left from this point to note that it corresponds to a value of 0.80.

Look at the graph to the point right from the value 0.20. Look down to note that it corresponds to 49 years. .

1.15. Match histogram *wild1f* to *wilddaf*. Note that both show the range is from 2 to 5 microns and that about half of the data lies between 2 and 3 microns. Match histogram *wild2f* with *wildcf*. The data is relatively uniform from 3.5 to 6.5 microns. Finally, match histogram *wild3f* with *wildbf*. The range is from 2 to 8 microns with most of the data between 3 and 6 microns. .

1.22. The fluctuation are due to the many bombardments with other molecules in the cell, most frequently, water molecules.

As force increases, we expect the velocity to increase - to a point. If the force is too large, then the kinesin is ripped away from the microtubule. As ATP concentration increases, we expect the velocity to increase - again, to a point. If ATP concentration is sufficiently large, then the biochemical processes are saturated.

Topic 2

Describing Distributions with Numbers

There are three kinds of lies: lies, damned lies, and statistics. - Benjamin Disraeli

It is easy to lie with statistics. It is hard to tell the truth without it. - Andrejs Dunkels

We next look at **quantitative data**. Recall that in this case, these data can be subject to the operations of arithmetic. In particular, we can add or subtract observation values, we can sort them and rank them from lowest to highest.

We will look at two fundamental properties of these observations. The first is a measure of the center value for the data, i.e., the **median** or the **mean**. Associated to this measure, we add a second value that describes how these observations are spread about this given measure of center.

The median is the central observation of the data after it is sorted from the lowest to highest observations. In addition, to give a sense of the spread in the data, we often give the smallest and largest observations as well as the observed value that is 1/4 and 3/4 of the way up this list, known at the **first** and **third quartiles**. For the mean, we commonly use the **standard deviation** to describe the spread of the data.

These concepts will be described in more detail in this section.

2.1 Measuring Center

2.1.1 Medians

The **median** take the middle value for $x_1, x_2, \ldots, x_n$ after the data has been sorted from smallest to largest,

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

($x_{(k)}$ is called the k-th **order statistic**. Sorting can be accomplished in R by using the `sort` command.)

If n is odd, then this is just the value of the middle observation $x_{((n+1)/2)}$. If n is even, then the two values closest to the center are averaged.

$$\frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}).$$

If we store the data in R in a vector `x`, we can write `median(x)` to compute the median.

2.1.2 Means

For a collection of numeric data, $x_1, x_2, \ldots, x_n$, the **sample mean** is the numerical average

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Alternatively, if the value x occurs $n(x)$ times in the data, then use the distributive property to see that

$$\bar{x} = \frac{1}{n}\sum_x xn(x) = \sum_x xp(x), \quad \text{where} \quad p(x) = \frac{n(x)}{n}.$$

So the mean $\bar{x}$ depends only on the proportion of observations $p(x)$ for each value of x.

Example 2.1. *For the data set* $\{1, 2, 2, 2, 3, 3, 4, 4, 4, 5\}$, *we have* $n = 10$ *and the sum*

$$\begin{aligned} 1+2+2+2+3+3+4+4+4+5 &= 1n(1) + 2n(2) + 3n(3) + 4n(4) + 5n(5) \\ &= 1(1) + 2(3) + 3(2) + 4(3) + 5(1) = 30 \end{aligned}$$

Thus, $\bar{x} = 30/10 = 3$.

Example 2.2. *For the data on the length in microns of wild type* Bacillus subtilis *data, we have*

length x	frequency $n(x)$	proportion $p(x)$	product $xp(x)$
1.5	18	0.090	0.135
2.0	71	0.355	0.710
2.5	48	0.240	0.600
3.0	37	0.185	0.555
3.5	16	0.080	0.280
4.0	6	0.030	0.120
4.5	4	0.020	0.090
sum	200	1	2.490

So the sample mean $\bar{x} = 2.49$.

If we store the data in R in a vector `x`, we can write `mean(x)` which is equal to `sum(x)/length(x)` to compute the mean.

To extend this idea a bit, we can take a real-valued function h and instead consider the observations $h(x_1), h(x_2), \ldots, h(x_n)$, then

$$\overline{h(x)} = \frac{1}{n}(h(x_1) + h(x_2) + \ldots + h(x_n)) = \frac{1}{n}\sum_{i=1}^{n} h(x_i) = \frac{1}{n}\sum_x h(x)n(x) = \sum_x h(x)p(x).$$

Exercise 2.3. *Let* $\bar{x}_n$ *be the sample mean for the quantitative data* $x_1, x_2, \ldots, x_n$. *For an additional observation* x_{n+1}, *use* $\bar{x}$ *to give a formula for* $\bar{x}_{n+1}$, *the mean of* $n+1$ *observations. Generalize this formula for the case of* k *additional observations* $x_{n+1} \ldots, x_{n+k}$

Many times, we do not want to give the same **weight** to each observation. For example, in computing a student's grade point average, we begin by setting values x_i corresponding to grades (A $\mapsto$ 4, B $\mapsto$ 3 and so on) and giving weights $w_1, w_2, \ldots, w_n$ equal to the number of units in a course. We then compute the **grade point average** as a **weighted mean**. To do this:

- Multiply the value of each course by its weight $x_i w_i$. This is called the number of **quality points** for the course.
- Add up the quality points:

$$x_1w_1 + x_2w_2 + \ldots + x_nw_n = \sum_{i=1}^{n} x_iw_i$$

- Add up the weights, i. e., the number of units attempted:

$$w_1 + w_2 + \ldots + w_n = \sum_{i=1}^{n} w_i$$

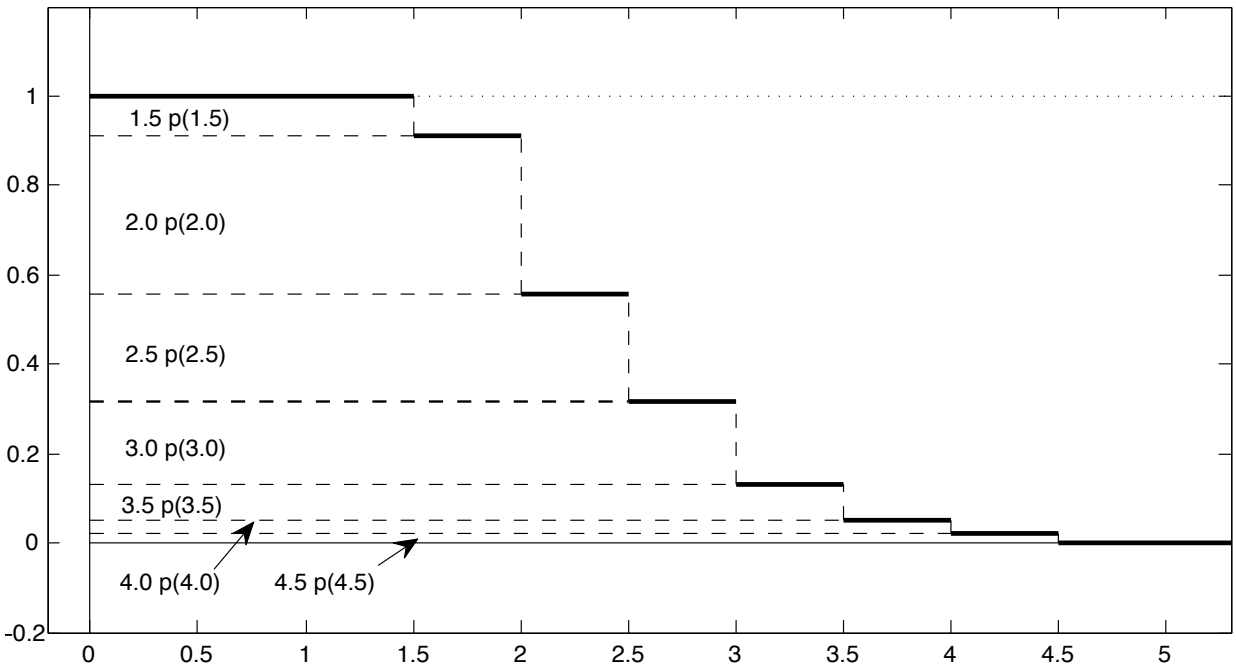


Figure 2.1: Empirical Survival Function for the Bacterial Data. This figure displays how the area under the survival function to the right of the y-axis and above the x-axis is the mean value $\bar{x}$ for non-negative data. For $x = 1.5, 2.0, 2.5, 3.0, 3.5, 4.0$, and 4.5. This area is the sum of the area of the retangles displayed. The width of each of the rectangles is x and the height is equal to $p(x)$. Thus, the area is the product $xp(x)$. The sum of these areas are presented in Example 2.2 to compute the sample mean.

- Divide the total quality points by the number of units attempted:

$$\frac{x_1w_1 + x_2w_2 + \ldots + x_nw_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^n x_iw_i}{\sum_{i=1}^n w_i}. \tag{2.1}$$

  If we let

$$p_j = w_j / \sum_{i=1}^n w_i$$

  be the **proportion** or **fraction** of the weight given to the j-th observation, then we can rewrite (2.1) as

$$\sum_{i=1}^n x_ip_i.$$

If we store the weights in a vector `w`, then we can compute the weighted mean using `weighted.mean(x,w)`

If an extremely high observation is changed to be even higher, then the mean follows this change while the median does not. For this reason, the mean is said to be *sensitive to outliers* while the median is not. To reduce the impact of extreme outliers on the mean as a measure of center, we can also consider a **truncated mean** or **trimmed mean**. The p trimmed mean is obtained by discarding both the lower and the upper $p \times 100\%$ of the data and taking the arithmetic mean of the remaining data.

In R, we write `mean(x, trim = p)` where `p`, a number between 0 and 0.5, is the fraction of observations to be trimmed from each end before the mean is computed.

Note that the median can be regarded as the 50% trimmed mean. The median does not change with a changes in the extreme observations. Such a property is called a **resistant measure**. On the other hand, the mean is *not* a resistant measure.

Exercise 2.4. *Give the relationship between the median and the mean for a (a) left skewed, (b) symmetric, or (c) right skewed distribution.*

2.2 Measuring Spread

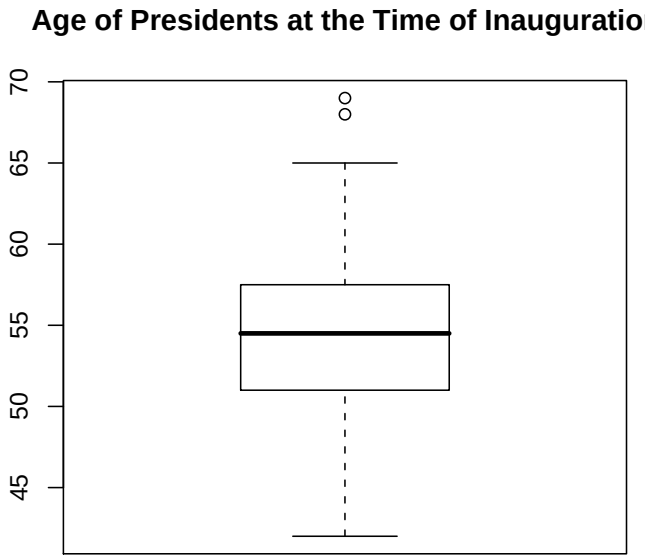
2.2.1 Five Number Summary

The **first** and **third quartile**, Q_1 and Q_3, are, respectively, the median of the lower half and the upper half of the data. The **five number summary** of the data are the values of the minimum, Q_1, the median, Q_3 and the maximum. These values, along with the mean, are given in R using `summary(x)`. Returning to the data set on the age of presidents:

```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  42.00   51.00   54.50   54.64   57.25   69.00
```

We can display the five number summary using a **boxplot**.

```
> boxplot(age, main = c("Age of Presidents at the Time of Inauguration"))
```



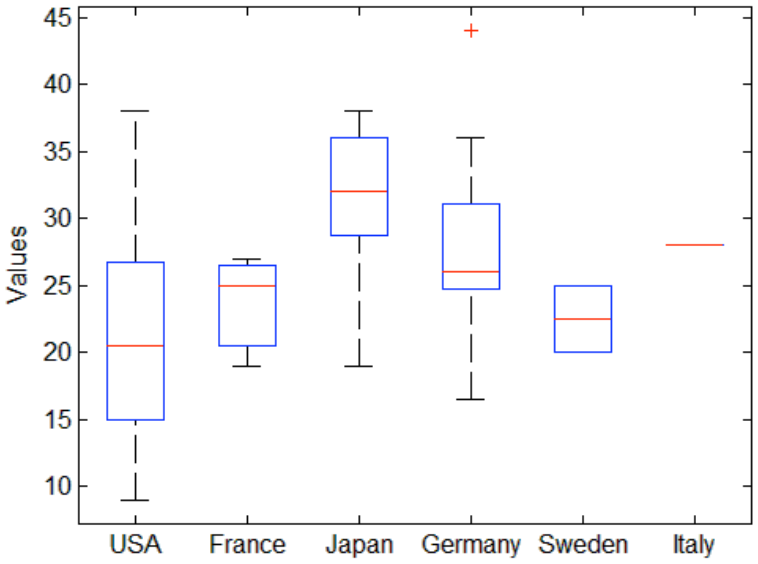
The value $Q_3 - Q_1$ is called the **interquartile range** and is denoted by IQR. It is found in R with the command `IQR`. Outliers are somewhat arbitrarily chosen to be those above $Q_3 + \frac{3}{2}IQR$ and below $Q_1 - \frac{3}{2}IQR$. With this criterion, the ages of William Henry Harrison and Ronald Reagan, considered outliers, are displayed by the two circles at the top of the boxplot. The `boxplot` command has the default value `range = 1.5` in the choice of displaying outliers. This can be altered to loosen or tighten this criterion.

Exercise 2.5. *Create a boxplot for the age of the presidents at the time of their inauguration using as outliers any value above $Q_3 + IQR$ and below $Q_1 - IQR$ as the criterion for outliers. How many outliers does this boxplot have?*

Example 2.6. *Consider a two column data set. Column 1 -* `MPH` *- gives car gas milage. Column 2 -* `origin` *- gives the country of origin for the car. We can create side by side boxplots with the command*

```
> boxplot(MPG,Origin)
```

to produce



2.2.2 Sample Variance and Standard Deviation

The **sample variance** averages the square of the differences from the mean

$$\text{var}(x) = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The **sample standard deviation**, s_x, is the square root of the sample variance. We shall soon learn the rationale for the decision to divide by $n-1$. However, we shall also encounter circumstances in which division by n is preferable. We will routinely drop the subscript x and write s to denote standard deviation if there is no ambiguity.

Example 2.7. *For the data set on* Bacillus subtilis *data, we have* $\bar{x} = 498/200 = 2.49$

length, x	frequency, $n(x)$	$x - \bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2 n(x)$
1.5	18	-0.99	0.9801	17.6418
2.0	71	-0.49	0.2401	17.0471
2.5	48	0.01	0.0001	0.0048
3.0	37	0.51	0.2601	9.6237
3.5	16	1.01	1.0201	16.3216
4.0	6	1.51	2.2801	13.6806
4.5	4	2.01	4.0401	16.1604
sum	200			90.4800

So the sample variance $s_x^2 = 90.48/199 = 0.4546734$ *and standard deviation* $s_x = 0.6742947$. *To accomplish this in* R

```
> bacteria<-c(rep(1.5,18),rep(2.0,71),rep(2.5,48),rep(3,37),rep(3.5,16),rep(4,6),
+ rep(4.5,4))
> length(bacteria)
[1] 200
> mean(bacteria)
[1] 2.49
> var(bacteria)
[1] 0.4546734
> sd(bacteria)
[1] 0.6742947
```

Exercise 2.8. *Show that* $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

We now begin to describe the rationale for the division by $n-1$ rather than n in the definition of the variance. To introduce the next exercise, define the sum of squares about the value α,

$$SS(\alpha) = \sum_{i=1}^{n}(x_i - \alpha)^2.$$

Exercise 2.9. *Flip a fair coin 16 times, recording the number of heads. Repeat this activity 20 times, giving* $x_1, \ldots, x_{20}$ *heads. Our instincts say that the mean should be 8. Compute* $SS(8)$. *Next find* $\bar{x}$ *for the data you generated and compute* $SS(\bar{x})$. *Notice that* $SS(8) > SS(\bar{x})$.

Note that in repeating the experiment of flipping a fair coin 16 times and recording the number of heads, we would like to compute the variation about 8, the value that our intuition tells us is the true mean. In many circumstances, we do not have such intuition. Thus, we doing the best we can by computing $\bar{x}$, the mean from the data. In this case, the variation about the sample mean is smaller than the variation about what may be called a *true mean*. Thus, division of $\sum_{i=1}^{n}(x_i - \bar{x})^2$ by n systematically underestimates the variance. The definition of sample variance is based on the fact that this can be compensated for this by dividing by something small than n. We will learn why the appropriate choice is $n-1$ when we investigate Unbiased Estimation in Topic 13.

To show that the phenomena in Exercise 2.9 is true more broadly, we next perform a little algebra. This is similar to the computation of the parallel axis theorem in physics. The parallel axis theorem is used to determine the moment of inertia of a rigid body about any axis, given the moment of inertia of the object about the parallel axis through the object's center of mass ($\bar{x}$) and the perpendicular distance between the axes. In this case, we a looking at the rigid motion of a finite number of equal point masses.

In the formula for $SS(\alpha)$, divide the difference in the value of each observation x_i to the value α into the difference to the sample mean $\bar{x}$ and then the distance from the sample mean to α (i.e. $\bar{x} - \alpha$).

$$\begin{aligned} SS(\alpha) &= \sum_{i=1}^{n}((x_i - \bar{x}) + (\bar{x} - \alpha))^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \alpha) + \sum_{i=1}^{n}(\bar{x} - \alpha)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - \alpha)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \alpha)^2. \end{aligned}$$

By Exercise 2.8, the cross term above $2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \alpha)$ equals to zero. Thus, we have *partitioned the sums of squares* into two levels. The first term gives the sums of squares about the sample mean $\bar{x}$. The second gives square of the difference between $\bar{x}$ and the chosen value α. We shall see this idea of partitioning in other contexts.

Note that the minimum value of $SS(\alpha)$ can be obtained by minimizing the second term. This takes place at $\alpha = \bar{x}$. Thus,

$$\min_{\alpha} SS(\alpha) = SS(\bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Our second use for this identity provides an alternative method to compute the variance. Take $\alpha = 0$ to see that

$$SS(0) = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\bar{x}^2. \quad \text{Thus, } \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Divide by $n - 1$ to see that

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right). \tag{2.2}$$

Exercise 2.10. *The following formulas may be useful in aggregating data. Suppose you have data sets collected on two consecutive days with the following summary statistics.*

day	number of observations	mean	standard deviation
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

Now combine the observations of the two days and use this to show that the combined mean

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

and the combined variance

$$s^2 = \frac{1}{n_1 + n_2 - 1} \left((n_1 - 1)s_1^2 + n_1\bar{x}_1^2 + (n_2 - 1)s_2^2 + n_2\bar{x}_2^2 - (n_1 + n_2)\bar{x}^2 \right).$$

*(*Hint*: Use (2.2)).*

Exercise 2.11. *For the data set $x_1, x_2, \ldots, x_n$, let*

$$y_i = a + bx_i.$$

Give the summary statistics for the y data set given the corresponding values of the x data set. (Consider carefully the consequences of the fact that a might be less than 0.)

Among these, the **quadratic identity**

$$\text{var}(x + bx) = b^2\text{var}(x)$$

is one of the most frequently used and useful in all of statistics.

2.3 Quantiles and Standardized Variables

A single observation, say 87 on a exam, gives little information about the performance on the exam. One way to include more about this observation would be to give the value of the empirical cumulative distribution function. Thus,

$$F_n(87) = 0.7223$$

tells us that about 72% of the exam scores were below 87. This is sometimes reported by saying that 87 is the 0.7223 **quantile** for the exam scores.

We can determine this value using the R command `quantile`. For the ages of presidents at inauguration, we have that the 72% quantile is 57 year old.

```
> quantile(age,0.72)
72%
 57
```

Thus, for example, for the ages of the president, we have that `IQR(age)` can also be computed using the command `quantile(age,3/4) - quantile(age,1/4)`. R returns the value 7.

Another, and perhaps more common use of the term quantiles is a general term for partitioning ranked data into equal parts. For example, quartiles partitions the data into 4 equal parts. Percentiles partitions the data into 100 equal parts. Thus, the k-th q-tile is the value in the data for which k/q of the values are below the given value. This naturally leads to some rounding issues which leads to a large variety of small differences in the definition of quantiles.

Exercise 2.12. *For the example above, describe the quintile, decile, and percentile of the observation 87.*

A second way to evaluate a score of 87 is to related it to the mean. Thus, if the mean $\bar{x} = 76$. Then, we might say that the exam score is 11 points above the mean. If the scores are quite spread out, then 11 points above the mean is just a little above average. If the scores are quite tightly spread, then 11 points is quite a bit above average. Thus, for comparisons, we will sometimes use the **standardized version** of x_i,

$$z_i = \frac{x_i - \bar{x}}{s_x}.$$

The observations z_i have mean 0 and standard deviation 1. The value z_i is also called the **standard score** , the z-**value**, the z-**score**, and the **normal score**. An individual z-score, z_i, gives the number of standard deviations an observation x_i is above (or below) the mean.

The R command `scale` transforms the data to the standard score. For the ages of the presidents, we use the scale command to show the standardized ages. The `head` command show the first 6 rows of the output for presidents from George Washington to John Qunicy Adams.

```
> head(data.frame(scale(age),(age-mean(age))/sd(age)))
  scale.age. X.age...mean.age...sd.age.
1  0.3792565                 0.3792565
2  1.0210751                 1.0210751
3  0.3792565                 0.3792565
4  0.3792565                 0.3792565
5  0.5397112                 0.5397112
6  0.3792565                 0.3792565
```

Exercise 2.13. *What are the units of the standard score? What is the relationship of the standard score of an observation x_i and $y_i = ax_i + b$?*

2.4 Quantile-Quantile Plots

In addition to side by side boxplots or histograms, we can also compare two cumulative distribution function directly with the **quantile-quantile** or **Q-Q plot**. If the quantitative data sets `x` and `y` have the same number of observations, then this is simply `plot(sort(x),sort(y))`. In this case the Q-Q plot matches each of the quantiles for the two data sets. If the data sets have an unequal number of observations, then observations from the larger data are reduced by interpolation to create data sets of equal length and the Q-Q plot is `plot(sort(xred),sort(yred))` for the reduced data sets `xred` and `yred`.

Example 2.14. *Dravet syndrome, also known as Severe Myoclonic Epilepsy of Infancy (SMEI), is a rare and catastrophic form of intractable epilepsy that begins in infancy. A recent study looks at* de novo *mutations in the DNA sequence SCN1A that codes for a sodium channel protein. An improperly functioning sodium channel can have severe consequences for brain function.*

The two basic types of mutations under study are point mutations, called missense *and* nonsense *mutations. A missense mutation results in a change in an amino acid in the SCN1A protein, whereas a nonsense mutation results in a truncated, incomplete, and usually nonfunctional protein segment that is degraded.*

Here is the age of first seizure in a study of 264 Japanese children. Age is in months.

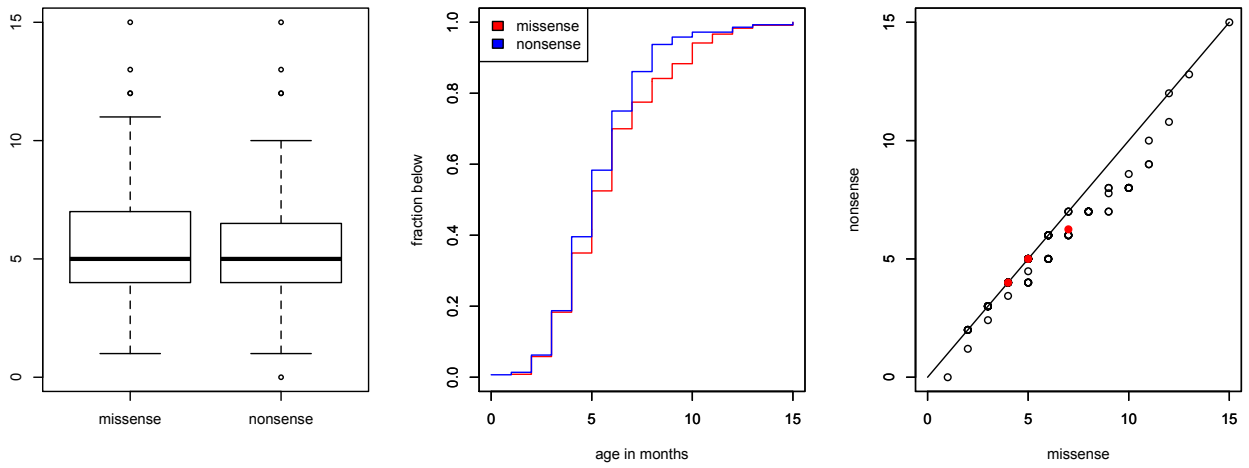


Figure 2.2: age of first seizure (**left**) side-by-side boxplots, (**center**) empirical cumulative distribution functions, (**right**) Q-Q plot with Q_1, the median, and Q_3 indicated by the solid red dots. The solid line on the plot has intercept 0 and slope 1. (missense age=nonsense age)

age	7	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
missense	0	1	6	15	20	21	21	9	8	5	7	3	2	1	0	1
nonsense	1	1	7	18	30	27	24	16	11	3	2	0	2	1	0	1

We enter the data into R *using* `rep`*, the repeat command and prepare a summary.*

```
> missense<-c(1,rep(2,6),rep(3,15),rep(4,20),rep(5,21),rep(6,21),rep(7,9),rep(8,8),
  rep(9,5),rep(10,7),11,11,11,12,12,13,15)
> nonsense<-c(0,1,rep(2,7),rep(3,18),rep(4,30),rep(5,27),rep(6,24),rep(7,16),
  rep(8,11),rep(9,3),10,10,12,12,13,15)
> summary(missense)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     4.0     5.0     5.8     7.0    15.0
> summary(nonsense)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   4.000   5.000   5.326   6.250  15.000
```

The side-by-side boxplots, empirical cumulative distribution functions, and the Q-Q plot . The R *code is below.*

```
> par(mfrow=c(1,3))
> boxplot(missense,nonsense,names=c("missense","nonsense"))
> plot(sort(missense),1:length(missense)/length(missense),type="s",
   xlab=c("age in months"),ylab=c("fraction below"),xlim=c(0,15),ylim=c(0,1),col="red")
> par(new=TRUE)
> plot(sort(nonsense),1:length(nonsense)/length(nonsense),type="s",
   xlab=c(""), ylab=c(""),xlim=c(0,15),ylim=c(0,1),col="blue")
> legend("topleft",c("missense","nonsense"),fill=c("red","blue"))
> qqplot(missense,nonsense,xlim=c(0,15),ylim=c(0,15))
> par(new=TRUE)
> plot(c(0,15),c(0,15),type="l",xlab=c(""), ylab=c(""),xlim=c(0,15),ylim=c(0,15))
```

The points on the Q-Q plot indicate values having equal quantiles for male and female scores. To see the first and third quartiles, Q_1 and Q_3 as well as the median, we use the `points` and the `quantile` commands.

```
> q<-c(0.25,0.50,0.75)
> points(quantile(missense,q),quantile(nonsense,q),col="red",pch=19)
```

The `points` command was used to add the three solid red dots. Moving from lower left to upper right, dots coordinates that are the the values for Q_1, $(4, 4)$, the median, $(5, 5)$, and Q_3, $(7, 6.25)$, for the two data sets.

Exercise 2.15. *The mean time of first seizure is slightly higher for patients with a missense mutation. Explain how this can be seen for each of the plots that compare the two data sets.*

2.5 Answers to Selected Exercises

2.3. Check the formula

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}x_{n+1}.$$

For k additional observations, write

$$\bar{x}_{n,k+n} = \frac{1}{k}(x_{n+1} + \cdots + x_{n+k}).$$

Then the mean of the $n + k$ observations is

$$\bar{x}_{n+k} = \frac{n}{n+k}\bar{x}_n + \frac{k}{n+k}\bar{x}_{n,k+n}.$$

2.4 (a) If the distribution is skewed left, then the mean follows the tail and is less than the median. (b) For a symmetric distribution, the mean and the median are equal. (c) If the distribution is skewed right, then the mean is greater than the median.

2.5. The boxplot has 5 outliers. Three are above $Q_3 + IQR$ and two are below $Q_3 - IQR$.

2.8. Divide the sum into 2 terms.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}\right) = 0.$$

2.9. This can be obtained by flipping coins. In R, we shall learn that the command to simulate this is `rbinom(20,16,0.5)`. Here are the data. Focusing on the first three columns, we see a total of 166 heads in the 20 observations. Thus, $\bar{x} = 8.3$.

heads	counts	heats×counts	counts×(heads-8)^2	counts×(heads-$\bar{x}$)^2
4	1	4	$1 \cdot (4-8)^2 = 16$	$1 \cdot (4-8.3)^2 = 18.49$
5	1	5	$1 \cdot (5-8)^2 = 9$	$1 \cdot (5-8.3)^2 = 10.89$
6	2	12	$2 \cdot (6-8)^2 = 8$	$2 \cdot (6-8.3)^2 = 10.58$
7	2	14	$2 \cdot (7-8)^2 = 2$	$2 \cdot (7-8.3)^2 = 5.07$
8	5	40	$5 \cdot (8-8)^2 = 0$	$5 \cdot (8-8.3)^2 = 0.45$
9	3	27	$3 \cdot (9-8)^2 = 3$	$3 \cdot (9-8.3)^2 = 1.47$
10	3	30	$3 \cdot (10-8)^2 = 12$	$3 \cdot (10-8.3)^2 = 8.67$
11	2	22	$2 \cdot (11-8)^2 = 18$	$2 \cdot (11-8.3)^2 = 14.58$
12	1	12	$1 \cdot (12-8)^2 = 16$	$1 \cdot (12-8.3)^2 = 13.69$
sum	20	166	$SS(8) = 84$	$SS(\alpha) = 82.2$

Notice that $SS(8) > SS(\bar{x})$.

2.10. Let $x_{1,1}, x_{1,2} \ldots, x_{1,n_1}$ denote the observed values on day 1 and $x_{2,1}, x_{2,2} \ldots, x_{2,n_2}$ denote the observed values on day 2. The mean of the combined data

$$\bar{x} = \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{n_1} x_{1,i} + \sum_{i=1}^{n_2} x_{2,i} \right) = \frac{1}{n_1 + n_2} \left(n_1 \bar{x}_1 + n_2 \bar{x}_2 \right)$$

Using (2.2), we find that

$$s^2 = \frac{1}{n_1 + n_2 - 1} \left(\sum_{i=1}^{n_1} x_{1,i}^2 + \sum_{i=1}^{n_2} x_{2,i}^2 - (n_1 + n_2) \bar{x}^2 \right).$$

Use (2.2) twice more to see that

$$\sum_{i=1}^{n_1} x_{1,i}^2 = (n_1 - 1) s_1^2 + n_1 \bar{x}_1^2 \quad \text{and} \quad \sum_{i=1}^{n_2} x_{2,i}^2 = (n_2 - 1) s_2^2 + n_2 \bar{x}_2^2$$

Now substitute the sums in the line above into the equation for s^2.

2.11.

statistic	
median	If m_x is the median for the x observations, then $a + bm_x$ is the median of the y observations.
mean	$\bar{y} = a + b\bar{x}$
variance	$\text{var}(y) = b^2 \text{var}(x)$
standard deviation	$s_y = \lvert b \rvert s_x$
first quartile	If Q_1 is the first quartile of the x observations and if $b > 0$, then $a + bQ_1$ is the first quartile of the y observations. If $b < 0$, then $a + bQ_3$ is the first quartile of the y observations.
third quartile	If Q_3 is the third quartile of the x observations and if $b > 0$, then $a + bQ_3$ is the third quartile of the y observations. If $b < 0$, then $a = BQ_1$ is the third quartile of the y observations.
interquartile range	$IQR(y) = \lvert b \rvert IQR(x)$.

To verify the quadratic identity for the variance:

$$\text{var}(y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^{n} ((a + bx_i) - (a + b\bar{x}))^2 = \frac{1}{n-1} \sum_{i=1}^{n} (b(x_i - \bar{x}))^2 = b^2 \text{var}(x).$$

2.12.

$$S(\alpha) = \sum_{i=1}^{n} (x_i - \alpha)^2. \quad \text{Thus,} \quad S'(\alpha) = -2 \sum_{i=1}^{n} (x_i - \alpha)$$

and $S'(\bar{x}) = 0$. Next, $S''(\alpha) = 2n$ for all α and thus $S''(\bar{x}) = 2n > 0$. Consequently, $\bar{x}$ is a minimum.

2.13. 87 is between the 3-rd and the 4-th quintile, between the 7-th and the 8-th decile and the 72-nd and 73-rd percentile.

2.14. Both the numerator and the denominator of the z-score have the same units. Their ratio is thus unitless. The standard score for y,

$$z_i^y = \frac{y_i - \bar{y}}{s_y} = \frac{(ax_i + b) - (a\bar{x} + b)}{\lvert a \rvert s_x} = \frac{a(x_i - \bar{x})}{\lvert a \rvert s_x} = \frac{a}{\lvert a \rvert} z_i^x.$$

Thus, if $a > 0$, the two standard scores are the same. If $a < 0$, the two standard scores are the negative of one another.

2.15. (**left**) The boxplots agree for Q_1 and the median, but the missense children show a higher skew to the right, giving a higher mean.

(**center**) Recall that the area under the survival function and thus the area above the empirical cumulative distribution function is the mean. The cumulative distribution function rises more quickly for the nonsense distribution. Thus, the area above it is smaller and so the mean is smaller.

(**right**) The points are on or below the line missense age = nonsense age. Thus, for a given quantile, the missense value is at least as big ans the nonsense and thus the mean is bigger for the age of first seizure for missense children.

Topic 3

Correlation and Regression

Reflection soon made it clear to me that not only were the two new problems identical in principle with the old one of kinship which I had already solved, but that all three of them were no more than special cases of a much more general problem–namely, that of Correlation. - Sir Francis Galton

My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission ... that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. - Sir Francis Galton

In this section, we shall take a careful look at the nature of **linear relationships** found in the data used to construct a scatterplot. The first of these, **correlation**, examines this relationship in a symmetric manner. The second, **regression**, considers the relationship of a **response variable** as determined by one or more **explanatory variables**. Correlation focuses primarily on association, while regression is designed to help make predictions. Consequently, the first does not attempt to establish any cause and effect. The second is a often used as a tool to establish causality.

3.1 Covariance and Correlation

The **covariance** measures the linear relationship between a pair of quantitative measures

$$x_1, x_2, \ldots, x_n \quad \text{and} \quad y_1, y_2, \ldots, y_n$$

on the same sample of n individuals. Beginning with the definition of variance, the definition of covariance is similar to the relationship between the square of the norm $||v||^2$ of a vector v and the inner product $\langle v, w\rangle$ of two vectors v and w.

$$\text{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

vectors	**quantitative observations**
$\mathbf{v} = (v_1, \ldots, v_n)$ $\mathbf{w} = (w_1, \ldots, w_n)$	$\mathbf{x} = (x_1, \ldots, x_n)$ $\mathbf{y} = (y_1, \ldots, y_n)$
norm-squared $\|\|\mathbf{v}\|\|^2 = \sum_{i=1}^{n} v_i^2$	variance $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
norm $\|\|\mathbf{v}\|\|$	standard deviation s_x
inner product $\langle \mathbf{v}, \mathbf{w}\rangle = \sum_{i=1}^{n} v_i w_i$	covariance $\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
cosine $\cos\theta = \frac{\langle \mathbf{v}, \mathbf{w}\rangle}{\|\|\mathbf{v}\|\|\,\|\|\mathbf{w}\|\|}$	correlation $r = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{s_x s_y}$

Table I: Analogies between vectors and quantitative observations.

A positive covariance means that the terms $(x_i - \bar{x})(y_i - \bar{y})$ in the sum are more likely to be positive than negative. This occurs whenever the x and y variables are more often both above or below the mean in tandem than not. Just like the situation in which the inner product of a vector with itself yields the square of the norm, the covariance of x with itself $\text{cov}(x, x) = s_x^2$ is the variance of x.

Exercise 3.1. *Explain in words what a negative covariance signifies, and what a covariance near 0 signifies.*

We next look at several exercises that call for algebraic manipulations of the formula for covariance or closely related functions.

Exercise 3.2. *Derive the alternative expression for the covariance:*

$$\text{cov}(x, y) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right).$$

Exercise 3.3. $\text{cov}(ax+b, cy+d) = ac \cdot \text{cov}(x, y)$. *How does a change in units (say from centimeters to meters) affect the covariance?*

Thus, covariance as a measure of association has the drawback that its value depends on the units of measurement. This shortcoming is remedied by using the correlation.

Definition 3.4. *The* **correlation**, r, *is the covariance of the standardized versions of* x *and* y.

$$r(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\text{cov}(x, y)}{s_x s_y}.$$

The observations x *and* y *are called* **uncorrelated** *if* $r(x, y) = 0$.

Exercise 3.5. $r(ax + b, cy + d) = \pm r(x, y)$. *How does a change in units (say from centimeters to meters) affect the correlation? The plus sign occurs if* $a \cdot c > 0$ *and the minus sign occurs if* $a \cdot c < 0$.

Sometimes we will drop (x, y) if there is no ambiguity and simply write r for the correlation.

Exercise 3.6. *Show that*

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2\text{cov}(x, y) = s_x^2 + s_y^2 + 2r s_x s_y. \tag{3.1}$$

Give the analogy between this formula and the law of cosines.

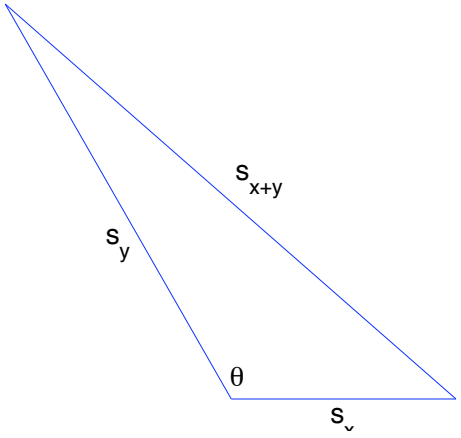


Figure 3.1: The analogy of the sample standard deviations and the law of cosines in equation (3.1). Here, the corrrelation $r = -\cos\theta$.

In particular if the two observations are uncorrelated we have the **Pythagorean identity**

$$s_{x+y}^2 = s_x^2 + s_y^2. \tag{3.2}$$

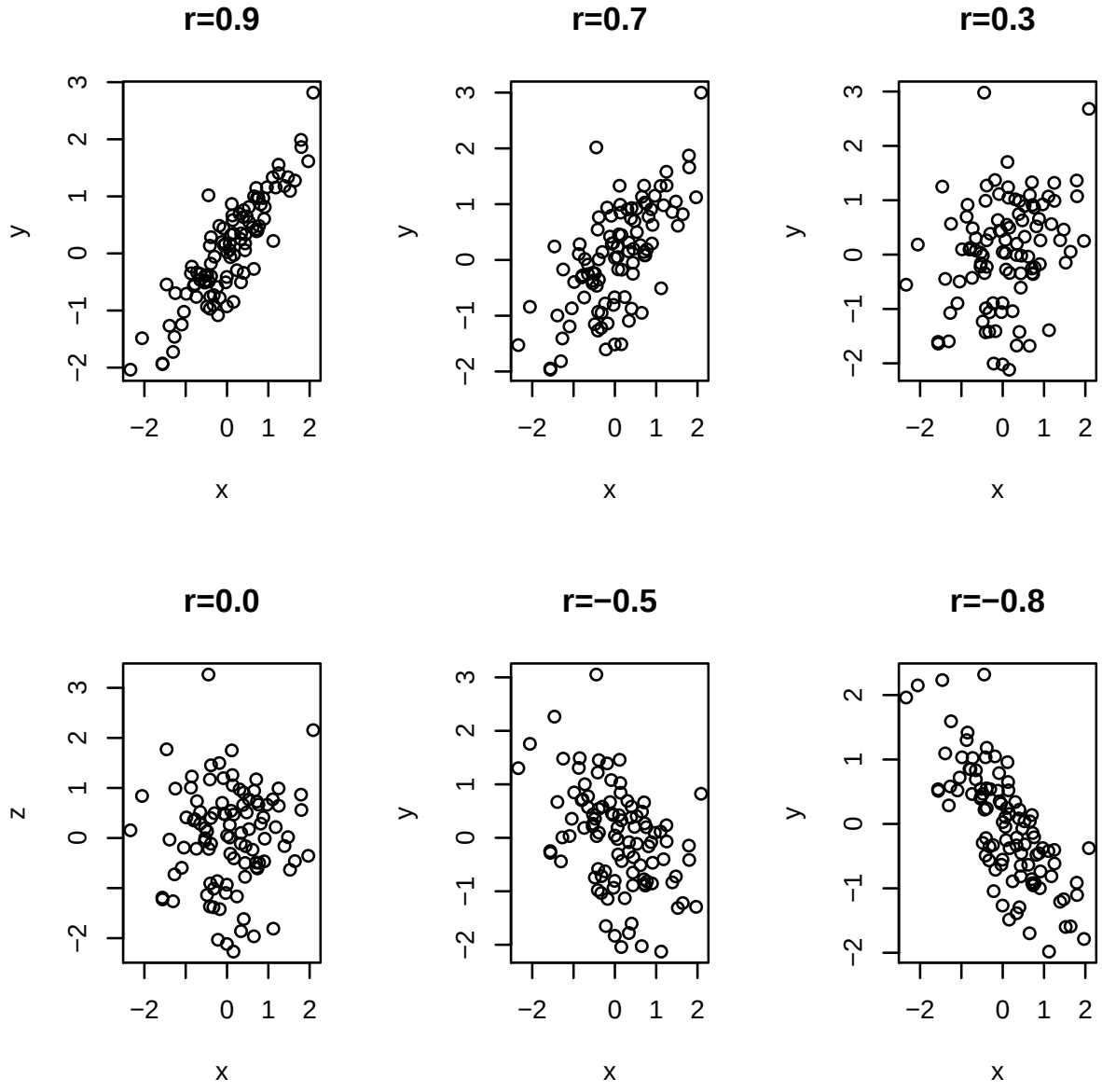
We will now look to uncover some of the properties of correlation. The next steps are to show that the correlation is always a number between -1 and 1 and to determine the relationship between the two variables in the case that the correlation takes on one of the two possible extreme values.

Exercise 3.7 (Cauchy-Schwarz inequality)**.** *For two sequences* $v_1, \cdots, v_n$ *and* $w_1, \ldots, w_n$, *show that*

$$\left(\sum_{i=1}^{n} v_i w_i\right)^2 \leq \left(\sum_{i=1}^{n} v_i^2\right)\left(\sum_{i=1}^{n} w_i^2\right). \tag{3.3}$$

Written in terms of norms and inner products, the Cauchy-Schwarz inequality becomes $\langle v, w\rangle^2 \leq ||v||^2||w||^2$.

(Hint: $\sum_{i=1}^{n}(v_i + w_i\zeta)^2$ *is a non-negative quadratic expression in the variable* ζ *and consider the discriminant in the quadratic formula.) If the discriminant is zero, then we have equality in (3.3) and we have that* $\sum_{i=1}^{n}(v_i + w_i\zeta)^2 = 0$ *for exactly one value of* ζ.



[t!]

Figure 3.2: Scatterplots showing differing levels of the correlation r

We shall use inequality (3.3) by choosing $v_i = x_i - \bar{x}$ and $w_i = y_i - \bar{y}$ to obtain

$$\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2 \le \left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right),$$

$$\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2 \le \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\right),$$

$$\mathrm{cov}(x,y)^2 \le s_x^2 s_y^2$$

$$\frac{\mathrm{cov}(x,y)^2}{s_x^2 s_y^2} \le 1$$

Consequently, we find that

$$r^2 \le 1 \quad \text{or} \quad -1 \le r \le 1.$$

When we have $|r| = 1$, then we have equality in (3.3). In addition, for some value of ζ we have that

$$\sum_{i=1}^{n}((x_i-\bar{x}) + (y_i-\bar{y})\zeta)^2 = 0.$$

The only way for a sum of nonnegative terms to add to give zero is for each term in the sum to be zero, i.e.,

$$(x_i-\bar{x}) + (y_i-\bar{y})\zeta = 0, \quad \text{for all } i = 1,\dots,n. \tag{3.4}$$

Thus x_i and y_i are linearly related.

$$y_i = \alpha + \beta x_i.$$

In this case, the sign of r is the same as the sign of β.

Exercise 3.8. *For an alternative derivation that $-1 \le r \le 1$. Use equation (3.1) with x and y standardized observations. Use this to determine ζ in equation (3.4) (*Hint: *Consider the separate cases s^2_{x+y} for the $r = -1$ and s^2_{x-y} for the $r = 1$.)*

We can see how this looks for simulated data. Choose a value for `r` between -1 and $+1$.

```
>x<-rnorm(100)
>z<-rnorm(100)
>y<-r*x + sqrt(1-r^2)*z
```

Example of plots of the output of this simulation are given in Figure 3.1. For the moment, the object of this simulation is to obtain an intuitive feeling for differing values for correlation. We shall soon see that this is the simulation of pairs of normal random variables with the desired correlation. From the discussion above, we can see that the scatterplot would lie on a straight line for the values $r = \pm 1$.

For the *Archeopteryx* data on bone lengths, we have the correlation

```
> cor(femur, humerus)
[1] 0.9941486
```

Thus, the data land very nearly on a line with positive slope.

For the banks in 1974, we have the correlation

```
> cor(income,assets)
[1] 0.9325191
```

3.2 Linear Regression

Covariance and correlation are measures of linear association. For the *Archeopteryx* measurements, we learn that the relationship in the length of the femur and the humerus is very nearly linear.

We now turn to situations in which the value of the first variable x_i will be considered to be **explanatory** or **predictive**. The corresponding observation y_i, taken from the input x_i, is called the **response**. For example, can we **explain** or **predict** the income of banks from its assets? In this case, *assets* is the explanatory variable and income is the response.

In **linear regression**, the response variable is linearly related to the explanatory variable (also known as the **covariate**), but is subject to deviation, discrepency, or to error. We write

$$y_i = \alpha + \beta x_i + \epsilon_i. \tag{3.5}$$

Our goal is, given the data, the x_i's and y_i's, to find α and β that determines the line having the best fit to the data. The principle of **least squares regression** states that the best choice of this linear relationship is the one that minimizes the square in the *vertical distance* from the y values in the data and the y values on the regression line. This choice reflects the fact that the values of x are set by the experimenter and are thus assumed known. Thus, the "error" appears in the value of the response variable y.

This principle leads to a minimization problem for

$$SS(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2. \tag{3.6}$$

In other words, given the data, determine the values for α and β that minimizes the **sum of squares** SS. Let's the denote by $\hat{\alpha}$ and $\hat{\beta}$ the value for α and β that minimize SS.

Take the partial derivative with respect to α.

$$\frac{\partial}{\partial \alpha} SS(\alpha, \beta) = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)$$

At the values $\hat{\alpha}$ and $\hat{\beta}$, this partial derivative is 0. Consequently

$$0 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (\hat{\alpha} - \hat{\beta} x_i).$$

Now, divide by both sides of the equation by n to obtain

$$\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}. \tag{3.7}$$

Thus, we see that the **center of mass point** $(\bar{x}, \bar{y})$ is on the regression line. To emphasize this fact, we rewrite (3.5) in **slope-point** form.

$$y_i - \bar{y} = \beta (x_i - \bar{x}) + \epsilon_i. \tag{3.8}$$

We then apply this to the sums of squares criterion (3.6) to obtain a condition that depends on β,

$$\tilde{SS}(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} ((y_i - \bar{y}) - \beta (x_i - \bar{x}))^2. \tag{3.9}$$

Now, differentiate with respect to β and set this equation to zero for the value $\hat{\beta}$.

$$\frac{d}{d\beta} \tilde{SS}(\beta) = -2 \sum_{i=1}^{n} ((y_i - \bar{y}) - \hat{\beta} (x_i - \bar{x}))(x_i - \bar{x}) = 0.$$

Thus,

$$\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) = \hat{\beta}\sum_{i=1}^{n}(x_i-\bar{x})^2$$
$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) = \hat{\beta}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$
$$\text{cov}(x,y) = \hat{\beta}\text{var}(x)$$

Now solve for $\hat{\beta}$.

$$\hat{\beta} = \frac{\text{cov}(x,y)}{\text{var}(x)}. \tag{3.10}$$

In summary, to determine the **regression line**.

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i,$$

we use (3.10) to determine $\hat{\beta}$ and then (3.7) to solve for

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

We call $\hat{y}_i$ the **fit** for the value x_i.

Example 3.9. *Let's begin with 6 points and derive by hand the equation for regression line.*

x	-2	-1	0	1	2	3
y	-3	-1	-2	0	4	2

Add the x and y values and divide by $n = 6$ to see that $\bar{x} = 0.5$ and $\bar{y} = 0$.

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \hat{y})$	$(x_i - \bar{x})^2$
-2	-3	-2.5	-3	7.5	6.25
-1	-1	-1.5	-1	1.5	2.25
0	-2	-0.5	-2	1.0	0.25
1	0	0.5	0	0.0	0.25
2	4	1.5	4	6.0	2.25
3	2	2.5	2	5.0	6.25
sum		0	0	$\text{cov}(x,y) = 21/5$	$\text{var}(x) = 17.50/5$

Thus,

$$\hat{\beta} = \frac{21/5}{17.5/5} = 1.2 \quad \textit{and} \hat{\alpha} = 0 - 1.2 \times 0.5 = -0.6$$

As seen in this example, fits, however rarely perfect. The difference between the fit and the data is an estimate $\hat{\epsilon}_i$ for the error ϵ_i. This difference is called the **residual**. So,

$$\text{RESIDUAL}_i = \text{DATA}_i - \text{FIT}_i = y_i - \hat{y}_i$$

or, by rearranging terms,

$$\text{DATA}_i = \text{FIT}_i + \text{RESIDUAL}_i, \quad \textit{or} \quad y_i = \hat{y}_i + \hat{\epsilon}_i.$$

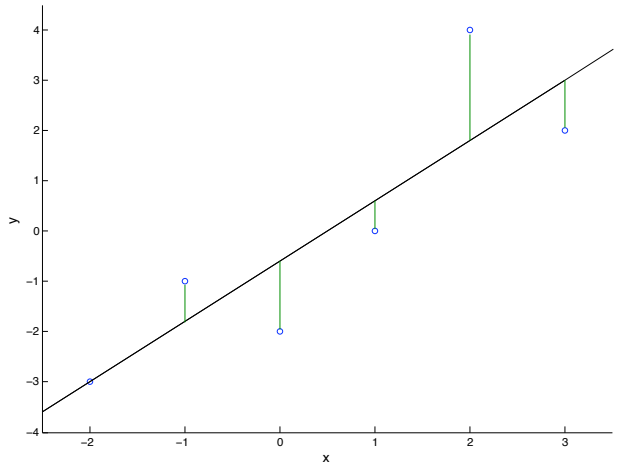
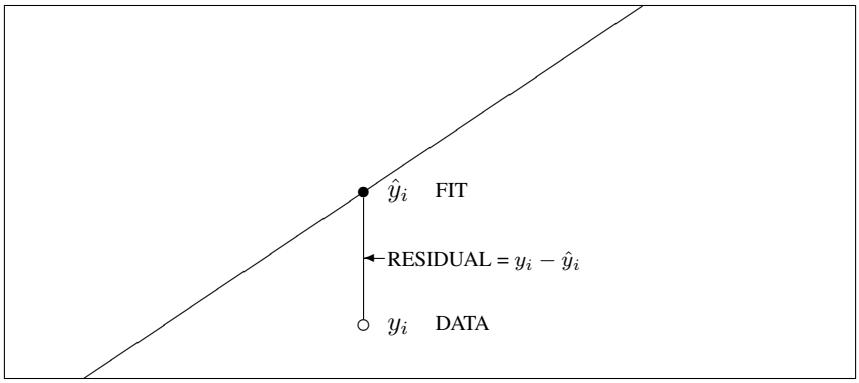


Figure 3.3: Scatterplot and the regression line for the six point data set below. The regression line is the choice that minimizes the square of the vertical distances from the observation values to the line, indicated here in green. Notice that the total length of the positive residuals (the lengths of the green line segments above the regression line) is equal to the total length of the negative residuals. This property is derived in equation (3.11).



We can rewrite equation (3.6) with $\hat{\epsilon}_i$ estimating the error in (3.5).

$$0 = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}\hat{\epsilon}_i \tag{3.11}$$

to see that the sum of the residuals is 0. Thus, we started with a criterion for a line of best fit, namely, least squares, and discover that a consequence of this criterion the regression line has the property that the sum of the residual values is 0. This is illustrated in Figure 3.3.

Let's check this property for the example above.

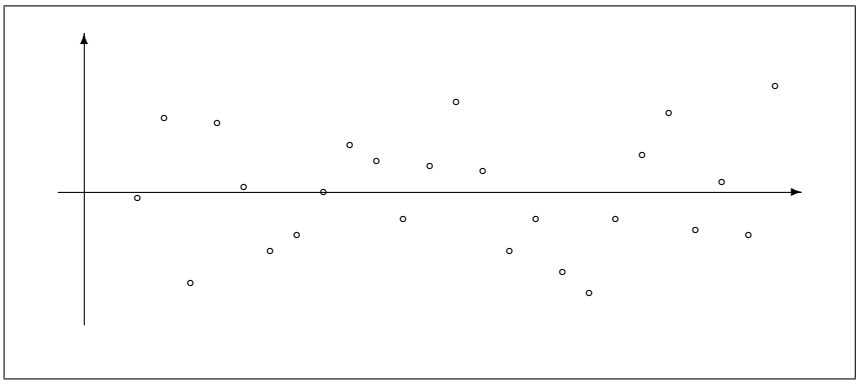
	DATA	FIT	RESIDUAL
x_i	y_i	$\hat{y}_i$	$\hat{y}_i - y_i$
-2	-3	-3.0	0
-1	-1	-1.8	0.8
0	-2	-0.6	-1.4
1	0	0.6	-0.6
2	4	1.8	2.2
3	2	3.0	-1.0
	total		0

Generally speaking, we will look at a **residual plot**, the plot of the residuals versus the explanatory variable, to assess the appropriateness of a regression line. Specifically, we will look for circumstances in which the explanatory variable and the residuals have no systematic pattern.

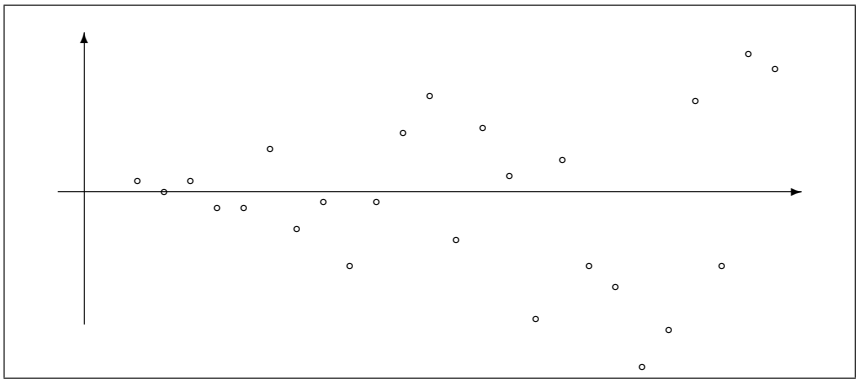
Exercise 3.10. *Use* R *to perform the following operations on the data set in Example 3.9.*

1. *Enter the data and make a scatterplot.*
2. *Use the* `lm` *command to find the equation of the regression line.*
3. *Use the* `abline` *command to draw the regression line on the scatterplot.*
4. *Use the* `resid` *and the* `predict` *command command to find the residuals and place them in a* `data.frame` *with* `x` *and* `y`
5. *Draw the residual plot and use* `abline` *to add the horizontal line at 0.*

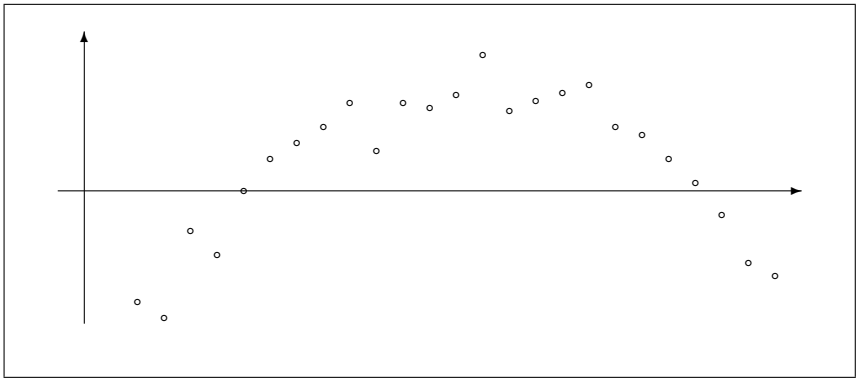
We next show three examples of the residuals plotting against the value of the explanatory variable.



Regression fits the data well - homoscedasticity



Prediction is less accurate for large x, an example of heteroscedasticity



Data has a curve. A straight line fits the data poorly.

For any value of x, we can use the regression line to estimate or **predict** a value for y. We must be careful in using this prediction outside the range of x. This **extrapolation** will not be valid if the relationship between x and y is not known to be linear in this extended region.

Example 3.11. *For the 1974 bank data set, the regression line*

$$\widehat{income} = 7.680 + 4.975 \cdot assets.$$

So, each dollar in assets brings in about $5 income.

For a bank having 10 billion dollars in assets, the predicted income is 56.430 billion dollars. However, if we extrapolate this down to very small banks, we would predict nonsensically that a bank with no *assets would have an income of 7.68 billion dollars. This illustrates the caution necessary to perform a reliable prediction through an extrapolation.*

In addition for this data set, we see that three banks have assets much greater than the others. Thus, we should consider examining the regression lines omitting the information from these three banks. If a small number of observations has a large impact on our results, we call these points **influential**.

Obtaining the regression line in R *is straightforward:*

```
> lm(income~assets)

Call:
lm(formula = income ~ assets)

Coefficients:
(Intercept)       assets
      7.680        4.975
```

Example 3.12 (regression line in standardized coordinates)**.** *Sir Francis Galton was the first to use the term* **regression** *in his study* Regression towards mediocrity in hereditary stature. *The rationale for this term and the relationship between regression and correlation can be best seen if we convert the observations into a standardized form.*

First, write the regression line to point-slope form.

$$\hat{y}_i - \bar{y} = \hat{\beta}(x_i - \bar{x}).$$

Because the slope

$$\hat{\beta} = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{r s_x s_y}{s_x^2} = \frac{r s_y}{s_x},$$

we can rewrite the point-slope form as

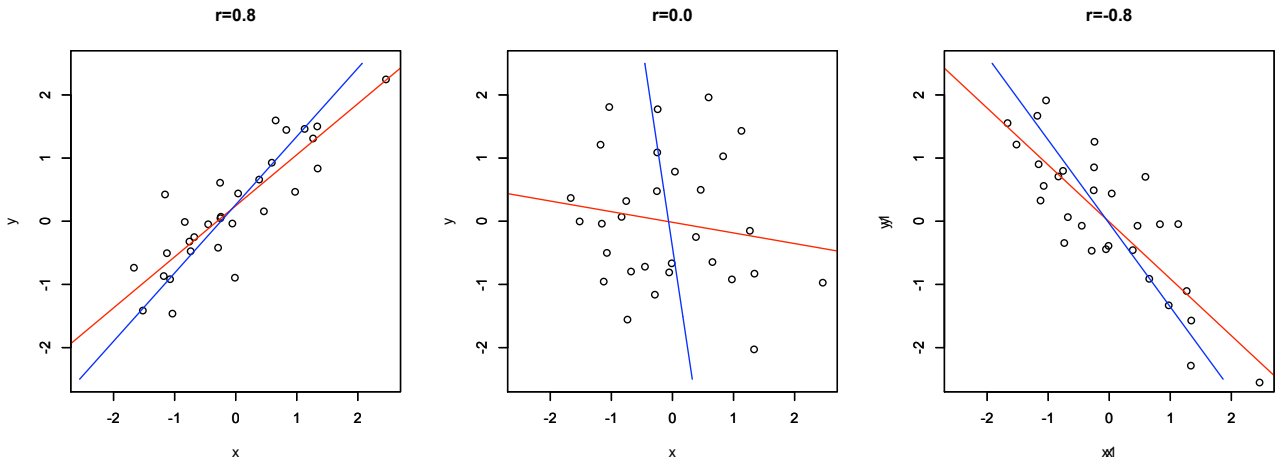


Figure 3.4: Scatterplots of standardized variables and their regression lines. The red lines show the case in which x is the explanatory variable and the blue lines show the case in which y is the explanatory variable.

$$\hat{y}_i - \bar{y} = \frac{rs_y}{s_x}(x_i - \bar{x}) \quad or \quad \frac{\hat{y}_i - \bar{y}}{s_y} = r\frac{x_i - \bar{x}}{s_x}, \quad \hat{y}_i^* = rx_i^*. \tag{3.12}$$

where the asterisk is used to indicate that we are stating our observations in standardized form. In words, if we use this standardized form, then the slope of the regression line is the correlation.

For Galton's example, let's use the height of a male as the explanatory variable and the height of his adult son as the response. If we observe a correlation $r = 0.6$ and consider a man whose height is 1 standard deviation above the mean, then we predict *that the son's height is 0.6 standard deviations above the mean. If a man whose height is 0.5 standard deviation below the mean, then we* predict *that the son's height is 0.3 standard deviations below the mean. In either case, our prediction for the son is a height that is closer to the mean then the father's height. This is the "regression" that Galton had in mind.*

From the discussion above, we can see that if we reverse the role of the explanatory and response variable, then we change the regression line. This should be intuitively obvious since in the first case, we are minimizing the total square vertical distance and in the second, we are minimizing the total square horizontal distance. In the most extreme circumstance, $\text{cov}(x, y) = 0$. In this case, the value x_i of an observation is no help in predicting the response variable. Thus, as the formula states, when x is the explanatory variable the regression line has slope 0 - it is a horizontal line through $\bar{y}$. Correspondingly, when y is the explanatory variable, the regression is a vertical line through $\bar{x}$. Intuitively, if x and y are uncorrelated, then the best prediction we can make for y_i given the value of x_i is just the sample mean $\bar{y}$ and the best prediction we can make for x_i given the value of y_i is the sample mean $\bar{x}$.

More formally, the two regression equations are

$$\hat{y}_i^* = rx_i^* \quad \text{and} \quad \hat{x}_i^* = ry_i^*.$$

These equations have slopes r and $1/r$. This is shown by example in Figure 3.2.

Exercise 3.13. *Compute the regression line for the 6 pairs of observations above assuming that y is the explanatory variable. Show that the two region lines differ by showing that the product of the slopes in not equal to one.*

Exercise 3.14. *Continuing the previous example, let $\hat{\beta}_x$ be the slope of the regression line obtained from regressing y on x and $\hat{\beta}_y$ be the slope of the regression line obtained from regressing x on y. Show that the product of the slopes $\hat{\beta}_x\hat{\beta}_y = r^2$, the square of the correlation.*

Because the point $(\bar{x}, \bar{y})$ is on the regression line, we see from the exercise above that two regression lines coincide precisely when the slopes are reciprocals, namely precisely when $r^2 = 1$. This occurs for the values $r = 1$ and $r = -1$.

Exercise 3.15. *Show that the FIT, $\hat{y}$, and the RESIDUALS, $y - \hat{y}$ are uncorrelated.*

Let's again write the regression line in point slope form

$$\text{FIT}_i - \bar{y} = \hat{y}_i - \bar{y} = r\frac{s_y}{s_x}(x_i - \bar{x}).$$

Using the quadratic identity for variance we find that

$$s^2_{FIT} = r^2\frac{s^2_y}{s^2_x}s^2_x = r^2 s^2_y = r^2 s^2_{DATA}.$$

Thus, the variance in the FIT is reduced from the variance in the DATA by a factor of r^2 and

$$r^2 = \frac{s^2_{FIT}}{s^2_{DATA}}.$$

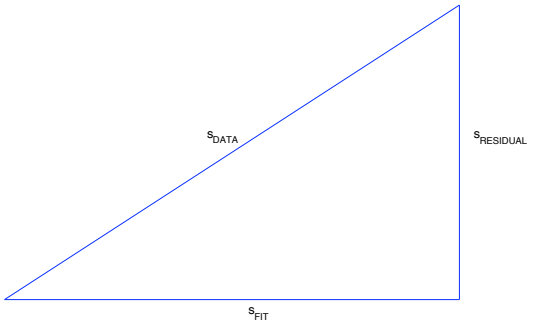


Figure 3.5: The relationship of the standard deviations of the DATA, the FIT, and the RESIDUALS. $s^2_{DATA} = s^2_{FIT} + s^2_{RESID}$. We call r^2 the **coefficient of determination** and say that r^2 of the variation in the response variable is due to the fit and the rest $1 - r^2$ is due to the residuals.

Exercise 3.16. *Use the equation above to show that*

$$r^2 = \frac{SS_{FIT}}{SS_{DATA}}$$

where the **sums of squares of the fit** *or the* **variation of the fit**, $SS_{FIT} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ *,and the* **sums of squares of the data** *or the* **variation of the data**, $SS_{DATA} = \sum_{i=1}^n (y_i - \bar{y})^2$.

Because the fit and the residuals are uncorrelated, the Pythagorean identity (3.2) applies and we see that that

$$s^2_{DATA} = s^2_{FIT} + s^2_{RESIDUAL} = r^2 s^2_{DATA} + s^2_{RESIDUAL}$$

$$s^2_{RESIDUAL} = (1 - r^2)s^2_{DATA}.$$

This leads to the expression

$$r^2 = 1 - \frac{s^2_{RESIDUAL}}{s^2_{DATA}} = 1 - \frac{SS_{RESIDUAL}}{SS_{DATA}}$$

where the **sums of squares of the residuals** or the **variation of the residuals**, $SS_{RESIDUAL} = \sum_{i=1}^n (y_i - \hat{y})^2$.

Thus, r^2 of the variance in the data can be explained by the fit. As a consequence of this computation, many statistical software tools report r^2 as a part of the linear regression analysis. For this reason, r^2 is sometimes called the **coefficient of determination**. The remaining $1 - r^2$ of the variance in the data is found in the residuals. We saw a similar partitioning of the variation in *Topic 2: Describing Distributions with Numbers* when we first introduced the concept of variance. We shall see it again in *Topic 22: Analysis of Variance*.

Exercise 3.17. *For some situations, the circumstances dictate that the line contain the origin ($\alpha = 0$). Use a least squares criterion to show that the slope of the regression line*

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

R accommodates this circumstance with the commands `lm(y~x-1)` or `lm(y~0+x)`. Note that in this case, the sum of the residuals is not necessarily equal to zero. For least squares regression, this property followed from $\partial SS(\alpha, \beta)/\partial\alpha = 0$ where α is the y-intercept.

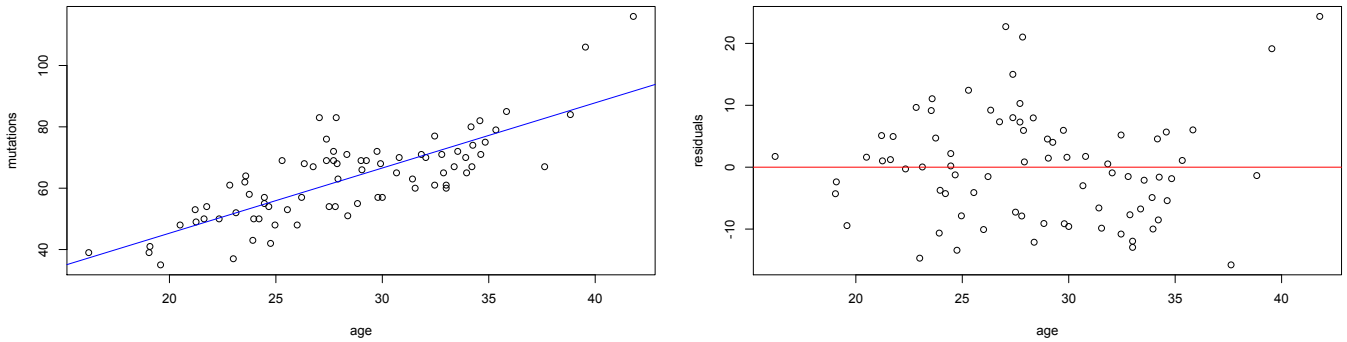


Figure 3.6: (left) scatterplot with regression line in blue (right) residual plot with horizontal axis in red

Example 3.18. *We continue to investigate the relationship of age of parents to the* de novo *mutations in the offspring for the 78 Icelandic trios. We use the average age of the parents to* predict *the number of mutations in the offspring. Thus, age is on the horizontal axis. We can quickly obtain the regression line using* R.

```
> mutations.lm<- lm(mutations~age)
> summary(mutations.lm)
Call:
lm(formula = mutations ~ age)
Residuals:
     Min       1Q   Median       3Q      Max
-15.7849  -7.1364  -0.1244   5.1745  24.3591
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8145     5.5034   0.511    0.611
age           2.1255     0.1904  11.164   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 8.79 on 76 degrees of freedom
Multiple R-squared:  0.6212,Adjusted R-squared:  0.6162
F-statistic: 124.6 on 1 and 76 DF,  p-value: < 2.2e-16
```

Thus, the regression line has the equation.

$$\widehat{mutations} = 2.815 + 2.125\, age.$$

The value for the coefficient of determination, r^2, is 0.6212, the variance in the data explained by the fit.

Next, we plot the data and add the regression line to the plot.

```
> mutations.lm<-lm(mutations~age)
> plot(age,mutation)
> abline(mutations.lm,col="blue")
```

We continue our analysis in calling for the residuals, making a residual plot, and creating a horizontal line at 0.

```
> residuals<-resid(mutations.lm)
> plot(age,residuals)
> abline(h=0,col="red")
```

Finally, we can use the regression line to predict the number of mutations for parents whose average age is 20, 30, or 40.

```
> agepredict<-c(20,30,40)
> predictions<-predict(mutations.lm,newdata=data.frame(age=agepredict))
> data.frame(agepredict,predictions)
  agepredict predictions
1         20    45.32367
2         30    66.57824
3         40    87.83281
```

3.2.1 Transformed Variables

For pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$, the linear relationship may exist not with these variables, but rather with transformation of the variables. In this case we have,

$$g(y_i) = \alpha + \beta\psi(x_i) + \epsilon_i. \tag{3.13}$$

A common choice to to perform linear regression on the variables $\tilde{y} = g(y)$ and $\tilde{x} = \psi(x)$ using the least squares criterion. In this case, g is called the **link function**

For example, if

$$y_i = Ae^{kx_i+\epsilon_i},$$

we take logarithms,

$$\ln y_i = \ln A + kx_i + \epsilon_i$$

So, in (3.13), the link function $g(y_i) = \ln y_i$. The parameters are $\alpha = \ln A$ and $\beta = k$.

Before we look at an example, let's review a few basic properties of **logarithms**

Remark 3.19 (logarithms)**.** *We will use both* $\log$*, the base 10* **common logarthm***, and* $\ln$*, the base* e **natural logarithm***. Common logarithms more readily help us see orders of magnitude. For example, if* $\log y = 5$*, then we know that* $y = 10^5 = 100,000$*. if* $\log y = -1$*, then we know that* $y = 10^{-1} = 1/10$*. Typically, we will use natural logarithms when we want to emphasize instantaneous rates of growth. To understand how this works, consider the differential equation*

$$\frac{dy}{dt} = ky.$$

We are saying that the instantaneous rate of growth of y *is proportional to* y *with constant of proportionality* k*. The solution to this equation is*

$$y = y_0e^{kt} \qquad \textit{or} \qquad \ln y = \ln y_0 + kt$$

where y_0 *is the initial value for* y*. This gives a linear relationship between* $\ln y$ *and* t*. The two values of logarithm have a simple relationship. If we write*

$$x = 10^a. \textit{ Then } \log x = a \textit{ and } \ln x = a \ln 10.$$

Thus, by substituting for a*, we find that*

$$\ln x = \log x \cdot \ln 10 = 2.3026 \log x.$$

In R*, the command for the natural logarithm of* `x` *is* `log(x)`*. For the common logarithm, it is* `log(x,10)`*.*

Example 3.20. *In the data on world oil production, the relationship between the explanatory variable and response variable is nonlinear but can be made to be linear with a simple transformation, the common logarithm. Call the new*

response variable logbarrel. *The explanatory variable remains* year. *With these variables, we can use a regression line to help describe the data. Here the model is*

$$\log y_i = \alpha + \beta x_i + \epsilon_i. \tag{3.14}$$

Regression is the first example of a class of statistical models called **linear models**. *At this point we emphasize that* **linear** *refers to the appearance of the parameters α and β linearly in the function (3.14). This acknowledges that, in this circumstance, the values x_i and y_i are known. Indeed, they are the data. Using the principle of least squares, our goal is to give an estimate using the principle of least squares to $\hat{\alpha}$ and $\hat{\beta}$ for the values of α and β.* R *accomplishes this gaol with the command* `lm` *(for* l*inear* m*odel. Here is the output.*

```
> summary(lm(logbarrel~year))

Call:
lm(formula = logbarrel ~ year)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25562 -0.03390  0.03149  0.07220  0.12922

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.159e+01  1.301e+00  -39.64   <2e-16 ***
year         2.675e-02  6.678e-04   40.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1115 on 27 degrees of freedom
Multiple R-Squared: 0.9834,Adjusted R-squared: 0.9828
F-statistic:  1604 on 1 and 27 DF,  p-value: < 2.2e-16
```

Note that the output outputs the five number summary for the residuals and reports a coefficient of determination $r^2 = 0.9828$. Thus, the correlation is $r = 0.9914$ is very nearly one and so the data lies very close to the regression line.

For world oil production, we obtained the relationship

$$\widehat{\log(barrel)} = -51.59 + 0.02675 \cdot year$$

between the common logarithm of the number of millions of barrels of oil and the year. If we rewrite the equation in exponential form, we obtain, for some constant value, A,

$$\widehat{barrel} = A10^{0.02675 \cdot year} = Ae^{\hat{k} \cdot year}.$$

Thus, $\hat{k}$ gives the instantaneous growth rate that best fits the data. This is obtained by converting from a common logarithm to a natural logarithm.

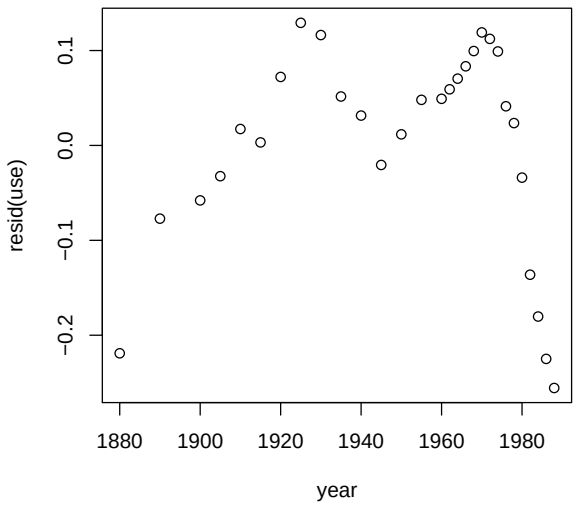
$$\hat{k} = 0.02675 \ln 10 = 0.0616$$

.

Consequently, the use of oil sustained a growth of 6% per year over a span of a hundred years.

Next, we will look for finer scale structure in the scatterplot by examining the residual plot.

```
> use<-lm(logbarrel~year)
> plot(year,resid(use))
```



Exercise 3.21. *Remove the data points after the oil crisis of the mid 1970s, find the regression line and the instantaneous growth rate that best fits the data. Look at the residual plot and use fact about American history to explain why the residuals increase until 1920's, decrease until the early 1940's and increase again until the early 1970's.*

Example 3.22 (Michaelis-Menten Kinetics)**.** *In this example, we will have to use a more sophisticated line of reasoning to create a linear relationship between a explanatory and response variable. Consider the chemical reaction in which an enzyme catalyzes the action on a substrate.*

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \overset{k_2}{\to} E + P \tag{3.15}$$

Here

- E_0 *is the total amount of enzyme.*
- E *is the free enzyme.*
- S *is the substrate.*
- ES *is the substrate-bound enzyme.*
- P *is the product.*
- $V = d[P]/dt$ *is the production rate.*

The numbers above or below the arrows gives the reaction rates. Using the symbol $[\,\cdot\,]$ *to indicate* **concentration***, notice that the enzyme,* E_0*, is either free or bound to the substrate. Its total concentration is, therefore,*

$$[E_0] = [E] + [ES], \quad \textit{and, thus} \quad [E] = [E_0] - [ES] \tag{3.16}$$

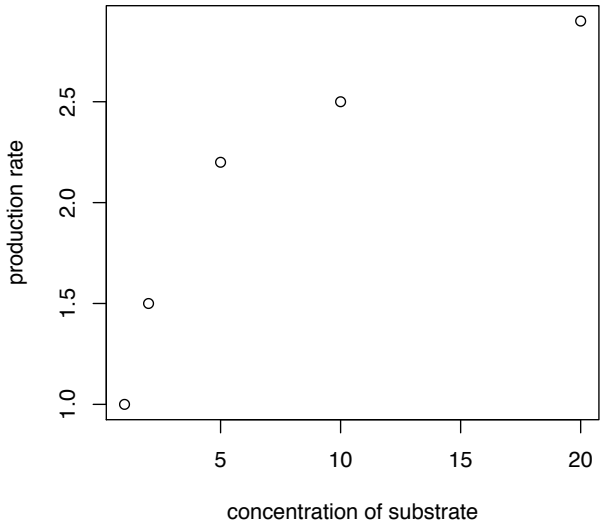
Our goal is to relate the production rate V *to the substrate concentration* $[S]$*. Because the total concentration* $[E_0]$ *is set by the experimenter, we can assume that it is a known quantity.*

The **law of mass action** *turns the chemical reactions in (3.15) into differential equations. In particular, the reactions, focusing on the substrate-bound enzyme and the product, gives the equations*

$$\frac{d[ES]}{dt} = k_1[E][S] - [ES](k_{-1} + k_2) \quad \textit{and} \quad V = \frac{d[P]}{dt} = k_2[ES] \tag{3.17}$$

We can meet our goal if we can find an equation for $V = k_2[ES]$ that depends only on $[S]$, the substrate concentration. Let's look at data,

$[S]$ (mM)	1	2	5	10	20
V (nmol/sec)	1.0	1.5	2.2	2.5	2.9



If we wish to use linear regression, then we will have to transform the data. In this case, we will develop the Michaelis-Menten transformation applied to situations in which the concentration of the substrate-bound enzyme (and hence also the unbound enzyme) changes much more slowly than those of the product and substrate.

$$0 \approx \frac{d[ES]}{dt}$$

In words, the substrate-bound enzyme is nearly in steady state. Using the law of mass action equation (3.17) for $d[ES]/dt$, we can rearrange terms to conclude that

$$[ES] \approx \frac{k_1[E][S]}{k_{-1} + k_2} = \frac{[E][S]}{K_m}. \tag{3.18}$$

The ratio $K_m = (k_{-1} + k_2)/k_1$ of the rate of loss of the substrate-bound enzyme to its production is called the **Michaelis constant**. *We have now met our goal part way, V is a function of $[S]$, but it is also stated as a function of $[E]$.*

Thus, we have shown in 3.16 that $[E]$ as a function of $[ES]$. Now, we combine this with (3.18) and solve for $[ES]$ to obtain

$$[ES] \approx \frac{([E_0] - [ES])[S]}{K_m}, \quad [ES] \approx [E_0]\frac{[S]}{K_m + [S]}$$

Under this approximation, known as the **Michaelis-Metens kinetic equation** *the production rate of the product is:*

$$V = \frac{d[P]}{dt} = k_2[ES] = k_2[E_0]\frac{[S]}{K_m + [S]} = V_{\max}\frac{[S]}{K_m + [S]} \tag{3.19}$$

Here, $V_{\max} = k_2[E_0]$ is the maximum production rate. (To see this, let the substrate concentration $[S] \to \infty$.) To perform linear regression, we need to have a function of V be linearly related to a function of $[S]$. This is achieved via taking the reciprocal of both sides of this equation.

$$\frac{1}{V} = \frac{K_m + [S]}{V_{\max}[S]} = \frac{1}{V_{\max}} + \frac{K_m}{V_{\max}}\frac{1}{[S]} \tag{3.20}$$

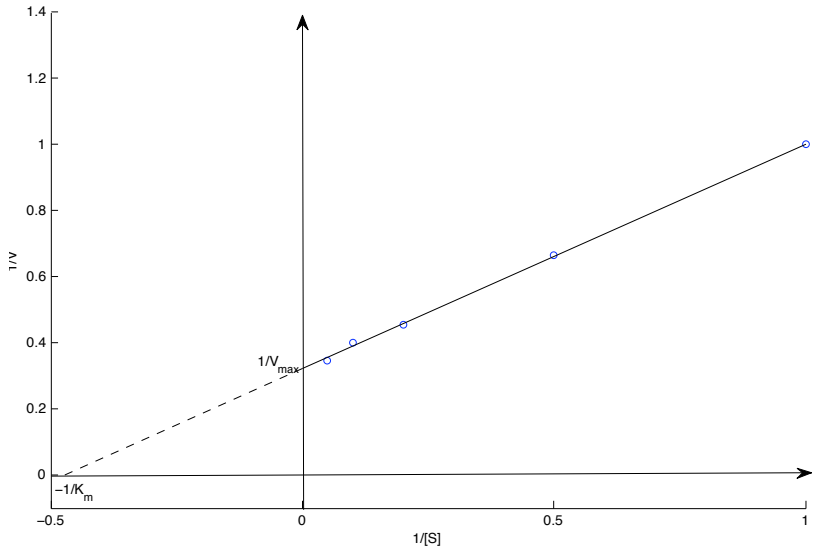


Figure 3.7: Lineweaver-Burke double reciprocal plot for the data presented above. The y-intercept gives the reciprocal of the maximum production. The dotted line indicates that negative concentrations are not physical. Nevertheless, the x-intercept give the negative reciprocal of the Michaelis constant.

Thus, we have a linear relationship between

$$\frac{1}{V}, \text{the response variable, and } \frac{1}{[S]}, \text{the explanatory variable}$$

subject to experimental error. The **Lineweaver-Burke double reciprocal plot** *provides a useful method for analysis of the Michaelis-Menten equation. See Figure 3.6.*

For the data,

$[S]$ (mM)	1	2	5	10	20
V (nmol/sec)	1.0	1.5	2.2	2.5	2.9

The regression line is

$$\frac{1}{V} = 0.3211 + \frac{1}{[S]}0.6813.$$

Here are the R *commands. (Both* `1` *and the slash* `/` *have a specific meaning for the* `lm` *command and so we set variables* `Vinv` *and* `Sinv` *for the inverses.)*

```
> S<-c(1,2,5,10,20)
> V<-c(1.0,1.5,2.2,2.5,2.9)
> Sinv<-1/S
> Vinv<-1/V
> lm(Vinv~Sinv)

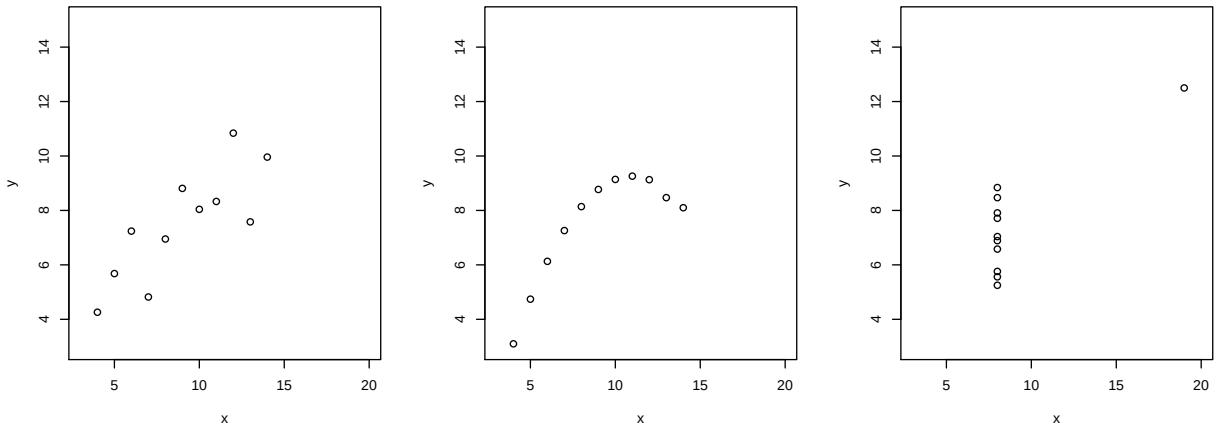
Call:
lm(formula = Vinv ~ Sinv)

Coefficients:
(Intercept)         Sinv
     0.3211       0.6813
```

Using (3.20), we find that $V_{\max} = 3.1141$ and $K_m = 2.122$. With more access to computational software, this method is not used as much as before. The measurements for small values of the concentration (and thus large value of $1/[S]$) are more variable and consequently the residuals are likely to be heteroscedastic. We look in the next section for an alternative approach, namely nonlinear regression.

Example 3.23 (Frank Amscombe)**.** *Consider the three data sets:*

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.47	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50



Each of these data sets has a regression line $\hat{y} = 3 + 0.5x$ and correlations between 0.806 and 0.816. However, only the first is a suitable data set for linear regression. This example is meant to emphasize the point that software will happily compute a regression line and an coefficient of determination value, but further examination of the data is required to see if this method is appropriate for any given data set.

3.3 Extensions

We will discuss briefly two extensions - the first is a least squares criterion between x and y that is **nonlinear** in the parameters $\beta = (\beta_0, \ldots, \beta_k)$. Thus, the model is

$$y_i = g(x_i|\beta) + \epsilon_i$$

for g, a nonlinear function of the parameters.

The second considers situations with more than one explanatory variable.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i. \tag{3.21}$$

This brief discussion does not have the detail necessary to begin to use these methods. It serves primarily as an invitation to begin to consult resources that more fully develop these ideas.

3.3.1 Nonlinear Regression

Here, we continue using estimation of parameters using a **least squares** criterion.

$$SS(\beta) = \sum_{i=1}^{n} (y_i - g(x_i|\beta))^2.$$

For most choices of $g(x|\beta)$ the solution to the nonlinear least square criterion cannot be expressed in closed form. Thus, a numerical strategy for the solution $\hat{\beta}$ is necessary. This generally begins with some initial guess of parameter values and an iteration scheme to minimize $SS(\beta)$. Such a scheme is likely to use local information about the first and second partial derivatives of g with respect to the parameters β_i. For example, **gradient descent** (also known as **steepest descent**, or the **method of steepest descent**) is an iterative method in which produces a sequence of parameter values. The increment of the parameter values for an iteration is proportional to the negative of the gradient of $SS(\beta)$ evaluated at the current point. The hope is that the sequence converges to give the desired minimum value for $SS(\beta)$. The R command `gnls` for **general nonlinear least squares** is used to accomplish this. As above, you should examine the residual plot to see that it has no structure. For, example, if the Lineweaver-Burke method for Michaelis-Mentens kinetics yields structure in the residuals, then linear regression is not considered a good method. Under these circumstances, one can next try to use the parameter estimates derived from Lineweaver-Burke as an initial guess in a nonlinear least squares regression using a least square criterion based on the sum of squares

$$SS(V_{max}, K_m) = \sum_{j=1}^{n} \left(V_j - V_{max} \frac{[S]_j}{K_m + [S]_j} \right)^2$$

for data $(V_1, [S]_1), (V_2, [S]_2), \ldots (V_n, [S]_n)$.

To use the `gnls` command we need to install the R package `nmle`. The command requires a model equation, here, the Michaelis-Metens equation (3.19), written as response variable model equation, the data, and a starting point for the numerical method for finding the parameter values that minimize $SS(V_{max}, K_m)$. Here, we begin with the values obtained from the regression equation for the Lineweaver-Burke double reciprocal plot.

```
> gnls(V~Vmax*S/(Km+S),data=data.frame(V,S),start=list(Vmax=3.1141,Km=2.1216))
Generalized nonlinear least squares fit
  Model: V ~ Vmax * S/(Km + S)
  Data: data.frame(V, S)
  Log-likelihood: 8.212577

Coefficients:
    Vmax       Km
3.154482 2.210205

Degrees of freedom: 5 total; 3 residual
Residual standard error: 0.06044381
```

We see a small change in the estimates $\hat{V}_{\max}$ and $\hat{K}_m$ from the previous estimates.

3.3.2 Multiple Linear Regression

Before we start with multiple linear regression, we first recall a couple of concepts and results from linear algebra.

- Let C_{ij} denote the entry in the i-th row and j-th column of a matrix C.
- A matrix A with r_A rows and c_A and a matrix B with r_B rows and c_B columns can be multiplied to form a matrix AB provide that $c_A = r_B$, the number of columns in A equals the number of rows in B. In this case

$$(AB)_{ij} = \sum_{k=1}^{c_A} A_{ik} B_{kj}.$$

- The d-dimensional **identity matrix** I is the matrix with the value 1 for all entries on the diagonal ($I_{jj} = 1, j = 1 \ldots, d$) and 0 for all other entries. Notice for and d-dimensional vector x,
$$Ix = x.$$
- A $d \times d$ matrix C is called invertible with **inverse** C^{-1} provided that
$$CC^{-1} = C^{-1}C = I.$$
Only one matrix can have this property.
- Suppose we have a d-dimensional vector a of known values and a $d \times d$ matrix C and we want to determine the vectors x that satisfy
$$a = Cx.$$
This equation could have no solutions, a single solution, or an infinite number of solutions. If the matrix C is invertible, then we have a single solution
$$x = C^{-1}a.$$
- The **transpose** of a matrix is obtained by reversing the rows and columns of a matrix. We use a superscript T to indicate the transpose. Thus, the ij entry of a matrix C is the ji entry of its transpose, C^T.

Example 3.24.
$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 2 & 7 \end{pmatrix}^T = \begin{pmatrix} 2 & 4 \\ 1 & 2 \\ 3 & 7 \end{pmatrix}$$
- A square matrix C is invertible if and only if its determinant $\det(C) \neq 0$. For a 2×2 matrix
$$C = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$
$\det(C) = ad - bc$ and the matrix inverse
$$C^{-1} = \frac{1}{\det(C)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Exercise 3.25. $(Cx)^T = x^T C^T$

Exercise 3.26. *For*
$$C = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix},$$
find C^T, $\det(C)$ *and* C^{-1} *by hand and using* R

In multiple linear regression, we have more than one predictor or explanatory random variable. Thus can write (3.21) in matrix form
$$y = X\beta + \epsilon \tag{3.22}$$
- $y = (y_1, y_2, \ldots, y_n)^T$ is a column vector of responses,
- X is a matrix of predictors,
$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}. \tag{3.23}$$
The column of ones on the left give the constant term in a multilinear equation. This matrix X is an an example of what is know as a **design matrix**.

- $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is a column vector of parameters, and
- $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is a column vector of "errors".

Exercise 3.27. *Show that the least squares criterion*

$$SS(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - \beta_k x_{ik})^2. \tag{3.24}$$

can be written in matrix form as

$$SS(\beta) = (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta).$$

To minimize SS, we take the gradient and set it equal to 0.

Exercise 3.28. *Check that the gradient is*

$$\nabla_\beta SS(\beta) = -2(\mathbf{y} - X\beta)^T X. \tag{3.25}$$

Based on the exercise above, the value $\hat{\beta}$ that minimizes SS is

$$(\mathbf{y} - X\hat{\beta})^T X = 0, \quad \mathbf{y^T} X = \hat{\beta}^T X^T X.$$

The transpose of this last equation is sometimes known at the **normal equations**

$$X^T X\hat{\beta} = X^T \mathbf{y}. \tag{3.26}$$

If $X^T X$ is invertible, then we can multiply both sides of the equation above by $(X^T X)^{-1}$ to obtain an equation for the parameter values $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n)$ in the least squares regression.

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}. \tag{3.27}$$

Thus the estimates $\hat{\beta}$ are a linear transformation of the repsonses $\mathbf{y}$ through the so-called **hat matrix** $H = (X^T X)^{-1} X^T$, i.e. $\hat{\beta} = H\mathbf{y}$.

Exercise 3.29. *Verify that the hat matrix H is a left inverse of the design matrix X.*

This gives the regression equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

Example 3.30 (ordinary least squares regression)**.** *In this case,*

$$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

and

$$X^T y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

The determinant of $X^T X$ is

$$n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 = n(n-1)\text{var}(x).$$

and thus

$$(X^TX)^{-1} = \frac{1}{n(n-1)\text{var}(x)} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

and

$$\hat{\beta} = (X^TX)^{-1}X^T\mathbf{y} = \frac{1}{n(n-1)\text{var}(x)} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_iy_i \end{pmatrix}.$$

For example, for the second row, we obtain

$$\frac{1}{n(n-1)\text{var}(x)} \left(\left(-\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) + n \sum_{i=1}^n x_iy_i \right) = \frac{n(n-1)\text{cov}(x,y)}{n(n-1)\text{var}(x)} = \frac{\text{cov}(x,y)}{\text{var}(x)}$$

as seen in equation (3.10).

Example 3.31. *We can estimate the number of mutations from each parent by using two explanatory variables, namely both the father's and mother's age at the time of the conception of the offspring to the* de novo *mutations in the offspring for the 78 Icelandic trios.*

```
> iceland.lm<-lm(mutations~paternal+maternal)
> summary(iceland.lm)

Call:
lm(formula = mutations ~ paternal + maternal)

Residuals:
     Min       1Q   Median       3Q      Max
-18.6769  -6.5272   0.7632   4.8546  21.0775

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5901     5.2829   0.490    0.625
paternal      1.8414     0.2987   6.165 3.25e-08 ***
maternal      0.2220     0.3203   0.693    0.491
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.437 on 75 degrees of freedom
Multiple R-squared:  0.6556,Adjusted R-squared:  0.6465
F-statistic:  71.4 on 2 and 75 DF,  p-value: < 2.2e-16
```

Here, we estimate that, on average, each year of the father's age adds 1.84 mutations. The mother adds 0.22 mutations per year of age. Thus, it is the father's age that dominates the number of de novo *mutations in the offspring. The coefficient of determination is now 0.6556, increased from 0.6212 with a single explanatory variable. Because the source of mutations is unknown and the parents' ages are highly correlated (~ 0.8), estimation is difficult.*

Example 3.32. *The choice of $x_{ij} = x_i^j$ in (3.23) results in* **polynomial regression**

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i.$$

in equation (3.21).

Example 3.33 (US population)**.** *Below are the census populations*

year	census population	year	census population	year	census population	year	census population
1790	3,929,214	1850	23,191,876	1910	92,228,496	1970	203,211,926
1800	5,236,631	1860	31,443,321	1920	106,021,537	1980	226,545,805
1810	7,239,881	1870	38,558,371	1930	123,202,624	1990	248,709,873
1820	9,638,453	1880	49,371,340	1940	132,164,569	2000	281,421,906
1830	12,866,020	1890	62,979,766	1950	151,325,798	2010	308,745,538
1840	17,069,453	1900	76,212,168	1960	179,323,175		

To analyze this in R *we enter the data:*

```
> uspop<-c(3929214,5236631,7239881,9638453,12866020,17069453,23191876,31443321,
+ 38558371,49371340,62979766,76212168,92228496,106021537,123202624,132164569,
+ 151325798,179323175,203211926,226545805,248709873,281421906,308745538)
> year<-c(0:22)*10+1790
> plot(year,uspop)
> loguspop<-log(uspop,10)
> plot(year,loguspop)
```

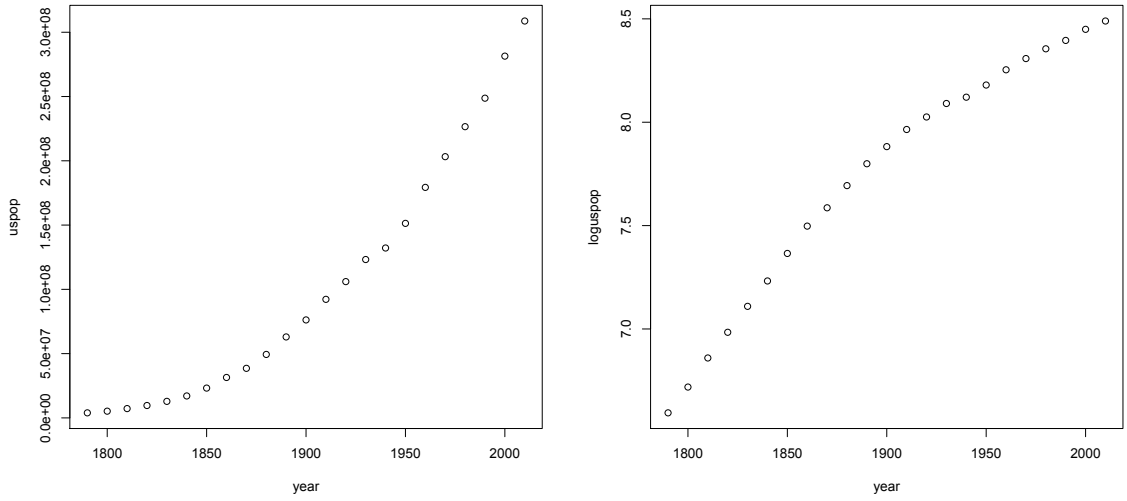


Figure 3.8: (a) United States census population from 1790 to 2010 and (b) its base 10 logarithm.

Note that the logarithm of the population still has a bend to it, so we will perform a quadratic regression on the logarithm of the population. In order to keep the numbers smaller, we shall give the year minus 1790, the year of the first census for our explanatory variable.

$$\log(uspopulation) = \beta_0 + \beta_1(year - 1790) + \beta_2(year - 1790)^2.$$

```
> year1<-year-1790
> year2<-year1^2
```

Thus, `loguspop` *is the reponse variable. The* + *sign is used in the case of more than one explanatory variable and here is placed between the response variables* `year1` *and* `year2`*.*

```
> lm.uspop<-lm(loguspop~year1+year2)
> summary(lm.uspop)

Call:
lm(formula = loguspop ~ year1 + year2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.037387 -0.013453 -0.000912  0.015281  0.029782

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.582e+00  1.137e-02  578.99   <2e-16 ***
year1        1.471e-02  2.394e-04   61.46   <2e-16 ***
year2       -2.808e-05  1.051e-06  -26.72   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.01978 on 20 degrees of freedom
Multiple R-squared: 0.999,Adjusted R-squared: 0.9989
F-statistic:  9781 on 2 and 20 DF,  p-value: < 2.2e-16
```

The R *output shows us that*

$$\hat{\beta}_0 = 6.587 \quad \hat{\beta}_1 = 0.01471 \; quad\hat{\beta}_2 = -0.00002808.$$

So, taking the the regression line to the power 10, we have that

$$\widehat{uspopulation} = 3863670 \times 10^{0.0147(year-1790)-0.00002808(year-1790)^2}$$

In Figure 3.8, we show the residual plot for the logarithm of the US population.

```
> resid.uspop<-resid(lm.uspop)
> plot(year,resid.uspop)
```

3.4 Answers to Selected Exercises

3.1. Negative covariance means that the terms $(x_i - \bar{x})(y_i - \bar{y})$ in the sum are more likely to be negative than positive. This occurs whenever one of the x and y variables is above the mean, then the other is likely to be below.

3.2. We expand the product inside the sum.

$$\begin{aligned}\text{cov}(x,y) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y}\right)\\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}\right)\end{aligned}$$

The change in measurements from centimeters to meters would divide the covariance by 10,000.

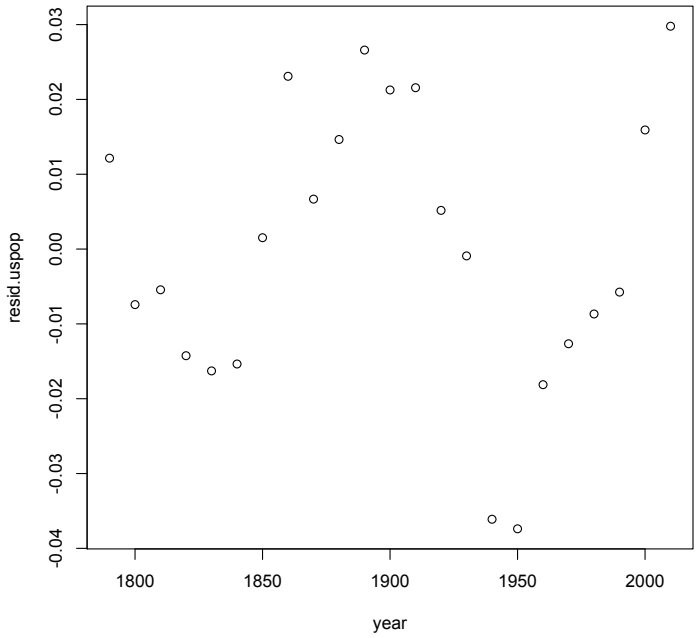


Figure 3.9: Residual plot for US population regression.

3.3. We rearrange the terms and simplify.

$$\begin{aligned}\text{cov}(ax+b, cy+d) &= \frac{1}{n-1}\sum_{i=1}^{n}((ax_i+b)-(a\bar{x}+b)((cy_i+d)-(c\bar{y}-d)) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i-a\bar{x})(cy_i-c\bar{y}) = ac\cdot\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = ac\cdot\text{cov}(x,y)\end{aligned}$$

3.5. Assume that $a \neq 0$ and $c \neq 0$. If $a = 0$ or $c = 0$, then the covariance is 0 and so is the correlation.

$$r(ax+b, cy+d) = \frac{\text{cov}(ax+b, cy+d)}{s_{ax+b}s_{cy+d}} = \frac{ac\cdot\text{cov}(x,y)}{|a|s_x\cdot|c|s_y} = \frac{ac}{|ac|}\frac{\cdot\text{cov}(x,y)}{s_x\cdot s_y} = \pm r(x,y)$$

We take the plus sign if the sign of a and c agree and the minus sign if they differ.

3.6. First we rearrange terms

$$\begin{aligned}s_{x+y}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}((x_i+y_i)-(\bar{x}+\bar{y}))^2 = \frac{1}{n-1}\sum_{i=1}^{n}((x_i-\bar{x})+(y_i-\bar{y}))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 + 2\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) + \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\ &= s_x^2 + 2\text{cov}(x,y) + s_y^2 = s_x^2 + 2rs_xs_y + s_y^2\end{aligned}$$

For a triangle with sides a, b and c, the law of cosines states that

$$c^2 = a^2 + b^2 - 2ab\cos\theta$$

where θ is the measure of the angle opposite side c. Thus the analogy is

$$s_x \text{ corresponds to } a, \quad s_y \text{ corresponds to } b, \quad s_{x+y} \text{ corresponds to } c, \quad \text{and} \quad r \text{ corresponds to } -\cos\theta$$

Notice that both r and $\cos\theta$ take values between -1 and 1.

3.7. Using the hint,

$$0 \leq \sum_{i=1}^{n}(v_i + w_i\zeta)^2 = \sum_{i=1}^{n} v_i^2 + 2\left(\sum_{i=1}^{n} v_i w_i\right)\zeta + \left(\sum_{i=1}^{n} w_i^2\right)\zeta^2 = A + B\zeta + C\zeta^2$$

For a quadratic equation to always take on non-negaitive values, we must have a non-positive discriminant, i. e.,

$$0 \geq B^2 - 4AC = 4\left(\sum_{i=1}^{n} v_i w_i\right)^2 - 4\left(\sum_{i=1}^{n} v_i^2\right)\left(\sum_{i=1}^{n} w_i^2\right).$$

Now, divide by 4 and rearrange terms.

$$\left(\sum_{i=1}^{n} v_i^2\right)\left(\sum_{i=1}^{n} w_i^2\right) \geq \left(\sum_{i=1}^{n} v_i w_i\right)^2.$$

3.8. The value of the correlation is the same for pairs of observations and for their standardized versions. Thus, we take x and y to be standardized observations. Then $s_x = s_y = 1$. Now, using equation (3.1), we have that

$$0 \leq s_{x+y}^2 = 1 + 1 + 2r = 2 + 2r. \text{ Simplifying, we have } -2 \leq 2r \text{ and } r \geq -1.$$

For the second inequality, use the similar identity to (3.1) for the difference in the observations

$$s_{x-y}^2 = s_x^2 + s_y^2 - 2r s_x s_y.$$

Then,

$$0 \leq s_{x-y} = 1 + 1 - 2r = 2 - 2r. \text{ Simplifying, we have } 2r \leq 2 \text{ and } r \leq 1.$$

Thus, correlation must always be between -1 and 1.

In the case that $r = -1$, we that that $s_{x+y}^2 = 0$ and thus using the standardized coordinates

$$\frac{x_i - \bar{x}}{s_x} + \frac{y_i - \bar{y}}{s_y} = 0.$$

Thus, $\zeta = s_y/s_x$.

In the case that $r = 1$, we that that $s_{x-y}^2 = 0$ and thus using the standardized coordinates

$$\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y} = 0.$$

Thus, $\zeta = -s_y/s_x$.

3.10.

1. First the data and the scatterplot, preparing by using `mfrow`to have side-by-side plots

```
> x<-c(-2:3)
> y<-c(-3,-1,-2,0,4,2)
> par(mfrow=c(1,2))
> plot(x,y)
```

2. Then the regression line and its summary.

```
> regress.lm<-lm(y~x)
> summary(regress.lm)

Call:
lm(formula = y ~ x)

Residuals:
         1          2          3          4          5          6
-2.776e-16  8.000e-01 -1.400e+00 -6.000e-01  2.200e+00 -1.000e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6000     0.6309  -0.951   0.3955
x             1.2000     0.3546   3.384   0.0277 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.483 on 4 degrees of freedom
Multiple R-squared:  0.7412,Adjusted R-squared:  0.6765
F-statistic: 11.45 on 1 and 4 DF,  p-value: 0.02767
```

3. Add the regression line to the scatterplot.

```
> abline(regress.lm)
```

4. Make a data frame to show the predictions and the residuals.

```
> residuals<-resid(regress.lm)
> predictions<-predict(regress.lm,newdata=data.frame(x=c(-2:3)))
> data.frame(x,y,predictions,residuals)
   x  y predictions     residuals
1 -2 -3        -3.0 -2.775558e-16
2 -1 -1        -1.8  8.000000e-01
3  0 -2        -0.6 -1.400000e+00
4  1  0         0.6 -6.000000e-01
5  2  4         1.8  2.200000e+00
6  3  2         3.0 -1.000000e+00
```

5. FInally, the residual plot and a horizontal line at 0.

```
> plot(x,residuals)
> abline(h=0)
```

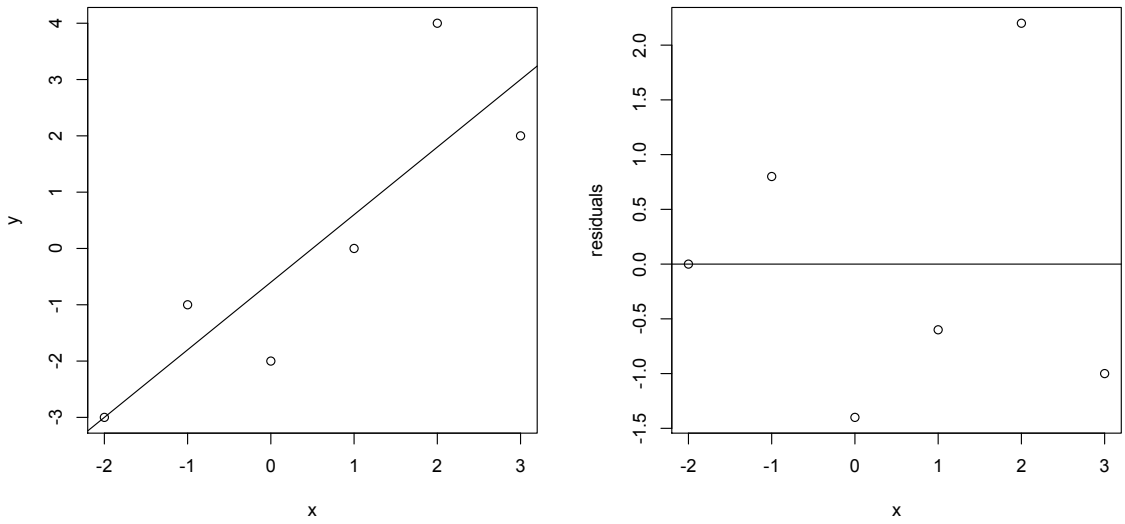


Figure 3.10: (left) scatterplot and regression line (right) residual plot and horizontal line at 0

3.13. Use the subscript y in $\hat{\alpha}_y$ and $\hat{\beta}_y$ to emphasize that y is the explanatory variable. We still have $\bar{x} = 0.5, \bar{y} = 0$.

y_i	x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \hat{y})$	$(y_i - \bar{y})^2$
-3	-2	-3	-2.5	7.5	9
-1	-1	-1	-1.5	1.5	1
-2	0	-2	-0.5	1.0	4
0	1	0	0.5	0.0	0
4	2	4	1.5	6.0	16
2	3	2	2.5	5.0	4
total		0	0	$\text{cov}(x, y) = 21/5$	$\text{var}(y) = 34/5$

So, the slope $\hat{\beta}_y = 21/34$ and

$$\bar{x} = \hat{\alpha}_y + \hat{\beta}_y \bar{y}, \quad 1/2 = \hat{\alpha}_y.$$

Thus, to predict x from y, the regression line is $\hat{x}_i = 1/2 + 21/34 y_i$. Because the product of the slopes

$$\frac{6}{5} \times \frac{21}{34} = \frac{63}{85} \neq 1,$$

this line differs from the line used to predict y from x.

3.14. Recall that the covariance of x and y is symmetric, i.e., $\text{cov}(x, y) = \text{cov}(y, x)$. Thus,

$$\hat{\beta}_x \cdot \hat{\beta}_y = \frac{\text{cov}(x, y)}{s_x^2} \cdot \frac{\text{cov}(y, x)}{s_y^2} = \frac{\text{cov}(x, y)^2}{s_x^2 s_y^2} = \left(\frac{\text{cov}(x, y)}{s_x s_y} \right)^2 = r^2.$$

In the example above, the coefficient of determination,

$$r^2 = \frac{\text{cov}(x, y)^2}{s_x^2 x_y^2} = \frac{(21/5)^2}{(17.5/5) \cdot (34/5)} = \frac{21^2}{17.5 \cdot 34} = \frac{21}{35} \cdot \frac{21}{17} = \frac{3}{5} \cdot \frac{21}{17} = \frac{63}{85}$$

3.15 To show that the correlation is zero, we show that the numerator in the definition, the covariance is zero. First,

$$\text{cov}(\hat{y}, y - \hat{y}) = \text{cov}(\hat{y}, y) - \text{cov}(\hat{y}, \hat{y}).$$

The first term in this difference,

$$\text{cov}(\hat{y}, y) = \text{cov}\left(\frac{\text{cov}(x,y)}{s_x^2}x, y\right) = \frac{\text{cov}(x,y)^2}{s_x^2} = \frac{r^2 s_x^2 s_y^2}{s_x^2} = r^2 s_y^2.$$

For the second,

$$\text{cov}(\hat{y}, \hat{y}) = s_{\hat{y}}^2 = r^2 s_y^2.$$

So, the difference is 0.

3.16. For the denominator

$$s_{DATA}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

For the numerator, recall that $(\bar{x}, \bar{y})$ is on the regression line. Consequently, $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$. Thus, the mean of the fits

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}x_i) = \hat{\alpha} + \hat{\beta}\bar{x} = \bar{y}.$$

This could also be seen by using the fact (3.11) that the sum of the residuals is 0. For the denominator,

$$s_{FIT}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

Now, take the ratio and notice that the fractions $1/(n-1)$ in the numerator and denominator cancel.

3.17. The least squares criterion becomes

$$SS(\beta) = \sum_{i=1}^{n}(y_i - \beta x_i)^2.$$

The derivative with respect to β is

$$SS'(\beta) = -2\sum_{i=1}^{n} x_i(y_i - \beta x_i).$$

$SS'(\beta) = 0$ for the value

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

3.22. The i-th component of $(Cx)^T$ is

$$\sum_{j=1}^{n} C_{ij}x_j.$$

Now the i-th component of $x^T C^T$ is

$$\sum_{j=1}^{n} x_j C_{ji}^T = \sum_{j=1}^{n} x_j C_{ij}.$$

3.25. The transpose

$$C^T = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

the determinant $\det(C) = 4 - 6 = -2$ and

$$C^{-1} = \frac{1}{2}\begin{pmatrix} 4 & -3 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 3/2 \\ 1 & -1/2 \end{pmatrix}.$$

Using R,

```
> C<-matrix(c(1,2,3,4),nrow=2)
> C
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> t(C)
     [,1] [,2]
[1,]    1    2
[2,]    3    4
> det(C)
[1] -2
> chol2inv(C)
        [,1]    [,2]
[1,]  1.5625 -0.1875
[2,] -0.1875  0.0625
```

3.26. Using equation (3.22), the i-th component of $\mathbf{y} - X\beta$,

$$(\mathbf{y} - X\beta)_i = y_i - \sum_{j=0}^{n} \beta_j x_{jk} = y_i - \beta_0 - x_{i1}\beta_1 - \cdots - \beta_k x_{in}.$$

Now, $(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)$ is the dot product of $\mathbf{y} - X\beta$ with itself. This gives (3.24).

3.27. Write $x_{i0} = 1$ for all i, then we can write (3.24) as

$$SS(\beta) = \sum_{i=1}^{n} (y_i - x_{i0}\beta_0 - x_{i1}\beta_1 - \cdots - \beta_k x_{ik})^2.$$

Then,

$$\begin{aligned}\frac{\partial}{\partial \beta_j} S(\beta) &= -2\sum_{i=1}^{n} (y_i - x_{i0}\beta_0 - x_{i1}\beta_1 - \cdots - \beta_k x_{ik}) x_{ij} \\ &= -2\sum_{i=1}^{n} (y_i - (X\beta)_i) x_{ij} = -2((y - X\beta)^T X))_j.\end{aligned}$$

This is the j-th coordinate of (3.25).

3.28. $HX = (X^TX)^{-1}X^TX = (X^TX)^{-1}(X^TX) = I$, the identity matrix.

Topic 4

Producing Data

Statistics has been the handmaid of science, and has poured a flood of light upon the dark questions of famine and pestilence, ignorance and crime, disease and death. - James A. Garfield, December 16, 1867

Our health care is too costly; our schools fail too many; and each day brings further evidence that the ways we use energy strengthen our adversaries and threaten our planet.

These are the indicators of crisis, subject to data and statistics. Less measurable but no less profound is a sapping of confidence across our land a nagging fear that America's decline is inevitable, and that the next generation must lower its sights. - Barack Obama, January 20, 2009

4.1 Preliminary Steps

Many questions begin with an anecdote or an unexplained occurrence in the lab or in the field. This can lead to fact-finding interviews or easy to perform experimental assays. The next step will be to review the literature and begin an **exploratory data analysis**. At this stage, we are looking, on the one hand, for patterns and associations, and, on the other hand, apparent inconsistencies occurring in the scientific literature. Next we will examine the data using quantitative methods - summary statistics for quantitative variables, tables for categorical variables - and graphical methods - boxplots, histograms, scatterplots, time plots for quantitative data - bar charts for categorical data.

The strategy of these investigations is frequently the same - look at a **sample** in order to learn something about a **population** or to take a **census** or the total population.

Designs for producing data begin with some basic questions:

- What can I measure?
- What shall I measure?
- How shall I measure it?
- How frequently shall I measure it?
- What obstacles do I face in obtaining a reliable measure?

The frequent goal of a statistical study is to investigate the nature of **causality**. In this way we try to explain the values of some **response variables** based on knowing the values of one or more **explanatory variables**. The major issue is that the associated phenomena could be caused by a third, previously unconsidered factor, called a **lurking variable** or **confounding variable**.

Two approaches are generally used to mitigate the impact of confounding. The first, primarily statistical, involves subdividing the population under study into smaller groups that are more similar. This subdivision is called **cross tabulation** or **stratification**. For human studies, this could mean subdivision by gender, by age, by economic class,

by geographic region, or by level of education. For laboratory, this could mean subdivision by temperature, by pH, by length of incubation, or by concentration of certain compounds (e.g. ATP). For field studies, this could mean subdivision by soil type, by average winter temperature or by total rainfall. Naturally, as the number of subgroups increase, the size of these groups can decrease to the point that chance effects dominate the data.

The second is mathematical or probabilistic modeling. These models often take the form of a mechanistic model that takes into an account the variables in the cross tabulation and builds a **parametric model**.

The best methodologies, of course, make a comprehensive use of both of these types of approaches.

4.2 Professional Ethics

As a citizen, we should participate in public discourse. Those with particular training have a special obligation to bring to the public their special knowledge. Such public statements can take several forms. We can speak out as a member of society with no particular basis in our area of expertise. We can speak out based on the wisdom that comes with this specialized knowledge. Finally, we can speak out based on a formal procedure of gathering information and reporting carefully the results of our analysis. In each case, it is our obligation to be clear about the nature of that communication and that the our statements follow the highest ethical standards. In the same vein, as consumers of information, we should have a clear understanding of the perspective in any document that presents statistical information.

Professional statistical societies have provided documents that provide guidance on what can be sometimes be difficult judgements and decisions. Two sources of guidance are the *Ethical Guidelines for Statistical Practice* from the American Statistical Society.

`http://www.amstat.org/about/ethicalguidelines.cfm`

and the International Statistical Institute *Declaration on Professional Ethics*

`http://www.isi-web.org/about-isi/professional-ethics`

4.3 Formal Statistical Procedures

The formal procedures that will be described in this section presume that we will have a sufficiently well understood mathematical model to support the analysis of data obtained under a given procedure. Thus, this section anticipates some of the concepts in probability theory like independence, conditional probability, distributions under different sampling protocols and expected values. It also will rely fundamentally on some of the consequences of this theory as seen, for example, in the law of large numbers and the central limit theorem. These are topics that we shall soon explore in greater detail.

4.3.1 Observational Studies

The goal is to learn about a population by observing a sample with as little disturbance as possible to the sample.

Sometimes the selection of treatments is not under the control of the researcher. For example, if we suspect that a certain mutation would render a virus more or less virulent, we cannot ethically perform the genetic engineering and infect humans with the viral strains.

For an observational study, effects are often confounded and thus causation is difficult to assert. The link between smoking and a variety of diseases is one very well known example. We have seen the data set relating student smoking habits in Tucson to their parents. We can see that children of smokers are more likely to smoke. This is more easily described if we look at **conditional distributions**.

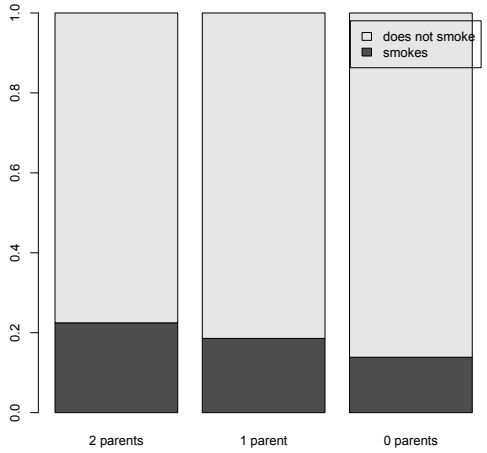
0 parents smoke	
student smokes	student does not smoke
0.1386	0.8614

1 parent smoke	
student smokes	student does not smoke
0.1858	0.8142

2 parents smoke	
student smokes	student does not smoke
0.2247	0.7753

To display these conditional distributions in R:

```
> smoking<-matrix(c(400,1380,416,1823,188,1168),ncol=3)
> smoking
     [,1] [,2] [,3]
[1,]  400  416  188
[2,] 1380 1823 1168
> condsmoke<-matrix(rep(0,6),ncol=3)
> for (i in 1:3)
  {condsmoke[,i]=smoking[,i]/sum(smoking[,i])}
> colnames(condsmoke)
  <-c("2 parents","1 parent", "0 parents")
> rownames(condsmoke)
  <-c("smokes","does not smoke")
> condsmoke
               2 parents  1 parent 0 parents
smokes         0.2247191 0.1857972 0.1386431
does not smoke 0.7752809 0.8142028 0.8613569
> barplot(condsmoke,legend=rownames(condsmoke))
```



Even though we see a trend - children are more likely to smoke in households with parents who smoke, we cannot assert causation, i.e., children smoke because their parents smoke. An alternative explanation might be, for example, people may have a genetic predisposition to smoking.

4.3.2 Randomized Controlled Experiments

In a controlled experiment, the researcher imposes a treatment on the **experimental units** or **subjects** in order to observe a response. Great care and knowledge must be given to the design of an effective experiment. A University of Arizona study on the impact of diet on cancers in women had as its goal specific recommendations on diet. Such recommendations were set to encourage lifestyle changes for millions of American women. Thus, enormous effort was taken in the design of the experiment so that the research team was confident in its results.

A good experimental design is one that is based on a solid understanding of both the science behind the study and the probabilistic tools that will lead to the inferential techniques used for the study. This study is often set to assess some hypothesis - *Do parents smoking habits influence their children?* or estimate some value - *What is the mean length of a given strain of bacteria?*

Principles of Experimental Design

1. **Control** for the effects of lurking variables by comparing several treatments.

2. **Randomize** the assignment of subjects to treatments to eliminate bias due to systematic differences among categories.
3. **Replicate** the experiment on many subjects to reduce the impact of chance variation on the results.

Issues with Control

The desired control can sometimes be quite difficult to achieve. For example;

- In medical trials, some individuals may display a **placebo effect**, the favorable response to any treatment.
- Overlooking or introducing a lurking variable can introduce a **hidden bias**.
- The time and money invested can lead to a subconscious effect by the experimenter. Use an appropriate **blind** or **double blind** procedure. In this case, neither the experimenter nor the subject are aware of which treatment is being used.
- Changes in the wording of questions can lead to different outcomes.
- Transferring discoveries from the laboratory to a genuine living situation can be difficult to make.
- The data may suffer from undercoverage of difficult to find groups. For example, mobile phone users are less accessible to pollsters.
- Some individuals leave the experimental group, especially in longitudinal studies.
- In some instances, a control is not possible. The outcomes of the absence of the enactment of an economic policy, for example, a tax cut or economic stimulus plan, cannot be directly measured. Thus, economists are likely to use a mathematical model of different policies and examine the outcomes of computer simulations as a proxy for control.
- Social desirability bias describes the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others. Thus surveys on medical issues, religious practices, sexual practices, political preferences, personal achievement typically use specialized techniques to obtain more truthful responses. The **Bradley effect** is a theory proposed to explain observed discrepancies between voter opinion polls and election outcomes in some US government elections where a white candidate and a non-white candidate run against each other. The theory proposes that some voters tend to tell pollsters that they are undecided or likely to vote for a black candidate, and yet, on election day, vote for his white opponent. It was named after Tom Bradley, an African-American who lost the 1982 California governor's race despite being ahead in voter polls going into the elections.

Setting a Design

Before data are collected, we must consider some basic questions:

- Decide on the number of explanatory variables or **factors**.
- Decide on the values or **levels** that will be used in the treatment.

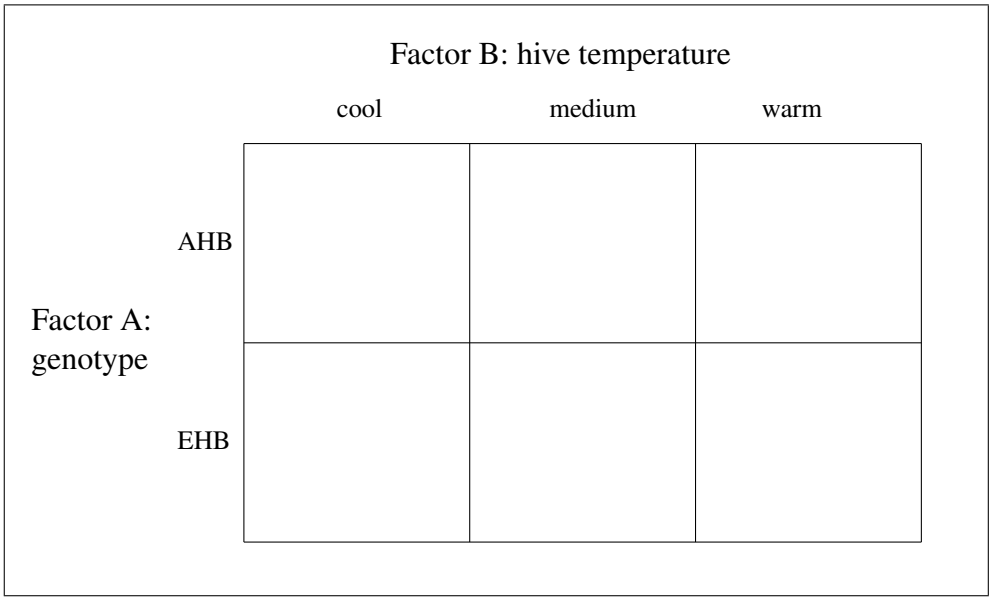
Example 4.1. *For over a century, beekeepers have attempted to breed honey bees belonging to different races to take advantage of the effects of hybrid vigor to create a better honey producer. No less a figure than Gregor Mendel failed in this endeavor because he could not control the matings of queens and drones.*

A more recent failure, a breeding experiment using African and European bees, occurred in 1956 in an apiary in the southeast of Brazil. The hybrid Africanized honey bees escaped, and today, in the western hemisphere, all Africanized honey bees are descended from the 26 Tanzanian queen bees that resided in this apiary. By the mid-1990s, Africanized bees have spread to Texas, Arizona, New Mexico, Florida and southern California.

When the time arrives for replacing the mother queen in a colony (a process known as **supercedure***), the queen will lay about ten queen eggs. The first queen that completes her development and emerges from her cell is likely to become the next queen. Suppose we have chosen to investigate the question of whether a shorter time for development for Africanized bee queens than for the resident European bee queens is the mechanism behind the replacement by Africanized subspecies in South and Central American and in the southwestern United States. The development time will depend upon hive temperature, so we will determine a range of hive temperatures by looking through the literature and making a few of our own measurements. From this, we will set a cool, medium, and warm hive temperature. We will use European honey bee (EHB) queens as a control. Thus, we have two factors.*

- *Queen type - European or Africanized*
- *Hive temperature - cool, medium, or warm.*

Thus, this experiment has **6 treatment groups***.*



The response variable is the queen development time - the length of time from the depositing of the egg from the mother queen to the time that the daughter queen emerges from the hive. The immature queen is kept in the hive to be fed during the egg and larval stages. At that point the cell containing the larval queen is capped by the worker bees. We then transfer the cell to an incubator for the pupal stage. The hive where the egg is laid and the incubator that houses the queen is checked hourly. A few queens are chosen and their genotypes are determined to verify the genetic designations of the groups. To reduce hidden biases, the queens in the incubator are labeled in such a way that their genotype is unknown.

We will attempt to rear 120 queens altogether and use 20 in each treatment group. The determination how the number of samples in the study is necessary to have the desired confidence in our results is called a **power analysis***. We will investigate this aspect of experimental design when we study hypothesis testing.*

Random Samples

A **simple random sample (SRS)** of size n consists of n individuals chosen in such a way that every set of n individuals has an equal chance to be in the sample actually selected. This is easy to accomplish in R. First, give labels to the individuals in the population and then use the command `sample` to make the random choice. For the experiment above, we rear 90 Africanized queens and choose a sample of 60. (Placing the command in parenthesis calls on R to print the output.)

```
> population<-c(1:90)
> (subjects<-sample(population,60))
 [1] 61 16 65 73 13 25 10 82 24 62 28 66 55  8 26 72 67 17 58 69  6 27 41 20
[25] 87 68 22 11  5 48 33 63 50 88 35 37 84 12  4 59 90 86  2 60 19 18 74 23
[49] 78 49 45  7 64  3 42 57 81 56 46 32
```

If your experimental design call for grouping similar individuals, called **strata**, then a **stratified random sample** from the full sample by choosing a separate random sample from each stratum. If one or more of the groups forms a small fraction of the population, then a stratified random sample ensures the desired number of sample from these groups is included in the sample.

If we mark the 180 queens 1 through 180 with 1 through 90 being Africanized bees and 91 through 180 being European, then we can enter

```
> population<-c(1:180)
> subjectsAHB<-sample(population[1:90],60)
> subjectsEHB<-sample(population[91:180],60)
```

to ensure that 60 come from each group.

For the example above, we divide the sampled Africanized queens into 3 treatment groups based on hive temperature. Here `dim=c(3,20)` signifies that the array has 3 rows and 20 columns. Let the first row be the choice of queen bees for the cool hive, the second row for the medium temperature hive, and row three for the warm hive.

```
> groups<-array(subjectsAHB,dim=c(3,20))
> groups
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,]   61   73   10   62   55   72   58   27   87    11    33    88    84
[2,]   16   13   82   28    8   67   69   41   68     5    63    35    12
[3,]   65   25   24   66   26   17    6   20   22    48    50    37     4
     [,14] [,15] [,16] [,17] [,18] [,19] [,20]
[1,]    59     2    18    78     7    42    56
[2,]    90    60    74    49    64    57    46
[3,]    86    19    23    45     3    81    32
```

Most of the data sets that we shall encounter in this book have a modest size with hundreds and perhaps thousands of observations based on a small number of variables. In these situation, we can be careful in assuring that the experimental design was followed. we can make the necessary visual and numerical summaries of the data set to assess its quality and make appropriate corrections to ethically clean the data from issues of mislabeling and poorly collected observations. This will prepare us for the more formal procedures that are the central issues of the second half of this book.

We are now in a world of massive datasets, collected, for example, from genomic, astronomical observations or social media. Data collection, management and analysis require new and more sophisticated approaches that maintain data integrity and security. These considerations form a central issue in modern statistics.

4.3.3 Natural experiments

In this situation, a naturally occurring instance of the observable phenomena under study approximates the situation found in a controlled experiment. For example, during the oil crisis of the mid 1970s, President Nixon imposed a 55 mile per hour speed limit as a strategy to reduce gasoline consumption. This action had a variety of consequences from reduced car accidents to the economic impact of longer times for the transportation of goods. In this case, the *status quo ante* served as the control and the imposition of new highway laws became the natural experiment.

Helena, Montana during the six-month period from June 2002 to December 2002 banned smoking ban in all public spaces including bars and restaurants. This becomes the natural experiment with the control groups being Helena before and after the ban or other Montana cities during the ban.

4.4 Case Studies

4.4.1 Observational Studies

Governments and private consortia maintain databases to assist the public and researchers obtain data both for exploratory data analysis and for formal statistical procedures. We present several examples below.

United States Census

The official United States Census is described in Article I, Section 2 of the Constitution of the United States.

> *The actual enumeration shall be made within three years after the first meeting of the Congress of the United States, and within every subsequent term of 10 years, in such manner as they shall by Law direct.*

It calls for an actual enumeration to be used for apportionment of seats in the House of Representatives among the states and is taken in years that are multiples of 10 years.

`http://2010.census.gov/2010census/`

U.S. Census figures are based on actual counts of persons dwelling in U.S. residential structures. They include citizens, non-citizen legal residents, non-citizen long-term visitors, and undocumented immigrants. In recent censuses, estimates of uncounted housed, homeless, and migratory persons have been added to the directly reported figures.

In addition, the Censsu Bureau provides a variety of interactive internet data tools:

`https://www.census.gov/main/www/access.html`

Current Population Survey

The Current Population Survey (CPS) is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years.

`http://www.census.gov/cps/`

Selecting a random sample requires a current database of every household. The random sample is mutistage.

1. Take a sample from the 3000 counties in the United States.
2. Take a sample of townships from each county.
3. Take a sample of blocks from each township.
4. Take a sample of households from each block.

A household is interviewed for 4 successive months, then not interviewed for 8 months, then returned to the sample for 4 months after that. An adult member of each household provides information for all members of the household.

World Heatlh Organization Global Health Observatory (GHO)

The Global Health Observatory is the World Heatlh Organization's internet gateway to health-related statistics. The GHO compiles and verifies major sources of health data to provide easy access to scientifically sound information. GHO covers global health priorities such as the health-related Millennium Development Goals, women and health, mortality and burden of disease, disease outbreaks, and health equity and health systems.

`http://www.who.int/gho/en/`

The Women's Health Initiative

The Women's Health Initiative (WHI) was a major 15-year research program to address the most common causes of death, disability and poor quality of life in postmenopausal women.

`http://www.nhlbi.nih.gov/whi/`

The WHI observational study had several goals. These goals included:

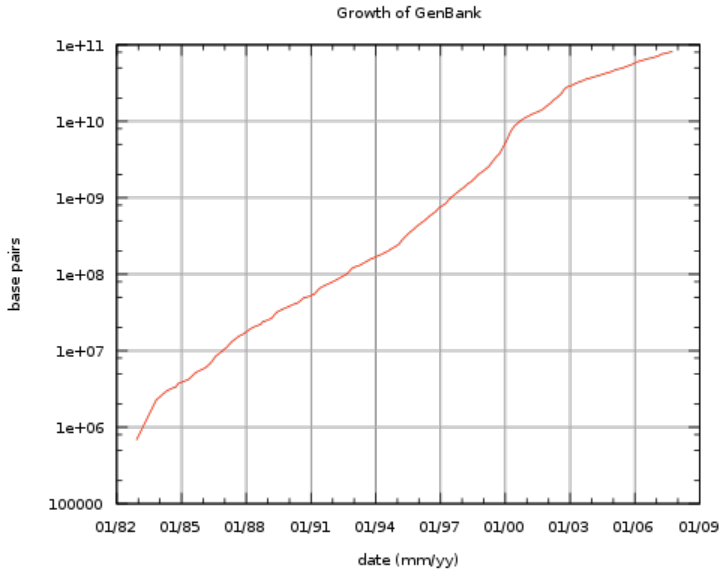
- To give reliable estimates of the extent to which known risk factors to predict heart disease, cancers and fractures.
- To identify "new" risk factors for these and other diseases in women.
- To compare risk factors, presence of disease at the start of the study, and new occurrences of disease during the WHI across all study components.
- To create a future resource to identify biological indicators of disease, especially substances and factors found in blood.

The observational study enlisted 93,676 postmenopausal women between the ages of 50 to 79. The health of participants was tracked over an average of eight years. Women who joined this study filled out periodic health forms and also visited the clinic three years after enrollment. Participants were not required to take any medication or change their health habits.

GenBank

The GenBank sequence database is an open access of nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration, or INSDC. GenBank has approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the whole genome sequence (WGS) division as of April, 2011.

`http://www.ncbi.nlm.nih.gov/genbank/`



4.4.2 Experiments

The history of science has many examples of experiments whose results strongly changed our view of the nature of things. Here we highlight two very important examples.

Light: Its Speed and Medium of Propagation

For many centuries before the seventeenth, a debate continued as to whether light travelled instantaneously or at a finite speed. In ancient Greece, Empedocles maintained that light was something in motion, and therefore must take some time to travel. Aristotle argued, to the contrary, that "light is due to the presence of something, but it is not a movement." Euclid and Ptolemy advanced the emission theory of vision, where light is emitted from the eye. Consequently, Heron of Alexandria argued, the speed of light must be infinite because distant objects such as stars appear immediately upon opening the eyes.

In 1021, Islamic physicist Alhazen (Ibn al-Haytham) published the Book of Optics, in which he used experiments related to the camera obscura to support the now accepted intromission theory of vision, in which light moves from an object into the eye. This led Alhazen to propose that light must therefore have a finite speed. In 1574, the Ottoman astronomer and physicist Taqi al-Din also concluded that the speed of light is finite, correctly explained refraction as the result of light traveling more slowly in denser bodies, and suggested that it would take a long time for light from distant stars to reach the Earth. In the early 17th century, Johannes Kepler believed that the speed of light was infinite since empty space presents no obstacle to it.

In 1638, Galileo Galilei finally proposed an *experiment* to measure the speed of light by observing the delay between uncovering a lantern and its perception some distance away. In 1667, Galileo's experiment was carried out by the Accademia del Cimento of Florence with the lanterns separated by about one mile. No delay was observed. The experiment was not well designed and led to the conclusion that if light travel is not instantaneous, it is very fast. A more powerful experimental design to estimate of the speed of light was made in 1676 by Ole Christensen Romer, one of a group of astronomers of the French Royal Academy of Sciences. From his observations, the periods of Jupiter's innermost moon Io appeared to be shorter when the earth was approaching Jupiter than when receding from it, Romer concluded that light travels at a finite speed, and was able to estimate that would it take light 22 minutes to cross the diameter of Earth's orbit. Christiaan Huygens combined this estimate with an estimate for the diameter of the Earth's orbit to obtain an estimate of speed of light of 220,000 km/s, 26% lower than the actual value.

With the finite speed of light established, nineteenth century physicists, noting that both water and sound waves required a medium for propagation, postulated that the vacuum possessed a "luminiferous aether", the medium for light waves. Because the Earth is in motion, the flow of aether across the Earth should produce a detectable "aether wind". In addition, because the Earth is in orbit about the Sun and the Sun is in motion relative to the center of the Milky Way, the Earth cannot remain at rest with respect to the aether at all times. Thus, by analysing the speed of light in different directions at various times, scientists could measure the motion of the Earth relative to the aether.

In order to detect aether flow, Albert Michelson designed a light interferometer sending a single source of white light through a half-silvered mirror that split the light into two beams travelling at right

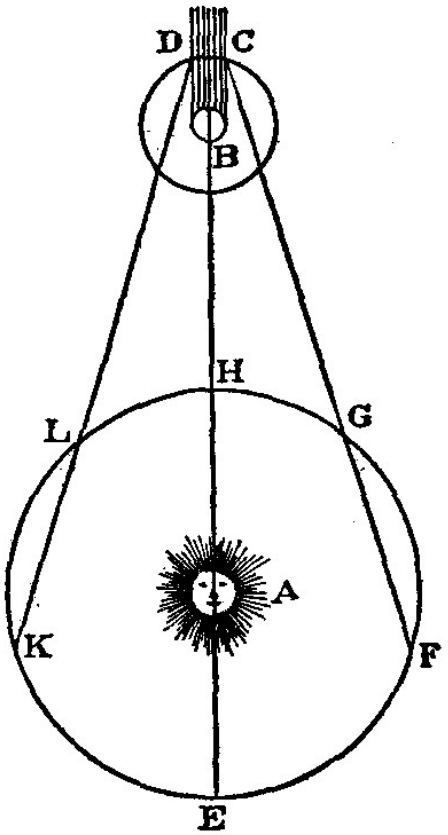


FIG. 70.

Figure 4.1: Romer's diagram of Jupiter (B) eclipsing its moon Io (DC) as viewed from different points in earth's orbit around the sun

angles to one another. The split beams were recombined producing
a pattern of constructive and destructive interference based on the travel time in transit. If the Earth is traveling through aether, a beam reflecting back and forth parallel to the flow of ether would take longer than a beam reflecting perpendicular to the aether because the time gained from traveling with the aether is less than that lost traveling against the ether. The result would be a delay in one of the light beams that could be detected by their interference patterns resulting for the recombined beams. Any slight change in the travel time would then be observed as a shift in the positions of the interference fringes. While Michaelson's prototype apparatus showed promise, it produced far too large experimental errors.

In 1887, Edward Morley joined the effort to create a new device with enough accuracy to detect the aether wind. The new apparatus had a longer path length, it was built on a block of marble, floated in a pool of mercury, and located in a closed room in the basement of a stone building to eliminate most thermal and vibrational effects. The mercury pool allowed the device to be turned, so that it could be rotated through the entire range of possible angles to the hypothesized aether wind. Their results were the first strong evidence against the aether theory and formed a basic contribution to the foundation of the theory of relativity. Thus, two natural questions - how fast does light travel and does it need a medium - awaited elegant and powerful experiments to achieve the understanding we have today and set the stage for the theory of relatively, one of the two great theories of modern physics.

Principles of Inheritance and Genetic Material

Patterns of inheritance have been noticed for millenia. Because of the needs for food, domesticated plants and animals have been bred according to deliberate patterns for at least 5000 years. Progress towards the discovery of the laws for inheritance began with a good set of model organisms. For example, annual flowering plants had certainly been used successfully in the 18th century by Josef Gottlieb Kölreuter. His experimental protocols took the advantage of the fact that these plants are easy to grow, have short generation times, have individuals that possess both male and female reproductive organs, and have easily controlled mating through artificial pollination. Kölreuter established a principle of equal parental contribution. The nature of inheritance remained unknown with a law of blending becoming a leading hypothesis.

In the 1850s and 1860s, the Austrian monk Gregor Mendel used pea plants to work out the basic principles of genetics as we understand them today. Through careful inbreeding, Mendel found 7 true-breeding traits - traits that remained present through many generations and persisted from parent to offspring. By this process, Mendel was sure that potential parent plants were from a true-breeding strain. Mendel's explanatory variables were the traits of the **parental generation**, G. His response variables were the traits of the individual plants in the **first filial generation,** F_1 and **second filial generation,** F_2.

Mendel noted that only one trait was ever expressed in the F_1 generation and called it **dominant**. The alternative trait was called **recessive**. The most striking result is that in the F_2 generation the fraction expressing the dominant trait was very close to 3/4 for each of the seven traits. (See the table below summarizing Mendel's data.) These results in showing *no* intermediate traits disprove the blending hypothesis. Also, the blending theory could not explain the appearance of a pea plant expressing the recessive trait that is the offspring of two plants each expressing the dominant trait. This lead to the hypothesis that each plant has two units of inheritance and transmits one of them to each of its offspring. Mendel could check this hypothesis by crossing, in modern terms, heterozygous plants with those that are dominant homozygous. Mendel went on to examine the situation in which two traits are examined simultaneously and showed that the two traits sort independently. We now use the squares devised in 1905 by Reginald Punnett to compute the probabilities of a particular cross or breeding experiment.

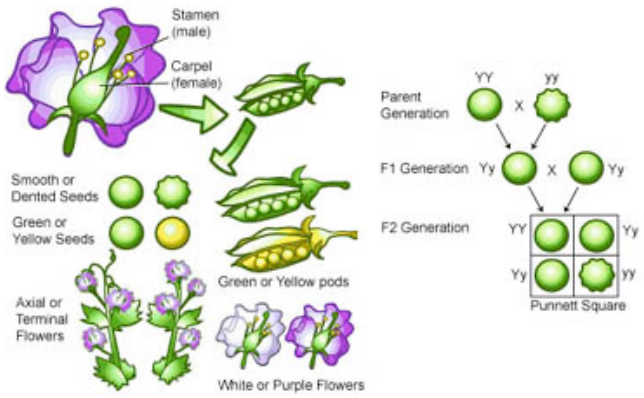


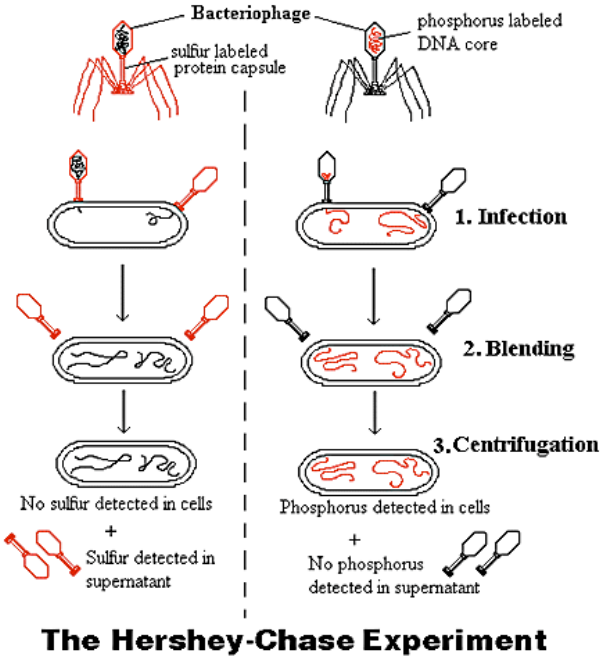
Figure 4.2: Mendel's traits and experiments.

parental phenotypes			F_2 generation phenotypes			
dominant		recessive	dominant	recessive	total	fraction dominant
spherical seeds	×	wrinkled seeds	5474	1850	7324	0.747
yellow seeds	×	green seeds	6022	2001	8023	0.751
purple flowers	×	white flowers	705	224	929	0.758
inflated pods	×	constricted pods	882	299	1181	0.747
green pods	×	yellow pods	428	152	580	0.738
axial flowers	×	terminal flowers	651	207	858	0.759
tall stems	×	dwarf stems	787	277	1064	0.740

We now know that many traits whose expression depends on environment can vary continuously. We can also see that some genes are linked by their position and do not sort independently. (A pea plant has 7 pairs of chromosomes.) The effects can sometimes look like blending. But thanks to Mendel's work, we can see how these expressions are built from the expression of several genes.

Now we know that inheritance is given in "packets". The next question is what material in the living cell is the source of inheritance. Theodor Boveri using sea urchins and Walter Sutton using grasshoppers independently developed the **chromosome theory of inheritance** in 1902. From their work, we know that all the chromosomes had to be present for proper embryonic development and that chromosomes occur in matched pairs of maternal and paternal chromosomes which separate during meiosis. Soon thereafter, Thomas Hunt Morgan, working with the fruit fly *Drosophila melanogaster* as a model system, noticed that a mutation resulting in white eyes was linked to sex - only males had white eyes. Microscopy revealed a dimorphism in the sex chromosome and with this information, Morgan could predict the inheritance of sex linked traits. Morgan continued to learn that genes must reside on a particular chromosomes.

We now think of chromosomes as composed of DNA, but it is in reality an organized structure of DNA and protein. Thus, which of the two formed the inheritance material was in doubt. Phoebus Levene, who identified the components of DNA, declared that it could not store the genetic code because it was chemically far too simple. At that time, DNA was wrongly thought to be made up of regularly repeated tetranucleotides and so could not be the carrier of genetic information. Indeed, in 1944 when Oswald Avery, Colin MacLeod, and Maclyn McCarty found that DNA to be the substance that causes bacterial transformation, the scientific community was reluctant to accept the result despite the care taken in the experiments. These researchers considered several organic molecules - proteins, nucleic acids, carbohydrates, and lipids. In each case, if the DNA was destroyed, the ability to continue heritability ended.



Alfred Hershey and Martha Chase continued the search for the genetic material with an experiment using bacteriophage. This virus that infects bacteria is made up of liitle more than DNA inside a protein shell. The virus introduces material into the bacterium that co-opts the host, producing dozens of viruses that emerge from the lysed bacterium. Their experiment begins with growing one culture of phage in a medium containing radioactive phosphorus (that appears in DNA but not in proteins) and another culture in a medium containing radioactive sulfur (that appears in proteins but not in DNA). Afterwards they agitated the bacteria in a blender to strip away the parts of the virus that did not enter the cell in a way that does minimal damage to the bacteria. They then isolated the bacteria finding that the sulfur separated from the bacteria and that the phosphorus had not. By 1952 when Hershey and Chase confirmed that DNA was the genetic material with

their experiment using bacteriophage, scientists were more prepared to accept the result. This, of course, set the stage for the importance of the dramatic discovery by Watson, Crick, and Franklin of the double helix structure of DNA.

Again, for both of these fundamental discoveries, the principles of inheritance and DNA as the carrier of inheritance information, the experimental design was key. In the second case, we learned that even though Avery, MacLeod, and McCarty had designed their experiment well, they did not, at that time, have a scientific community prepared to acknowledge their findings.

Salk Vaccine Field Trials


All Local Teams Are
Journal American
SALK'S VACCINE WORKS!
Official Count
All Local Teams Are Raised Out

Poliomyelitis, often called polio or infantile paralysis, is an acute viral infectious disease spread from person to person, primarily via the fecal-oral route. The overwhelming majority of polio infections have no symptoms. However, if the virus enters the central nervous system, it can infect motor neurons, leading to symptoms ranging from muscle weakness and paralysis. The effects of polio have been known since prehistory; Egyptian paintings and carvings depict otherwise healthy people with withered limbs, and children walking with canes at a young age. The first US epidemic was in 1916. By 1950, polio had claimed hundreds of thousands of victims, mostly children.

In 1950, the Public Health Service (PHS) organized a field trial of a vaccine developed by Jonas Salk.

Polio is an epidemic disease with

- 60,000 cases in 1952, and
- 30,000 cases in 1953.

So, a low incidence without control could mean

- the vaccine works, or
- no epidemic in 1954.

Some basic facts were known before the trial started:

- Higher income parents are more likely to consent to allow children to take the vaccine.
- Children of lower income parents are thought to be less susceptible to polio. The reasoning is that these children live in less hygienic surroundings and so are more likely to contract very mild polio and consequently more likely to have polio antibodies.

To reduce the role of chance variation dominating the results, the United States Public Health Service (PHS) decided on a study group of two million people. At the same time, a parents advocacy group, the National Foundation for Infantile Paralysis (NFIP) set out its own design. Here are the essential features of the NFIP design:

- Vaccinate all grade 2 children with parental consent.
- Use grades 1 and 3 as controls.

This design fails to have some of essential features of the principles of experimental design. Here is a critique:

- Polio spreads through contact, so infection of one child in a class can spread to the classmates.
- The treatment group is biased towards higher income.

Thus, the treatment group and the control group have several differences beyond the fact that the treatment group receives the vaccine and the control group does not. This leaves the design open to having lurking variables be the primary cause in the differences in outcomes between the treatment and control groups. The Public Health Service design is intended to take into account these shortcomings. Their design has the following features:

- Flip a coin for each child. (randomized control)
- Children in the control group were given an injection of salt water. (placebo)
- Diagnosticians were not told whether a child was in treatment or control group. (double blind)

The results:

	PHS		NFIP	
	Size	Rate	Size	Rate
Treatment	200,000	28	225,000	25
Control	200,000	71	725,000	54
No consent	350,000	46	125,000	44

Rates are per 100,000

We shall learn later that the evidence is overwhelming that the vaccine reduces the risk of contracting polio. As a consequence of the study, universal vaccination was undertaken in the United States in the early 1960s. A global effort to eradicate polio began in 1988, led by the World Health Organization, UNICEF, and The Rotary Foundation. These efforts have reduced the number of annual diagnosed from an estimated 350,000 cases in 1988 to 1,310 cases in 2007. Still, polio persists. The world now has four polio endemic countries - Nigeria, Afghanistan, Pakistan, and India. One goal of the Gates Foundation is to eliminate polio.

The National Foundation for Infantile Paralysis was founded in 1938 by Franklin D. Roosevelt. Roosevelt was diagnosed with polio in 1921, and left him unable to walk. The Foundation is now known as the March of Dimes. The expanded mission of the March of Dimes is to improve the health of babies by preventing birth defects, premature birth and infant mortality. Its initiatives include rubella (German measles) and pertussis (whooping cough) vaccination, maternal and neonatal care, folic acid and spin bifida, fetal alcohol syndrome, newborn screening, birth defects and prematurity.

The INCAP Study

The World Health Organization cites malnutrition as the gravest single threat to the world's public health. Improving nutrition is widely regarded as the most effective form of aid. According to Jean Ziegler (the United Nations Special Rapporteur on the Right to Food from 2000 to 2008) mortality due to malnutrition accounted for 58% of the total mortality in 2006. In that year, more than 36 million died of hunger or diseases due to deficiencies in micronutrients.

Malnutrition is by far the biggest contributor to child mortality, present in half of all cases. Underweight births and inter-uterine growth restrictions cause 2.2 million child deaths a year. Poor or non-existent breastfeeding causes another 1.4 million. Other deficiencies, such as lack of vitamins or minerals, for example, account for 1 million deaths. According to The Lancet, malnutrition in the first two years is irreversible. Malnourished children grow up with worse health and lower educational achievements.

Thus, understanding the root causes of malnutrition and designing remedies is a major global health care imperative. As the next example shows, not every design sufficiently considers the necessary aspects of human behavior to allow for a solid conclusion.



Figure 4.3: The orange ribbon is often used as a symbol to promote awareness of malnutrition,

The Instituto de Nutrición de Centro Americo y Panama (INCAP) conducted a study on the effects of malnutrition. This 1969 study took place in Guatemala and was administered by the World Health Organization, and supported by the United States National Institute of Health.

Growth deficiency is thought to be mainly due to protein deficiency. Here are some basic facts known in advance of the study:

- Guatemalan children eat 2/3 as much as children in the United States.
- Age 7 Guatemalan children are, on average, 5 inches shorter and 11 pounds lighter than children in the United States.

What are the confounding factors that might explain these differences?

- Genetics
- Prevalence of disease
- Standards of hygiene.
- Standards of medical care.

The experimental design: Measure the effects in four very similar Guatemalan villages. Here are the criterion used for the Guatemalan villages chosen for the study..

- The village size is 150 families, 700 inhabitants with 100 under 6 years of age.
- The village is culturally Latino and not Mayan
- Village life consists of raising corn and beans for food and tomatoes for cash.
- Income is approximately $200 for a family of five.
- The literacy rate is approximately 30% for individuals over age 7.

For the experiment:

- Two villages received the treatment, a drink called *atole*, rich in calories and protein.
- Two villages received the control, a drink called *fresca*, low in calories and no protein.
- Both drinks contain missing vitamins and trace elements. The drinks were served at special cafeterias. The amount consumed by each individual was recorded, but *the use of the drinks was unrestricted*.
- Free medical care was provided to compensate for the burden on the villagers.

The lack of control in the amount of the special drink consumed resulted in enormous variation in consumption. In particular, much more fresca was consumed. Consequently, the design fails in that differences beyond the specific treatment and control existed among the four villages.

The researchers were able to salvage some useful information from the data. They found a linear relationship between a child's growth and the amount of protein consumed:

$$\text{child's growth rate} = 0.04 \text{ inches/pound protein}$$

North American children consume an extra 100 pounds of protein by age 7. Thus, the protein accounts for 4 of the 5 inches in the average difference in heights between Latino Guatemalans and Americans.

Part II

Probability

Topic 5

The Basics of Probability

The theory of probability as mathematical discipline can and should be developed from axioms in exactly the same way as Geometry and Algebra. - Andrey Kolmogorov, 1933, Foundations of the Theory of Probability

5.1 Introduction

Mathematical structures like Euclidean geometry or algebraic fields are defined by a set of axioms. "Mathematical reality" is then developed through the introduction of concepts and the proofs of theorems. These axioms are inspired, in the instances introduced above, by our intuitive understanding, for example, of the nature of parallel lines or the real numbers. Probability is a branch of mathematics based on three axioms inspired originally by calculating chances from card and dice games.

Statistics, in its role as a facilitator of science, begins with the collection of data. From this collection, we are asked to make inference on the **state of nature**, that is to determine the conditions that are likely to produce these data. Probability, in undertaking the task of investigating differing states of nature, takes the complementary perspective. It begins by examining **random** phenomena, i.e., those whose exact outcomes are uncertain. Consequently, in order to determine the "scientific reality" behind the data, we must spend some time working with the concepts of the theory of probability to investigate properties of the data arising from the possible states of nature to assess which are most useful in making inference.

We will motivate the axioms of probability through the case of equally likely outcomes for some simple games of chance and look at some of the direct consequences of the axioms. In order to extend our ability to use the axioms, we will learn counting techniques, e.g, permutations and combinations, based on the **fundamental principle of counting**.

A **probability model** has two essential pieces of its description.

- Ω, the **sample space**, the set of possible outcomes.
  - An **event** is a collection of **outcomes**. We can define an event by explicitly giving its outcomes,

    $$A = \{\omega_1, \omega_2, \cdots, \omega_n\}$$

    or with a description

    $$A = \{\omega; \omega \text{ has property } \mathcal{P}\}.$$

    In either case, A is subset of the sample space, $A \subset \Omega$.
- P, the **probability** assigns a number to each event.

Thus, a probability is a function. We are familiar with functions in which both the domain and range are subsets of the real numbers. The domain of a probability function is the collection of all events. The range is still a number. We will see soon which numbers we will accept as probabilities of events.

You may recognize these concepts from a basic introduction to sets. In talking about sets, we use the term **universal set** instead of sample space, **element** instead of outcome, and **subset** instead of event. At first, having two words for the same concept seems unnecessarily redundant. However, we will later consider more complex situations which will combine ideas from sets and from probability. In these cases, having two expression for a concept will facilitate our understanding. A *Set Theory - Probability Theory Dictionary* is included at the end of this topic to relate to the new probability terms with the more familiar set theory terms.

5.2 Equally Likely Outcomes and the Axioms of Probability

The essential relationship between events and the probability are described through the three **axioms of probability**. These axioms can be motivated through the first uses of probability, namely the case of equal likely outcomes.

If Ω is a finite sample space, then if each outcome is equally likely, we define the probability of A as the fraction of outcomes that are in A. Using $\#(A)$ to indicate the number of elements in an event A, this leads to a simple formula

$$P(A) = \frac{\#(A)}{\#(\Omega)}.$$

Thus, computing $P(A)$ means counting the number of outcomes in the event A and the number of outcomes in the sample space Ω and dividing.

Exercise 5.1. *Find the probabilities under equal likely outcomes.*

(a) Toss a coin.

$$P\{\text{heads}\} = \frac{\#(A)}{\#(\Omega)} = \text{——}.$$

(b) Toss a coin three times.

$$P\{\text{toss at least two heads in a row}\} = \frac{\#(A)}{\#(\Omega)} = \text{——}$$

(c) Roll two dice.

$$P\{\text{sum is 7}\} = \frac{\#(A)}{\#(\Omega)} = \text{——}$$

Because we always have $0 \leq \#(A) \leq \#(\Omega)$, we always have

$$P(A) \geq 0 \tag{5.1}$$

and

$$P(\Omega) = 1 \tag{5.2}$$

This gives us 2 of the three axioms. The third will require more development.

Toss a coin 4 times. $\#(\Omega) = 16$

A = {exactly 3 heads} = {HHHT, HHTH, HTHH, THHH} $\#(A) = 4$

$$P(A) = \frac{4}{16} = \frac{1}{4}$$

B = {exactly 4 heads} = {HHHH} $\#(B) = 1$

$$P(B) = \frac{1}{16}$$

Now let's define the set C = {at least three heads}. If you are asked the supply the probability of C, your intuition is likely to give you an immediate answer.

$$P(C) = \frac{5}{16}.$$

Let's have a look at this intuition. The events A and B have no outcomes in common,. We say that the two events are **disjoint** or **mutually exclusive** and write $A \cap B = \emptyset$. In this situation,

$$\#(A \cup B) = \#(A) + \#(B).$$

If we take this **addition principle** and divide by $\#(\Omega)$, then we obtain the following identity:

If $A \cap B = \emptyset$, then

$$\frac{\#(A \cup B)}{\#(\Omega)} = \frac{\#(A)}{\#(\Omega)} + \frac{\#(B)}{\#(\Omega)}.$$

or

$$P(A \cup B) = P(A) + P(B). \tag{5.3}$$

Using this property, we see that

$$P\{\text{at least 3 heads}\} = P\{\text{exactly 3 heads}\} + P\{\text{exactly 4 heads}\} = \frac{4}{16} + \frac{1}{16} = \frac{5}{16}.$$

We are saying that any function P that accepts events as its domain and returns numbers as its range and satisfies Axioms 1, 2, and 3 as defined in (5.1), (5.2), and (5.3) can be called a **probability**.

If we iterate the procedure in Axiom 3, we can also state that if the events, $A_1, A_2, \cdots, A_n$, are mutually exclusive, then

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n). \tag{5.3$'$}$$

This is a sufficient definition for a probability if the sample space Ω is finite. However, we will want to examine infinite sample spaces and to use the idea of limits. This introduction of limits is the pathway that allows to bring in calculus with all of its powerful theory and techniques as a tool in the development of the theory of probability.

Example 5.2. *For the random experiment, consider a rare event - a lightning strike at a given location, winning the lottery, finding a planet with life - and look for this event repeatedly until it occurs, we can write*

$$A_j = \{\text{the first occurrence appears on the } j\text{-th observation}\}.$$

Then, each of the A_j are mutually exclusive and

$$\{\text{event occurs eventually}\} = A_1 \cup A_2 \cup \cdots \cup A_n \cup \cdots = \bigcup_{j=1}^{\infty} A_j = \{\omega; \omega \in A_j \text{ for some } j\}.$$

We would like to say that

$$P\{\text{event occurs ventually}\} = P(A_1) + P(A_2) + \cdots + P(A_n) + \cdots = \sum_{j=1}^{\infty} P(A_j) = \lim_{n\to\infty} \sum_{j=1}^{n} P(A_j).$$

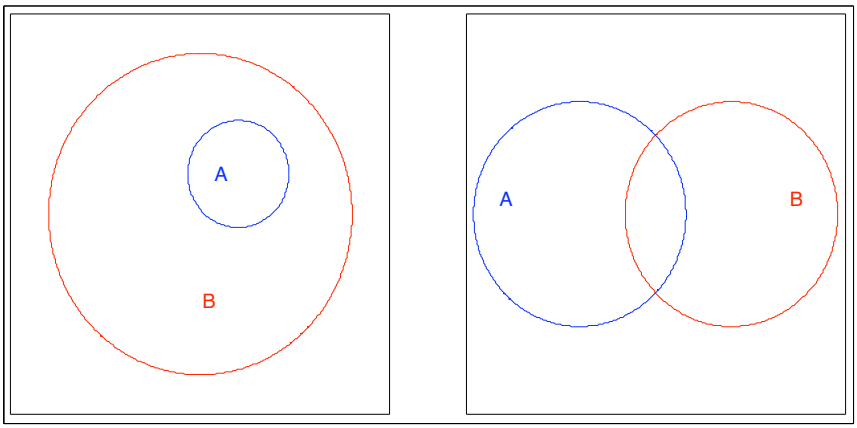


Figure 5.1: (left) **Difference** and **Monotonicity Rule**. If $A \subset B$, then $P(B \setminus A) = P(B) - P(A)$. (right) **The Inclusion-Exclusion Rule**. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Using area as an analogy for probability, $P(B \setminus A)$ is the area between the circles and the area $P(A) + P(B)$ double counts the lens shaped area $P(A \cap B)$.

This would call for an extension of Axiom 3 to an infinite number of mutually exclusive events. This is the general version of Axiom 3 we use when we want to use calculus in the theory of probability:

For mutually exclusive events, $\{A_j; j \geq 1\}$, then

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j) \tag{5.3''}$$

Thus, statements (5.1), (5.2), and (5.3") give us the complete axioms of probability.

5.3 Consequences of the Axioms

Other properties that we associate with a probability can be derived from the axioms.

1. **The Complement Rule**. Because A and its **complement** $A^c = \{\omega; \omega \notin A\}$ are mutually exclusive

   $$P(A) + P(A^c) = P(A \cup A^c) = P(\Omega) = 1$$

   or

   $$P(A^c) = 1 - P(A).$$

   For example, if we toss a *biased* coin. We may want to say that $P\{\text{heads}\} = p$ where p is not necessarily equal to 1/2. By necessity,

   $$P\{\text{tails}\} = 1 - p.$$

   Example 5.3. *Toss a coin 4 times.*

   $$P\{\text{fewer than 3 heads}\} = 1 - P\{\text{at least 3 heads}\} = 1 - \frac{5}{16} = \frac{11}{16}.$$

2. **The Difference Rule**. Write $B \setminus A$ to denote the outcomes that are in B but **not** in A. If $A \subset B$, then

   $$P(B \setminus A) = P(B) - P(A).$$

   (The symbol $\subset$ denotes "contains in". A and $B \setminus A$ are mutually exclusive and their union is B. Thus $P(B) = P(A) + P(B \setminus A)$.) See Figure 5.1 (left).

Exercise 5.4. *Give an example for which* $P(B \setminus A) \neq P(B) - P(A)$

Because $P(B \setminus A) \geq 0$, we have the following:

3. **Monotonicity Rule**. If $A \subset B$, then $P(A) \leq P(B)$
   We already know that for any event A, $P(A) \geq 0$. The monotonicity rule adds to this the fact that
   $$P(A) \leq P(\Omega) = 1.$$
   Thus, the range of a probability is a subset of the interval $[0, 1]$.

4. **The Inclusion-Exclusion Rule**. For any two events A and B,
   $$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{5.4}.$$
   ($P(A) + P(B)$ accounts for the outcomes in $A \cap B$ twice, so remove $P(A \cap B)$.) See Figure 5.1 (right).

Exercise 5.5. *Show that the inclusion-exclusion rule follows from the axioms. Hint:* $A \cup B = (A \cap B^c) \cup B$ *and* $A = (A \cap B^c) \cup (A \cap B)$.

Exercise 5.6. *Give a generalization of the inclusion-exclusion rule for three events.*

Deal two cards.

A = {ace on the second card}, B = {ace on the first card}

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P\{\text{at least one ace}\} = \frac{1}{13} + \frac{1}{13} - \quad ?$$

To complete this computation, we will need to compute $P(A \cap B) = P\{\text{both cards are aces}\} = \frac{\#(A \cap B)}{\#(\Omega)}$

We will learn a strategy for this when we learn the fundamental principles of counting. We will also learn a simpler strategy in the next topic where we learn about conditional probabilities.

5. **The Bonferroni Inequality.** For any two events A and B,
   $$P(A \cup B) \leq P(A) + P(B).$$

6. **Continuity Property.** If events satisfy
   $$B_1 \subset B_2 \subset \cdots \text{ and } B = \bigcup_{i=1}^{\infty} B_i$$
   Then, by the monotonicity rule, $P(B_i)$ is an increasing sequence. In addition, they satisfy
   $$P(B) = \lim_{i \to \infty} P(B_i). \tag{5.5}$$
   Similarly, use the symbol $\supset$ to denote "contains". If events satisfy
   $$C_1 \supset C_2 \supset \cdots \text{ and } C = \bigcap_{i=1}^{\infty} C_i$$
   Again, by the monotonicity rule, $P(C_i)$ is a decreasing sequence. In addition, they satisfying
   $$P(C) = \lim_{i \to \infty} P(C_i). \tag{5.6}$$

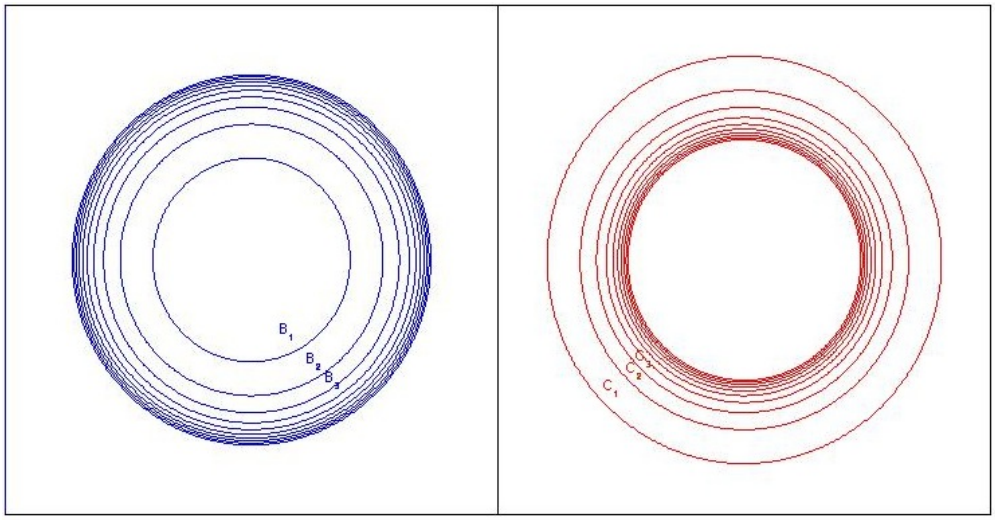


Figure 5.2: Continuity Property. (left) B_i increasing to an event B. Here, equation (5.5) is satisfied. (right) C_i decreasing to an event C. Here, equation (5.6) is satisfied.

Exercise 5.7. *Establish the continuity property.* Hint: *For the first, let* $A_1 = B_1$ *and* $A_i = B_i \setminus B_{i-1},\ i > 1$ *in axiom (5.3"). For the second, use the complement rule and* **de Morgan's law**

$$C^c = \bigcup_{i=1}^{\infty} C_i^c$$

Exercise 5.8 (odds)**.** *The statement of* $a : b$ *odds for an event* A *indicates that*

$$\frac{P(A)}{P(A^c)} = \frac{a}{b}$$

Show that

$$P(A) = \frac{a}{a+b}.$$

So, for example, $1 : 2$ *odds means* $P(A) = 1/3$ *and* $5 : 3$ *odds means* $P(A) = 5/8$.

5.4 Counting

In the case of equally likely outcomes, finding the probability of an event A is the result of two counting problems - namely finding $\#(A)$, the number of outcomes in A and finding $\#(\Omega)$, the number of outcomes in the sample space. These counting problems can become quite challenging and many advanced mathematical techniques have been developed to address these issues. However, having some facility in counting is necessary to have a sufficiently rich number of examples to give meaning to the axioms of probability. Consequently, we shall develop a few counting techniques leading to the concepts of **permutations** and **combinations**.

5.4.1 Fundamental Principle of Counting

We start with the **fundamental principle of counting**.

Suppose that two experiments are to be performed.

- Experiment 1 can have n_1 possible outcomes and
- for each outcome of experiment 1, experiment 2 has n_2 possible outcomes.

Then together there are $n_1 \times n_2$ possible outcomes.

Example 5.9. *For a group of n individuals, one is chosen to become the president and a second is chosen to become the treasurer. By the multiplication principle, if these position are held by different individuals, then this task can be accomplished in*

$$n \times (n-1)$$

possible ways

Exercise 5.10. *Find the number of ways to draw two cards and the number of ways to draw two aces.*

Exercise 5.11. *Generalize the fundamental principle of counting to k experiments.*

Assume that we have a collection of n objects and we wish to make an **ordered arrangement** of k of these objects. Using the generalized multiplication principle, the number of possible outcomes is

$$n \times (n-1) \times \cdots \times (n-k+1).$$

We will write this as $(n)_k$ and say n **falling** k. Note that this symbol, used here for the *falling factorial* is also used as a symbol for a **rising factorial** , also referred to as the **Pochhammer symbol**.

5.4.2 Permutations

Example 5.12 (birthday problem)**.** *In a list the birthday of k people, there are 365^k possible lists (ignoring leap year day births) and*

$$(365)_k$$

possible lists with no date written twice. Thus, the probability, under equally likely outcomes, that no two people on the list have the same birthday is

$$\frac{(365)_k}{365^k} = \frac{365 \cdot 364 \cdots (365-k+1)}{365^k}$$

and, by the complement rule,

$$P\{\text{at least one pair of individuals share a birthday}\} = 1 - \frac{(365)_k}{365^k} \tag{5.1}$$

Here is a short table of these probabilities. A graph is given in Figure 5.3.

k	5	10	15	18	20	22	23	25	30	40	50	100
probability	0.027	0.117	0.253	0.347	0.411	0.476	0.507	0.569	0.706	0.891	0.970	0.994

The R *code and output follows. We can create an iterative process by noting that*

$$\frac{(365)_k}{365^k} = \frac{(365)_{k-1}}{365^{k-1}} \frac{(365-k+1)}{365}$$

Thus, we can find the probability that no pair in a group of k individuals has the same birthday by taking the probability that no pair in a group of $k-1$ individuals has the same birthday and multiplying by $(365-k+1)/365$. Here is the output for $k = 1$ to 45.

```
> prob=rep(1,45)
> for (k in 2:45){prob[k]=prob[k-1]*(365-k+1)/365}
> data.frame(c(1:15),1-prob[1:15],c(16:30),1-prob[16:30],c(31:45),1-prob[31:45])
```

and the output

	c.1.15.	X1...prob.1.15.	c.16.30.	X1...prob.16.30.	c.31.45.	X1...prob.31.45.
1	1	0.000000000	16	0.2836040	31	0.7304546
2	2	0.002739726	17	0.3150077	32	0.7533475
3	3	0.008204166	18	0.3469114	33	0.7749719
4	4	0.016355912	19	0.3791185	34	0.7953169
5	5	0.027135574	20	0.4114384	35	0.8143832
6	6	0.040462484	21	0.4436883	36	0.8321821
7	7	0.056235703	22	0.4756953	37	0.8487340
8	8	0.074335292	23	0.5072972	38	0.8640678
9	9	0.094623834	24	0.5383443	39	0.8782197
10	10	0.116948178	25	0.5686997	40	0.8912318
11	11	0.141141378	26	0.5982408	41	0.9031516
12	12	0.167024789	27	0.6268593	42	0.9140305
13	13	0.194410275	28	0.6544615	43	0.9239229
14	14	0.223102512	29	0.6809685	44	0.9328854
15	15	0.252901320	30	0.7063162	45	0.9409759

Definition 5.13. *The number of ordered arrangements of all n objects (also called permutations) is*

$$(n)_n = n \times (n-1) \times \cdots \times 1 = n!,$$

n **factorial**. *We take* $0! = 1$

Exercise 5.14.

$$(n)_k = \frac{n!}{(n-k)!}.$$

5.4.3 Combinations

Write

$$\binom{n}{k}$$

for the number of different groups of k objects that can be chosen from a collection of size n.

We will next find a formula for this number by counting the number of possible outcomes in two different ways. To introduce this with a concrete example, suppose 3 cities will be chosen out of 8 under consideration for a vacation. If we think of the vacation as visiting three cities in a particular **order**, for example,

New York then Boston then Montreal.

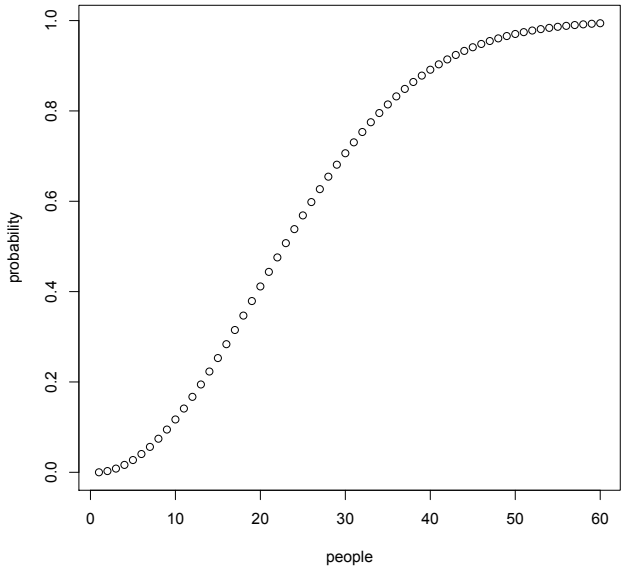


Figure 5.3: The Birthday Problem. For a room of containing k individuals. Using (5.1), a plot of k versus $P_k\{\text{at least one pair of individuals share a birthday}\}$.

Then we are counting the number of ordered arrangements. This results in

$$(8)_3 = 8 \cdot 7 \cdot 6$$

choices.

If we are just considering the 3 cities we visit, irrespective of order, then these **unordered** choices are combinations. The number of ways of doing this is written

$$\binom{8}{3},$$

a number that we do not yet know how to determine. After we have chosen the three cities, we will also have to pick an order to see the cities and so using the fundamental principle of counting, we have

$$\binom{8}{3} \times 3 \cdot 2 \cdot 1 = \binom{8}{3} 3!$$

possible vacations if the order of the cities is included in the choice.

These two strategies are counting the same possible outcomes and so must be equal.

$$(8)_3 = 8 \cdot 7 \cdot 6 = \binom{8}{3} \times 3 \cdot 2 \cdot 1 = \binom{8}{3} 3! \quad \text{or} \quad \binom{8}{3} = \frac{8 \cdot 7 \cdot 6}{3 \cdot 2 \cdot 1} = \frac{(8)_3}{3!}.$$

Thus, we have a formula for $\binom{8}{3}$. Let's do this more generally.

Theorem 5.15.

$$\binom{n}{k} = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!}.$$

The second equality follows from the previous exercise.

The number of ordered arrangements of k objects out of n is

$$(n)_k = n \times (n-2) \times \cdots \times (n-k+1).$$

Alternatively, we can form an ordered arrangement of k objects from a collection of n by:

1. First choosing a group of k objects.
   The number of possible outcomes for this experiment is $\binom{n}{k}$.
2. Then, arranging this k objects in order.
   The number of possible outcomes for this experiment is $k!$.

So, by the fundamental principle of counting,

$$(n)_k = \binom{n}{k} \times k!.$$

Now complete the argument by dividing both sides by $k!$.

Exercise 5.16 (binomial theorem)**.**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Exercise 5.17. *Verify the identities*

$$\binom{n}{1} = \binom{n}{n-1} = n \quad \textit{and} \quad \binom{n}{k} = \binom{n}{n-k}.$$

Thus, we set

$$\binom{n}{n} = \binom{n}{0} = 1.$$

The number of combinations is computed in R using `choose`. In the vacation example above, $\binom{8}{3}$ is determined by entering

```
> choose(8,3)
[1] 56
```

Theorem 5.18 (Pascal's triangle)**.**

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

To see this using the example on vacations,

$$\binom{8}{3} = \binom{7}{2} + \binom{7}{3}.$$

Assume that New York is one of 8 vacation cities. Then of the $\binom{8}{3}$ possible vacations, Then of the $\binom{8}{3}$ vacations, if New York is on the list, then we must choose the remaining 2 cities from the remaining 7. If New York in *not* on the list, then all 3 choices must be from the remaining 7. Because New York is either on the list or off the list, but never both, the two types of choices have no overlap.

To establish this identity in general, distinguish one of the n objects in the collection. Say that we are looking at a collection of n marbles, $n-1$ are blue and 1 is red.

1. For outcomes in which the red marble is chosen, we must choose $k-1$ marbles from the $n-1$ blue marbles. (The red marble is the k-th choice.) Thus, $\binom{n-1}{k-1}$ different outcomes have the red marble.

2. If the red marble is not chosen, then we must choose k blue marbles. Thus, $\binom{n-1}{k}$ outcomes do not have the red marbles.

3. These choices of groups of k marbles have no overlap. And so $\binom{n}{k}$ is the sum of the values in 1 and 2.

This gives us an iterative way to compute the values of $\binom{n}{k}$. Let's build a table of values for n (vertically) and $k \leq n$ (horizontally). Then, by the Pascal's triangle formula, a given table entry is the sum of the number directly above it and the number above and one column to the left. We can get started by noting that $\binom{n}{0} = \binom{n}{n} = 1$.

Pascal's triangle

	$k-1$	k	
$n-1$	$\binom{n-1}{k-1}$	$\binom{n-1}{k}$	← the sum of these two numbers
n		$\binom{n}{k}$	← equals this number

		k								
		0	1	2	3	4	5	6	7	8
	0	1								
	1	1	1							
	2	1	2	1						
	3	1	3	3	1					
n	4	1	4	6	4	1				
	5	1	5	10	10	5	1			
	6	1	6	15	20	15	6	1		
	7	1	7	21	35	35	21	7	1	
	8	1	8	28	56	70	56	28	8	1

Example 5.19. *For the experiment on honey bee queen - if we rear 60 of the 90 queen eggs, the we have*

```
> choose(90,60)
[1] 6.73133e+23
```

more than 10^{23} *different possible simple random samples.*

Example 5.20. *Deal out three cards. There are*

$$\binom{52}{3}$$

possible outcomes. Let x *be the number of hearts. Then we have chosen* x *hearts out of* 13 *and* $3-x$ *cards that are not hearts out of the remaining 39. Thus, by the multiplication principle there are*

$$\binom{13}{x} \cdot \binom{39}{3-x}$$

possible outcomes.

If we assume equally likely outcomes, the probability of x hearts is the ratio of these two numbers. To compute these numbers in R *for $x = 0, 1, 2, 3$, the possible values for x, we enter*

```
> x<-c(0:3)
> prob<-choose(13,x)*choose(39,3-x)/choose(52,3)
> data.frame(x,prob)
  x       prob
1 0 0.41352941
2 1 0.43588235
3 2 0.13764706
4 3 0.01294118
```

Notice that

```
> sum(prob)
[1] 1
```

We can simulate this activity repeatedly and see how it matches the result of the computation. First, we create a vector of length 52 - `c(rep(1,13),rep(0,39))`*. The 13 ones for the hearts and 39 zeros for the other cards. We can sample 3 with the* `sample` *command and use the* `sum` *command to add the ones (and hence the hearts). The first line in* R *below creates a space for our simulation and the* `table` *allows for a quick display of the results.*

```
> hearts<-numeric(10000)
> for (i in 1:10000){hearts[i]<-sum(sample(c(rep(1,13),rep(0,39)),3))}
> table(hearts)
hearts
   0    1    2    3
4133 4382 1335  150
```

Now divide by 10,000 to see that these simulated probabilities match closely the computed values above. Because it is a random simulation, we will have different results with a second simulation

Alternatively, we can use the `replicate` command.

```
> hearts<-replicate(10000,sum(sample(c(rep(1,13),rep(0,39)),3)))
> table(hearts)
hearts
   0    1    2    3
4164 4284 1414  138
```

Exercise 5.21. *Deal out 5 cards. Let x be the number of fours. What values can x take? Find the probability of x fours for each possible value. Repeat this with 6 cards. Compare your answers to a simulation.*

5.5 Answers to Selected Exercises

5.1. (a) 1/2, (b) 3/8, (c) 6/36 = 1/6

5.3. Toss a coin 6 times. Let $A = \{\text{at least 3 heads}\}$ and Let $B = \{\text{at least 3 tails}\}$. Then

$$P(A) = P(B) = \frac{42}{64} = \frac{21}{32}.$$

Thus, $P(B) - P(A) = 0$. However, the event

$$B \setminus A = \{\text{exactly 3 tails}\} = \{\text{exactly 3 heads}\}$$

and $P(B \setminus A) = 20/64 = 5/16 \neq 0$.

5.5. Using the hint, we have that

$$\begin{aligned} P(A \cup B) &= P(A \cap B^c) + P(B) \\ P(A) &= P(A \cap B^c) + P(A \cup B) \end{aligned}$$

Subtract these two equations

$$P(A \cup B) - P(A) = P(B) - P(A \cup B).$$

Now add $P(A)$ to both sides of the equation to obtain (5.4).

5.6. Use the associativity property of unions to write $A \cup B \cup C = (A \cup B) \cup C$ and use (5.4), the inclusion-exclusion property for the 2 events $A \cup B$ and C and then to the 2 events A and B,

$$\begin{aligned} P((A \cup B) \cup C) &= P(A \cup B) + P(C) - P((A \cup B) \cap C) \\ &= (P(A) + P(B) - P(A \cap B)) + P(C) - P((A \cap C) \cup (B \cap C)) \end{aligned}$$

For the final expression, we use one of De Morgan's Laws. Now rearrange the other terms and apply inclusion-exlcusion to the final expression.

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) - P(A \cap B) + P(C) - P((A \cap C) \cup (B \cap C)) \\ &= P(A) + P(B) + P(C) - P(A \cap B) - (P(A \cap C) + P(B \cap C) - P((A \cap C) \cap (B \cap C))) \\ &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

The last expression uses the identity $(A \cap C) \cap (B \cap C)) = A \cap B \cap C$.

5.7. Using the hint and writing $B_0 = \emptyset$, we have that $P(A_i) = P(B_i) - P(B_{i-1})$ and that

$$\bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i$$

Because the A_i are disjoint, we have by (5.3')

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} B_i\right) &= P\left(\bigcup_{i=1}^{n} A_i\right) \\ &= P(A_n) + P(A_{n-1}) + \cdots + P(A_2) + P(A_1) \\ &= (P(B_n) - P(B_{n-1})) + (P(B_{n-1}) - P(B_{n-2})) + \cdots + (P(B_2) - P(B_1)) + (P(B_1) - P(B_0)) \\ &= P(B_n) - (P(B_{n-1}) - (P(B_{n-1})) - P(B_{n-2})) + \cdots + P(B_2) - (P(B_1) - (P(B_1)) - P(\emptyset) \\ &= P(B_n) \end{aligned}$$

because all of the other terms cancel. This is an example of a **telescoping sum**. Now use (5.3") to obtain

$$P\left(\bigcup_{i=1}^{\infty} B_i\right) = \lim_{n \to \infty} P(B_n).$$

For the second part. Write $B_i = C_i^c$. Then, the B_i satisfy the required conditions and that $B = C^c$. Thus,

$$1 - P(C) = P(C^c) = \lim_{i \to \infty} P(C_i^c) = \lim_{i \to \infty} (1 - P(C_i)) = 1 - \lim_{i \to \infty} P(C_i)$$

and

$$P(C) = \lim_{i \to \infty} P(C_i)$$

5.8. If

$$\frac{a}{b} = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}.$$

Then,

$$a - aP(A) = bP(A), \quad a = (a+b)P(A), \quad P(A) = \frac{a}{a+b}.$$

5.10. The number of ways to obtain two cards is $52 \cdot 51$. The number of ways to obtain two aces is $4 \cdot 3$.

5.11. Suppose that k experiments are to be performed and experiment i can have n_i possible outcomes irrespective of the outcomes on the other $k-1$ experiments. Then together there are $n_1 \times n_2 \times \cdots \times n_k$ possible outcomes.

5.14.

$$\begin{aligned}(n)_k &= n \times (n-1) \times \cdots \times (n-k+1) \times \frac{(n-k)!}{(n-k)!} \\ &= \frac{n \times (n-1) \times \cdots \times (n-k+1)(n-k)!}{(n-k)!} \\ &= \frac{n!}{(n-k)!}.\end{aligned}$$

5.15. Expansion of $(x+y)^n = (x+y) \times (x+y) \times \cdots \times (x+y)$ will result in 2^n terms. Each of the terms is achieved by one choice of x or y from each of the factors in the product $(x+y)^n$. Each one of these terms will thus be a result in n factors - some of them x and the rest of them y. For a given k from $0, 1, \ldots, n$, we will see choices that will result in k factors of x and $n-k$ factors of y, i. e., $x^k y^{n-k}$. The number of such choices is the combination

$$\binom{n}{k}$$

Add these terms together to obtain

$$\binom{n}{k} x^k y^{n-k}.$$

Next adding these values over the possible choices for k results in

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

5.17. The formulas are easy to work out. One way to consider

$$\binom{n}{1} = \binom{n}{n-1}$$

is to note that $\binom{n}{1}$ is the number of ways to choose 1 out of a possible n. This is the same as $\binom{n}{n-1}$, the number of ways to exclude 1 out of a possible n. A similar reasoning gives

$$\binom{n}{k} = \binom{n}{n-k}.$$

(Replace 1 by k in the argument above.)

5.21. The possible values for x are 0, 1, 2, 3, and 4. When we have chosen x fours out of 4, we also have $5 - x$ cards that are not fours out of the remaining 48. Thus, by the multiplication principle, the probability of x fours is

$$\frac{\binom{4}{x} \cdot \binom{48}{5-x}}{\binom{52}{5}}.$$

Similarly for 6 cards, the probability of x fours is

$$\frac{\binom{4}{x} \cdot \binom{48}{6-x}}{\binom{52}{6}}.$$

To compute the numerical values for the probability of x fours:

```
> x<-c(0:4)
> prob5<-choose(4,x)*choose(48,5-x)/choose(52,5)
> sum(prob5)
[1] 1
> prob6<-choose(4,x)*choose(48,6-x)/choose(52,6)
> sum(prob6)
[1] 1
> data.frame(x,prob5,prob6)
  x        prob5        prob6
1 0 6.588420e-01 6.027703e-01
2 1 2.994736e-01 3.364300e-01
3 2 3.992982e-02 5.734602e-02
4 3 1.736079e-03 3.398282e-03
5 4 1.846893e-05 5.540678e-05
```

For the simulation of 5 cards,

```
> fours<-replicate(10000,sum(sample(c(rep(1,4),rep(0,48)),5)))
> table(fours)
fours
   0    1    2    3    4
6596 3024  365   14    1
```

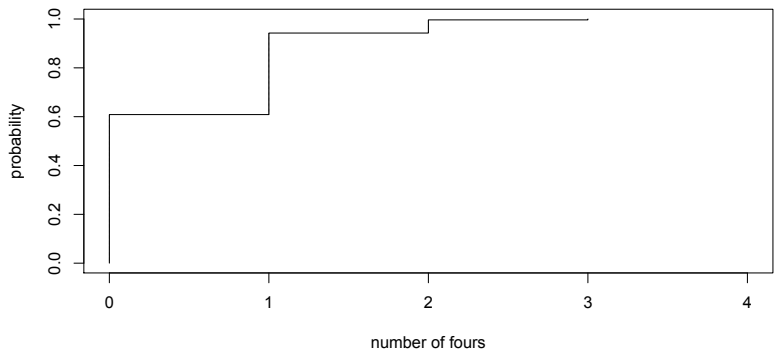


Figure 5.4: Empirical cumulative distribution functions for the number of fourss in a draw of 6 cards.

The simulated values are a bit lower than computed values for 0 fours (0.6522 vs 0.6558), and a bit higher for 1, 2 and 3 fours. Notice that 4 fours appeared only once. For 6 cards,

```
> fours<-replicate(10000,sum(sample(c(rep(1,4),rep(0,48)),6)))
> table(fours)
fours
   0    1    2    3
6035 3382  543   40
```

The simulated values are a bit higher than computed values for 0 fours (0.6086 vs 0.6027), a bit lower for 1, 2, and 3 fours. In this simulation, 4 fours never appeared.

We can add an empirical cumulative distribution functions for the number of fours in a draw of 6 cards.

```
> plot(sort(fours),1:length(fours)/length(fours),xlim=c(0,4),ylim=c(0,1),
   xlab=c("number of fours"),ylab=c("probability"),type="s")
```

5.6 Set Theory - Probability Theory Dictionary

Event Language	Set Language	Set Notation
sample space	universal set	Ω
event	subset	$A, B, C, \cdots$
outcome	element	ω
impossible event	empty set	$\emptyset$
not A	A complement	A^c
A or B	A union B	$A \cup B$
A and B	A intersect B	$A \cap B$
A and B are mutually exclusive	A and B are disjoint	$A \cap B = \emptyset$
if A then B	A is a subset of B	$A \subset B$

Topic 6

Conditional Probability and Independence

Under Bayes' theorem, no theory is perfect. Rather, it is a work in progress, always subject to further refinement and testing. - Nate Silver

One of the most important concepts in the theory of probability is based on the question: *How do we modify the probability of an event in light of the fact that something new is known?* What is the chance that we will win the game now that we have taken the first point? What is the chance that I am a carrier of a genetic disease now that my first child does not have the genetic condition? What is the chance that a child smokes if the household has two parents who smoke? This question leads us to the concept of **conditional probability.**

6.1 Restricting the Sample Space - Conditional Probability

Toss a fair coin 3 times. Let winning be "at least two heads out of three"

HHH	HHT	HTH	HTT
THH	THT	TTH	TTT

Figure 6.1: Outcomes on three tosses of a coin, with the winning event indicated.

If we now know that the first coin toss is heads, then only the top row is possible and we would like to say that the probability of winning is

$$\frac{\#(\text{outcomes that result in a win and also have a heads on the first coin toss})}{\#(\text{outcomes with heads on the first coin toss})}$$
$$= \frac{\#\{\text{HHH, HHT, HTH}\}}{\#\{\text{HHH, HHT, HTH, HTT}\}} = \frac{3}{4}.$$

We can take this idea to create a formula in the case of equally likely outcomes for the statement the **conditional probability of** A **given** B.

$$\begin{aligned} P(A|B) &= \text{the proportion of outcomes in } A \text{ that are also in } B \\ &= \frac{\#(A \cap B)}{\#(B)} \end{aligned}$$

We can turn this into a more general statement using only the probability, P, by dividing both the numerator and the denominator in this fraction by $\#(\Omega)$.

$$P(A|B) = \frac{\#(A \cap B)/\#(\Omega)}{\#(B)/\#(\Omega)} = \frac{P(A \cap B)}{P(B)} \tag{6.1}$$

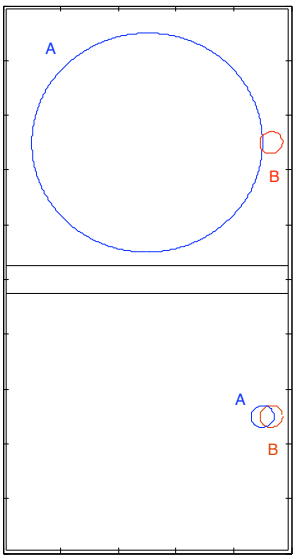


Figure 6.2: Two Venn diagrams to illustrate conditional probability. For the top diagram $P(A)$ is large but $P(A|B)$ is small. For the bottom diagram $P(A)$ is small but $P(A|B)$ is large.

We thus take this version (6.1) of the identity as the general definition of conditional probability for any pair of events A and B as long as the denominator $P(B) > 0$.

Exercise 6.1. *Pick an event B so that $P(B) > 0$. Define, for every event A,*

$$Q(A) = P(A|B).$$

Show that Q satisfies the three axioms of a probability. In words, a conditional probability is a probability.

Exercise 6.2. *Roll two dice. Find $P\{$sum is 8$|$first die shows 3$\}$, and $P\{$sum is 8$|$first die shows 1$\}$*

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Figure 6.3: Outcomes on the roll of two dice. The event $\{$first roll is 3$\}$ is indicated.

Exercise 6.3. *Roll two four-sided dice. With the numbers 1 through 4 on each die, the value of the roll is the number on the side facing downward. Assuming all 16 outcomes are equally likely, find $P\{$sum is at least 5$\}$, $P\{$first die is 2$\}$ and $P\{$sum is at least 5$|$first die is 2$\}$*

6.2 The Multiplication Principle

The defining formula (6.1) for conditional probability can be rewritten to obtain **the multiplication principle**,

$$P(A \cap B) = P(A|B)P(B). \tag{6.2}$$

Now, we can complete an earlier problem:

$$\begin{aligned} P\{\text{ace on first two cards}\} &= P\{\text{ace on second card}|\text{ace on first card}\}P\{\text{ace on first card}\} \\ &= \frac{3}{51} \times \frac{4}{52} = \frac{1}{17} \times \frac{1}{13}. \end{aligned}$$

We can continue this process to obtain a **chain rule**:

$$P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C).$$

Thus,

$P\{$ace on first three cards$\}$
$= P\{$ace on third card$|$ace on first and second card$\}P\{$ace on second card$|$ace on first card$\}P\{$ace on first card$\}$
$= \frac{2}{50} \times \frac{3}{51} \times \frac{4}{52} = \frac{1}{25} \times \frac{1}{17} \times \frac{1}{13}.$

Extending this to 4 events, we consider the following question:

Example 6.4. *In an urn with b blue balls and g green balls, the probability of green, blue, green, blue (in that order) is*

$$\frac{g}{b+g} \cdot \frac{b}{b+g-1} \cdot \frac{g-1}{b+g-2} \cdot \frac{b-1}{b+g-3} = \frac{(g)_2(b)_2}{(b+g)_4}.$$

Notice that any choice of 2 green and 2 blue would result in the same probability. There are $\binom{4}{2} = 6$ *such choices. Thus, with 4 balls chosen without replacement*

$$P\{2 \text{ blue and } 2 \text{ green}\} = \binom{4}{2}\frac{(g)_2(b)_2}{(b+g)_4}.$$

Exercise 6.5. *Show that*

$$\binom{4}{2}\frac{(g)_2(b)_2}{(b+g)_4} = \frac{\binom{b}{2}\binom{g}{2}}{\binom{b+g}{4}}.$$

Explain in words why $P\{2 \text{ blue and } 2 \text{ green}\}$ *is the expression on the right.*

We will later extend this idea when we introduce sampling without replacement in the context of the hypergeometric random variable.

6.3 The Law of Total Probability

If we know the fraction of the population in a given state of the United States that has a given attribute - is diabetic, over 65 years of age, has an income of $100,000, owns their own home, is married - then how do we determine what fraction of the total population of the United States has this attribute? We address this question by introducing a concept - **partitions** - and an identity - **the law of total probability**.

Definition 6.6. *A* **partition** *of the sample space* Ω *is a finite collection of pairwise mutually exclusive events*

$$\{C_1, C_2, \ldots, C_n\}$$

whose union is Ω.

Thus, every outcome $\omega \in \Omega$ belongs to *exactly* one of the C_i. In particular, distinct members of the partition are mutually exclusive. ($C_i \cap C_j = \emptyset$, if $i \neq j$)

If we know the fraction of the population from 18 to 25 that has been infected by the H1N1 influenza A virus in each of the 50 states, then we cannot just average these 50 values to obtain the fraction of this population infected in the whole country. This method fails because it give equal weight to California and Wyoming. The **law of total probability** shows that we should weigh these conditional probabilities by the probability of residence in a given state and then sum over all of the states.

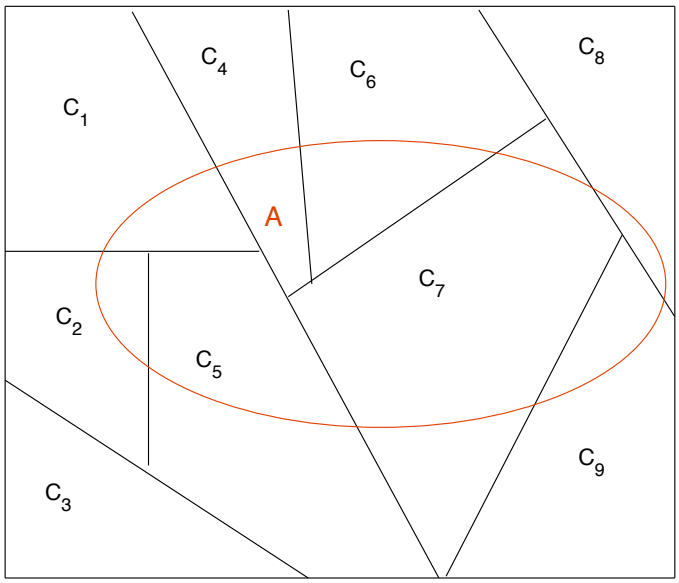


Figure 6.4: A partition $\{C_1 \ldots, C_9\}$ of the sample space Ω. The event A can be written as the union $(A \cap C_1) \cup \cdots \cup (A \cap C_9)$ of mutually exclusive events.

Theorem 6.7 (law of total probability)**.** *Let* P *be a probability on* Ω *and let* $\{C_1, C_2, \ldots, C_n\}$ *be a partition of* Ω *chosen so that* $P(C_i) > 0$ *for all* i*. Then, for any event* $A \subset \Omega$*,*

$$P(A) = \sum_{i=1}^{n} P(A|C_i)P(C_i). \tag{6.3}$$

Because $\{C_1, C_2, \ldots, C_n\}$ is a partition, $\{A \cap C_1, A \cap C_2, \ldots, A \cap C_n\}$ are pairwise mutually exclusive events. By the distributive property of sets, their union is the event A. (See Figure 6.4.)

To refer the example above the C_i are the residents of state i, $A \cap C_i$ are those residents who are from 18 to 25 years old and have been been infected by the H1N1 influenza A virus. Thus, distinct $A \cap C_i$ are mutually exclusive - individuals cannot reside in 2 different states. Their union is A, all individuals in the United States between the ages of 18 and 25 years old who have been been infected by the H1N1 virus.

Thus,

$$P(A) = \sum_{i=1}^{n} P(A \cap C_i). \tag{6.4}$$

Finish by using the multiplication identity (6.2),

$$P(A \cap C_i) = P(A|C_i)P(C_i), \quad i = 1, 2, \ldots, n$$

and substituting into (6.4) to obtain the identity in (6.3).

The most frequent use of the law of total probability comes in the case of a partition of the sample space into two events, $\{C, C^c\}$. In this case the law of total probability becomes the identity

$$P(A) = P(A|C)P(C) + P(A|C^c)P(C^c). \tag{6.5}$$

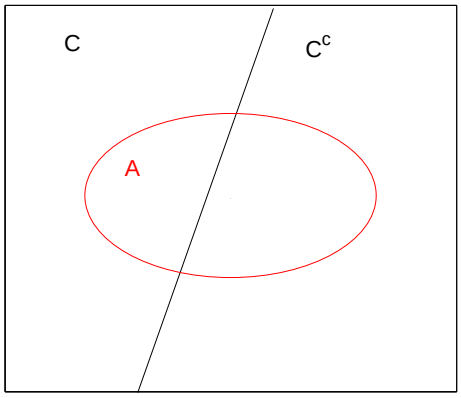


Figure 6.5: A partition into two events C and C^c.

Exercise 6.8. *The problem of points is a classical problem in probability theory. The problem concerns a series of games with two sides who have equal chances of winning each game. The winning side is one that first reaches a given number n of wins. Let $n = 4$ for a* best of seven *playoff. Determine*

$$p_{ij} = P\{\text{winning the playoff after } i \text{ wins vs } j \text{ opponent wins}\}$$

(Hint: $p_{ii} = \frac{1}{2}$ for $i = 0, 1, 2, 3$.)

6.4 Bayes formula

Let A be the event that an individual tests positive for some disease and C be the event that the person actually has the disease. We can perform clinical trials to estimate the probability that a randomly chosen individual tests positive given that they have the disease,

$$P\{\text{tests positive}|\text{has the disease}\} = P(A|C),$$

by taking individuals with the disease and applying the test. However, we would like to use the test as a method of *diagnosis* of the disease. Thus, we would like to be able to give the test and assert the chance that the person has the disease. That is, we want to know the probability with the reverse conditioning

$$P\{\text{has the disease}|\text{tests positive}\} = P(C|A).$$

Example 6.9. *The Public Health Department gives us the following information.*

- *A test for the disease yields a positive result 90% of the time when the disease is present.*
- *A test for the disease yields a positive result 1% of the time when the disease is not present.*
- *One person in 1,000 has the disease.*

Let's first think about this intuitively and then look to a more formal way using Bayes formula to find the probability of

$$P(C|A).$$

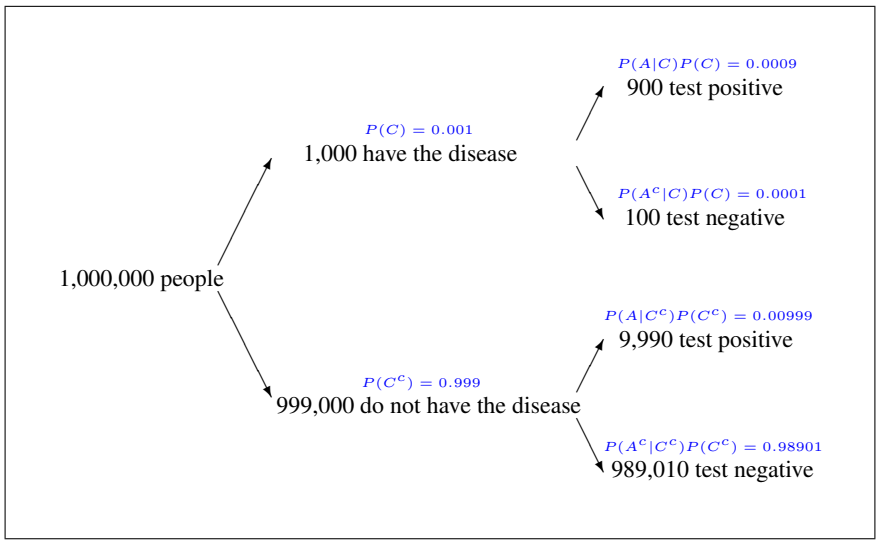


Figure 6.6: Tree diagram. We can use a tree diagram to indicate the number of individuals, on average, in each group (in black) or the probablity (in blue). Notice that in each column the number of individuals adds to give 1,000,000 and the probabilities add to give 1. In addition, each pair of arrows divides an events into two mutually exclusive subevents. Thus, both the numbers and the probabilities at the tip of the arrows add to give the respective values at the head of the arrow.

- In a city with a population of 1 million people, on average,

  1,000 have the disease and 999,000 do not

- Of the 1,000 that have the disease, on average,

  900 test positive and 100 test negative

- Of the 999,000 that do not have the disease, on average,

  999,000 $\times$ 0.01 = 9990 test positive and 989,010 test negative.

Consequently, among those that test positive, the *odds* of having the disease is

#(have the disease):#(does not have the disease)

900:9990

and converting odds to probability we see that

$$P\{\text{have the disease}|\text{test is positive}\} = \frac{900}{900 + 9990} = 0.0826.$$

We now derive Bayes formula. First notice that we can flip the order of conditioning by using the multiplication formula (6.2) **twice**

$$P(A \cap C) = \begin{cases} P(A|C)P(C) \\ \\ P(C|A)P(A) \end{cases}$$

researcher			public health worker	clinician			
	has disease C	does not have disease C^c	$\longrightarrow$		has disease C	does not have disease C^c	sum
tests positive A	$P(A\|C)$ 0.90	$P(A\|C^c)$ 0.01	$P(C) = 0.001$ $P(C^c) = 0.999$	tests positive A	$P(C\|A)$ 0.0826	$P(C^c\|A)$ 0.9174	1
tests negative A^c	$P(A^c\|C)$ 0.10	$P(A^c\|C^c)$ 0.99	$\nwarrow$	tests negative A^c	$P(C\|A^c)$ 0.0001	$P(C^c\|A^c)$ 0.9999	1
sum	1	1					

Table I: Using Bayes formula to evaluate a test for a disease. Successful analysis of the results of a clinical test require researchers to provide results on the quality of the test and public health workers to provide information on the prevalence of a disease. The conditional probabilities, provided by the researchers, and the probability of a person having the disease, provided by the public health service (shown by the east arrow), are necessary for the clinician, using Bayes formula (6.6), to give the probability of the conditional probability of having the disease given the test result. Notice, in particular, that the order of the conditioning needed by the clinician is the reverse of that provided by the researcher. If the clinicians provide reliable data to the public health service, then this information can be used to update the probabilities for the prevalence of the disease (indicated by the northeast arrow). The numbers in gray can be computed from the numbers in black by using the complement rule. In particular, the column sums for the researchers and the row sums for the clinicians much be .

Now we can create a formula for $P(C|A)$ as desired in terms of $P(A|C)$.

$$P(C|A)P(A) = P(A|C)P(C) \quad \text{or} \quad P(C|A) = \frac{P(A|C)P(C)}{P(A)}.$$

Thus, given A, the probability of C changes by the **Bayes factor**

$$\frac{P(A|C)}{P(A)}.$$

Example 6.10. *Both autism A and epilepsy C exists at approximately 1% in human populations. In this case*

$$P(A|C) = P(C|A)$$

Clinical evidence shows that this common value is about 30%. The Bayes factor is

$$\frac{P(A|C)}{P(A)} = \frac{0.3}{0.01} = 30.$$

Thus, the knowledge of one disease increases the chance of the other by a factor of 30.

From this formula we see that in order to determine $P(C|A)$ from $P(A|C)$, we also need to know $P(C)$, the fraction of the population with the disease and $P(A)$. We can find $P(A)$ using the law of total probability in (6.5) and write Bayes formula as

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|C^c)P(C^c)}. \tag{6.6}$$

This shows us that we can determine $P(A)$ if, in addition, we collect information from our clinical trials on $P(A|C^c)$, the fraction that test positive who do not have the disease.

Let's now compute $P(C|A)$ using Bayes formula directly and use this opportunity to introduce some terminology. We have that $P(A|C) = 0.90$. We call this **true positive** probability or the **sensitivity**. If one tests negative for the disease (the outcome is in A^c) given that one has the disease, (the outcome is in C), then we call this a **false negative**. In this case, the false negative probability is $P(A^c|C) = 0.10$

If one tests positive for the disease (the outcome is in A) given that one does not have the disease, (the outcome is in C^c), then we call this a **false positive**. In this case, the false positive probability is $P(A|C^c) = 0.01$. The **true negative** probability $P(A^c|C^c) = 0.99$ is also called the **specificity**.

The probability of having the disease is $P(C) = 0.001$ and so the probability of being disease free is $P(C^c) = 0.999$. Now, we apply the law of total probability (6.5) as the first step in Bayes formula (6.6),

$$P(A) = P(A|C)P(C) + P(A|C^c)P(C^c) = 0.90 \cdot 0.001 + 0.01 \cdot 0.999 = 0.0009 + 0.009999 = 0.01089.$$

Thus, the probability of having the disease given that the test was positive is

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} = \frac{0.0009}{0.01089} = 0.0826.$$

Notice that the numerator is one of the terms that was summed to compute the denominator.

The answer in the previous example may be surprising. Only 8% of those who test positive actually have the disease. This example underscores the fact that good predictions based on intuition are hard to make in this case. To determine the probability, we must weigh the odds of two terms, each of them itself a product.

- $P(A|C)P(C)$, a big number (the true positive probability) times a small number (the probability of having the disease) **versus**
- $P(A|C^c)P(C^c)$, a small number (the false positive probability) times a large number (the probability of being disease free).

We do not need to restrict Bayes formula to the case of C, *has the disease*, and C^c, *does not have the disease*, as seen in (6.5), but rather to any partition of the sample space. Indeed, Bayes formula can be generalized to the case of a partition $\{C_1, C_2, \ldots, C_n\}$ of Ω chosen so that $P(C_i) > 0$ for all i. Then, for any event $A \subset \Omega$ and any j

$$P(C_j|A) = \frac{P(A|C_j)P(C_j)}{\sum_{i=1}^{n} P(A|C_i)P(C_i)}. \tag{6.7}$$

To understand why this is true, use the law of total probability to see that the denominator is equal to $P(A)$. By the multiplication identity for conditional probability, the numerator is equal to $P(C_j \cap A)$. Now, make these two substitutions into (6.7) and use one more time the definition of conditional probability.

Example 6.11. *We begin with a simple and seemingly silly example involving fair and two sided coins. However, we shall soon see that this leads us to a question in the vertical transmission of a genetic disease.*

A box has a two-headed coin and a fair coin. It is flipped n times, yielding heads each time. What is the probability that the two-headed coin is chosen?

To solve this, note that

$$P\{\text{two-headed coin}\} = \frac{1}{2}, \qquad P\{\text{fair coin}\} = \frac{1}{2}.$$

and

$$P\{n \text{ heads}|\text{two-headed coin}\} = 1, \qquad P\{n \text{ heads}|\text{fair coin}\} = 2^{-n}.$$

By the law of total probability,

$$\begin{aligned} P\{n \text{ heads}\} &= P\{n \text{ heads}|\text{two-headed coin}\}P\{\text{two-headed coin}\} + P\{n \text{ heads}|\text{fair coin}\}P\{\text{fair coin}\} \\ &= 1 \cdot \frac{1}{2} + 2^{-n} \cdot \frac{1}{2} = \frac{2^n + 1}{2^{n+1}}. \end{aligned}$$

Next, we use Bayes formula.

$$P\{\text{two-headed coin}|n \text{ heads}\} = \frac{P\{n \text{ heads}|\text{two-headed coin}\}P\{\text{two-headed coin}\}}{P\{n \text{ heads}\}} = \frac{1 \cdot (1/2)}{(2^n + 1)/2^{n+1}} = \frac{2^n}{2^n + 1} < 1.$$

Notice that as n increases, the probability of a two headed coin approaches 1 - with a longer and longer sequence of heads we become increasingly suspicious (but, because the probability remains less than one, are never completely certain) that we have chosen the two headed coin.

This is the related genetics question: Based on the pedigree of her past, a female knows that she has in her history a allele on her X chromosome that indicates a genetic condition. The allele for the condition is recessive. Because she does not have the condition, she knows that she cannot be homozygous for the recessive allele. Consequently, she wants to know her chance of being a carrier (heteorzygous for a recessive allele) or not a carrier (homozygous for the common genetic type) of the condition. The female is a mother with n male offspring, none of which show the recessive allele on their single X chromosome and so do not have the condition. What is the probability that the female is not a carrier?

Let's look at the computation above again, based on her pedigree, the female estimates that

$$P\{\text{mother is not a carrier}\} = p, \qquad P\{\text{mother is a carrier}\} = 1 - p.$$

Then, from the law of total probability

$$\begin{aligned} & P\{n \text{ male offspring condition free}\} \\ = \; & P\{n \text{ male offspring condition free}|\text{mother is not a carrier}\}P\{\text{mother is not a carrier}\} \\ & +P\{n \text{ male offspring condition free}|\text{mother is a carrier}\}P\{\text{mother is a carrier}\} \\ = \; & 1 \cdot p + 2^{-n} \cdot (1-p). \end{aligned}$$

and Bayes formula

$$\begin{aligned} & P\{\text{mother is not a carrier}|n \text{ male offspring condition free}\} \\ = \; & \frac{P\{n \text{ male offspring condition free}|\text{mother is not a carrier}\}P\{\text{mother is not a carrier}\}}{P\{n \text{ male offspring condition free}\}} \\ = \; & \frac{1 \cdot p}{1 \cdot p + 2^{-n} \cdot (1-p)} = \frac{p}{p + 2^{-n}(1-p)} = \frac{2^n p}{2^n p + (1-p)}. \end{aligned}$$

Again, with more sons who do not have the condition, we become increasingly more certain that the mother is not a carrier.

One way to introduce **Bayesian statistics** *is to consider the situation in which we do not know the value of p and replace it with a probability distribution. Even though we will concentrate on classical approaches to statistics, we will take the time in later sections to explore the Bayesian approach*

6.5 Independence

An event A is **independent of** B if its Bayes factor is 1, i.e.,

$$1 = \frac{P(A|B)}{P(A)}, \quad P(A) = P(A|B).$$

In words, the occurrence of the event B does not alter the probability of the event A. Multiply this equation by $P(B)$ and use the multiplication rule to obtain

$$P(A)P(B) = P(A|B)P(B) = P(A \cap B).$$

The formula

$$P(A)P(B) = P(A \cap B) \tag{6.8}$$

is the usual definition of independence and is symmetric in the events A and B. If A is independent of B, then B is independent of A. Consequently, when equation (6.8) is satisfied, we say that A and B are **independent**.

Example 6.12. *Roll two dice.*

$$\begin{aligned}\frac{1}{36} &= P\{a \text{ on the first die}, b \text{ on the second die}\} \\ &= \frac{1}{6} \times \frac{1}{6} = P\{a \text{ on the first die}\}P\{b \text{ on the second die}\}\end{aligned}$$

and, thus, the outcomes on two rolls of the dice are independent.

Exercise 6.13. *If A and B are independent events, then show that A^c and B, A and B^c, A^c and B^c are also independent.*

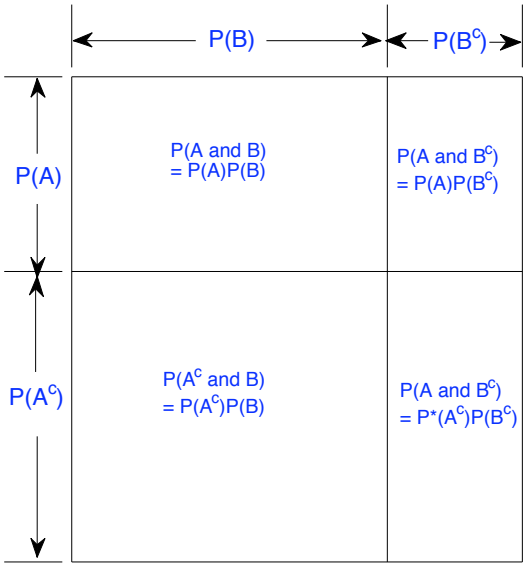


Figure 6.7: The Venn diagram for independent events is represented by the horizontal strip A and the vertical strip B is shown above. The identity $P(A \cap B) = P(A)P(B)$ is now represented as the area of the rectangle. Other aspects of Exercise 6.12 are indicated in this Figure.

We can also use this to extend the definition to n independent events:

Definition 6.14. *The events $A_1, \cdots, A_n$ are called* **independent** *if for any choice $A_{i_1}, A_{i_2}, \cdots, A_{i_k}$ taken from this collection of n events, then*

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}). \quad (6.9)$$

A similar product formula holds if some of the events are replaced by their complement.

Exercise 6.15. *Flip 10 biased coins. Their outcomes are independent with the i-th coin turning up heads with probability p_i. Find*

$$P\{\text{first coin heads, third coin tails, seventh \& ninth coin heads}\}.$$

Example 6.16. *Mendel studied inheritance by conducting experiments using a garden peas. Mendel's First Law, the* **law of segregation** *states that every diploid individual possesses a pair of alleles for any particular trait and that each parent passes one randomly selected allele to its offspring.*

In Mendel's experiment, each of the 7 traits under study express themselves independently. This is an example of Mendel's Second Law, also known as the **law of independent assortment**. *If the dominant allele was present in the population with probability p, then the recessive allele is expressed in an individual when it receive this allele from both of its parents. If we assume that the presence of the allele is independent for the two parents, then*

$$\begin{aligned}P\{\text{recessive allele expressed}\} &= P\{\text{recessive allele paternally inherited}\} \times P\{\text{recessive allele maternally inherited}\} \\ &= (1-p) \times (1-p) = (1-p)^2.\end{aligned}$$

In Mendel's experimental design, p was set to be 1/2. Consequently,

$$P\{\text{recessive allele expressed}\} = (1 - 1/2)^2 = 1/4.$$

Using the complement rule,

$$P\{\text{dominant allele expressed}\} = 1 - (1-p)^2 = 1 - (1 - 2p + p^2) = 2p - p^2.$$

This number can also be computed by added the three alternatives shown in the Punnett square in Table 6.1.

$$p^2 + 2p(1-p) = p^2 + 2p - 2p^2 = 2p - p^2.$$

Next, we look at two traits - 1 and 2 - with the dominant alleles present in the population with probabilities p_1 and p_2. If these traits are expressed independently, then, we have, for example, that

$$\begin{aligned}&P\{\text{dominant allele expressed in trait 1, recessive trait expressed in trait 2}\}\\&= P\{\text{dominant allele expressed in trait 1}\} \times P\{\text{recessive trait expressed in trait 2}\}\\&= (1-(1-p_1)^2)(1-p_2)^2.\end{aligned}$$

Exercise 6.17. *Show that if two traits are genetically linked, then the appearance of one increases the probability of the other. Thus,*

$$P\{\text{individual has allele for trait 1}|\text{individual has allele for trait 2}\} > P\{\text{individual has allele for trait 1}\}.$$

implies

$$P\{\text{individual has allele for trait 2}|\text{individual has allele for trait 1}\} > P\{\text{individual has allele for trait 2}\}.$$

More generally, for events A and B,

$$P(A|B) > P(A) \quad \textit{implies} \quad P(B|A) > P(B) \tag{6.10}$$

then we way that A and B are **positively associated**.

Exercise 6.18. *A* **genetic marker** *B for a disease A is one in which $P(A|B) \approx 1$. In this case, approximate $P(B|A)$. More precisely, if $P(A|B) = 1$ if and only if $P(B|A) = P(A)/P(B)$.*

Definition 6.19. Linkage disequilibrium *is the non-independent association of alleles at two loci on single chromosome. To define linkage disequilibrium, let*

- *A be the event that a given allele is present at the first locus, and*
- *B be the event that a given allele is present at a second locus.*

Then the linkage disequilibrium,

$$D_{A,B} = P(A)P(B) - P(A \cap B).$$

Thus if $D_{A,B} = 0$, the the two events are independent.

Exercise 6.20. *Show that $D_{A,B^c} = -D_{A,B}$*

6.6 Answers to Selected Exercises

6.1. Let's check the three axioms;

1. For any event A,

$$Q(A) = P(A|B) = \frac{P(A \cap B)}{P(B)} \geq 0.$$

2. For the sample space Ω,

$$Q(\Omega) = P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

	S	s
S	SS p^2	Ss $p(1-p)$
s	sS $(1-p)p$	ss $(1-p)^2$

Table II: Punnett square for a monohybrid cross using a dominant trait S (say spherical seeds) that occurs in the population with probability p and a recessive trait s (wrinkled seeds) that occurs with probability $1-p$. Maternal genotypes are listed on top, paternal genotypes on the left. See Example 6.14. The probabilities of a given genotype are given in the lower right hand corner of the box.

3. For mutually exclusive events, $\{A_j; j \geq 1\}$, we have that $\{A_j \cap B; j \geq 1\}$ are also mutually exclusive and

$$Q\left(\bigcup_{j=1}^{\infty} A_j\right) = P\left(\bigcup_{j=1}^{\infty} A_j \Big| B\right) = \frac{P\left(\left(\bigcup_{j=1}^{\infty} A_j\right) \cap B\right)}{P(B)} = \frac{P(\bigcup_{j=1}^{\infty}(A_j \cap B))}{P(B)}$$
$$= \frac{\sum_{j=1}^{\infty} P(A_j \cap B)}{P(B)} = \sum_{j=1}^{\infty} \frac{P(A_j \cap B)}{P(B)} = \sum_{j=1}^{\infty} P(A_j|B) = \sum_{j=1}^{\infty} Q(A_j)$$

6.2. $P\{\text{sum is } 8|\text{first die shows } 3\} = 1/6$, and $P\{\text{sum is } 8|\text{first die shows } 1\} = 0$.

6.3 Here is a table of outcomes. The symbol $\times$ indicates an outcome in the event $\{\text{sum is at least } 5\}$. The rectangle indicates the event $\{\text{first die is } 2\}$. Because there are 10 $\times$'s,

	1	2	3	4
1				$\times$
2			$\times$	$\times$
3		$\times$	$\times$	$\times$
4	$\times$	$\times$	$\times$	$\times$

$$P\{\text{sum is at least } 5\} = 10/16 = 5/8.$$

The rectangle contains 4 outcomes, so

$$P\{\text{first die is } 2\} = 4/16 = 1/4.$$

Inside the event $\{\text{first die is } 2\}$, 2 of the outcomes are also in the event $\{\text{sum is at least } 5\}$. Thus,

$$P\{\text{sum is at least } 5\}|\text{first die is } 2\} = 2/4 = 1/2$$

Using the definition of conditional probability, we also have

$$P\{\text{sum is at least } 5\}|\text{first die is } 2\} = \frac{P\{\text{sum is at least 5 and first die is 2}\}}{P\{\text{first die is 2}\}} = \frac{2/16}{4/16} = \frac{2}{4} = \frac{1}{2}.$$

6.5. We modify both sides of the equation.

$$\binom{4}{2}\frac{(g)_2(b)_2}{(b+g)_4} = \frac{4!}{2!2!}\frac{(g)_2(b)_2}{(b+g)_4}$$

$$\frac{\binom{b}{2}\binom{g}{2}}{\binom{b+g}{4}} = \frac{(b)_2/2! \cdot (g)_2/2!}{(b+g)_4/4!} = \frac{4!}{2!2!}\frac{(g)_2(b)_2}{(b+g)_4}.$$

The sample space Ω is set of collections of 4 balls out of $b+g$. This has $\binom{b+g}{4}$ outcomes. The number of choices of 2 blue out of b is $\binom{b}{2}$. The number of choices of 2 green out of g is $\binom{g}{2}$. Thus, by the fundamental principle of counting, the total number of ways to obtain the event 2 blue and 2 green is $\binom{b}{2}\binom{g}{2}$. For equally likely outcomes, the probability is the ratio of $\binom{b}{2}\binom{g}{2}$, the number of outcomes in the event, and $\binom{b+g}{4}$, the number of outcomes in the sample space.

6.8. Let A_{ij} be the event of winning the series that has i *wins* versus j *wins for the opponent*. Then $p_{ij} = P(A_{ij})$. We know that

$$p_{0,4} = p_{1,4} = p_{2,4} = p_{3,4} = 0$$

because the series is lost when the opponent has won 4 games. Also,

$$p_{4,0} = p_{4,1} = p_{4,2} = p_{4,3} = 1$$

because the series is won with 4 wins in games. For a tied series, the probability of winning the series is 1/2 for both sides.

$$p_{0,0} = p_{1,1} = p_{2,2} = p_{3,3} = \frac{1}{2}.$$

These values are filled in blue in the table below. We can determine the remaining values of p_{ij} iteratively by looking forward one game and using the law of total probability to condition of the outcome of the $(i+j+1$-st) game. Note that $P\{\text{win game } i+j+1\} = P\{\text{lose game } i+j+1\} = \frac{1}{2}$.

$$\begin{aligned} p_{ij} &= P(A_{ij}|\text{win game } i+j+1\}P\{\text{win game } i+j+1\} + P(A_{ij}|\text{lose game } i+j-1\}P\{\text{lose game } i+j+1\} \\ &= \frac{1}{2}(p_{i+1,j} + p_{i,j+1}) \end{aligned}$$

This can be used to fill in the table above the diagonal. For example,

$$p_{23} = \frac{1}{2}(p_{33} + p_{42}) = \frac{1}{2}\left(\frac{1}{2} + 1\right) = \frac{3}{4}.$$

For below the diagonal, note that

$$p_{ij} = 1 - p_{ji}.$$

For example,

$$p_{23} = 1 - p_{32} = 1 - \frac{3}{4} = \frac{1}{4}.$$

Filling in the table, we have:

		i				
		0	1	2	3	4
j	0	1/2	21/32	13/16	15/16	1
	1	11/32	1/2	11/16	7/8	1
	2	3/16	5/16	1/2	3/4	1
	3	1/16	1/8	1/4	1/2	1
	4	0	0	0	0	-

6.13. We take the questions one at a time. Because A and B are independent $P(A \cap B) = P(A)P(B)$.

(a) B is the disjoint union of $A \cap B$ and $A^c \cap B$. Thus,

$$P(B) = P(A \cap B) + P(A^c \cap B)$$

Subtract $P(A \cap B)$ to obtain

$$P(A^c \cap B) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = (1 - P(A))P(B) = P(A^c)P(B)$$

and A^c and B are independent.

(b) Just switch the roles of A and B in part (a) to see that A and B^c are independent.

(c) Use the complement rule and inclusion-exclusion

$$\begin{aligned} P(A^c \cap B^c) &= P((A \cup B)^c) = 1 - P(A \cup B) = 1 - P(A) - P(B) - P(A \cap B) \\ &= 1 - P(A) - P(B) - P(A)P(B) = (1 - P(A))(1 - P(B)) \\ &= P(A^c)P(B^c) \end{aligned}$$

and A^c and B^c are independent.

6.15. Let A_i be the event $\{i$-th coin turns up heads$\}$. Then the event can be written $A_1 \cap A_3^c \cap A_7 \cap A_9$. Thus,

$$\begin{aligned} P(A_1 \cap A_3^c \cap A_7 \cap A_9) &= P(A_1)P(A_3^c)P(A_7)P(A_9) \\ &= p_1(1 - p_3)p_7p_9. \end{aligned}$$

6.17. Multiply both of the expressions in (6.10) by the appropriate probability to see that they are equivalent to

$$P(A \cap B) > P(A)P(B).$$

6.18. By using Bayes formula, if $P(A|B) = 1$, then

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(B)}{P(A)}.$$

On the other hand, if $P(B|A) = P(B)/P(A)$, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = 1.$$

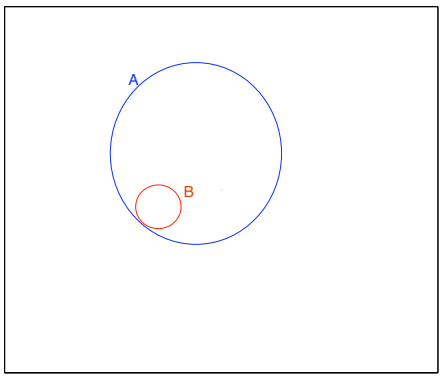


Figure 6.8: If $P(A|B) \approx 1$, then nearly all of B is inside A and the probability of $P(B|A) \approx P(B)/P(A)$ as shown in the figure.

6.20 Because A is the disjoint union of $A \cap B$ and $A \cap B^c$, we have $P(A) = P(A \cap B) + P(A \cap B^c)$ or $P(A \cap B^c) = P(A) - P(A \cap B)$. Thus,

$$D_{A,B^c} = P(A)P(B^c) - P(A \cap B^c) = P(A)(1 - P(B)) - (P(A) - P(A \cap B)) = -P(A)P(B) + P(A \cap B) = -D_{A,B}.$$

Topic 7

Random Variables and Distribution Functions

While writing my book I had an argument with Feller. He asserted that everyone said "random variable" and I asserted that everyone said "chance variable." We obviously had to use the same name in our books, so we decided the issue by a stochastic procedure. That is, we tossed for it and he won. – Joseph Doob, *Statistical Science*

While writing my book I had an argument with Feller. He asserted that everyone said "random variable" and I asserted that everyone said "chance variable." We obviously had to use the same name in our books, so we decided the issue by a stochastic procedure. That is, we tossed for it and he won.

statistics	**probability**
universe of information	sample space - Ω and probability - P
$\Downarrow$	$\Downarrow$
ask a question and collect data	define a random variable X
$\Downarrow$	$\Downarrow$
organize into the empirical cumulative distribution function	organize into the cumulative distribution function
$\Downarrow$	$\Downarrow$
compute sample means and variances	compute distributional means and variances

Table I: Corresponding notions between statistics and probability. Examining probabilities models and random variables will lead to strategies for the collection of data and inference from these data.

7.1 Introduction

From the universe of possible information, we ask a question. To address this question, we might collect quantitative data and organize it, for example, using the empirical cumulative distribution function. With this information, we are able to compute sample means, standard deviations, medians and so on.

Similarly, even a fairly simple probability model can have an enormous number of outcomes. For example, flip a coin 332 times. Then the number of outcomes is more than a google (10^{100}) – a number at least 100 quintillion times the number of elementary particles in the known universe. We may not be interested in an analysis that considers separately every possible outcome but rather some simpler concept like the number of heads or the longest run of tails. To focus our attention on the issues of interest, we take a given outcome and compute a number. This function is called a **random variable**.

Definition 7.1. *A* **random variable** *is a real valued function from the probability space.*

$$X : \Omega \to \mathbb{R}.$$

Generally speaking, we shall use capital letters near the end of the alphabet, e.g., X, Y, Z for random variables. The range S of a random variable is sometimes called the **state space**.

Exercise 7.2. *Roll a die twice and consider the sample space $\Omega = \{(i, j); i, j = 1, 2, 3, 4, 5, 6\}$ and give some random variables on Ω.*

Exercise 7.3. *Flip a coin 10 times and consider the sample space Ω, the set of 10-tuples of heads and tails, and give some random variables on Ω.*

We often create new random variables via composition of functions:

$$\omega \mapsto X(\omega) \mapsto f(X(\omega))$$

Thus, if X is a random variable, then so are

$$X^2, \quad \exp \alpha X, \quad \sqrt{X^2 + 1}, \quad \tan^2 X, \quad \lfloor X \rfloor$$

and so on. The last of these, rounding down X to the nearest integer, is called the **floor function**.

Exercise 7.4. *How would we use the floor function to round down a number x to n decimal places.*

7.2 Distribution Functions

Having defined a random variable of interest, X, the question typically becomes, "What are the chances that X lands in some subset of values B?" For example,

$$B = \{\text{odd numbers}\}, \quad B = \{\text{greater than 1}\}, \quad \text{or} \quad B = \{\text{between 2 and 7}\}.$$

We write

$$\{\omega \in \Omega; X(\omega) \in B\} \tag{7.1}$$

to indicate those outcomes ω which have $X(\omega)$, the value of the random variable, in the subset A. We shall often abbreviate (7.1) to the shorter statement $\{X \in B\}$. Thus, for the example above, we may write the events

$$\{X \text{ is an odd number}\}, \quad \{X \text{ is greater than 1}\} = \{X > 1\}, \quad \{X \text{ is between 2 and 7}\} = \{2 < X < 7\}$$

to correspond to the three choices above for the subset B.

Many of the properties of random variables are not concerned with the specific random variable X given above, but rather depends on the way X *distributes its values*. This leads to a definition in the context of random variables that we saw previously with quantitive data..

Definition 7.5. *A* **(cumulative) distribution function** *of a random variable X is defined by*

$$F_X(x) = P\{\omega \in \Omega; X(\omega) \le x\}.$$

Recall that with quantitative observations, we called the analogous notion the *empirical cumulative distribution function*. Using the abbreviated notation above, we shall typically write the less explicit expression

$$F_X(x) = P\{X \le x\}$$

for the distribution function.

Exercise 7.6. *Establish the following identities that relate a random variable the complement of an event and the union and intersection of events*

1. $\{X \in B\}^c = \{X \in B^c\}$

2. *For sets $B_1, B_2, \ldots$,*

$$\bigcup_i \{X \in B_i\} = \{X \in \bigcup_i B\} \quad \textit{and} \quad \bigcap_i \{X \in B_i\} = \{X \in \bigcap_i B\}.$$

3. *If $B_1, \ldots B_n$ form a partition of the sample space S, then $C_i = \{X \in B_i\}$, $i = 1, \ldots, n$ form a partition of the probability space Ω.*

Exercise 7.7. *For a random variable X and subset B of the sample space S, define*

$$P_X(B) = P\{X \in B\}.$$

Show that P_X is a probability.

For the complement of $\{X \le x\}$, we have the **survival function**

$$\bar{F}_X(x) = P\{X > x\} = 1 - P\{X \le x\} = 1 - F_X(x).$$

Choose $a < b$, then the event $\{X \le a\} \subset \{X \le b\}$. Their set theoretic difference

$$\{X \le b\} \setminus \{X \le a\} = \{a < X \le b\}.$$

In words, the event that X is less than or equal to b but not less than or equal to a is the event that X is greater than a and less than or equal to b. Consequently, by the difference rule for probabilities,

$$P\{a < X \le b\} = P(\{X \le b\} \setminus \{X \le a\}) = P\{X \le b\} - P\{X \le a\} = F_X(b) - F_X(a). \tag{7.2}$$

Thus, we can compute the probability that a random variable takes values in an interval by subtracting the distribution function evaluated at the endpoints of the intervals. Care is needed on the issue of the inclusion or exclusion of the endpoints of the interval.

Example 7.8. *To give the cumulative distribution function for X, the sum of the values for two rolls of a die, we start with the table*

x	2	3	4	5	6	7	8	9	10	11	12
$P\{X = x\}$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

and create the graph.

If we look at the graph of this cumulative distribution function, we see that it is constant in between the possible values for X and that the jump size at x is equal to $P\{X = x\}$. In this example, $P\{X = 5\} = 4/36$, the size of the jump at $x = 5$. In addition,

$$F_X(5) - F_X(2) = P\{2 < X \le 5\} = P\{X = 3\} + P\{X = 4\} + P\{X = 5\} = \sum_{2 < x \le 5} P\{X = x\}$$
$$= \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{9}{36}.$$

We shall call a random variable **discrete** if it has a finite or countably infinite state space. Thus, we have in general that:

$$P\{a < X \le b\} = \sum_{a < x \le b} P\{X = x\}.$$

Exercise 7.9. *Let X be the number of heads on three independent flips of a biased coin that turns ups heads with probability p. Give the cumulative distribution function F_X for X.*

Exercise 7.10. *Let X be the number of spades in a collection of three cards. Give the cumulative distribution function for X. Use* R *to plot this function.*

Exercise 7.11. *Find the cumulative distribution function of $Y = X^3$ in terms of F_X, the distribution function for X.*

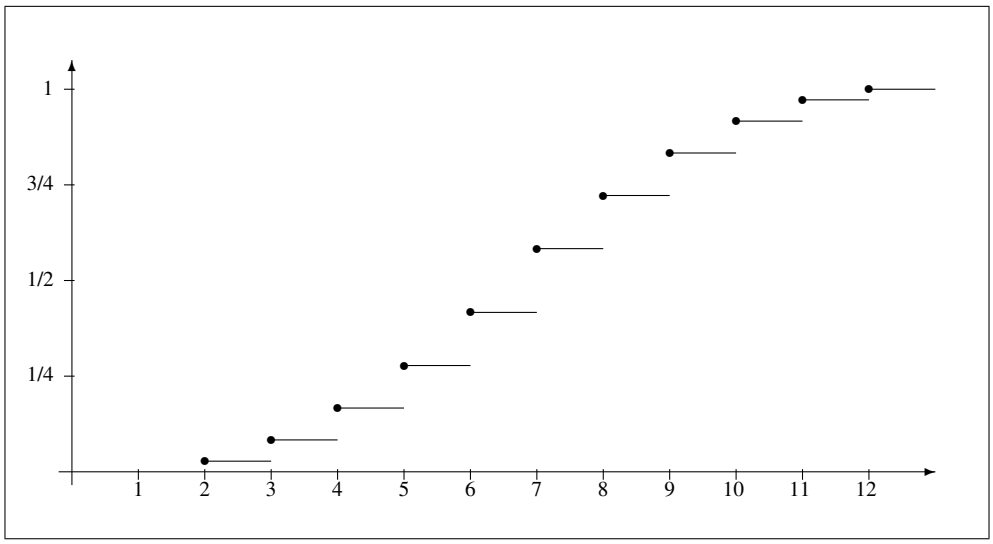


Figure 7.1: Graph of F_X, the cumulative distribution function for the sum of the values for two rolls of a die.

7.3 Properties of the Distribution Function

A distribution function F_X has the property that it starts at 0, ends at 1 and does not decrease with increasing values of x.. This is the content of the next exercise.

Exercise 7.12. *1.* $\lim_{x\to-\infty} F_X(x) = 0$.

2. $\lim_{x\to\infty} F_X(x) = 1$.

3. F_X *is nondecreasing.*

The cumulative distribution function F_X of a discrete random variable X is constant except for jumps. At the jump, F_X is **right continuous**,

$$\lim_{x\to x_0+} F_X(x) = F_X(x_0).$$

The next exercise ask that this be shown more generally.

Exercise 7.13. *Prove the statement concerning the right continuity of the distribution function from the continuity property of a probability.*

Exercise 7.14. *Show that for any* x_0*,*

$$P\{X < x_0\} = \lim_{x\to x-} F_X(x) = F_X(x_0-),$$

the left limit of F_X *at* x_0.

Putting the previous two exercises together, we find that

$$P\{X = x_0\} = P(\{X \le x_0\} \setminus \{X < x_0\}) = P\{X \le x_0\} - P\{X < x_0\} = F_X(x_0) - F_X(x_0-).$$

Definition 7.15. *A* **continuous random variable** *has a cumulative distribution function* F_X *that is differentiable.*

So, distribution functions for continuous random variables increase smoothly. To show how this can occur, we will develop an example of a continuous random variable.

Example 7.16. *Consider a dartboard having unit radius. Assume that the dart lands randomly uniformly on the dartboard.*

Let X be the distance from the center. For $x \in [0, 1]$,

$$F_X(x) = P\{X \le x\} = \frac{\text{area inside circle of radius } x}{\text{area of circle}} = \frac{\pi x^2}{\pi 1^2} = x^2.$$

Thus, we have the distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x \le 0, \\ x^2 & \text{if } 0 < x \le 1, \\ 1 & \text{if } x > 1. \end{cases}$$

The first line states that X cannot be negative. The third states that X is at most 1, and the middle lines describes how X distributes is values between 0 and 1. For example,

$$F_X\left(\frac{1}{2}\right) = \frac{1}{4}$$

indicates that with probability 1/4, the dart will land within 1/2 unit of the center of the dartboard.

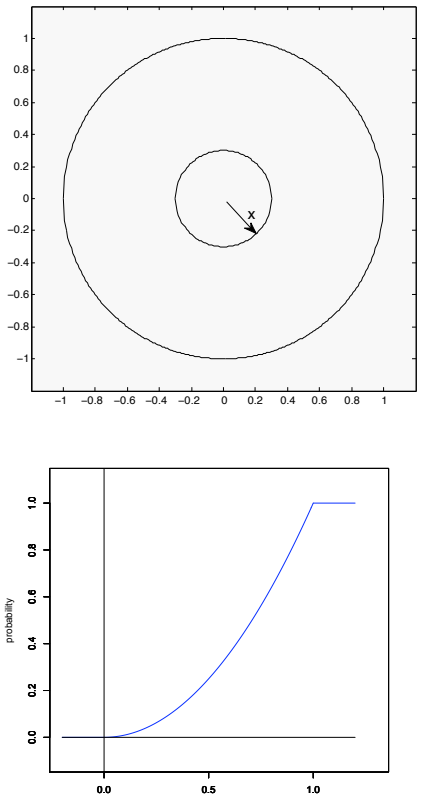


Figure 7.2: (top) Dartboard. (bottom) Cumulative distribution function for the dartboard random variable.

Exercise 7.17. *Find the probability that the dart lands between 1/3 unit and 2/3 unit from the center.*

Exercise 7.18. *Let the reward Y for throwing the dart be the inverse $1/X$ of the distance from the center. Find the cumulative distribution function for Y.*

Exercise 7.19. *An* **exponential random variable** *X has cumulative distribution function*

$$F_X(x) = P\{X \le x\} = \begin{cases} 0 & \text{if } x \le 0, \\ 1 - \exp(-\lambda x) & \text{if } x > 0. \end{cases} \tag{7.3}$$

for some $\lambda > 0$. Show that F_X has the properties of a distribution function.

We can create an expression and perform an evaluation using R.

```
> F<-expression(1-exp(-lambda*x))
```

We can then evaluate $F_X(3)$ and $F_X(1)$ with $\lambda = 2$ as follows.

```
> x<-c(10,30);lambda<-1/10
> (Feval<-eval(F))
[1] 0.6321206 0.9502129
> Feval[2]-Feval[1]
[1] [1] 0.3180924
```

The last expression gives the value for $F_X(30) - F_X(10) = P\{10 < X \le 30\}$.

This function is also stored in R and so its value at x can be computed in R using the command `pexp(x,0.1)` for $\lambda = 1/10$ and drawn using

```
> curve(pexp(x,0.1),0,80)
```

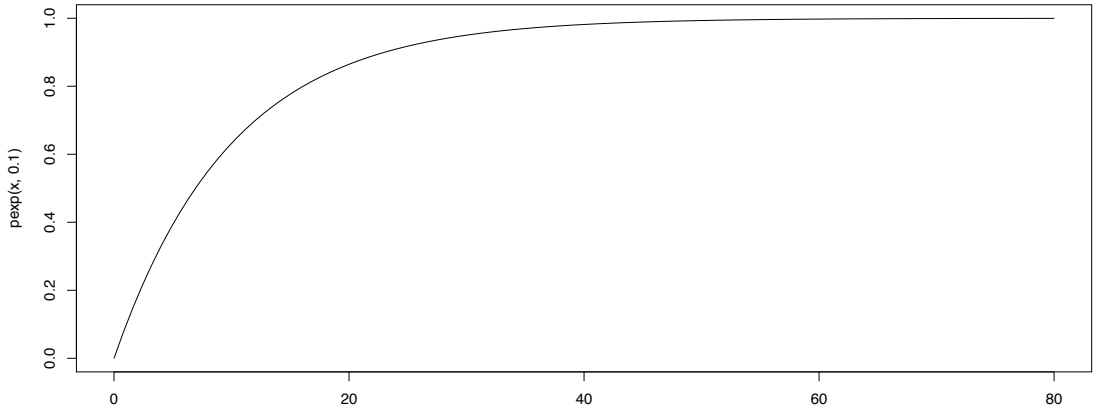


Figure 7.3: Cumulative distribution function for an exponential random variable with $\lambda = 1/10$.

Exercise 7.20. *The time until the next bus arrives is an exponential random variable with $\lambda = 1/10$ minutes. A person waits for a bus at the bus stop until the bus arrives, giving up when the wait reaches 20 minutes. Give the cumulative distribution function for T, the time that the person remains at the bus station and sketch a graph.*

Even though the cumulative distribution function is defined for every random variable, we will often use other characterizations, namely, the **mass function** for discrete random variable and the **density function** for continuous random variables. Indeed, we typically will introduce a random variable via one of these two functions. In the next two sections we introduce these two concepts and develop some of their properties.

7.4 Mass Functions

Definition 7.21. *The* **(probability) mass function** *of a discrete random variable X is*

$$f_X(x) = P\{X = x\}.$$

The mass function has a value at x equal to the size of the jump in the distribution function. In symbols,

$$f_X(x) = F_X(x) - F_X(x-)$$

where $F_X(x-)$ is the left limit of F_X at x.

The mass function has two basic properties:

- $f_X(x) \geq 0$ for all x in the state space.
- $\sum_x f_X(x) = 1$.

The first property is based on the fact that probabilities are non-negative. The second follows from the observation that the collection $C_x = \{\omega; X(\omega) = x\}$ for all $x \in S$, the state space for X, forms a partition of the probability space Ω. In Example 7.8, we saw the mass function for the random variable X that is the sum of the values on two independent rolls of a fair dice.

Example 7.22. *Let's make tosses of a biased coin whose outcomes are independent. We shall continue tossing until we obtain a toss of heads. Let X denote the random variable that gives the number of tails before the first head and p*

denote the probability of heads in any given toss. Then

$$\begin{aligned} f_X(0) &= P\{X=0\} = P\{H\} = p \\ f_X(1) &= P\{X=1\} = P\{TH\} = (1-p)p \\ f_X(2) &= P\{X=2\} = P\{TTH\} = (1-p)^2 p \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ f_X(x) &= P\{X=x\} = P\{T\cdots TH\} = (1-p)^x p \end{aligned}$$

So, the probability mass function $f_X(x) = (1-p)^x p$. Because the terms in this mass function form a geometric sequence, X is called a **geometric random variable***. Recall that a geometric sequence $c, cr, cr^2, \ldots, cr^n$ has sum*

$$s_n = c + cr + cr^2 + \cdots + cr^n = \frac{c(1-r^{n+1})}{1-r}$$

for $r \neq 1$. If $|r| < 1$, then $\lim_{n\to\infty} r^n = 0$ and thus s_n has a limit as $n \to \infty$. In this case, the infinite sum is the limit

$$c + cr + cr^2 + \cdots + cr^n + \cdots = \lim_{n\to\infty} s_n = \frac{c}{1-r}.$$

Exercise 7.23. *Establish the formula above for s_n.*

The mass function above forms a geometric sequence with the ratio $r = 1-p$. Consequently, for positive integers a and b,

$$\begin{aligned} P\{a < X \le b\} &= \sum_{x=a+1}^{b} (1-p)^x p = (1-p)^{a+1}p + \cdots + (1-p)^b p \\ &= \frac{(1-p)^{a+1}p - (1-p)^{b+1}p}{1-(1-p)} = (1-p)^{a+1} - (1-p)^{b+1} \end{aligned}$$

We can take $a = 0$ to find the distribution function for a geometric random variable.

$$F_X(b) = P\{X \le b\} = 1 - (1-p)^{b+1}. \tag{7.4}$$

The event $\{X \ge b+1\} = \{X > b\}$ is the same as having the first $b+1$ coin tosses turn up *tails*. Thus, the outcome is $b+1$ independent events each with probability $1-p$. Thus, $P\{X \ge b+1\} = P\{X > b\} = (1-p)^{b+1}$. By noting that the distribution function, $F_X(b) = 1 - P\{X > b\}$, this gives us another way to see (7.4).

Exercise 7.24. *Show that for a geometric random variable X,*

$$P\{X \ge a+b | X \ge b\} = P\{X \ge a\}. \tag{7.5}$$

This property is called **memorylessness***. In words, if the first b trials results in failures, then the probability of at least a additional failures is the same as the probability of at least a failures from the beginning. The fact that we begin with b failures does not impact the number of trials afterwards until a success.*

Conversely, if the memoryless property holds for an $\mathbb{N}$-valued random variable X, then X is a geometric random variable.

The mass function and the cumulative distribution function for the geometric random variable with parameter $p = 1/3$ can be found in R by writing

```
> x<-c(0:10)
> f<-dgeom(x,1/3)
> F<-pgeom(x,1/3)
```

The initial `d` indicates **d**ensity and `p` indicates the **p**robability from the distribution function.

```
> data.frame(x,f,F)
    x           f         F
1   0 0.333333333 0.3333333
2   1 0.222222222 0.5555556
3   2 0.148148148 0.7037037
4   3 0.098765432 0.8024691
5   4 0.065843621 0.8683128
6   5 0.043895748 0.9122085
7   6 0.029263832 0.9414723
8   7 0.019509221 0.9609816
9   8 0.013006147 0.9739877
10  9 0.008670765 0.9826585
11 10 0.005780510 0.9884390
```

Note that the difference in values in the distribution function $F_X(x) - F_X(x-1)$, giving the height of the jump in F_X at x, is equal to the value of the mass function. For example,

$$F_X(3) - F_X(2) = 0.8024691 - 0.7037037 = 0.0987654 = f_X(3).$$

Exercise 7.25. *Check that the jumps in the cumulative distribution function for the geometric random variable above is equal to the values of the mass function.*

Exercise 7.26. *For the geometric random variable above, find* $P\{X \leq 3\}$, $P\{2 < X \leq 5\}$. $P\{X > 4\}$.

We can simulate 100 geometric random variables with parameter $p = 1/3$ using the R command `rgeom(100,1/3)`. (See Figure 7.4.)

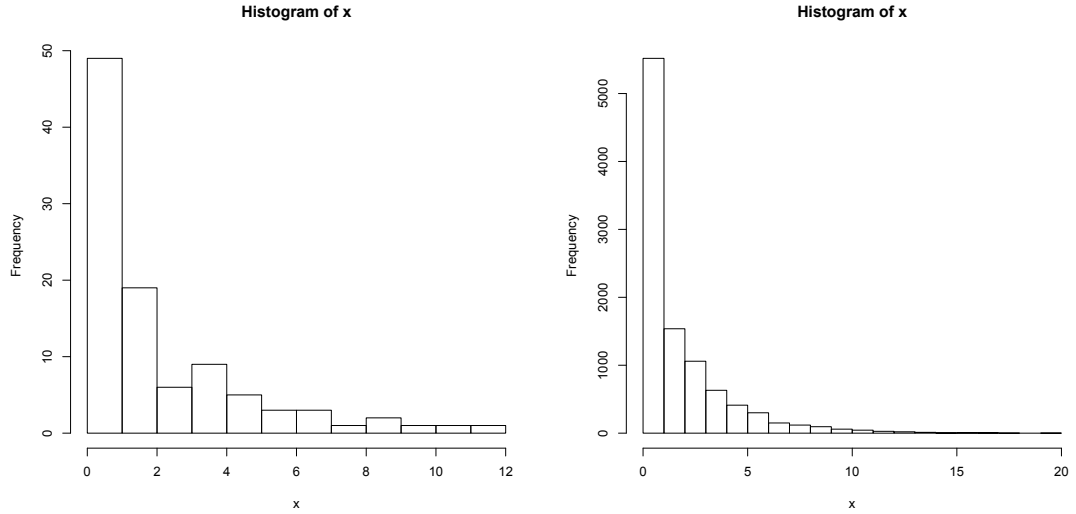


Figure 7.4: Histogram of 100 and 10,000 simulated geometric random variables with $p = 1/3$. Note that the histogram looks much more like a geometric series for 10,000 simulations. We shall see later how this relates to the law of large numbers.

7.5 Density Functions

Definition 7.27. *For X a random variable whose distribution function F_X has a derivative. The function f_X satisfying*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

is called the **probability density function** *and X is called a* **continuous random variable**.

By the fundamental theorem of calculus, the density function is the derivative of the distribution function.

$$f_X(x) = \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} = F_X'(x).$$

In other words,

$$F_X(x + \Delta x) - F_X(x) \approx f_X(x)\Delta x.$$

We can compute probabilities by evaluating definite integrals

$$P\{a < X \leq b\} = F_X(b) - F_X(a) = \int_a^b f_X(t)\, dt.$$

The density function has two basic properties that mirror the properties of the mass function:

- $f_X(x) \geq 0$ for all x in the state space.
- $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

Return to the dart board example, letting X be the distance from the center of a dartboard having unit radius. Then,

$$\begin{aligned} P\{x < X \leq x + \Delta x\} &= F_X(x + \Delta x) - F_X(x) \\ &\approx f_X(x)\Delta x = 2x\Delta \end{aligned}$$

and X has density

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 2x & \text{if } 0 \leq x \leq 1, \\ 0 & \text{if } x > 1. \end{cases}$$

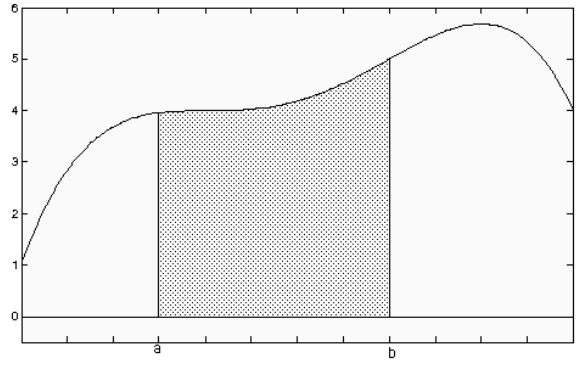


Figure 7.5: The probability $P\{a < X \leq b\}$ is the area under the density function, above the x axis between $y = a$ and $y = b$.

Exercise 7.28. *Let f_X be the density for a random variable X and pick a number x_0. Explain why $P\{X = x_0\} = 0$.*

Example 7.29. *For the exponential distribution function (7.3), we have the density function*

$$f_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \lambda e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

R *performs differentiation. We must first create an expression*

```
> F<-expression(1-exp(-lambda*x))
```

We then differentiate using the `D` *command, placing x, the variable of differentiation in quotes.*

```
> f<-D(F,"x")
> f
exp(-lambda * x) * lambda
```

Example 7.30. *Density functions do not need to be bounded, for example, if we take*

$$f_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{c}{\sqrt{x}} & \text{if } 0 < x < 1, \\ 0 & \text{if } 1 \leq x. \end{cases}$$

Then, to find the value of the constant c, we compute the integral

$$1 = \int_0^1 \frac{c}{\sqrt{t}}\, dt = 2c\sqrt{t}\Big|_0^1 = 2c.$$

So $c = 1/2$.

For $0 \le a < b \le 1$,

$$P\{a < X \le b\} = \int_a^b \frac{1}{2\sqrt{t}}\, dt = \sqrt{t}\Big|_a^b = \sqrt{b} - \sqrt{a}.$$

Exercise 7.31. *Give the cumulative distribution function for the random variable in the previous example.*

Exercise 7.32. *Let X be a continuous random variable with density f_X, then the random variable $Y = aX + b$ has density*

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

(Hint: Begin with the definition of the cumulative distribution function F_Y for Y. Consider the cases $a > 0$ and $a < 0$ separately.)

7.6 Mixtures

Exercise 7.33. *Let F_1 and F_2 be two cumulative distribution functions and let $\pi \in (0, 1)$, then*

$$F(x) = \pi F_1(x) + (1 - \pi)F_2(x)$$

is a cumulative distribution function.

Mixture distributions occur routinely. To see this, first flip a coin, heads occurring with probability π. In this case the random variable

$$X = \begin{cases} X_1 & \text{if the coin lands heads,} \\ X_2 & \text{if the coin lands tails.} \end{cases}$$

If X_i has distribution function F_i, $i = 1, 2$, then, by the law of total probability,

$$\begin{aligned} F_X(x) = P\{X \le x\} &= P\{X \le x|\text{coin lands heads}\}P\{\text{coin lands heads}\} \\ &\quad + P\{X \le x|\text{coin lands tails}\}P\{\text{coin lands tails}\} \\ &= P\{X_1 \le x\}\pi + P\{X_2 \le x\}(1 - \pi) = \pi F_1(x) + (1 - \pi)F_2(x) \end{aligned}$$

More generally, let $X_1, \ldots, X_n$ be random variables with distribution functions $F_1, \ldots, F_n$ and $\pi_1, \ldots, \pi_n$ be positive numbers with $\sum_{i=1}^n \pi_i = 1$. In this case, roll an n sided die, i showing with probability π_i. If the die shows i, then we use the random variable X_i. To be concrete, individuals arriving to take an airline flight are assigned to group i with probability π_i. Let X_i be the (random) time until individuals in group i are seated. Then the distribution function for the time to be seated

$$\begin{aligned} F_X(x) = P\{X \le x\} &= \sum_{i=1}^n P\{X \le x|\text{assigned group } i\}P\{\text{assigned group } i\} \\ &= \sum_{i=1}^n P\{X_i \le x\}\pi_i = \pi_1 F_1(x) + \cdots + \pi_n F_n(x). \end{aligned}$$

F is call the **mixture** of $F_1, \ldots, F_n$ with weights $\pi_1, \ldots, \pi_n$.

If the X_i are discrete random variables, then so is X. The **mass function** for X is

$$\begin{aligned} f_X(x) = F_X(x) - F_X(x-) &= \pi_1(F_1(x) - F_1(x-)) + \cdots + \pi_n(F_n(x) - F_n(x-)) \\ &= \pi_1 f_1(x) + \cdots + \pi_n f_n(x). \end{aligned}$$

Exercise 7.34. *Check that f_X is a mass function.*

Exercise 7.35. *Find the mass function for the mixture of the three mass functions*

x	$f_1(x)$	$f_2(x)$	$f_3(x)$
1	0.2	0.5	0.1
2	0.3	0.5	0.1
3	0.1	0	0.2
4	0.4	0	0.2
5	0	0	0.4

and weights $\pi = (1/4, 1/4, 1/2)$,

If the X_i are continuous random variables, then so is X. The **density function** for X is

$$\begin{aligned} f_X(x) = F'_X(x) &= \pi_1 F'_1(x) + \cdots + \pi_n F'_n(x) \\ &= \pi_1 f_1(x) + \cdots + \pi_n f_n(x). \end{aligned}$$

Checking that f_X is a density function is similar to the exercise above, Just replace the sum on x with an integral.

7.7 Joint Distributions

Because we will collect data on several observations, we must, as well, consider more than one random variable at a time in order to model our experimental procedures. Consequently, we will expand on the concepts above to the case of multiple random variables and their joint distribution. For the case of two random variables, X_1 and X_2, this means looking at the probability of events,

$$P\{X_1 \in B_1, X_2 \in B_2\}.$$

For discrete random variables, take $B_1 = \{x_1\}$ and $B_2 = \{x_2\}$ and define the **joint probability mass function**

$$f_{X_1,X_2}(x_1, x_2) = P\{X_1 = x_1, X_2 = x_2\}.$$

For continuous random variables, we consider $B_1 = (x_1, x_1 + \Delta x_1]$ and $B_2 = (x_2, x_2 + \Delta x_2]$ and ask that for some function f_{X_1,X_2}, the **joint probability density function** to satisfy

$$P\{x_1 < X_1 \leq x_1 + \Delta x_1, x_2 < X_2 \leq x_2 + \Delta x_2\} \approx f_{X_1,X_2}(x_1, x_2)\Delta x_1 \Delta x_2.$$

Example 7.36. *Generalize the notion of mass and density functions to more than two random variables.*

7.7.1 Independent Random Variables

Many of our experimental protocols will be designed so that observations are independent. More precisely, we will say that two random variables X_1 and X_2 are **independent** if any two events associated to them are independent, i.e.,

$$P\{X_1 \in B_1, X_2 \in B_2\} = P\{X_1 \in B_1\}P\{X_2 \in B_2\}.$$

In words, the probability that the two events $\{X_1 \in B_1\}$ and $\{X_2 \in B_2\}$ happen simultaneously is equal to the product of the probabilities that each of them happen individually.

For independent discrete random variables, we have that

$$f_{X_1,X_2}(x_1, x_2) = P\{X_1 = x_1, X_2 = x_2\} = P\{X_1 = x_1\}P\{X_2 = x_2\} = f_{X_1}(x_1)f_{X_2}(x_2).$$

In this case, we say that the joint probability mass function is the product of the **marginal mass functions**.

For continuous random variables,

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2)\Delta x_1\Delta x_2 &\approx P\{x_1 < X_1 \le x_1+\Delta x_1, x_2 < X_2 \le x_2+\Delta x_2\} \\ &= P\{x_1 < X_1 \le x_1+\Delta x_1\}P\{x_2 < X_2 \le x_2+\Delta x_2\} \approx f_{X_1}(x_1)\Delta x_1 f_{X_2}(x_2)\Delta x_2 \\ &= f_{X_1}(x_1)f_{X_2}(x_2)\Delta x_1\Delta x_2. \end{aligned}$$

Thus, for independent continuous random variables, the joint probability density function

$$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

is the product of the **marginal density functions**.

Exercise 7.37. *Generalize the notion of independent mass and density functions to more than two random variables.*

Soon, we will be looking at n independent observations $x_1, x_2, \ldots, x_n$ arising from an unknown density or mass function f. Thus, the joint density is

$$f(x_1)f(x_2)\cdots f(x_n).$$

Generally speaking, the density function f will depend on the choice of a parameter value θ. (For example, the unknown parameter in the density function for an exponential random variable that describes the waiting time for a bus.) Given the data arising from the n observations, the **likelihood function** arises by considering this joint density not as a function of $x_1, \ldots, x_n$, but rather as a function of the parameter θ. We shall learn how the study of the likelihood plays a major role in parameter estimation and in the testing of hypotheses.

7.8 Simulating Random Variables

One goal for these notes is to provide the tools needed to design inferential procedures based on sound principles of statistical science. Thus, one of the very important uses of statistical software is the ability to generate pseudo-data to simulate the actual data. This provides the opportunity to test and refine methods of analysis in advance of the need to use these methods on genuine data. This requires that we explore the properties of the data through simulation. For many of the frequently used families of random variables, R provides commands for their simulation. We shall examine these families and their properties in Topic 9, *Examples of Mass Functions and Densities*. For other circumstances, we will need to have methods for simulating sequence of independent random variables that possess a common distribution. We first consider the case of discrete random variables.

7.8.1 Discrete Random Variables and the `sample` Command

The `sample` command is used to create simple and stratified random samples. Thus, if we enter a sequence `x`, `sample(x,40)` chooses 40 entries from `x` in such a way that all choices of size 40 have the same probability.

This uses the default R command of **sampling without replacement**. We can use this command to simulate discrete random variables. To do this, we need to give the state space in a vector `x` and a mass function `f`. The call for `replace=TRUE` indicates that we are **sampling with replacement**. Then to give a sample of n independent random variables having common mass function `f`, we use `sample(x,n,replace=TRUE,prob=f)`.

Example 7.38. *Let X be described by the mass function*

x	1	2	3	4
$f_X(x)$	0.1	0.2	0.3	0.4

Then to simulate 50 independent observations from this mass function:

```
> x<-c(1,2,3,4)
> f<-c(0.1,0.2,0.3,0.4)
> sum(f)
```

```
[1] 1
> data<-sample(x,50,replace=TRUE,prob=f)
> data
  [1] 1 4 4 4 4 4 3 3 4 3 3 2 3 3 3 4 4 3 3 2 4 1 3 3 4 2 3 3 3 1 2 4 3 2 3 4 4 4 4 2 4 1
[43] 2 3 4 4 1 4 3 4
```

Notice that 1 is the least represented value and 4 is the most represented. If the command `prob=f` *is omitted, then* `sample` *will choose uniformly from the values in the vector* `x`*. Let's check our simulation against the mass function that generated the data. (Notice the double equal sign, ==.) First, recount the observations that take on each possible value for x. We can make a table.*

```
> table(data)
data
 1  2  3  4
 5  7 18 20
```

or use the counts to determine the simulated proportions.

```
> counts<-rep(0,max(x)-min(x)+1)
> for (i in min(x):max(x)){counts[i]<-length(data[data==i])}
> simprob<-counts/(sum(counts))
> data.frame(x,f,simprob)
  x   f simprob
1 1 0.1 0.10
2 2 0.2 0.14
3 3 0.3 0.36
4 4 0.4 0.40
```

Exercise 7.39. *Simulate the sums on each of 20 rolls of a pair of dice. Repeat this for 1000 rolls and compare the simulation with the appropriate mass function.*

Exercise 7.40. *Simulate the mixture in Exercise 7.35 and comment on how it matches the mixture mass function.*

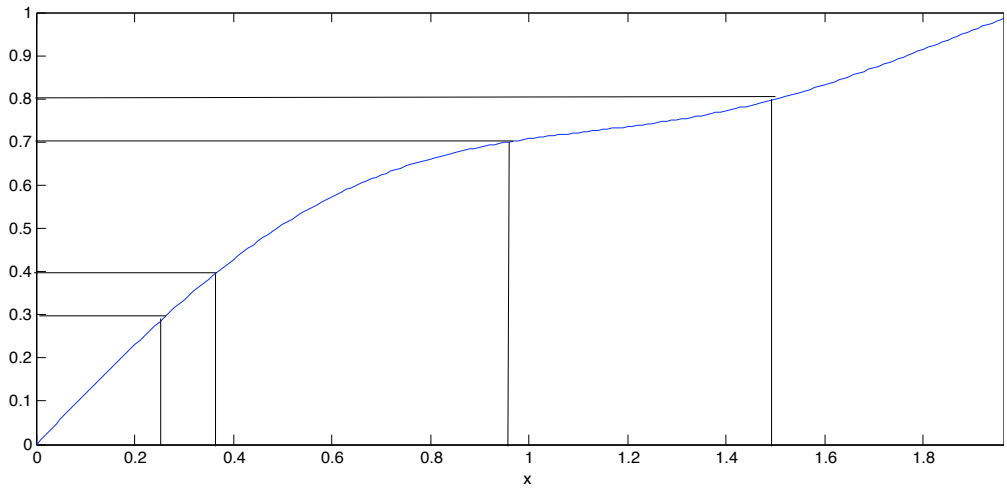


Figure 7.6: Illustrating the Probability Transform. First simulate uniform random variables $u_1, u_2, \ldots, u_n$ on the interval $[0, 1]$. About 10% of the random numbers should be in the interval $[0.3, 0.4]$. This corresponds to the 10% of the simulations on the interval $[0.28, 0.38]$ for a random variable with distribution function F_X shown. Similarly, about 10% of the random numbers should be in the interval $[0.7, 0.8]$ which corresponds to the 10% of the simulations on the interval $[0.96, 1.51]$ for a random variable with distribution function F_X, These values on the x-axis can be obtained from taking the inverse function of F_X, i.e., $x_i = F_X^{-1}(u_i)$.

7.8.2 Continuous Random Variables and the Probability Transform

If X a continuous random variable with a density f_X that is positive everywhere in its domain, then the distribution function $F_X(x) = P\{X \leq x\}$ is strictly increasing. In this case F_X has a inverse function F_X^{-1}, known as the **quantile function**.

Exercise 7.41. *$F_X(x) \leq u$ if and only if $x \leq F_X^{-1}(u)$.*

The **probability transform** follows from an analysis of the random variable

$$U = F_X(X)$$

Note that F_X has range from 0 to 1. It cannot take values below 0 or above 1. Thus, U takes on values between 0 and 1. Thus,

$$F_U(u) = 0 \text{ for } u < 0 \quad \text{and} \quad F_U(u) = 1 \text{ for } u \geq 1.$$

For values of u between 0 and 1, note that

$$P\{F_X(X) \leq u\} = P\{X \leq F_X^{-1}(u)\} = F_X(F_X^{-1}(u)) = u.$$

Thus, the distribution function for the random variable U,

$$F_U(u) = \begin{cases} 0 & u < 0, \\ u & 0 \leq u < 1, \\ 1 & 1 \leq u. \end{cases}$$

If we can simulate U, we can simulate a random variable with distribution F_X via the quantile function

$$X = F_X^{-1}(U). \tag{7.6}$$

Take a derivative to see that the density

$$f_U(u) = \begin{cases} 0 & u < 0, \\ 1 & 0 \leq u < 1, \\ 0 & 1 \leq u. \end{cases}$$

Because the random variable U has a constant density over the interval of its possible values, it is called **uniform** on the interval $[0, 1]$. It is simulated in R using the `runif` command. The identity (7.6) is called the **probability transform**. This transform is illustrated in Figure 7.6. We can see how the probability transform works in the following example.

Example 7.42. *For the dart board, for x between 0 and 1, the distribution function $u = F_X(x) = x^2$ and thus the quantile function*

$$x = F_X^{-1}(u) = \sqrt{u}.$$

We can simulate independent observations of the distance from the center $X_1, X_2, \ldots, X_n$ of the dart board by simulating independent uniform random variables $U_1, U_2, \ldots U_n$ and taking the quantile function

$$X_i = \sqrt{U_i}.$$

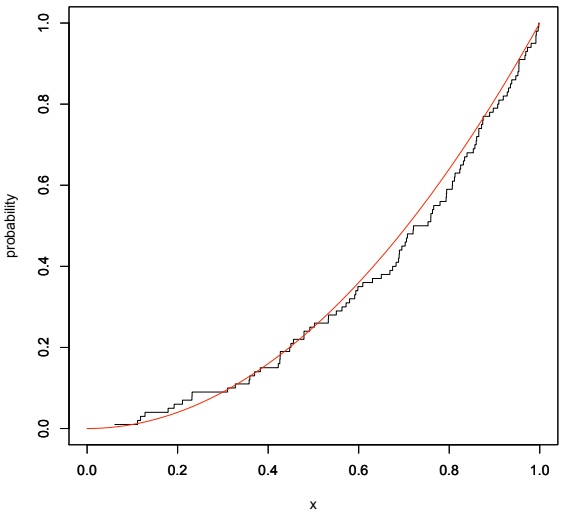


Figure 7.7: The distribution function (red) and the empirical cumulative distribution function (black) based on 100 simulations of the dart board distribution. R commands given below.

```
> u<-runif(100)
> xu<-sqrt(u)
> plot(sort(xu),1:length(xu)/length(xu),
+   type="s",xlim=c(0,1),ylim=c(0,1), xlab="x",ylab="probability")
> x<-seq(0,1,0.01)
> lines(x,x^2,col="red") #add the distribution function to the graph
```

Note that we have used the `lines` *command to ad the distribution function* $F_X(x) = x^2$*. Notice how it follows the empirical cumulative distribution function.*

Exercise 7.43. *If* U *is uniform on* $[0, 1]$*, then so is* $V = 1 - U$*.*

Sometimes, it is easier to simulate X using $F_X^{-1}(V)$.

Example 7.44. *For an exponential random variable, set*

$$u = F_X(x) = 1 - \exp(-\lambda x), \text{ and thus } x = -\frac{1}{\lambda}\ln(1-u)$$

Consequently, we can simulate independent exponential random variables $X_1, X_2, \ldots, X_n$ *by simulating independent uniform random variables* $V_1, V_2, \ldots V_n$ *and taking the transform*

$$X_i = -\frac{1}{\lambda}\ln V_i.$$

R *accomplishes this directly through the* `rexp` *command.*

7.9 Answers to Selected Exercises

7.2. The sum, the maximum, the minimum, the difference, the value on the first die, the product.

7.3. The roll with the first H, the number of T, the longest run of H, the number of Ts after the first H.

7.4. $\lfloor 10^n x \rfloor / 10^n$

7.6. A common way to show that two events A_1 and A_2 are equal is to pick an element $\omega \in A_1$ and show that it is in A_2. This proves $A_1 \subset A_2$. Then pick an element $\omega \in A_2$ and show that it is in A_1, proving that $A_2 \subset A_1$. Taken together, we have that the events are equal, $A_1 = A_2$. Sometimes the logic needed in showing $A_1 \subset A_2$ consist not solely of implications, but rather of equivalent statements. (We can indicate this with the symbol $\Longleftrightarrow$.) In this case we can combine the two parts of the argument. For this exercise, as the lines below show, this is a successful strategy.

We follow an arbitrary outcome $\omega \in \Omega$.

1. $\omega \in \{X \in B\}^c \Longleftrightarrow \omega \notin \{X \in B\} \Longleftrightarrow X(\omega) \notin B \Longleftrightarrow X(\omega) \in B^c \Longleftrightarrow \omega \in \{X \in B^c\}$. Thus, $\{X \in B\}^c = \{X \in B^c\}$.

2. $\omega \in \bigcup_i \{X \in B_i\} \Longleftrightarrow \omega \in \{X \in B_i\}$ for some $i \Longleftrightarrow X(\omega) \in B_i$ for some $i \Longleftrightarrow X(\omega) \in \bigcup_i B_i \Longleftrightarrow \omega \in \{X \in \bigcup_i B\}$. Thus, $\bigcup_i \{X \in B_i\} = \{X \in \bigcup_i B\}$. The identity with intersection is similar with *for all* instead of *for some*.

3. We must show that the union of the C_i is equal to the state space S and that each pair are mutually exclusive. For this

   (a) Because B_i are a partition of Ω, $\bigcup_i B_i = \Omega$, and

   $$\bigcup_i C_i = \bigcup_i \{X \in B_i\} = \{X \in \bigcup_i B_i\} = \{X \in \Omega\} = S,$$

   the state space.

(b) For $i \neq j$, $B_i \cap B_j = \emptyset$, and

$$C_i \cap C_j = \{X \in B_i\} \cap \{X \in B_j\} = \{X \in B_i \cap B_j\} = \{X \in \emptyset\} = \emptyset.$$

7.7. Let's check the three axioms. Each verification is based on the corresponding axiom for the probability P.

1. For any subset B, $P_X(B) = P\{X \in B\} \geq 0$.
2. For the sample space S, $P_X(S) = P\{X \in S\} = P(\Omega) = 1$.
3. For mutually exclusive subsets $B_i, i = 1, 2, \cdots$, we have by the exercise above the mutually exclusive events $\{X \in B_i\}, i = 1, 2, \cdots$. Thus,

$$P_X\left(\bigcup_{i=1}^{\infty} B_i\right) = P\left\{X \in \bigcup_{i=1}^{\infty} B_i\right\} = P\left(\bigcup_{i=1}^{\infty} \{X \in B_i\}\right) = \sum_{i=1}^{\infty} P\{X \in B_i\} = \sum_{i=1}^{\infty} P_X(B_i).$$

7.9. For three tosses of a biased coin, we have

x	0	1	2	3
$P\{X = x\}$	$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3

Thus, the cumulative distribution function,

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0, \\ (1-p)^3 & \text{for } 0 \leq x < 1, \\ (1-p)^3 + 3p(1-p)^2 = (1-p)^2(1+2p) & \text{for } 1 \leq x < 2, \\ (1-p)^2(1+2p) + 3p^2(1-p) = 1 - p^3 & \text{for } 2 \leq x < 3, \\ 1 & \text{for } 3 \leq x \end{cases}$$

7.10. From the example in the section *Basics of Probability*, we know that

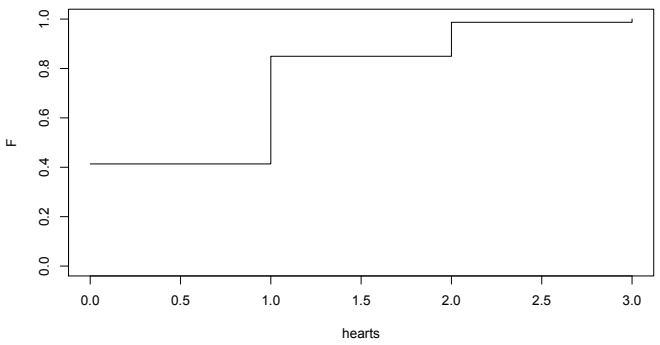
x	0	1	2	3
$P\{X = x\}$	0.41353	0.43588	0.13765	0.01294

To plot the distribution function, we use,

```
> hearts<-c(0:3)
> f<-choose(13,hearts)*choose(39,3-hearts)/choose(52,3)
> (F<-cumsum(f))
[1] 0.4135294 0.8494118 0.9870588 1.0000000
> plot(hearts,F,ylim=c(0,1),type="s")
```

Thus, the cumulative distribution function,

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0, \\ 0.41353 & \text{for } 0 \leq x < 1, \\ 0.84941 & \text{for } 1 \leq x < 2, \\ 0.98706 & \text{for } 2 \leq x < 3, \\ 1 & \text{for } 3 \leq x \end{cases}$$



7.11. The cumulative distribution function for Y,

$$\begin{aligned} F_Y(y) = P\{Y \leq y\} &= P\{X^3 \leq y\} \\ &= P\{X \leq \sqrt[3]{y}\} = F_X(\sqrt[3]{y}). \end{aligned}$$

7.12. To verify the three properties for the distribution function:

1. Let $x_n \to -\infty$ be a decreasing sequence. Then $x_1 > x_2 > \cdots$

$$\{X \le x_1\} \supset \{X \le x_2\} \supset \cdots$$

Thus,

$$P\{X \le x_1\} \ge P\{X \le x_2\} \ge \cdots$$

For each outcome ω, eventually, for some n, $X(\omega) > x_n$, and $\omega \notin \{X \le x_n\}$ and consequently no outcome ω is in *all* of the events $\{X \le x_n\}$ and

$$\bigcap_{n=1}^{\infty} \{X \le x_n\} = \emptyset.$$

Now, use the second continuity property of probabilities.

2. Let $x_n \to \infty$ be an increasing sequence. Then $x_1 < x_2 < \cdots$

$$\{X \le x_1\} \subset \{X \le x_2\} \subset \cdots .$$

Thus,

$$P\{X \le x_1\} \le P\{X \le x_2\} \le \cdots .$$

For each outcome ω, eventually, for some n, $X(\omega) \le x_n$, and

$$\bigcup_{n=1}^{\infty} \{X \le x_n\} = \Omega.$$

Now, use the first continuity property of probabilities.

3. Let $x_1 < x_2$, then $\{X \le x_1\} \subset \{X \le x_2\}$ and by the monotonicity rule for probabilities

$$P\{X \le x_1\} \le P\{X \le x_2\}, \quad \text{or written in terms of the distribution function,} \quad F_X(x_1) \le F_X(x_2)$$

7.13. Let $x_n \to x_0$ be a strictly decreasing sequence. Then $x_1 > x_2 > \cdots$

$$\{X \le x_1\} \supset \{X \le x_2\} \supset \cdots , \quad \bigcap_{n=1}^{\infty} \{X \le x_n\} = \{X \le x_0\}.$$

(Check this last equality.) Then $P\{X \le x_1\} \ge P\{X \le x_2\} \ge \cdots$. Now, use the second continuity property of probabilities to obtain $\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P\{X \le x_n\} = P\{X \le x_0\} = F_X(x_0)$. Because this holds for every strictly decreasing sequencing sequence with limit x_0, we have that

$$\lim_{x \to x_0+} F_X(x) = F_X(x_0).$$

7.14. Correspondingly from the previous exercise, let $x_n \to x_0$ be a strictly increasing sequence. Then $x_1 < x_2 < \cdots$

$$\{X \le x_1\} \subset \{X \le x_2\} \subset \cdots , \quad \bigcup_{n=1}^{\infty} \{X \le x_n\} = \{X < x_0\}.$$

(Again, check this last equality.) Then $P\{X \le x_1\} \le P\{X \le x_2\} \le \cdots$. Now, use the second continuity property of probabilities to obtain $= F_X(x_0-) = \lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P\{X \le x_n\} = P\{X < x_0\}$. Because this holds for every strictly increasing sequencing sequence with limit x_0, we have that

$$\lim_{x \to x_0-} F_X(x) = P\{X < x_0\}.$$

7.16. Using the identity in (7.2), we have

$$P\left\{\frac{1}{3} < X \le \frac{2}{3}\right\} = F_x\left(\frac{2}{3}\right) - F_x\left(\frac{1}{3}\right) = \frac{4}{9} - \frac{1}{9} = \frac{3}{9} = \frac{1}{3}.$$

Check Exercise 7.22 to see that the answer does not depend on whether or not the endpoints of the interval are included.

7.18. Using the relation $Y = 1/X$, we find that the distribution function for Y,

$$F_Y(y) = P\{Y \le y\} = P\{1/X \le y\} = P\{X \ge 1/y\} = 1 - P\{X < 1/y\} = 1 - \frac{1}{y^2}.$$

Thus uses the fact that $P\{X = 1/y\} = 0$.

7.19. We use the fact that the exponential function is increasing, and that $\lim_{u\to\infty} \exp(-u) = 0$. Using the numbering of the properties above

1. Because $F_X(x) = 0$ for all $x < 0$, $\lim_{x\to-\infty} F_X(x) = 0$.
2. $\lim_{x\to\infty} \exp(-\lambda x) = 0$. Thus, $\lim_{x\to\infty} F_X(x) = \lim_{x\to\infty} 1 - \exp(-\lambda x) = 1$.
3. For $x < 0$, F_X is constant, $F_X(0) = 0$. For $x \ge 0$, note that $\exp(-\lambda x)$ is decreasing. Thus, $F_X(x) = 1 - \exp(-\lambda x)$ is increasing. Consequenlty, the distribution function F_X is non-decreasing.

7.20. The distribution function has the graph shown in Figure 7.8.

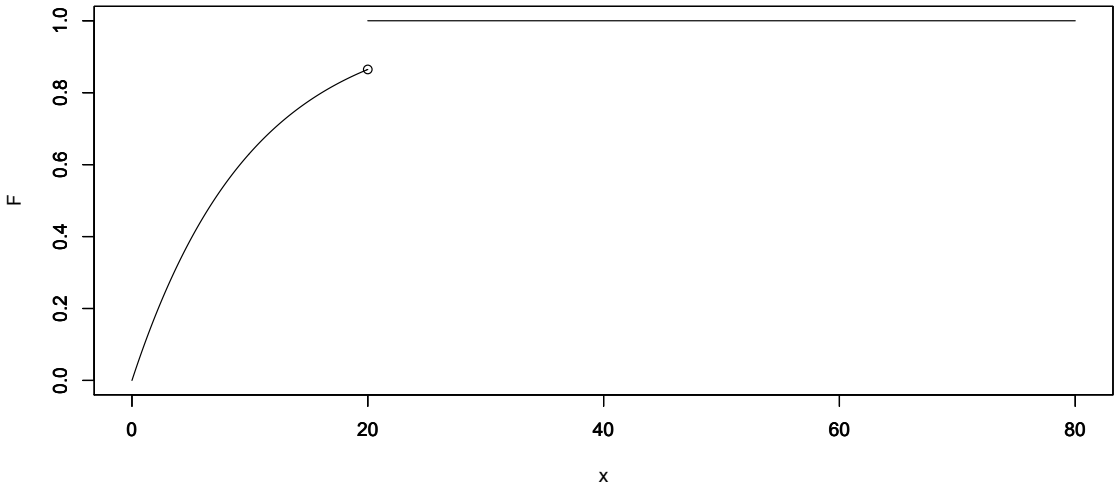


Figure 7.8: Cumulative distribution function for an exponential random variable with $\lambda = 1/10$ and a jump at $x = 20$.

The formula

$$F_T(x) = P\{X \le x\} = \begin{cases} 0 & \text{if } x < 0, \\ 1 - \exp(-x/10) & \text{if } 0 \le x < 20, \\ 1 & \text{if } 20 \le x. \end{cases}$$

7.23. For $r \neq 1$, write the expressions for s_n and rs_n and subtract. Notice that most of the terms cancel.

$$\begin{aligned} s_n &= c+ cr +cr^2+ \cdots + cr^n \\ rs_n &= \quad cr +cr^2+ \cdots + cr^n +cr^{n+1} \\ (1-r)s_n &= c \qquad\qquad\qquad\qquad -cr^{n+1} = c(1-r^{n+1}) \end{aligned}$$

Now divide by $1-r$ to obtain the formula.

7.24. First, $\{X \geq a+b\} \subset \{X \geq b\}$ (If $X \geq a+b$, then automatically $X \geq b$. Thus, $\{X \geq b+a, X \geq b\} = \{X \geq b+a\}$. By the definition of conditional expectation

$$P\{X \geq b+a | X \geq b\} = \frac{P\{X \geq b+a, X \geq b\}}{P\{X \geq b\}} = \frac{P\{X \geq b+a\}}{P\{X \geq b\}} = \frac{p^{b+a}}{p^b} = p^a = P\{X \geq a\}.$$

Conversely, taking $a = 1$, then vy the memorylessness property, the conditional probabilities

$$P\{X \geq b+1 | X \geq b\} = \frac{P\{X \geq b+1\}}{P\{X \geq b\}}$$

do not depend on b. Call their common value p. Then

$$P\{X > b\} = P\{X \geq b+1\} = pP\{X \geq b\} = p^2 P\{X \geq b-1\} = \cdots = p^{b+1} P\{X \geq 0\} = p^{b+1},$$

The cumulative distribution

$$F_X(b) = 1 - P\{X > b\} = 1 - p^{b+1},$$

and X is a geometric random variable.

7.26. $P\{X \leq 3\} = F_X(3) = .8024691$, $P\{2 < X \leq 5\} = F_X(5) - F_X(2) = 0.9122085 - 0.7037037 = 0.2085048$, and $P\{X > 4\} - 1 - F_X(4) = 1 - 0.8683128 = 0.1316872$.

7.28. Let f_X be the density. Then

$$0 \leq P\{X = x_0\} \leq P\{x_0 - \Delta x < X \leq x + \Delta x\} = \int_{x_0 - \Delta x}^{x_0 + \Delta x} f_X(x)\, dx.$$

Now the integral goes to 0 as $\Delta x \to 0$. So, we must have $P\{X = x_0\} = 0$.

7.29. Because the density is non-negative on the interval $[0, 1]$, $F_X(x) = 0$ if $x < 0$ and $F_X(x) = 1$ if $x \geq 1$. For x between 0 and 1,

$$\int_0^x \frac{1}{2\sqrt{t}}\, dt = \sqrt{t}\Big|_0^x = \sqrt{x}.$$

Thus,

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \sqrt{x} & \text{if } 0 < x < 1, \\ 1 & \text{if } 1 \leq x. \end{cases}$$

7.32. The random variable Y has distribution function

$$F_Y(y) = P\{Y \leq y\} = P\{aX + b \leq y\} = P\{aX \leq y - b\}.$$

For $a > 0$

$$F_Y(y) = P\left\{X \leq \frac{y-b}{a}\right\} = F_X\left(\frac{y-b}{a}\right).$$

Now take a derivative and use the chain rule to find the density

$$f_Y(y) = F'_Y(y) = f_X\left(\frac{y-b}{a}\right)\frac{1}{a} = \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right).$$

For $a < 0$

$$F_Y(y) = P\left\{X \geq \frac{y-b}{a}\right\} = 1 - F_X\left(\frac{y-b}{a}\right).$$

Now the derivative

$$f_Y(y) = F'_Y(y) = -f_X\left(\frac{y-b}{a}\right)\frac{1}{a} = \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right).$$

7.33. First, notice that the sum of right continuous functions is right continuous. Then, check the properties in Exercise 7.12 using the basic properties of limits and of right continuity.

7.34. Because the f_i are mass functions, $f_i(x) \geq 0$ for all x. Using the fact that the $\pi_i \geq 0$ for all i, we have that $\pi_i f_i(x) \geq 0$ and thus their sum, $f_X(x) \geq 0$. Also, for each i,

$$\sum_x f_i(x) = 1 \quad \text{and} \quad \sum_{i=1}^n \pi_i = 1.$$

Therefore,

$$\sum_x f(x) = \sum_x \left(\sum_{i=1}^n \pi_i f_i(x)\right) = \sum_{i=1}^n \pi_i \left(\sum_x f_i(x)\right) = \sum_{i=1}^n \pi_i(1) = 1.$$

7.35. We enter π and f_1, f_2, f_3 into R and use matrix multiplication.

```
> pi<-c(1/4,1/4,1/2)
> f<-matrix(c(0.2,0.3,0.1,0.4,0,0.5,0.5,0,0,0,0.1,0.1,0.2,0.2,0.4),ncol=3)
> f
     [,1] [,2] [,3]
[1,]  0.2  0.5  0.1
[2,]  0.3  0.5  0.1
[3,]  0.1  0.0  0.2
[4,]  0.4  0.0  0.2
[5,]  0.0  0.0  0.4
> f%*%pi
      [,1]
[1,] 0.225
[2,] 0.250
[3,] 0.125
[4,] 0.200
[5,] 0.200
```

So the mixture distribution is

x	1	2	3	4	5
$f_X(x)$	0.225	0.250	0.125	0.200	0.200

7.36. The joint density (mass function) for $X_1, X_2, \ldots, X_n$

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n)$$

is the product of the marginal densities (mass functions).

7.37. Here is the R code.

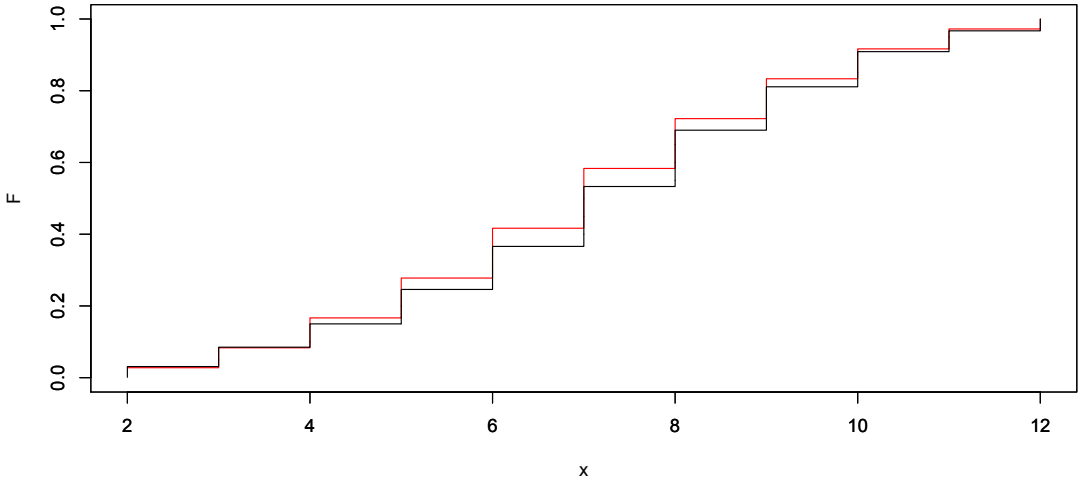


Figure 7.9: Sum on two fair dice. The empirical cumulative distribution function from the simulation (in black) and the cumulative distribution function (in red) are shown for Exercise 7.35.

```
> x<-c(2:12)
> x
 [1]  2  3  4  5  6  7  8  9 10 11 12
> f<-c(1,2,3,4,5,6,5,4,3,2,1)/36
> sum(f)
[1] 1
> (twodice<-sample(x,20,replace=TRUE,prob=f))
 [1]  9  7  3  9  3  6  9  5  5  5  5 10 10 12  9  8  6  8 11  8
> twodice<-sample(x,1000,replace=TRUE,prob=f)
> counts<-rep(0,max(x)-min(x)+1)
> for (i in min(x):max(x)){counts[i]<-length(twodice[twodice==i])}
> freq=counts/(sum(counts))
> data.frame(x,f,freq[min(x):max(x)])
    x          f freq.min.x..max.x..
1   2 0.02777778              0.031
2   3 0.05555556              0.054
3   4 0.08333333              0.065
4   5 0.11111111              0.096
5   6 0.13888889              0.120
6   7 0.16666667              0.167
7   8 0.13888889              0.157
8   9 0.11111111              0.121
9  10 0.08333333              0.098
10 11 0.05555556              0.058
11 12 0.02777778              0.033
```

We also have a plot to compare the empirical cumulative distribution function from the simulation with the cumulative distribution function.

```
> plot(sort(twodice),1:length(twodice)/length(twodice),type="s",xlim=c(2,12),
```

```
ylim=c(0,1),xlab="",ylab="")
> par(new=TRUE)
> plot(x,F,type="s",xlim=c(2,12),ylim=c(0,1),col="red")
```

7.40. Using the information from Exercise 7.35, we have

```
> data<-rep(0,10000)
> for (i in 1:10000){toss<-sample(1:3,1,prob=pi);
  data[i]<-sample(1:5,1,prob=f[,toss])}
> table(data)
data
   1    2    3    4    5
2260 2522 1249 2000 1969
```

As can be seen from the table below, all of the simulated probabilities are within 0.3% of the distributional values.

x	1	2	3	4	5
$f_X(x)$	0.225	0.250	0.125	0.200	0.200
simulated	0.2260	0.2522	0.1249	0.2000	0.1969

7.43. F_X is increasing and continuous, so the set $\{x; F_X(x) \leq u\}$ is the interval $(-\infty, F_X^{-1}(u)]$. In addition, x is in this inverval precisely when $x \leq F_X^{-1}(u)$.

7.44. Let's find F_V. If $v < 0$, then

$$P\{V \leq v\} = P\{1 - U \leq v\} = P\{1 - v \leq U\} = P\{U \geq 1 - v = 0$$

because U is never greater than $1 - v > 1$. Thus, $F_V(v) = 0$ Similarly, if $v \geq 1$,

$$P\{V \leq v\} = P\{1 - U \leq v\} = P\{1 - v \leq U\} = 1$$

because U is always greater than $1 - v < 0$. Thus, $F_V(v) = 1$. Finally, for $0 \leq v < 1$,

$$F_V(v) = P\{V \leq v\} = P\{1 - U \leq v\} = P\{1 - v \leq U\} = 1 - P\{U < 1 - v\} = 1 - (1 - v) = v.$$

This matches the distribution function of a uniform random variable on $[0, 1]$.

Topic 8

The Expected Value

> *Multiply each gain and loss by the probability of the event on which it depends; compare the total of the result of the gains with that of the losses; the balance is the average required, and is known by the name of the* mathematical expectation.- August de Morgan, *An Essay on Probabilities*, 1838

Among the simplest summaries of quantitative data is the sample mean. Given a random variable, the corresponding concept is given a variety of names, the **distributional mean**, the **expectation** or the **expected value**. We begin with the case of discrete random variables where this analogy is more apparent. The formula for continuous random variables is obtained by approximating with a discrete random variable and noticing that the formula for the expected value is a Riemann sum. Thus, expected values for continuous random variables are determined by computing an integral.

8.1 Definition and Properties

Recall for a data set taking numerical values $x_1, x_2, \ldots, x_n$, one of the methods for computing the sample mean of a real-valued function of the data is accomplished by evaluating the sum,

$$\overline{h(x)} = \sum_x h(x)p(x),$$

where $p(x)$ is the proportion of observations taking the value x.

For a finite sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ and a probability P on Ω, we can define the **expectation** or the **expected value** of a random variable X by an analogous average,

$$EX = \sum_{j=1}^{N} X(\omega_j)P\{\omega_j\}. \tag{8.1}$$

More generally for a real-valued function g of the random vector $X = (X_1, X_2, \ldots, X_n)$, we have the formula

$$Eg(X) = \sum_{j=1}^{N} g(X(\omega_j))P\{\omega_j\}. \tag{8.2}$$

Notice that even though we have this analogy, the two formulas come from very different starting points. The value of $\overline{h(x)}$ is derived from **data** whereas no data are involved in computing $Eg(X)$. The starting point for the expected value is a **probability model**.

Example 8.1. *Roll one die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let X be the value on the die. So, $X(\omega) = \omega$. If the die is fair, then the probability model has $P\{\omega\} = 1/6$ for each outcome ω. Using the formula (8.1), the expected value*

$$\begin{aligned} EX &= 1 \cdot P\{1\} + 2 \cdot P\{2\} + 3 \cdot P\{3\} + 4 \cdot P\{4\} + 5 \cdot P\{5\} + 6 \cdot P\{6\} \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = \frac{7}{2}. \end{aligned}$$

An example of an unfair dice would be the probability with $P\{1\} = P\{2\} = P\{3\} = 1/4$ and $P\{4\} = P\{5\} = P\{6\} = 1/12$. In this case, the expected value

$$EX = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{12} + 5 \cdot \frac{1}{12} + 6 \cdot \frac{1}{12} = \frac{11}{4}.$$

Exercise 8.2. *Use the formula (8.2) with $g(x) = x^2$ to find EX^2 for these two examples.*

Two properties of expectation are immediate from the formula for EX in (8.1):

1. If $X(\omega) \geq 0$ for every outcome $\omega \in \Omega$, then every term in the sum in (8.1) is nonnegative and consequently their sum $EX \geq 0$.

2. Let X_1 and X_2 be two random variables and c_1, c_2 be two real numbers, then by using $g(x_1, x_2) = c_1x_1 + c_2x_2$ and the distributive property to the sum in (8.2), we find out that

$$E[c_1X_1 + c_2X_2] = c_1EX_1 + c_2EX_2.$$

The first of these properties states that nonnegative random variables have nonnegative expected value. The second states that expectation is a linear operation. Taking these two properties together, we say that the operation of taking an expectation

$$X \mapsto EX$$

is a **positive linear functional**. We have studied extensively another example of a positive linear functional, namely, the definite integral

$$g \mapsto \int_a^b g(x)\, dx$$

that takes a continuous positive function and gives the area between the graph of g and the x-axis between the vertical lines $x = a$ and $x = b$. For this example, these two properties become:

1. If $g(x) \geq 0$ for every $x \in [a, b]$, then $\int_a^b g(x)\, dx \geq 0$.

2. Let g_1 and g_2 be two continuous functions and c_1, c_2 be two real numbers, then

$$\int_a^b (c_1g_1(x) + c_2g_2(x))\, dx = c_1 \int_a^b g_1(x)\, dx + c_2 \int_a^b g_2(x)\, dx.$$

This analogy will be useful to keep in mind when considering the properties of expectation.

Example 8.3. *If X_1 and X_2 are the values on two rolls of a fair die, then the expected value of the sum*

$$E[X_1 + X_2] = EX_1 + EX_2 = \frac{7}{2} + \frac{7}{2} = 7.$$

A	B	C	D	E	F	G
ω	$X(\omega)$	x	$P\{\omega\}$	$P\{X=x\}$	$X(\omega)P\{\omega\}$	$xP\{X=x\}$
HHH	3	3	$P\{HHH\}$	$P\{X=3\}$	$X(HHH)P\{HHH\}$	$3P\{X=3\}$
HHT	2		$P\{HHT\}$		$X(HHT)P\{HHT\}$	
HTH	2	2	$P\{HTH\}$	$P\{X=2\}$	$X(HTH)P\{HTH\}$	$2P\{X=2\}$
THH	2		$P\{THH\}$		$X(THH)P\{THH\}$	
HTT	1		$P\{HTT\}$		$X(HHT)P\{HHT\}$	
TTH	1	1	$P\{THT\}$	$P\{X=1\}$	$X(HTH)P\{HTH\}$	$1P\{X=1\}$
THT	1		$P\{TTH\}$		$X(THH)P\{THH\}$	
TTT	0	0	$P\{TTT\}$	$P\{X=0\}$	$X(TTT)P\{TTT\}$	$0P\{X=0\}$

Table I: Developing the formula for EX for the case of the coin tosses.

8.2 Discrete Random Variables

Because sample spaces can be extraordinarily large even in routine situations, we rarely use the probability space Ω as the basis to compute the expected value. We illustrate this with the example of tossing a coin three times. Let X denote the number of heads. To compute the expected value EX, we can proceed as described in (8.1). For the table below, we have grouped the outcomes ω that have a common value $x = 3, 2, 1$ or 0 for $X(\omega)$.

From the definition of expectation in (8.1), EX, the expected value of X is the sum of the values in column F. We want to now show that EX is also the sum of the values in column G.

Note, for example, that, three outcomes HHT, HTH and THH each give a value of 2 for X. Because these outcomes are disjoint, we can add probabilities

$$P\{HHT\} + P\{HTH\} + P\{THH\} = P\{HHT, HTH, THH\}$$

But, the event

$$\{HHT, HTH, THH\} \text{ can also be written as the event } \{X = 2\}.$$

This is shown for each value of x in column C, $P\{X = x\}$, the probabilities in column E are obtained as a sum of probabilities in column D.

Thus, by combining outcomes that result in the same value for the random variable, the sums in the boxes in column F are equal to the value in the corresponding box in column G. and thus their total sums are the same. In other words,

$$EX = 0 \cdot P\{X = 0\} + 1 \cdot P\{X = 1\} + 2 \cdot P\{X = 2\} + 3 \cdot P\{X = 3\}.$$

As in the discussion above, we can, in general, find $Eg(X)$. First, to build a table, denote the outcomes in the probability space Ω as $\omega_1, \ldots, \omega_k, \omega_{k+1}, \ldots, \omega_N$ and the state space for the random variable X as $x_1, \ldots, x_i, \ldots, x_n$.

Note that we have partitioned the sample space Ω into the outcomes ω that result in the same value x for the random variable $X(\omega)$. This is shown by the horizontal lines in the table above showing that $X(\omega_k) = X(\omega_{k+1}) = \cdots = x_i$. The equality of sum of the probabilities in a box in columns D and and the probability in column E can be written

$$\sum_{\omega; X(\omega)=x_i} P\{\omega\} = P\{X = x_i\}.$$

For these particular outcomes, $g(X(\omega)) = g(x_i)$ and the sum of the values in a boxes in column F,

$$\sum_{\omega; X(\omega)=x_i} g(X(\omega))P\{\omega\} = \sum_{\omega; X(\omega)=x_i} g(x_i)P\{\omega\} = g(x_i)P\{X = x_i\},$$

A	B	C	D	E	F	G
ω	$X(\omega)$	x	$P\{\omega\}$	$P\{X=x\}$	$g(X(\omega))P\{\omega\}$	$g(x)P\{X=x\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_k	$X(\omega_k)$		$P\{\omega_k\}$		$g(X(\omega_k))P\{\omega_k\}$	
ω_{k+1}	$X(\omega_{k+1})$	x_i	$P\{\omega_{k+1}\}$	$P\{X=x_i\}$	$g(X(\omega_{k+1}))P\{\omega_{k+1}\}$	$g(x_i)P\{X=x_i\}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table II: Establishing the identity (8.3) from (8.2). Arrange the rows of the table so that common values of $X(\omega_k), X(\omega_{k+1}), \ldots$ in the box in column B have the value x_i in column C. Thus, the probabilities in a box in column D sum to give the probability in the corresponding box in column E. Because the values for $g(X(\omega_k)), g(X(\omega_{k+1})), \ldots$ equal $g(x_i)$, the sum in a box in column F sums to the value in the corresponding box in column G. Thus, the sums in columns F and G are equal. The sum in column F is the definition in (8.2). The sum in column G is the identity (8.3).

the value in the corresponding box in column G. Now, sum over all possible value for X for each side of this equation.

$$Eg(X) = \sum_{\omega} g(X(\omega))P\{\omega\} = \sum_{i=1}^{n} g(x_i)P\{X=x_i\} = \sum_{i=1}^{n} g(x_i)f_X(x_i)$$

where $f_X(x_i) = P\{X=x_i\}$ is the probability mass function for X.

The identity

$$Eg(X) = \sum_{i=1}^{n} g(x_i)f_X(x_i) = \sum_{x} g(x)f_X(x) \tag{8.3}$$

is the most frequently used method for computing the expectation of discrete random variables. We will soon see how this identity can be used to find the expectation in the case of continuous random variables

Example 8.4. *Flip a biased coin twice and let X be the number of heads. Then, to compute the expected value of X and X^2 we construct a table to prepare to use (8.3).*

x	$f_X(x)$	$xf_X(x)$	$x^2f_X(x)$
0	$(1-p)^2$	0	0
1	$2p(1-p)$	$2p(1-p)$	$2p(1-p)$
2	p^2	$2p^2$	$4p^2$
sum	1	$2p$	$2p+2p^2$

Thus, $EX = 2p$ and $EX^2 = 2p + 2p^2$.

Exercise 8.5. *Draw 5 cards from a standard deck. Let X be the number of hearts. Use* R *to find EX and EX^2.*

A similar formula to (8.3) holds if we have a vector of random variables $X = (X_1, X_2, \ldots, X_n)$, f_X, the joint probability mass function and g a real-valued function of $x = (x_1, x_2, \ldots, x_n)$. In the two dimensional case, this takes the form

$$Eg(X_1, X_2) = \sum_{x_1}\sum_{x_2} g(x_1, x_2)f_{X_1,X_2}(x_1, x_2). \tag{8.4}$$

We will return to (8.4) in computing the covariance of two random variables.

8.3 Bernoulli Trials

Bernoulli trials are the simplest and among the most common models for an experimental procedure. Each trial has two possible outcomes, variously called,

heads-tails, yes-no, up-down, left-right, win-lose, female-male, green-blue, dominant-recessive, or **success-failure**

depending on the circumstances. We will use the principles of counting and the properties of expectation to analyze Bernoulli trials. From the point of view of statistics, the data have an **unknown** success parameter p. Thus, the goal of statistical inference is to make as precise a statement as possible for the value of p behind the production of the data. Consequently, any experimenter that uses Bernoulli trials as a model ought to mirror its properties closely.

Example 8.6 (Bernoulli trials)**.** *Random variables $X_1, X_2, \ldots, X_n$ are called a sequence of* **Bernoulli trials** *provided that:*

1. *Each X_i takes on two values, namely, 0 and 1. We call the value 1 a* **success** *and the value 0 a* **failure***.*
2. *Each trial has the same probability for success, i.e., $P\{X_i = 1\} = p$ for each i.*
3. *The outcomes on each of the trials is independent.*

For each trial i, the expected value

$$EX_i = 0 \cdot P\{X_i = 0\} + 1 \cdot P\{X_i = 1\} = 0 \cdot (1-p) + 1 \cdot p = p$$

is the same as the success probability. Let $S_n = X_1 + X_2 + \cdots + X_n$ be the total number of successes in n Bernoulli trials. Using the linearity of expectation, we see that

$$ES_n = E[X_1 + X_2 \cdots + X_n] = p + p + \cdots + p = np,$$

the expected number of successes in n Bernoulli trials is np.

In addition, we can use our ability to count to determine the probability mass function for S_n. Beginning with a concrete example, let $n = 8$, and the outcome

success, fail, fail, success, fail, fail, success, fail.

Using the independence of the trials, we can compute the probability of this outcome:

$$p \times (1-p) \times (1-p) \times p \times (1-p) \times (1-p) \times p \times (1-p) = p^3(1-p)^5.$$

Moreover, any of the possible $\binom{8}{3}$ particular sequences of 8 Bernoulli trials having 3 successes also has probability $p^3(1-p)^5$. Each of the outcomes are mutually exclusive, and, taken together, their union is the event $\{S_8 = 3\}$. Consequently, by the axioms of probability, we find that

$$P\{S_8 = 3\} = \binom{8}{3} p^3(1-p)^5.$$

Returning to the general case, we replace 8 by n and 3 by x to see that any particular sequence of n Bernoulli trials having x successes has probability

$$p^x(1-p)^{n-x}.$$

In addition, we know that we have

$$\binom{n}{x}$$

mutually exclusive sequences of n Bernoulli trials that have x successes. Thus, we have the mass function

$$f_{S_n}(x) = P\{S_n = x\} = \binom{n}{x} p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

The fact that the sum

$$\sum_{x=0}^{n} f_{S_n}(x) = \sum_{x=0}^{n} \binom{n}{x} p^x(1-p)^{n-x} = (p + (1-p))^n = 1^n = 1$$

follows from the binomial theorem. Consequently, S_n is called a **binomial random variable***.*

In the exercise above where X is the number of hearts in 5 cards, let $X_i = 1$ if the i-th card is a heart and 0 if it is not a heart. Then, the X_i are not Bernoulli trials because the chance of obtaining a heart on one card depends on whether or not a heart was obtained on other cards. Still,

$$X = X_1 + X_2 + X_3 + X_4 + X_5$$

is the number of hearts and

$$EX = EX_1 + EX_2 + EX_3 + EX_4 + EX_5 = 1/4 + 1/4 + 1/4 + 1/4 + 1/4 = 5/4.$$

8.4 Continuous Random Variables

For X a continuous random variable with density f_X, consider the discrete random variable $\tilde{X}$ obtained from X by rounding down. Say, for example, we give lengths by rounding down to the nearest millimeter. Thus, $\tilde{X} = 2.134$ meters for any lengths X satisfying 2.134 meters $< X \leq 2.135$ meters.

The random variable $\tilde{X}$ is discrete. To be precise about the rounding down procedure, let Δx be the spacing between values for $\tilde{X}$. Then, $\tilde{x}$, an integer multiple of Δx, represents a possible value for $\tilde{X}$, then this rounding becomes

$$\tilde{X} = \tilde{x} \quad \text{if and only if} \quad \tilde{x} < X \leq \tilde{x} + \Delta x.$$

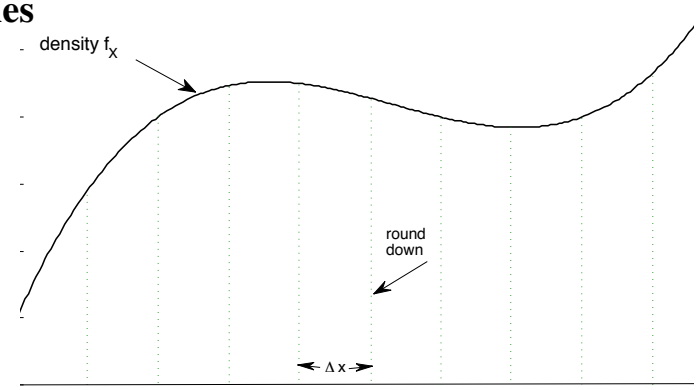


Figure 8.1: The discrete random variable $\tilde{X}$ is obtained by rounding down the continuous random variable X to the nearest multiple of Δx. The mass function $f_{\tilde{X}}(\tilde{x})$ is the integral of the density function from $\tilde{x}$ to $\tilde{x} + \Delta x$ indicated at the area under the density function between two consecutive vertical lines.

With this, we can give the mass function

$$f_{\tilde{X}}(\tilde{x}) = P\{\tilde{X} = \tilde{x}\} = P\{\tilde{x} < X \leq \tilde{x} + \Delta x\}.$$

Now, by the property of the density function,

$$P\{\tilde{x} \leq X < \tilde{x} + \Delta x\} \approx f_X(x)\Delta x. \tag{8.5}$$

In this case, we need to be aware of a possible source of confusion due to the similarity in the notation that we have for both the mass function $f_{\tilde{X}}$ for the discrete random variable $\tilde{X}$ and a density function f_X for the continuous random variable X.

For this discrete random variable $\tilde{X}$, we can use identity (8.3) and the approximation in (8.5) to approximate the expected value.

$$\begin{aligned} Eg(\tilde{X}) &= \sum_{\tilde{x}} g(\tilde{x}) f_{\tilde{X}}(\tilde{x}) = \sum_{\tilde{x}} g(\tilde{x}) P\{\tilde{x} \leq X < \tilde{x} + \Delta x\} \\ &\approx \sum_{\tilde{x}} g(\tilde{x}) f_x(\tilde{x}) \Delta x. \end{aligned}$$

This last sum is a Riemann sum and so taking limits as $\Delta x \to 0$, we have that $\tilde{X}$ converges to X and the Riemann sum converges to the definite integral. Thus,

$$Eg(X) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx. \tag{8.6}$$

As in the case of discrete random variables, a similar formula to (8.6) holds if we have a vector of random variables $X = (X_1, X_2, \ldots, X_n)$, f_X, the joint probability density function and g a real-valued function of the vector $x =$

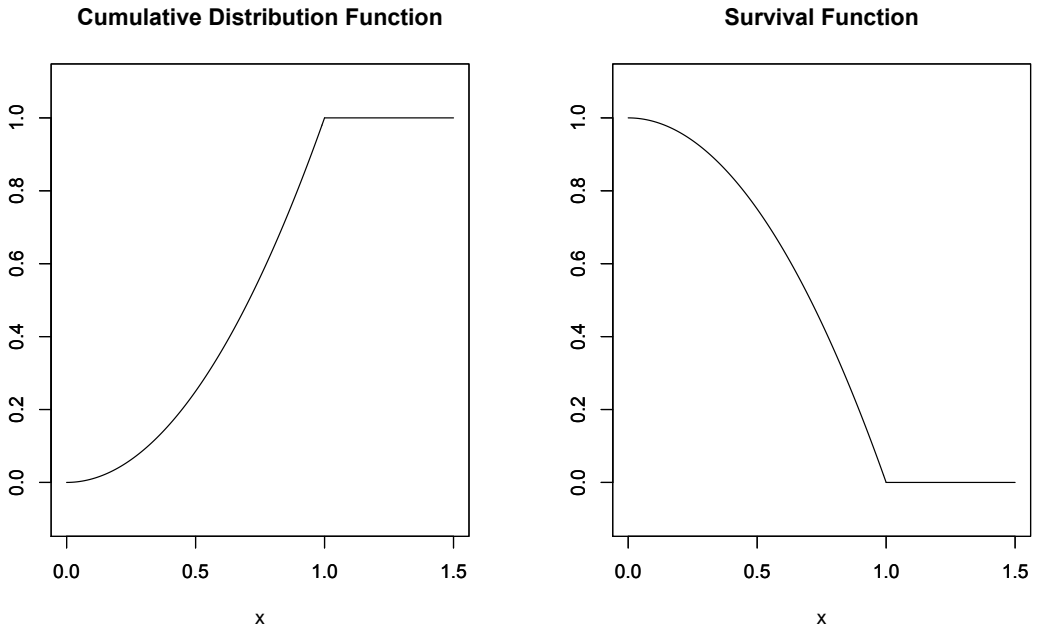


Figure 8.2: The cumulative distribution function $F_X(x)$ and the survival function $\bar{F}_X(x) = 1 - F_X(x)$ for the dart board example. Using the expression (8.7), we see that the expected value $EX = 2/3$ is the area under the survival function.

$(x_1, x_2, \ldots, x_n)$. The expectation in this case is an n-dimensional Riemann integral. For example, if X_1 and X_2 has joint density $f_{X_1,X_2}(x_1, x_2)$, then

$$Eg(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2) f_{X_1,X_2}(x_1, x_2)\, dx_2 dx_1$$

provided that the improper Riemann integral converges.

Example 8.7. *For the dart example, the density $f_X(x) = 2x$ on the interval $[0, 1]$ and 0 otherwise. Thus,*

$$EX = \int_0^1 x \cdot 2x\, dx = \int_0^1 2x^2\, dx = \frac{2}{3} x^3 \Big|_0^1 = \frac{2}{3}.$$

Exercise 8.8. *If X is a nonnegative random variable, then $F_X(0) = 0$.*

If we were to compute the mean of T, an exponential random variable,

$$ET = \int_0^{\infty} t f_T(t)\, dt = \int_0^{\infty} t\lambda e^{-\lambda t}\, dt,$$

then our first step is to integrate by parts. This situation occurs with enough regularity that we will benefit in making the effort to see how integration by parts gives an alternative to computing expectation. In the end, we will see an analogy between the mean with the **survival function** $P\{X > x\} = 1 - F_X(x) = \bar{F}_X(x)$, and the sample mean with the empirical survival function.

Let X be a positive random variable, then the expectation is the improper integral

$$EX = \int_0^{\infty} x f_X(x)\, dx$$

(The unusual choice for v is made to simplify some computations and to anticipate the appearance of the survival function.)

$$\begin{array}{ll} u(x) = x & v(x) = -(1 - F_X(x)) = -\bar{F}_X(x) \\ u'(x) = 1 & v'(x) = f_X(x) = -\bar{F}'_X(x). \end{array}$$

First integrate from 0 to b and take the limit as $b \to \infty$. Then, because $F_X(0) = 0$, $\bar{F}_X(0) = 1$ and

$$\begin{aligned} \int_0^b x f_X(x)\,dx &= -x\bar{F}_X(x)\Big|_0^b + \int_0^b \bar{F}_X(x)\,dx \\ &= -b\bar{F}_X(b) + \int_0^b \bar{F}_X(x)\,dx \end{aligned}$$

The product term in the integration by parts formula converges to 0 as $b \to \infty$. Thus, we can take a limit to obtain the identity,

$$EX = \int_0^\infty P\{X > x\}\,dx. \tag{8.7}$$

Exercise 8.9. *Show that the product term in the integration by parts formula does indeed converge to 0 as $b \to \infty$.*

In words, the expected value is the area between the cumulative distribution function and the line $y = 1$ or the area under the survival function. For the case of the dart board, we see that the area under the distribution function between $y = 0$ and $y = 1$ is $\int_0^1 x^2 dx = 1/3$, so the area below the survival function $EX = 2/3$. (See Figure 8.2.)

Example 8.10. *Let T be an exponential random variable, then for some λ, the survival function $\bar{F}_T(t) = P\{T > t\} = \exp(-\lambda t)$. Thus,*

$$ET = \int_0^\infty P\{T > t\}\,dt = \int_0^\infty \exp(-\lambda t)\,dt = -\frac{1}{\lambda}\exp(-\lambda t)\Big|_0^\infty = 0 - (-\frac{1}{\lambda}) = \frac{1}{\lambda}.$$

Exercise 8.11. *Generalize the identity (8.7) above to X be a positive random variable and g a* non-decreasing *function to show that the expectation*

$$Eg(X) = \int_0^\infty g(x) f_X(x)\,dx = g(0) + \int_0^\infty g'(x) P\{X > x\}\,dx.$$

The most important density function we shall encounter is

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2}), \quad z \in \mathbb{R}.$$

for Z, the standard normal random variable. Because the function ϕ has no simple antiderivative, we must use a numerical approximation to compute the cumulative distribution function, denoted Φ for a standard normal random variable.

Exercise 8.12. *Show that ϕ is increasing for $z < 0$ and decreasing for $z > 0$. In addition, show that ϕ is concave down for z between -1 and 1 and concave up otherwise.*

Example 8.13. *The expectation of a standard normal random variable,*

$$EZ = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty z\exp(-\frac{z^2}{2})\,dz = 0$$

because the integrand is an odd function. Next to evaluate

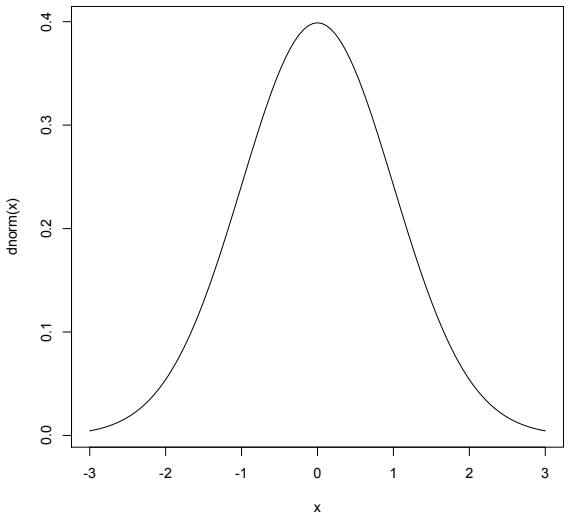


Figure 8.3: The density of a standard normal density, drawn in R using the command `curve(dnorm(x),-3,3)`.

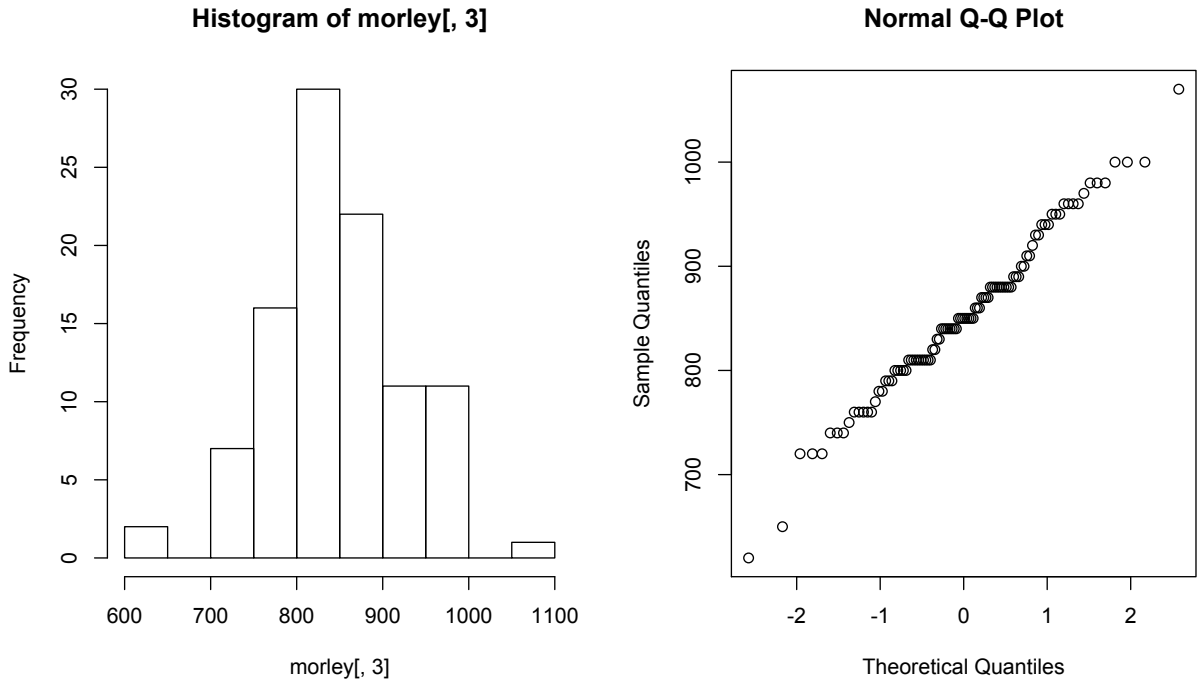


Figure 8.4: Histogram and normal probability plot of Morley's measurements of the speed of light.

$$EZ^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \exp(-\frac{z^2}{2})\, dz,$$

we integrate by parts. (Note the choices of u and v'.)

$$\begin{array}{ll} u(z) = z & v(z) = -\exp(-\frac{z^2}{2}) \\ u'(z) = 1 & v'(z) = z\exp(-\frac{z^2}{2}) \end{array}$$

Thus,

$$EZ^2 = \frac{1}{\sqrt{2\pi}} \left(-z\exp(-\frac{z^2}{2})\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \exp(-\frac{z^2}{2})\, dz \right) = 1.$$

Use l'Hôpital's rule to see that the first term is 0. The fact that the integral of a probability density function is 1 shows that the second term equals 1.

Exercise 8.14. *For Z a standard normal random variable, show that $EZ^3 = 0$ and $EZ^4 = 3$.*

8.5 Quantile Plots and Probability Plots

We have seen the quantile-quantile or Q-Q plot provides a visual method way to compare two quantitative data sets. A more common comparison is between quantitative data and the quantiles of the probability distribution of a continuous random variable. We will demonstrate the properties of these plots with an example.

Example 8.15. *As anticipated by Galileo, errors in independent accurate measurements of a quantity follow approximately a sample from a normal distribution with mean equal to the true value of the quantity. The standard deviation gives information on the precision of the measuring devise. We will learn more about this aspect of measurements when we study the central limit theorem. Our example is Morley's measurements of the speed of light, found in the*

third column of the data set `morley`*. The values are the measurements of the speed of light minus 299,000 kilometers per second.*

```
> length(morley[,3])
[1] 100
> mean(morley[,3])
[1] 852.4
> sd(morley[,3])
[1] 79.01055
> par(mfrow=c(1,2))
> hist(morley[,3])
> qqnorm(morley[,3])
```

The histogram has the characteristic bell shape of the normal density. We can obtain a clearer picture of the closeness of the data to a normal distribution by drawing a **Q-Q plot***. (In the case of the normal distribution, the Q-Q plot is often called the* **normal probability plot***.) One method of making this plot begins by ordering the measurements from smallest to largest:*

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

If these are independent measurements from a normal distribution, then these values should be close to the quantiles of the evenly space values

$$\frac{1}{n+1}, \frac{2}{n+1}, \cdots, \frac{n}{n+1}$$

(For the Morley data, $n = 100$*). Thus, the next step is to find the values in the standard normal distribution that have these quantiles. We can find these values by applying* Φ^{-1}*, the inverse distribution function for the standard normal (*`qnorm` *in* R*), applied to the* n *values listed above. Then the Q-Q plot is the scatterplot of the pairs*

$$\left(x_{(1)}, \Phi^{-1}\left(\frac{1}{n+1}\right)\right), \left(x_{(2)}, \Phi^{-1}\left(\frac{2}{n+1}\right)\right), \ldots, \left(x_{(n)}, \Phi^{-1}\left(\frac{n}{n+1}\right)\right)$$

Then a good fit of the data and a normal distribution can be seen in how well the plot follows a straight line. Such a plot can be seen in Figure 8.4.

Exercise 8.16. *Describe the normal probability plot in the case in which the data* X *are skewed right.*

8.6 Summary

	distribution function $F_X(x) = P\{X \le x\}$	
discrete	random variable	continuous
mass function $f_X(x) = P\{X = x\}$		density function $f_X(x)\Delta x \approx P\{x \le X < x + \Delta x\}$
$f_X(x) \ge 0$ $\sum_{\text{all } x} f_X(x) = 1$	properties	$f_X(x) \ge 0$ $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
$P\{X \in A\} = \sum_{x \in A} f_X(x)$	probability	$P\{X \in A\} = \int_A f_X(x)\, dx$
$Eg(X) = \sum_{\text{all } x} g(x) f_X(x)$	expectation	$Eg(X) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$

8.7 Names for $Eg(X)$.

Several choice for g have special names. We shall later have need for several of these expectations. Others are included to create a comprehensive reference list.

1. If $g(x) = x$, then $\mu = EX$ is called variously the **(distributional) mean**, and the **first moment.**
2. If $g(x) = x^k$, then EX^k is called the k-**th moment**. These names were made in analogy to a similar concept in physics. The second moment in physics is associated to the moment of inertia.
3. For integer valued random variables, if $g(x) = (x)_k$, where $(x)_k = x(x-1)\cdots(x-k+1)$, then $E(X)_k$ is called the k-**th factorial moment.** For random variable taking values in the natural numbers $x = 0, 1, 2, \ldots$, factorial moments are typically easier to compute than moments for these random variables.
4. If $g(x) = (x-\mu)^k$, then $E(X-\mu)^k$ is called the k-**th central moment.**
5. The most frequently used central moment is the second central moment $\sigma^2 = E(X-\mu)^2$ commonly called the **(distributional) variance.** Note that

$$\sigma^2 = \text{Var}(X) = E(X-\mu)^2 = EX^2 - 2\mu EX + \mu^2 = EX^2 - 2\mu^2 + \mu^2 = EX^2 - \mu^2.$$

   This gives a frequently used alternative to computing the variance. In analogy with the corresponding concept with quantitative data, we call σ the **standard deviation**.

   Exercise 8.17. *Find the variance of a single Bernoulli trial.*

   Exercise 8.18. *Compute the variance for the two types of dice in Exercise 8.2.*

   Exercise 8.19. *Compute the variance for the dart example.*

   If we subtract the mean and divide by the standard deviation, the resulting random variable

$$Z = \frac{X-\mu}{\sigma}$$

   has mean 0 and variance 1. Z is called the **standardized version** of X.
6. The third moment of the standardized random variable

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

   is called the **skewness**. Random variables with **positive skewness** have a more pronounced tail to the density on the right. Random variables with **negative skewness** have a more pronounced tail to the density on the left.

   Exercise 8.20. *Show that the skewness of X a Bernoulli random variable $Ber(p)$ is*

$$\frac{1-2p}{\sqrt{p(1-p)}}$$

   Thus, X is positively skewed if $p < 1/2$ and is negatively skewed if $p > 1/2$.
7. The fourth moment of the standard normal random variable is 3. The **kurtosis** compares the fourth moment of the standardized random variable to this value

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3.$$

   Random variables with a negative kurtosis are called **leptokurtic**. Lepto means slender. Random variables with a positive kurtosis are called **platykurtic**. Platy means broad.

8. For d-dimensional vectors $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_d)$ define the **standard inner product**,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{d} x_i y_i.$$

If X is $\mathbb{R}^d$-valued and $g(x) = e^{i\langle \theta, x \rangle}$, then $\chi_X(\theta) = Ee^{i\langle \theta, X \rangle}$ is called the **Fourier transform** or the **characteristic function**. The characteristic function receives its name from the fact that the mapping

$$F_X \mapsto \chi_X$$

from the distribution function to the characteristic function is one-to-one. Consequently, if we have a function that we know to be a characteristic function, then it can only have arisen from one distribution. In this way, χ_X characterizes that distribution.

9. Similarly, if X is $\mathbb{R}^d$-valued and $g(x) = e^{\langle \theta, x \rangle}$, then $M_X(\theta) = Ee^{\langle \theta, X \rangle}$ is called the **Laplace transform** or the **moment generating function**. The moment generating function also gives a one-to-one mapping. However, not every distribution has a moment generating function. To justify the name, consider the one-dimensional case $M_X(\theta) = Ee^{\theta X}$. Then, by noting that

$$\frac{d^k}{d\theta^k} e^{\theta x} = x^k e^{\theta x},$$

we substitute the random variable X for x, take expectation and evaluate at $\theta = 0$.

$$\begin{array}{cc} M_X'(\theta) = EXe^{\theta X} & M_X'(0) = EX \\ M_X''(\theta) = EX^2 e^{\theta X} & M_X''(0) = EX^2 \\ \vdots & \vdots \\ M_X^{(k)}(\theta) = EX^k e^{\theta X} & M_X^{(k)}(0) = EX^k. \end{array}$$

10. Let X have the natural numbers for its state space and $g(x) = z^x$, then $\rho_X(z) = Ez^X = \sum_{x=0}^{\infty} P\{X = x\} z^x$ is called the **(probability) generating function**. For these random variables, the probability generating function allows us to use ideas from the analysis of the complex variable power series.

Exercise 8.21. *Show that the moment generating function for an exponential random variable is*

$$M_X(t) = \frac{\lambda}{\lambda - t}.$$

Use this to find $\text{Var}(X)$.

Exercise 8.22. *For the probability generating function, show that* $\rho_X^{(k)}(1) = E(X)_k$. *This gives an instance that shows that falling factorial moments are easier to compute for natural number valued random variables.*

Particular attention should be paid to the next exercise.

Exercise 8.23. Quadratic identity for variance $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

The variance is meant to give a sense of the spread of the values of a random variable. Thus, the addition of a constant b should not change the variance. If we write this in terms of standard deviation, we have that

$$\sigma_{aX+b} = |a| \sigma_X.$$

Thus, multiplication by a factor a spreads the data, as measured by the standard deviation, by a factor of $|a|$. For example

$$\text{Var}(X) = \text{Var}(-X).$$

These identities are identical to those for a sample variance s^2 and sample standard deviation s.

8.8 Independence

Expected values in the case of more than one random variable is based on the same concepts as for a single random variable. For example, for two discrete random variables X_1 and X_2, the expected value is based on the **joint mass function** $f_{X_1,X_2}(x_1, x_2)$. In this case the expected value is computed using a double sum seen in the identity (8.4).

We will not investigate this in general, but rather focus on the case in which the random variables are independent. Here, we have the factorization identity $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$ for the joint mass function. Now, apply identity (8.4) to the product of functions $g(x_1, x_2) = g_1(x_1)g_2(x_2)$ to find that

$$\begin{aligned} E[g_1(X_1)g_2(X_2)] &= \sum_{x_1}\sum_{x_2} g_1(x_1)g_2(x_2)f_{X_1,X_2}(x_1,x_2) = \sum_{x_1}\sum_{x_2} g_1(x_1)g_2(x_2)f_{X_1}(x_1)f_{X_2}(x_2) \\ &= \left(\sum_{x_1} g_1(x_1)f_{X_1}(x_1)\right)\left(\sum_{x_2} g_2(x_2)f_{X_2}(x_2)\right) = E[g_1(X_1)]\cdot E[g_2(X_2)] \end{aligned}$$

A similar identity holds for continuous random variables - the expectation of the product of two independent random variables equals to the product of the expectation.

8.9 Covariance and Correlation

A very important example begins by taking X_1 and X_2 random variables with respective means μ_1 and μ_2. Then by the definition of variance

$$\begin{aligned} \text{Var}(X_1 + X_2) &= E[((X_1 + X_2) - (\mu_1 + \mu_2))^2] \\ &= E[((X_1 - \mu_1) + (X_2 - \mu_2))^2] \\ &= E[(X_1 - \mu_1)^2] + 2E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &\quad + E[(X_2 - \mu_2)^2] \\ &= \text{Var}(X_1) + 2\text{Cov}(X_1, X_2) + \text{Var}(X_2). \end{aligned}$$

where the **covariance** $\text{Cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$.

As you can see, the definition of covariance is analogous to that for a sample covariance. The analogy continues to hold for the **correlation** ρ, defined by

$$\rho(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)}\sqrt{\text{Var}(X_2)}}.$$

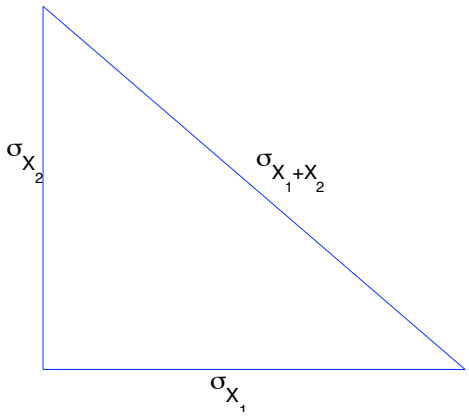


Figure 8.5: For independent random variables, the standard deviations σ_{X_1} and σ_{X_2} satisfy the Pythagorean theorem $\sigma^2_{X_1+X_2} = \sigma^2_{X_1} + \sigma^2_{X_2}$.

We can also use the computation for sample covariance to see that distributional covariance is also between -1 and 1. Correlation 1 occurs only when X and Y have a perfect positive linear association. Correlation -1 occurs only when X and Y have a perfect negative linear association.

If X_1 and X_2 are independent, then $\text{Cov}(X_1, X_2) = E[X_1 - \mu_1]\cdot E[X_2 - \mu_2] = 0$ and the variance of the sum is the sum of the variances. This identity and its analogy to the Pythagorean theorem is shown in Figure 8.5.

The following exercise is the basis in Topic 3 for the simulation of scatterplots having correlation ρ.

Exercise 8.24. *Let X and Z be independent random variables mean 0, variance 1. Define $Y = \rho_0 X + \sqrt{1-\rho_0^2}Z$. Then Y has mean 0, variance 1. Moreover, X and Y have correlation ρ_0*

We can extend this to a generalized Pythagorean identity for n independent random variable $X_1, X_2, \ldots, X_n$ each having a finite variance. Then, for constants $c_1, c_2, \ldots, c_n$, we have the identity

$$\text{Var}(c_1X_1 + c_2X_2 + \cdots c_nX_n) = c_1^2\text{Var}(X_1) + c_2^2\text{Var}(X_2) + \cdots + c_n^2\text{Var}(X_n).$$

We will see several opportunities to apply this identity. For example, if we take $c_1 = c_2 \cdots = c_n = 1$, then we have that for independent random variables

$$\text{Var}(X_1 + X_2 + \cdots X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n),$$

the variance of the sum is the sum of the variances.

Exercise 8.25. *Find the variance of a binomial random variable based on n trials with success parameter p.*

Exercise 8.26. *For random variables $X_1, X_2, \ldots, X_n$ with finite variance and constants $c_1, c_2, \ldots, c_n$*

$$\text{Var}(c_1X_1 + c_2X_2 + \cdots c_nX_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_j\text{Cov}(X_i, X_j).$$

Recall that $\text{Cov}(X_i, X_i) = \text{Var}(X_i)$. If the random variables are independent, then $\text{Cov}(X_i, X_j) = 0$ and the identity above give the generalized Pythagorean identity.

We can write this identity more compactly in matrix form. Let $\mathbf{c}$ be the vector $c_1, c_2, \ldots, c_n$, $X = (X_1, X_2, \ldots X_n$, and define the **covariance matrix** $\text{Cov}(X)$ with i, j entry $\text{Cov}(X_i, X_j)$. Then

$$\text{Var}(\mathbf{c}^T X) = \mathbf{c}^T\text{Cov}(X)\mathbf{c}.$$

8.9.1 Equivalent Conditions for Independence

We can summarize the discussions of independence to present the following 4 equivalent conditions for independent random variables $X_1, X_2, \ldots, X_n$.

1. For events $A_1, A_2, \ldots A_n$,

$$P\{X_1 \in A_1, X_2 \in A_2, \ldots X_n \in A_n\} = P\{X_1 \in A_1\}P\{X_2 \in A_2\} \cdots P\{X_n \in A_n\}.$$

2. The joint distribution function equals to the product of marginal distribution function.

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2) \cdots F_{X_n}(x_n).$$

3. The joint density (mass) function equals to the product of marginal density (mass) functions.

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

4. For bounded functions $g_1, g_2, \ldots, g_n$, the expectation of the product of the random variables equals to the product of the expectations.

$$E[g_1(X_1)g_2(X_2) \cdots g_n(X_n)] = Eg_1(X_1) \cdot Eg_2(X_2) \cdots Eg_n(X_n).$$

We will have many opportunities to use each of these conditions.

8.10 Answers to Selected Exercises

8.2. For the fair die

$$EX^2 = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = (1 + 4 + 9 + 16 + 25 + 36) \cdot \frac{1}{6} = \frac{91}{6}.$$

For the unfair dice

$$EX^2 = 1^2 \cdot \frac{1}{4} + 2^2 \cdot \frac{1}{4} + 3^2 \cdot \frac{1}{4} + 4^2 \cdot \frac{1}{12} + 5^2 \cdot \frac{1}{12} + 6^2 \cdot \frac{1}{12} = (1 + 4 + 9) \cdot \frac{1}{4} + (16 + 25 + 36) \cdot \frac{1}{12} = \frac{119}{12}.$$

8.5. The random variable X can take on the values 0, 1, 2, 3, 4, and 5. Thus,

$$EX = \sum_{x=0}^{5} x f_X(x) \text{ and } EX^2 = \sum_{x=0}^{5} x^2 f_X(x).$$

The R commands and output follow.

```
> hearts<-c(0:5)
> f<-choose(13,hearts)*choose(39,5-hearts)/choose(52,5)
> sum(f)
[1] 1
> prod<-hearts*f
> prod2<-hearts^2*f
> data.frame(hearts,f,prod,prod2)
  hearts           f       prod      prod2
1      0 0.2215336134 0.00000000 0.00000000
2      1 0.4114195678 0.41141957 0.41141957
3      2 0.2742797119 0.54855942 1.09711885
4      3 0.0815426170 0.24462785 0.73388355
5      4 0.0107292917 0.04291717 0.17166867
6      5 0.0004951981 0.00247599 0.01237995
> sum(prod);sum(prod2)
[1] 1.25
[1] 2.426471
```

Look in the text for an alternative method to find EX.

8.8. If X is a non-negative random variable, then $P\{X > 0\} = 1$. Taking complements, we find that

$$F_X(0) = P\{X \leq 0\} = 1 - P\{X > 0\} = 1 - 1 = 0.$$

8.9. The convergence can be seen by the following argument.

$$0 \leq b(1 - F_X(b)) = b \int_b^\infty f_X(x)\,dx = \int_b^\infty b f_X(x)\,dx \leq \int_b^\infty x f_X(x)\,dx$$

Use the fact that $x \geq b$ in the range of integration to obtain the inequality in the line above.. Because, $\int_0^\infty x f_X(x)\,dx < \infty$ (The improper Riemann integral converges.) we have that $\int_b^\infty x f_X(x)\,dx \to 0$ as $b \to \infty$. Consequently, $0 \leq b(1 - F_X(b)) \to 0$ as $b \to \infty$ by the squeeze theorem.

8.11. The expectation is the integral

$$Eg(X) = \int_0^\infty g(x) f_X(x)\,dx.$$

It will be a little easier to look at $h(x) = g(x) - g(0)$. Then

$$Eg(X) = g(0) + Eh(X).$$

For integration by parts, we have

$$\begin{array}{ll} u(x) = h(x) & v(x) = -(1 - F_X(x)) = -\bar{F}_X(x) \\ u'(x) = h'(x) = g'(x) & v'(x) = f_X(x) = -\bar{F}'_X(x). \end{array}$$

Again, because $F_X(0) = 0$, $\bar{F}_X(0) = 1$ and

$$\begin{aligned} Eh(X) = \int_0^b h(x) f_X(x)\,dx &= -h(x)\bar{F}_X(x)\Big|_0^b + \int_0^b h'(x)(1 - F_X(x))\,dx \\ &= -h(b)\bar{F}_X(b) + \int_0^b g'(x)\bar{F}_X(x)\,dx \end{aligned}$$

To see that the product term in the integration by parts formula converges to 0 as $b \to \infty$, note that, similar to Exercise 8.9,

$$0 \le h(b)(1 - F_X(b)) = h(b)\int_b^\infty f_X(x)\,dx = \int_b^\infty h(b) f_X(x)\,dx \le \int_b^\infty h(x) f_X(x)\,dx$$

The first inequality uses the assumption that $h(b) \ge 0$. The second uses the fact that h is non-decreaasing. Thus, $h(x) \ge h(b)$ if $x \ge b$. Now, because $\int_0^\infty h(x) f_X(x)\,dx < \infty$, we have that $\int_b^\infty h(x) f_X(x)\,dx \to 0$ as $b \to \infty$. Consequently, $h(b)(1 - F_X(b)) \to 0$ as $b \to \infty$ by the squeeze theorem.

8.12. For the density function ϕ, the derivative

$$\phi'(z) = \frac{1}{\sqrt{2\pi}}(-z)\exp(-\frac{z^2}{2}).$$

Thus, the sign of $\phi'(z)$ is opposite to the sign of z, i.e.,

$$\phi'(z) > 0 \text{ when } z < 0 \quad \text{and} \quad \phi'(z) < 0 \text{ when } z > 0.$$

Consequently, ϕ is increasing when z is negative and ϕ is decreasing when z is positive. For the second derivative,

$$\phi''(z) = \frac{1}{\sqrt{2\pi}}\left((-z)^2\exp(-\frac{z^2}{2}) - 1\exp(-\frac{z^2}{2})\right) = \frac{1}{\sqrt{2\pi}}(z^2 - 1)\exp(-\frac{z^2}{2}).$$

Thus,

$$\phi \text{ is concave down} \quad \text{if and only if} \quad \phi''(z) < 0 \quad \text{if and only if} \quad z^2 - 1 < 0.$$

This occurs if and only if z is between -1 and 1.

8.14. As argued above,

$$EZ^3 = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty z^3\exp(-\frac{z^2}{2})\,dz = 0$$

because the integrand is an odd function. For EZ^4, we again use integration by parts,

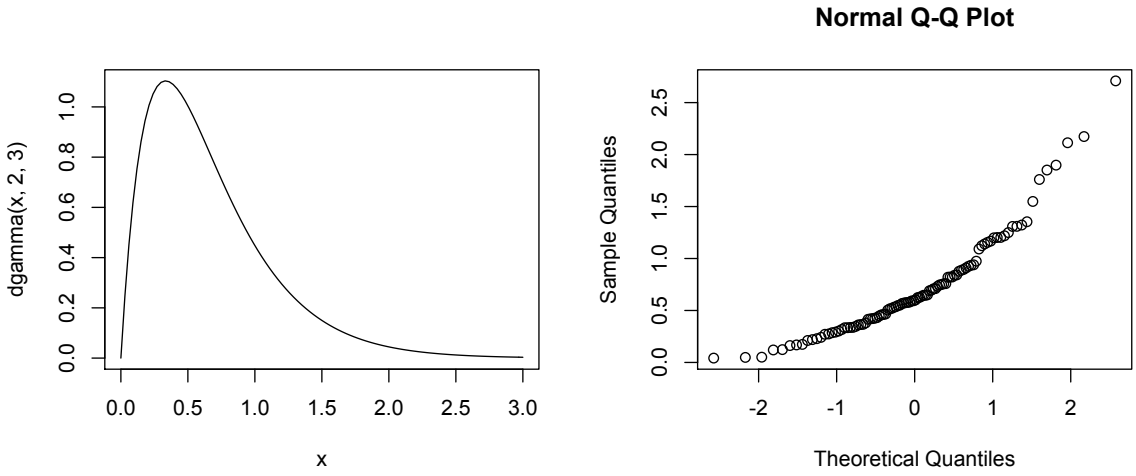
$$\begin{aligned} u(z) &= z^3 & v(z) &= -\exp(-\tfrac{z^2}{2}) \\ u'(z) &= 3z^2 & v'(z) &= z\exp(-\tfrac{z^2}{2}) \end{aligned}$$

Thus,

$$EZ^4 = \frac{1}{\sqrt{2\pi}}\left(-z^3\exp(-\frac{z^2}{2})\Big|_{-\infty}^\infty + 3\int_{-\infty}^\infty z^2\exp(-\frac{z^2}{2})\,dz\right) = 3EZ^2 = 3.$$

Use l'Hôpital's rule several times to see that the first term is 0. The integral is EZ^2 which we have previously found to be equal to 1.

8.16. For the larger order statistics, $z_{(k)}$ for the standardized version of the observations, the values are larger than what one would expect when compared to observations of a standard normal random variable. Thus, the probability plot will have a concave upward shape. As an example, we let X have the density shown below. Beside this is the probability plot for X based on 100 samples. (X is a gamma $\Gamma(2, 3)$ random variable. We will encounter these random variables soon.)



8.17 For a single Bernoulli trial with success probability p, $EX = EX^2 = p$. Thus, $\text{Var}(X) = p - p^2 = p(1-p)$.

8.18. For the fair die, the mean $\mu = EX = 7/2$ and the second moment $EX^2 = 91/6$. Thus,

$$\text{Var}(X) = EX^2 - \mu^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{182 - 147}{12} = \frac{35}{12}.$$

For the unfair die, the mean $\mu = EX = 11/4$ and the second moment $EX^2 = 119/12$. Thus,

$$\text{Var}(X) = EX^2 - \mu^2 = \frac{119}{12} - \left(\frac{11}{4}\right)^2 = \frac{476 - 363}{48} = \frac{113}{48}.$$

8.19. For the dart, we have that the mean $\mu = EX = 2/3$.

$$EX^2 = \int_0^1 x^2 \cdot 2x\, dx = \int_0^1 2x^3\, dx = \frac{2}{4}x^4\Big|_0^1 = \frac{1}{2}.$$

Thus,

$$\text{Var}(X) = EX^2 - \mu^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}.$$

8.20. The third central moment

$$\begin{aligned} E[(X-p)^3] &= (-p)^3 P\{X=0\} + (1-p)^2 P\{X=1\} = -p^3(1-p) + (1-p)^2 p \\ &= p(1-p)(-p^2 + (1-p)^2) = p(1-p)(-p^2 + 1 + 2p + p^2) = p(1-p)(1-2p). \end{aligned}$$

Now, $\sigma^2 = \text{Var}(X) = p(1-p)$. Thus, the skewness,

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = E\left[\left(\frac{X-p}{\sqrt{p(1-p)}}\right)^3\right] = \frac{p(1-p)(1-2p)}{(p(1-p))^{3/2}} = \frac{1-2p}{\sqrt{p(1-p)}}.$$

8.21. If $t < \lambda$, we have that $e^{(t-\lambda)x} \to 0$ as $x \to \infty$ and so

$$M_X(t) = Ee^{tX} = \lambda \int_0^\infty e^{tx} e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{(t-\lambda)x}\, dx = \frac{\lambda}{t-\lambda} e^{(t-\lambda)x}\Big|_0^\infty = \frac{\lambda}{\lambda - t}$$

Thus,

$$M'(t) = \frac{\lambda}{(\lambda - t)^2}, \quad EX = M'(0) = \frac{1}{\lambda},$$

and

$$M''(t) = \frac{2\lambda}{(\lambda - t)^3}, \quad EX = M''(0) = \frac{2}{\lambda^2}.$$

Thus,

$$\text{Var}(X) = EX^2 - (EX)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

8.22. $\rho_X(z) = Ez^X = \sum_{x=0}^\infty P\{X = x\} z^x$ The k-th derivative of z^x with respect to z is

$$\frac{d^k}{dz^k} z^x = (x)_k z^{x-k}.$$

Evaluating at $z = 1$, we find that

$$\frac{d^k}{dz^k} z^x \Big|_{z=1} = (x)_k.$$

Thus the k-th derivative of ρ,

$$\rho_X^{(k)}(z) = \sum_{x=0}^\infty (x)_k P\{X = x\} z^{x-k} \text{ and, thus,}$$

$$\rho_X^{(k)}(1) = \sum_{x=0}^\infty (x)_k P\{X = x\} = E(X)_k.$$

8.23. Let $EX = \mu$. Then the expected value $E[aX + b] = a\mu + b$ and the variance

$$\text{Var}(aX + b) = E[((aX + b) - (a\mu + b))^2] = E[(a(X - \mu))^2] = a^2 E[(X - \mu)^2] = a^2 \text{Var}(X).$$

8.24. By the linearity property of the mean

$$EY = \rho_0 EX + \sqrt{1 - \rho_0^2} EZ = 0.$$

By the Pythorean identity and then the quadratic identity for the variance,

$$\text{Var}(Y) = \text{Var}(\rho_0 X) + \text{Var}(\sqrt{1 - \rho_0^2}\, Z) = \rho_0^2 \text{Var}(X) + (1 - \rho_0^2)\text{Var}(Z) = \rho_0^2 + (1 - \rho_0^2) = 1.$$

Because X and Y both have variance 1, their correlation is equal to their covariance. Now use the linearity property of covariance

$$\rho(X, Y) = \text{Cov}(X, Y) = \text{Cov}\left(X, \rho_0 X + \sqrt{1 - \rho_0^2}\, Z\right) = \rho_0 \text{Cov}(X, X) + \sqrt{1 - \rho_0^2}\, \text{Cov}(X, Z)$$

$$= \rho_0 \cdot 1 + \sqrt{1 - \rho_0^2} \cdot 0 = \rho_0$$

8.25. This binomial random variable is the sum of n independent Bernoulli random variable. Each of these random variables has variance $p(1 - p)$. Thus, the binomial random variable has variance $np(1 - p)$.

Topic 9

Examples of Mass Functions and Densities

For a given **state space**, S, we will describe several of the most frequently encountered parameterized families of both discrete and continuous random variables

$$X : \Omega \to S.$$

indexed by some parameter θ. We will add the subscript θ to the notation P_θ to indicate the parameter value used to compute probabilities. This section is meant to serve as an introduction to these families of random variables and not as a comprehensive development. The section should be considered as a reference to future topics that rely on this information.

We shall use the notation

$$f_X(x|\theta)$$

both for a family of mass functions for discrete random variables and the density functions for continuous random variables that depend on the parameter θ. After naming the family of random variables, we will use the expression $Family(\theta)$ as shorthand for this family followed by the R command and state space S. A table of R commands, parameters, means, and variances is given at the end of this situation.

9.1 Examples of Discrete Random Variables

Incorporating the notation introduced above, we write

$$f_X(x|\theta) = P_\theta\{X = x\}$$

for the mass function of the given family of discrete random variables.

1. (Bernoulli) $Ber(p)$, $S = \{0, 1\}$

$$f_X(x|p) = \left\{ \begin{array}{ll} 0 & \text{with probability } (1-p), \\ 1 & \text{with probability } p, \end{array} \right\} = p^x(1-p)^{1-x}.$$

   This is the simplest random variable, taking on only two values, namely, 0 and 1. Think of it as the outcome of a Bernoulli trial, i.e., a single toss of an unfair coin that turns up heads with probability p.

2. (binomial) $Bin(n, p)$ (R command `binom`) $S = \{0, 1, \ldots, n\}$

$$f_X(x|p) = \binom{n}{x} p^x(1-p)^{n-x}.$$

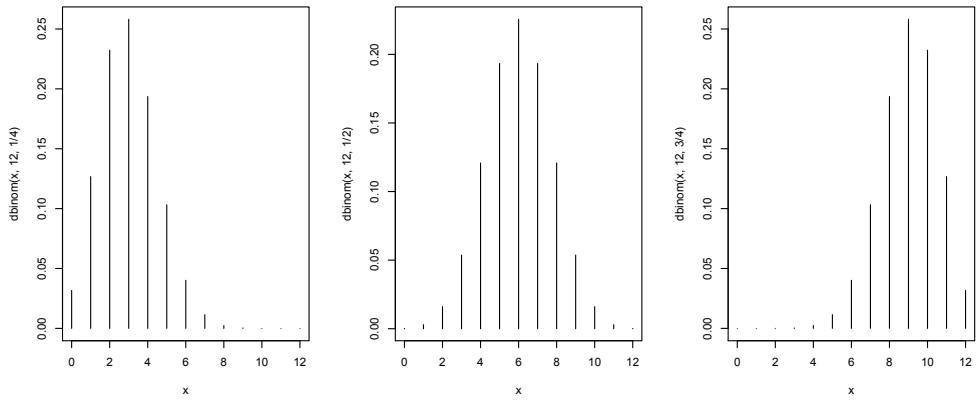


Figure 9.1: Binomial mass function $f(x|p)$ for $n = 12$ and $p = 1/4, 1/2, 3/4$.

We gave a more extensive introduction to Bernoulli trials and the binomial distribution in the discussion on *The Expected Value*. Here we found that the binomial distribution arises from computing the probability of x successes in n Bernoulli trials. Considered in this way, the family $Ber(p)$ is also $Bin(1, p)$.

Notice that by its definition if X_i is $Bin(n_i, p), i = 1, 2$ and are independent, then $X_1 + X_2$ is $Bin(n_1 + n_2, p)$

3. (geometric) $Geo(p)$ (R command `geom`) $S = \mathbb{N} = \{0, 1, 2, \ldots\}$.

$$f_X(x|p) = p(1-p)^x.$$

We previous described random variable as the number of failed Bernoulli trials before the first success. The name geometric random variable is also to the number of Bernoulli trials Y until the first success. Thus, $Y = X + 1$. As a consequence of these two choices for a geometric random variable, care should be taken to be certain which definition is under considertion.

Exercise 9.1. *Give the mass function for Y.*

4. (negative binomial) $Negbin(n, p)$ (R command `nbinom`) $S = \mathbb{N}$

$$f_X(x|p) = \binom{n+x-1}{x} p^n (1-p)^x.$$

This random variable is the number of failed Bernoulli trials before the n-th success. Thus, the family of geometric random variable $Geo(p)$ can also be denoted $Negbin(1, p)$. As we observe in our consideration

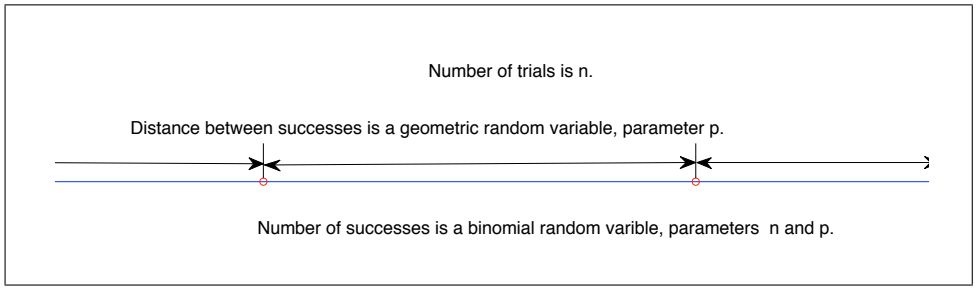


Figure 9.2: The relationship between the binomial and geometric random variable in Bernoulli trials.

of Bernoulli trials, we see that the number of failures between consecutive successes is a geometric random variable. In addition, the number of failures between any two pairs of successes (say, for example, the 2nd and 3rd success and the 6th and 7th success) are independent. In this way, we see that $Negbin(n, p)$ is the sum of n independent $Geo(p)$ random variables.

To find the mass function, note that in order for X to take on a given value x, then the n-th success must occur on the $n + x$-th trial. In other words, we must have $n - 1$ successes and x failures in first $n + x - 1$ Bernoulli trials followed by success on the last trial. The first $n+x-1$ trials and the last trial are independent and so their probabilities multiply.

$$\begin{aligned} P_p\{X = x\} &= P_p\{n-1 \text{ successes in } n+x-1 \text{ trials, success in the } n-x\text{-th trial}\} \\ &= P_p\{n-1 \text{ successes in } n+x-1 \text{ trials}\}P_p\{\text{success in the } n-x\text{-th trial}\} \\ &= \binom{n+x-1}{n-1} p^{n-1}(1-p)^x \cdot p = \binom{n+x-1}{x} p^n(1-p)^x \end{aligned}$$

The first factor is computed from the binomial distribution, the second from the Bernoulli distribution. Note the use of the identity

$$\binom{m}{k} = \binom{m}{m-k}$$

in giving the final formula.

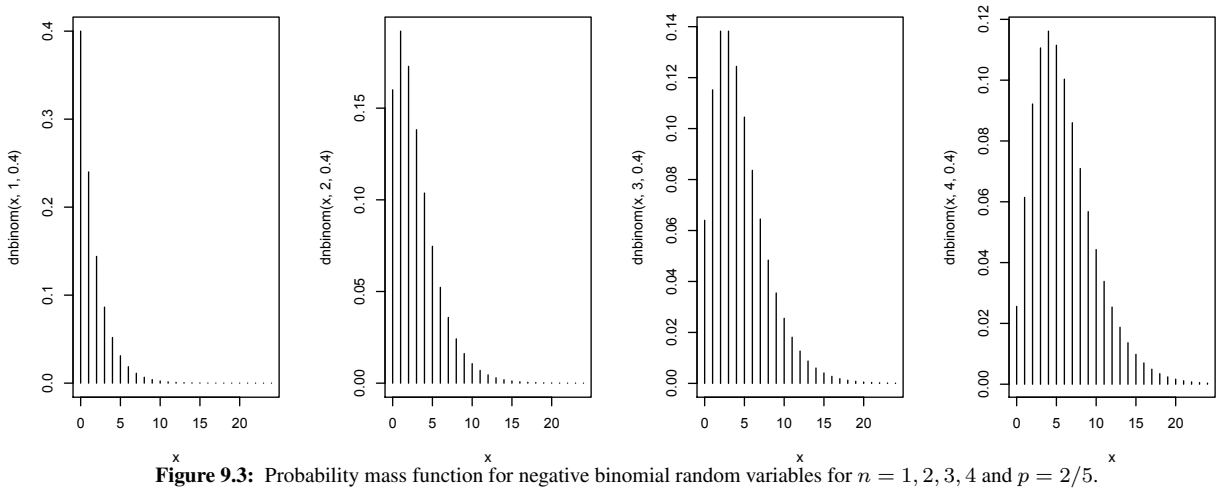


Figure 9.3: Probability mass function for negative binomial random variables for $n = 1, 2, 3, 4$ and $p = 2/5$.

Exercise 9.2. *Use the fact that a negative binomial random variable $Negbin(r, p)$ is the sum of independent geometric random variable $Geo(p)$ to find its mean and variance. Use the fact that a geometric random variable has mean $(1 - p)/p$ and variance $(1 - p)/p^2$.*

5. (Poisson) $Pois(\lambda)$ (R command `pois`) $S = \mathbb{N}$,

$$f_X(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}.$$

The Poisson distribution approximates of the binomial distribution when n is large, p is small, but the product $\lambda = np$ is moderate in size. One example for this can be seen in bacterial colonies. Here, n is the number of bacteria and p is the probability of a mutation and λ, the mean number of mutations is moderate. A second is the

number of recombination events occurring during meiosis. In this circumstance, n is the number of nucleotides on a chromosome and p is the probability of a recombination event occurring at a particular nucleotide.

The approximation is based on the limit

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda} \tag{9.1}$$

We now compute binomial probabilities, replace p by λ/n and take a limit as $n\to\infty$. In this computation, we use the fact that for a fixed value of x,

$$\frac{(n)_x}{n^x}\to 1 \quad \text{and} \quad \left(1-\frac{\lambda}{n}\right)^{-x}\to 1 \quad \text{as } n\to\infty$$

$$\begin{aligned}
P\{X=0\} &= \tbinom{n}{0}p^0(1-p)^n = \left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}\\
P\{X=1\} &= \tbinom{n}{1}p^1(1-p)^{n-1} = n\frac{\lambda}{n}\left(1-\frac{\lambda}{n}\right)^{n-1} \approx \lambda e^{-\lambda}\\
P\{X=2\} &= \tbinom{n}{2}p^2(1-p)^{n-2} = \frac{n(n-1)}{2}\left(\frac{\lambda}{n}\right)^2\left(1-\frac{\lambda}{n}\right)^{n-2} = \frac{n(n-1)}{n^2}\frac{\lambda^2}{2}\left(1-\frac{\lambda}{n}\right)^{n-2} \approx \frac{\lambda^2}{2}e^{-\lambda}\\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots\\
P\{X=x\} &= \tbinom{n}{x}p^x(1-p)^{n-x} = \frac{(n)_x}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} = \frac{(n)_x}{n^x}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^{n-x} \approx \frac{\lambda^x}{x!}e^{-\lambda}.
\end{aligned}$$

The Taylor series for the exponential function

$$\exp\lambda = \sum_{x=0}^{\infty}\frac{\lambda^x}{x!}.$$

shows that

$$\sum_{x=0}^{\infty} f_X(x) = 1.$$

Exercise 9.3. *Take logarithms and use l'Hôpital's rule to establish the limit (9.1) above.*

Exercise 9.4. *We saw that the sum of independent binomial random variables with a common value for p, the success probability, is itself a binomial random variable. Show that the sum of independent Poisson random variables is itself a Poisson random variable. In particular, if X_i are $Pois(\lambda_i), i = 1, 2$, then $X_1 + X_2$ is $Pois(\lambda_1 + \lambda_2)$.*

6. (uniform) $U(a,b)$ (R command `sample`) $S = \{a, a+1, \ldots, b\}$,

$$f_X(x|a,b) = \frac{1}{b-a+1}.$$

Thus each value in the designated range has the same probability.

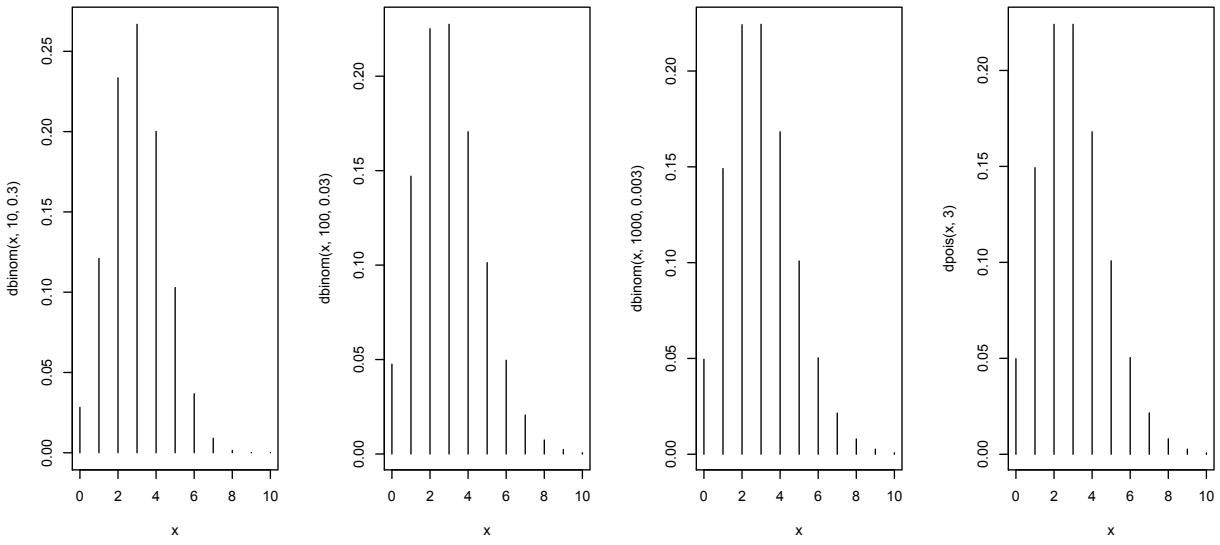


Figure 9.4: Probability mass function for binomial random variables for (a) $n = 10, p = 0.3$, (b) $n = 100, p = 0.03$, (c) $n = 1000, p = 0.003$ and for (d) the Poisson random varialble with $\lambda = np = 3$. This displays how the Poisson random variable approximates the binomial random variable with n large, p small, and their product $\lambda = np$ moderate.

7. (hypergeometric) $Hyper(m, n, k)$ (R command `hyper`). The hypergeometric distribution will be used in computing probabilities under circumstances that are associated with sampling without replacement. We will use the analogy of an urn containing balls having one of two possible colors.

   Begin with an urn holding m white balls and n black balls. Remove k and let the random variable X denote the number of white balls. The value of X has several restrictions. X cannot be greater than either the number of white balls, m, or the number chosen k. In addition, if $k > n$, then we must consider the possibility that all of the black balls were chosen. If $X = x$, then the number of black balls, $k - x$, cannot be greater than the number of black balls, n, and thus, $k - x \leq n$ or $x \geq k - n$.

   If we are considering equally likely outcomes, then we first compute the total number of possible outcomes, $\#(\Omega)$, namely, the number of ways to choose k balls out of an urn containing $m + n$ balls. This is the number of combinations

   $$\binom{m+n}{k}.$$

   This will be the denominator for the probability. For the numerator of $P\{X = x\}$, we consider the outcomes that result in x white balls from the total number m in the urn. We must also choose $k - x$ black balls from the total number n in the urn. By the multiplication property, the number of ways $\#(A_x)$ to accomplish this is product of the number of outcomes for these two combinations,

   $$\binom{m}{x}\binom{n}{k-x}.$$

   The mass function for X is the ratio of these two numbers.

   $$f_X(x|m, n, k) = \frac{\#(A_x)}{\#(\Omega)} = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}, \quad x = \max\{0, k - n\}, \ldots, \min\{m, k\}.$$

Exercise 9.5. *Show that we can rewrite this probability as*

$$f_X(x|m,n,k) = \frac{k!}{x!(k-x)!}\frac{(m)_x(n)_{k-x}}{(m+n)_k} = \binom{k}{x}\frac{(m)_x(n)_{k-x}}{(m+n)_k}. \tag{9.2}$$

This gives probabilities using sampling **without replacement**. If we were to choose the balls one-by-one returning the balls to the urn after each choice, then we would be sampling **with replacement**. This returns us to the case of k Bernoulli trials with success parameter $p = m/(m+n)$, the probability for choosing a white ball. In the case the mass function for Y, the number of white balls, is

$$f_Y(x|m,n,k) = \binom{k}{x}p^x(1-p)^{k-x} = \binom{k}{x}\left(\frac{m}{m+n}\right)^x\left(\frac{n}{m+n}\right)^{k-x} = \binom{k}{x}\frac{m^x n^{k-x}}{(m+n)^k}. \tag{9.3}$$

Note that the difference in the formulas between sampling with replacement in (9.3) and without replacement in (9.2) is that the powers are replaced by the falling function, e.g., m^x is replaced by $(m)_x$.

Let X_i be a Bernoulli random variable indicating whether or not the color of the i-th ball is white. Thus, its mean

$$EX_i = \frac{m}{m+n}.$$

The random variable for the total number of white balls $X = X_1 + X_2 + \cdots + X_k$ and thus its mean

$$EX = EX_1 + EX_2 + \cdots + EX_k = k\frac{m}{m+n}.$$

Because the selection of white for one of the marbles decreases the chance for black for another selection, the trials are not independent. One way to see this is by noting the variance (not derived here) of the sum $X = X_1 + X_2 + \cdots + X_k$

$$\text{Var}(X) = k\frac{m}{m+n}\frac{n}{m+n}\cdot\frac{m+n-k}{m+n-1}$$

is not the sum of the variances.

If we write $N = m + n$ for the total number of balls in the urn and $p = m/(m+n)$ as above, then

$$\text{Var}(X) = kp(1-p)\frac{N-k}{N-1}$$

Thus the variance of the hypergeometric random variable is reduced by a factor of $(N-k)/(N-1)$ from the case of the corresponding binomial random variable. In the cases for which k is much smaller than N, then sampling with and without replacement are nearly the same process - any given ball is unlikely to be chosen more than once under sampling with replacement. We see this situation, for example, in a opinion poll with k at 1 or 2 thousand and n, the population of a country, typically many millions.

On the the other hand, if k is a significant fraction of N, then the variance is significantly reduced under sampling without replacement. We are much less uncertain about the fraction of white and black balls. In the extreme case of $k = N$, we have chosen every ball and know that $X = m$ with probability 1. In the case, the variance formula gives $\text{Var}(X) = 0$, as expected.

Exercise 9.6. *Draw two balls without replacement from the urn described above. Let X_1, X_2 be the Bernoulli random indicating whether or not the ball is white. Find Cov(X_1, X_2).*

Exercise 9.7. *Check that $\sum_{x\in S} f_X(x|\theta) = 1$ in the examples above.*

9.2 Examples of Continuous Random Variables

For continuous random variables, we have for the density

$$f_X(x|\theta) \approx \frac{P_\theta\{x < X \le x + \Delta x\}}{\Delta x}.$$

1. (uniform) $U(a, b)$ (R command `unif`) on $S = [a, b]$,

$$f_X(x|a, b) = \frac{1}{b-a}.$$

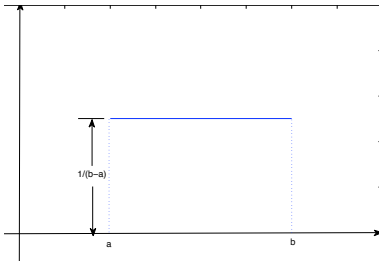


Figure 9.5: Uniform density

Independent $U(0, 1)$ are the most common choice for generating random numbers. Use the R command `runif(n)` to simulate `n` independent random numbers.

Exercise 9.8. *Find the mean and the variance of a $U(a, b)$ random variable.*

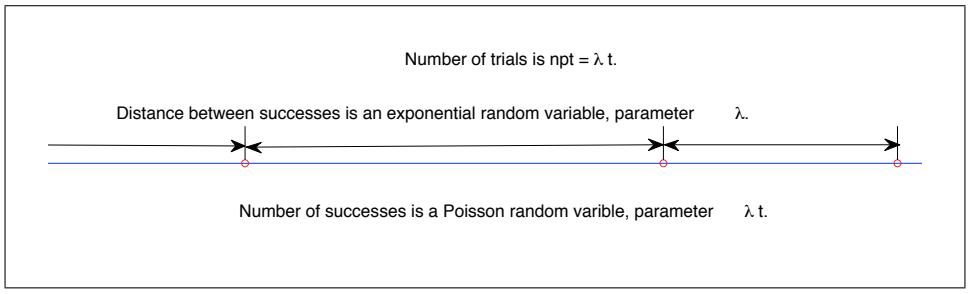


Figure 9.6: The relationship between the Poission and exponential random variable in Bernoulli trials with large n, small p and moderate size product $\lambda = np$. Notice the analogies from Figure 9.2. Imagine a bacterial colony with individual bacterium produced at a constant rate n per unit time. Then, the times between mutations can be approximated by independent exponential random variables and the number of mutations is approximately a Poisson random variable.

2. (exponential) $Exp(\lambda)$ (R command `exp`) on $S = [0, \infty)$,

$$f_X(x|\lambda) = \lambda e^{-\lambda x}.$$

To see how an exponential random variable arises, consider Bernoulli trials arriving at a rate of n trials per time unit and take the approximation seen in the Poisson random variable. Again, the probability of success p is small, ns the number of trials up to a given time s is large, and $\lambda = np$. Let T be the time of the first success. This random time exceeds a given time s if we begin with ns consecutive failures. Thus, the survival function

$$\bar{F}_T(s) = P\{T > s\} = (1-p)^{ns} = \left(1 - \frac{\lambda}{n}\right)^{ns} \approx e^{-\lambda s}.$$

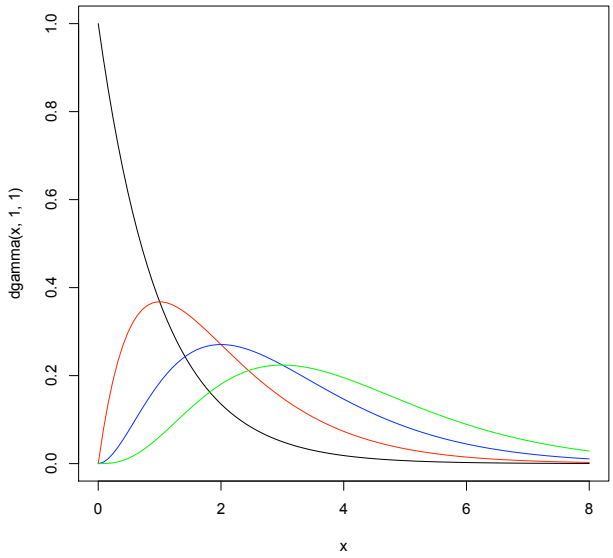


Figure 9.7: Density for a gamma random variable. Here, $\beta = 1$, $\alpha = 1$ (black), 2 (red), 3 (blue) and 4 (green)

The cumulative distribution function

$$F_T(s) = P\{T \leq s\} = 1 - P\{T > s\} \approx 1 - e^{-\lambda s}.$$

The density above can be found by taking a derivative of $F_T(s)$.

Exercise 9.9. *Shoe that the exponential distribution also has the* **memorylessness property**, *namely*

$$P\{T > t + s | T > t\} = P\{T > s\}.$$

In words, given that the wait for an event has taken t time units, then the probability of waiting an additional s time units is the same as the probability of waiting s time units from the beginning.

3. (gamma) $\Gamma(\alpha, \beta)$ (R command `gamma`) on $S = [0, \infty)$,

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Observe that $Exp(\lambda)$ is $\Gamma(1, \lambda)$. A $\Gamma(n, \lambda)$ can be seen as an approximation to the negative binomial random variable using the ideas that leads from the geometric random variable to the exponential. Alternatively, for a natural number n, $\Gamma(n, \lambda)$ is the sum of n independent $Exp(\lambda)$ random variables. This special case of the gamma distribution is sometimes called the **Erlang distribution** and was originally used in models for telephone traffic.

The **gamma function** Γ appears in the definition of the gamma density

$$\Gamma(s) = \int_0^\infty x^s e^{-x} \frac{dx}{x}$$

This is computed in R using `gamma(s)`.

For the graphs of the densities in Figure 9.7,

```
> curve(dgamma(x,1,1),0,8)
> curve(dgamma(x,2,1),0,8,add=TRUE,col="red")
> curve(dgamma(x,3,1),0,8,add=TRUE,col="blue")
> curve(dgamma(x,4,1),0,8,add=TRUE,col="green")
```

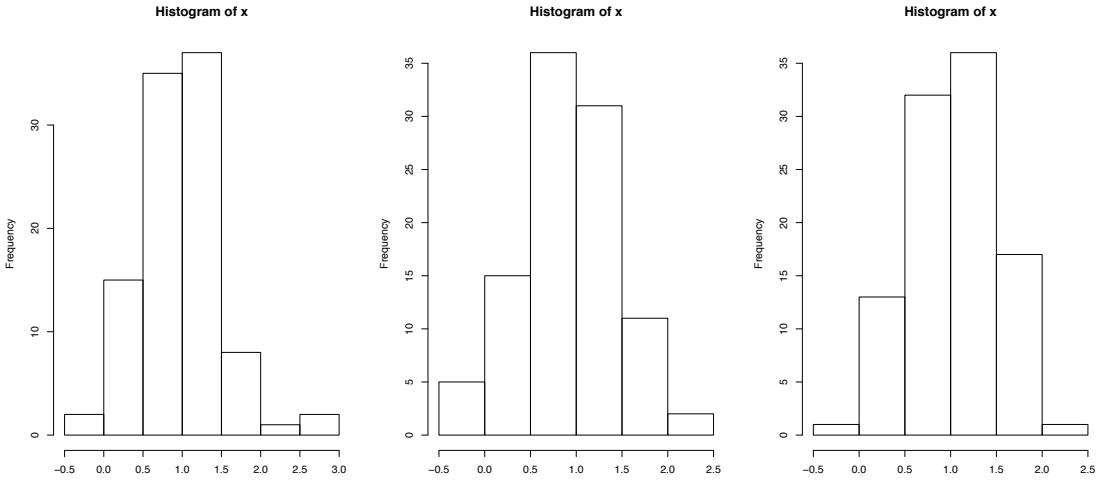


Figure 9.8: Histrogram of three simulations of 200 normal random variables, mean 1, standard deviation 1/2

Exercise 9.10. *Use integration by parts to show that*

$$\Gamma(t+1) = t\Gamma(t). \tag{9.4}$$

If n is a non-negative integer, show that $\Gamma(n) = (n-1)!$.

4. (beta) $Beta(\alpha, \beta)$ (R command `beta`) on $S = [0, 1]$,

$$f_X(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

Beta random variables appear in a variety of circumstances. One common example is the **order statistics**. Beginning with n observations, $X_1, X_2, \cdots, X_n$, of independent uniform random variables on the interval $[0, 1]$ and rank them

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$$

from smallest to largest. Then, the k-th order statistic $X_{(k)}$ is $Beta(k, n-k+1)$.

In the definition of the density, we can also use the **beta function**

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx.$$

As we can see ,using the fact that the integral of a density function equals to 1, that

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The R command for the beta function is `beta`.

Exercise 9.11. *Use the identity (9.4) for the gamma function to find the mean and variance of the beta distribution.*

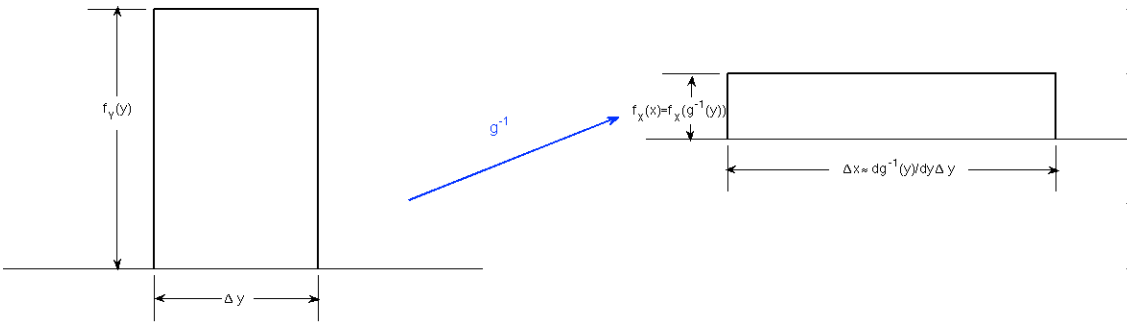


Figure 9.9: Finding the density of $Y = g(X)$ from the density of X. The areas of the two rectangles should be the same. Consequently, $f_Y(y)\Delta y \approx f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)\Delta y$.

5. (normal) $N(\mu, \sigma)$ (R command `norm`) on $S = \mathbb{R}$,

$$f_X(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Thus, a standard normal random variable is $N(0, 1)$. Other normal random variables are linear transformations of Z, the standard normal. In particular, $X = \sigma Z + \mu$ has a $N(\mu, \sigma)$ distribution. To simulate 200 normal random variables with mean 1 and standard deviation 1/2, use the R command `x<-rnorm(200,1,0.5)`. Histograms of three simulations are given in the Figure 9.8.

We often move from a random variable X to a function g of the random variable. Thus, $Y = g(X)$. The following exercise give the density f_Y in terms of the density f_X for X in the case that g is a monotone function.

Exercise 9.12. *Show that for g differentiable and monotone then g has a differentiable inverse g^{-1} and the density*

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|. \tag{9.5}$$

We can find this density geometrically by noting that for g increasing,

Y is between y and $y + \Delta y$ if and only if X is between $g^{-1}(y)$ and $g^{-1}(y + \Delta y) \approx g^{-1}(y) + \frac{d}{dy}g^{-1}(y)\Delta y$.

Thus,

$$f_Y(y)\Delta y = P\{y < Y \le y + \Delta y\} \approx P\{g^{-1}(y) < X \le g^{-1}(y) + \frac{d}{dy}g^{-1}(y)\Delta y\}$$
$$\approx f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)\Delta y.$$

Dropping the factors Δy gives (9.5). (See Figure 9.9.)

6. (log-normal) $\ln N(\mu, \sigma)$ (R command `lnorm`) on $S = (0, \infty)$. A **log-normal** random variable is the exponential of a normal random variable. Thus, the logarithm of a log-normal random variable is normal. The density of this family is

$$f_X(x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right).$$

Exercise 9.13. *Use the exercise above to find the density of a log-normal random variable.*

7. (Pareto) $Pareto(\alpha, \beta)$ (R command `pareto`) on $S = (\alpha, \infty)$. The Pareto distribution is used as a **power law distribution** used by a variety of disciplines. The density of this family is

$$f_X(x|\alpha, \beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}}.$$

The `pareto` command is not a part of the main R package, but can be used after downloading, for example, the `actuar` package.

Exercise 9.14. *Let X be $Exp(\lambda)$. Use the exercise above to show that $\exp(X)$ has a $Pareto(1, \lambda)$ distribution.*

As we stated previously, the normal family of random variables will be the most important collection of distributions we will use. For example, the final three examples, each of them derived from normal random variables, will be seen as densities of statistics used in hypothesis testing. Even though their densities are given explicitly, in practice, these formulas are rarely used directly. Probabilities are generally computed using software.

8. (chi-squared) χ^2_ν (R command `chisq`) on $S = [0, \infty)$

$$f_X(x|\nu) = \frac{x^{\nu/2-1}}{2^{\nu/2}\Gamma(\nu/2)}e^{-x/2}.$$

The value ν is called the number of **degrees of freedom**. For ν a positive integer, let $Z_1, Z_2, \ldots, Z_\nu$ be independent standard normal random variables. Then,

$$Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$$

has a χ^2_ν distribution.

Exercise 9.15. *Modify the solution to the exercise above to find the density of a χ^2_1 random variable. (*Hint*: For $Y = X^2$, $P\{Y \le y\} = P\{-\sqrt{y} \le X \le \sqrt{y}\}$.)*

Exercise 9.16. *Maxwell-Boltzmann distribution models the distribution of particle speeds in an ideal gas in thermal equilibrium. The density function is*

$$f_S(s) = \sqrt{\left(\frac{m}{2\pi kT}\right)^3}\, 4\pi s^2 e^{-ms^2/(2kT)},$$

where m is the particle mass and k is Boltzmann's constant and T the absolute temperature in kelvins.

(a) Let $Y = S\sqrt{m/kT}$, Then, Y has density

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}y^2 e^{-y^2/2,}$$

(b) Show that $X = Y^2$ is χ^2_3, a chi-square random variable with three degrees of freedom.

Therefore, the particle speed

$$S = \sqrt{\frac{kT}{m}}Y = \sqrt{\frac{kTX}{m}}$$

where X is χ^2_3.

9. (Student's t) $t_\nu(\mu, \sigma)$ (R command `t`) on $S = \mathbb{R}$,

$$f_X(x|\nu, \mu.\sigma) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)\sigma}\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}.$$

The value ν is also called the number of **degrees of freedom**. If $\bar{Z}$ is the sample mean of n standard normal random variables and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

is the sample variance, then

$$T = \frac{\sqrt{n}\bar{Z} + a}{S}.$$

has a $t_{n-1}(a, 1)$ distribution. In this case, a is called the **noncentrality parameter**. We shall see this distribution when we consider alternatives to hypotheses whose tests are based on the t distribution. The case $a = 0$

$$T = \frac{\sqrt{n}\bar{Z}}{S} = \frac{\bar{Z}}{S/\sqrt{n}}.$$

is the **classical** t distribution.

10. (Fisher's F) F_{ν_1,ν_2} (R command `f`) on $S = [0, \infty)$,

$$f_X(x|\nu_1, \nu_2) = \frac{\Gamma((\nu_1+\nu_2)/2)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}x^{\nu_1/2-1}(\nu_2 + \nu_1 x)^{-(\nu_1+\nu_2)/2}.$$

The F distribution will make an appearance when we see the analysis of variance test. It arises as the ratio of independent chi-square random variables. The chi-square random variable in the numerator has ν_1 degrees of freedom; the chi-square random variable in the denominator has ν_2 degrees of freedom

9.3 R Commands

R can compute a variety of values for many families of distributions.

- `dfamily(x, parameters)` is the mass function (for discrete random variables) or probability density (for continuous random variables) of `family` evaluated at x.
- `qfamily(p, parameters)` returns x satisfying $P\{X \le x\} = p$, the p-**th quantile** where X has the given distribution,
- `pfamily(x, parameters)` returns $P\{X \le x\}$ where X has the given distribution.
- `rfamily(n, parameters)` generates n random variables having the given distribution.

9.4 Summary of Properties of Random Variables

We have noted that several families of random variables, for example, the gamma random variables, were motivated by having one of its parameters taking on integral values. However, the density function makes sense for a range of real numbers. One additional case where this holds is the **negative binomial** family. Like the gamma family, we use the fact that the gamma function is a generalization of the factorial function to all non-negative numbers. Then the mass function for $Negbin(\alpha, p)$ on the state space $S = \mathbb{N}$ is

$$f_X(x|\alpha, p) = \frac{\Gamma(\alpha + x)}{\Gamma(\alpha)x!}p^\alpha(1-p)^x$$

For the tables below, the parameters are presented in the order required by R.

9.4.1 Discrete Random Variables

random variable	R	parameters	mean	variance	generating function
Bernoulli	*	p	p	$p(1-p)$	$(1-p)+pz$
binomial	`binom`	n, p	np	$np(1-p)$	$((1-p)+pz)^n$
geometric	`geom`	p	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$\frac{p}{1-(1-p)z}$
hypergeometric	`hyper`	m, n, k	$\frac{km}{m+n}$	$k\frac{m}{m+n}\cdot\frac{n}{m+n}\cdot\frac{m+n-k}{m+n-1}$	
negative binomial	`nbinom`	α, p	$\alpha\frac{1-p}{p}$	$\alpha\frac{1-p}{p^2}$	$\left(\frac{p}{1-(1-p)z}\right)^{\alpha}$
Poisson	`pois`	λ	λ	λ	$\exp(-\lambda(1-z))$
uniform	`sample`	a, b	$\frac{b-a+1}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{z^a}{b-a+1}\frac{1-z^{b-a+1}}{1-z}$

*For a Bernoulli random variable, use the binomial commands with `n=1` trial.

Example 9.17. *We give several short examples that use the* R *commands for discrete random variables.*

- *To find the values of the mass function $f_X(x|4, 0.7)$ for a binomial random variable 4 trials with probability of success $p = 0.7$.*

```
> x<-c(0:4)
> binomprob<-dbinom(x,4,0.7)
> data.frame(x,binomprob)
  x binomprob
1 0    0.0081
2 1    0.0756
3 2    0.2646
4 3    0.4116
5 4    0.2401
```

- *We can compare this to simulations of 10,000 independent $Binom(4, 0.7)$ random variables.*

```
> x<-rbinom(10000,4,0.7)
> table(x)/10000
x
     0      1      2      3      4
0.0092 0.0809 0.2562 0.4115 0.2422
> x<-rbinom(10000,4,0.7)
> table(x)/10000
x
     0      1      2      3      4
0.0072 0.0704 0.2720 0.4116 0.2388
```

- *To find the probability $P\{X \leq 3\}$ for X, a geometric random variable with probability of success $p = 0.3$ enter* `pgeom(3,0.3)`. R *returns the answer 0.7599.*

- *To give independent observations uniformly on a set S, use the* `sample` *command using* `replace=TRUE`. *Here is an example using 50 repeated rolls of a die*

```
> S<-c(1:6)
> x<-sample(S,50,replace=TRUE)
> x
 [1] 3 3 4 4 1 3 6 4 3 5 5 1 3 4 6 3 5 1 5 2 6 3 4 4 1 3 1 6 5 4 2 1
[33] 2 3 4 1 2 1 1 6 5 1 2 3 4 5 1 3 6 5
```

9.4.2 Continuous Random Variables

random variable	R	parameters	mean	variance	characteristic function
beta	`beta`	α, β	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$F_{1,1}(a, b; \frac{i\theta}{2\pi})$
chi-squared	`chisq`	ν	ν	2ν	$\frac{1}{(1-2i\theta)^{\nu/2}}$
exponential	`exp`	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{i\lambda}{\theta+i\lambda}$
log-normal	`lnorm`	μ, σ	$\exp(\mu + \sigma^2/2)$	$(e^{\sigma^2} - 1)\exp(2\mu + \sigma^2)$	
F	`f`	ν_1, ν_2	$\frac{\nu_2}{\nu_2-2}, \nu_2 > 2$	$2\nu_2^2 \frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)(\nu_2-2)^2}$	
gamma	`gamma`	α, β	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\left(\frac{i\beta}{\theta+i\beta}\right)^{\alpha}$
normal	`norm`	μ, σ^2	μ	σ^2	$\exp(i\mu\theta - \frac{1}{2}\sigma^2\theta^2)$
Pareto	`pareto`	α, β	$\frac{\alpha\beta}{\beta-1}, (\beta > 1)$	$\frac{\alpha^2\beta}{(\beta-1)^2(\beta-2)}, (\beta > 2)$	
t	`t`	ν, a, σ	$a, (\nu > 1)$	$\sigma^2\frac{a}{a-2}, (\nu > 2)$	
uniform	`unif`	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$-i\frac{\exp(i\theta b)-\exp(i\theta a)}{\theta(b-a)}$

Example 9.18. *We continue with examples that use the* R *commands for continuous random variables.*

- *The standard normal random variable has mean 0 and standard deviation. The value of the distribution function for three standard deviations below and blow the mean.*

```
> data.frame(z,pnorm(z))
   z    pnorm.z.
1 -3 0.001349898
2 -2 0.022750132
3 -1 0.158655254
4  0 0.500000000
5  1 0.841344746
6  2 0.977249868
7  3 0.998650102
```

- *To find the deciles of a gamma random variable with* $\alpha = 4$ *and* $\beta = 5$

```
> decile<-seq(0,0.9,0.1)
> value<-qgamma(decile,4,5)
> data.frame(decile,value)
   decile    value
1     0.0 0.0000000
2     0.1 0.3489539
3     0.2 0.4593574
4     0.3 0.5527422
5     0.4 0.6422646
6     0.5 0.7344121
7     0.6 0.8350525
8     0.7 0.9524458
9     0.8 1.1030091
10    0.9 1.3361566
```

- *The command* `rnorm(200,1,0.5)` *was used to create the histograms in Figure 9.8.*

- *Use the* `curve` *command to plot density and distribution functions. Thus was accomplished in Figure 9.7 using* `dgamma` *for the density of a gamma random variable. For cumulative distribution functions use* `pdist` *and substitute for* `dist` *the appropriate command from the table above.*

To add points for the deciles on the plot of the $\Gamma(4,5)$ density, we use the following R *commands.*

```
> curve(dgamma(x,4,5),0,2.0)
> points(value,dgamma(value,4,5),
   pch=19,col="blue")
```

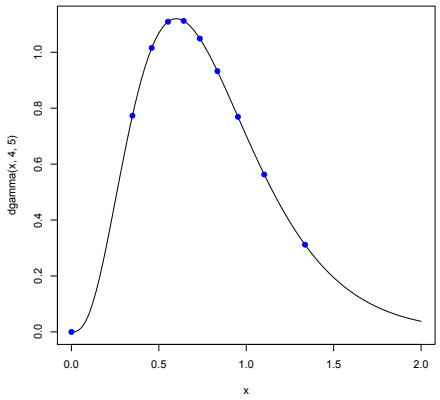


Figure 9.10: Density for a $\Gamma(4,5)$ random variable. The plots contain the values Indicated in blue) on the density plot matching the deciles.

Exercise 9.19. *Add point on the graph of the distribution function for the standard normal corresponding to the standard deviations in the example above.*

9.5 Answers to Selected Exercises

9.1. For $y = 1, 2, \ldots$,

$$f_Y(y) = P\{Y = y\} = P\{X + 1 = y\} = P\{X = y - 1\} = p(1-p)^{y-1}.$$

9.2. Write X a $Negbin(n,p)$ random variable as $X = Y_1 + \cdots + Y_n$ where the Y_i are independent random variable. Then,

$$EX = EY_1 + \cdots + EY_n = \frac{1-p}{p} + \cdots + \frac{1-p}{p} = \frac{n(1-p)}{p}$$

and because the Y_i are independent

$$\text{Var}(X) = \text{Var}(Y_1) + \cdots + \text{Var}(Y_n) = \frac{1-p}{p^2} + \cdots + \frac{1-p}{p^2} = \frac{n(1-p)}{p^2}$$

9.3. By taking the logarithm, the limit above is equivalent to

$$\lim_{n\to\infty} n \ln\left(1 - \frac{\lambda}{n}\right) = -\lambda.$$

Now change variables letting $\epsilon = 1/n$, then the limit becomes

$$\lim_{\epsilon\to 0} \frac{\ln(1-\epsilon\lambda)}{\epsilon} = -\lambda.$$

The limit has the indeterminant form 0/0. Thus, by l'Hôpital's rule, we can take the derivative of the numerator and denominator to obtain the equivalent problem

$$\lim_{\epsilon\to 0} \frac{-\lambda}{1-\epsilon\lambda} = -\lambda.$$

9.4. We have mass functions

$$f_{X_1}(x_1) = \frac{\lambda_1^{x_1}}{x_1!} e^{-\lambda_1} \quad f_{X_2}(x_2) = \frac{\lambda_2^{x_2}}{x_2!} e^{-\lambda_2}$$

Thus,

$$f_{X_1+X_2}(x) = P\{X_1 + X_2 = x\} = \sum_{x_1=0}^{x} P\{X_1 = x_1, X_2 = x - x_1\} = \sum_{x_1=0}^{x} P\{X_1 = x_1\}P\{X_2 = x - x_1\}$$
$$= \sum_{x_1=0}^{x} \frac{\lambda_1^{x_1}}{x_1!} e^{-\lambda_1} \frac{\lambda_2^{x-x_1}}{(x-x_1)!} e^{-\lambda_2} = \left(\sum_{x_1=0}^{x} \frac{1}{x_1!(x-x_1)!} \lambda_1^{x_1} \lambda_2^{x-x_1} \right) e^{-(\lambda_1+\lambda_2)}$$
$$= \left(\sum_{x_1=0}^{x} \frac{x!}{x_1!(x-x_1)!} \lambda_1^{x_1} \lambda_2^{x-x_1} \right) \frac{1}{x!} e^{-(\lambda_1+\lambda_2)} = \frac{(\lambda_1+\lambda_2)^x}{x!} e^{-(\lambda_1+\lambda_2)}.$$

This is the probability mass function for a $Pois(\lambda_1+\lambda_2)$ random variable. The last equality uses the binomial theorem.

9.5. Using the definition of the choose function

$$f_X(x|m,n,k) = \frac{\binom{b}{x}\binom{n}{k-x}}{\binom{m+n}{k}} = \frac{\frac{(m)_x}{x!}\frac{(n)_{k-x}}{(k-x)!}}{\frac{(m+n)_k}{k!}} = \frac{k!}{x!(k-x)!} \frac{(m)_x (n)_{k-x}}{(m+n)_k} = \binom{k}{x} \frac{(m)_x (n)_{k-x}}{(m+n)_k}.$$

9.6. $\text{Cov}(X_1, X_2) = EX_1X_2 - EX_1EX_2$. Now,

$$EX_1 = EX_2 = \frac{m}{m+n} = p$$

and

$$EX_1X_2 = P\{X_1X_2 = 1\} = P\{X_2 = 1, X_1 = 1\} = P\{X_2 = 1|X_1 = 1\}P\{X_1 = 1\} = \frac{m-1}{m+n-1} \cdot \frac{m}{m+n}.$$

Thus,

$$\text{Cov}(X_1, X_2) = \frac{m-1}{m+n-1} \cdot \frac{m}{m+n} - \left(\frac{m}{m+n}\right)^2 = \frac{m}{m+n}\left(\frac{m-1}{m+n-1} - \frac{m}{m+n}\right)$$
$$= \frac{m}{m+n}\left(\frac{(m+n)(m-1) - m(m+n-1)}{(m+n)(m+n-1)}\right) = \frac{m}{m+n}\left(\frac{-n}{(m+n)(m+n-1)}\right) = -\frac{np}{N(N-1)}$$

where, as above, $N = m + n$.

9.8. For the mean

$$E_{(a,b)}X = \int_a^b x f_X(x|a,b)\, dx = \frac{1}{b-a}\int_a^b x\, dx = \frac{1}{2(b-a)} x^2 \Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2},$$

the average of endpoints a and b. For the variance, we first find the second moment

$$E_{(a,b)}X^2 = \int_a^b x^2 f_X(x|a,b)\, dx = \frac{1}{b-a}\int_a^b x^2\, dx = \frac{1}{3(b-a)} x^3 \Big|_a^b = \frac{b^3-a^3}{3(b-a)} \frac{(b-a)(a^2+ab+b^2)}{3(b-a)} = \frac{a^2+ab+b^2}{3}.$$

Thus,

$$\text{Var}_{(a,b)}(X) = \frac{a^2+ab+b^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{4b^2+4ab+4b^2}{12} - \frac{3a^2+6ab+3a^2}{12} = \frac{a^2-2ab+b^2}{12} = \frac{(a-b)^2}{12}.$$

9.9. Because $\{T > t+s\} \subset \{T > t\}$, we have for $T \sim Exp(\lambda)$,

$$P\{T > t+s | T > t\} = \frac{P\{T > t+s, T > t\}}{P\{T > t\}} = \frac{P\{T > t+s\}}{P\{T > t\}} = \frac{\exp(-\lambda(t+s))}{\exp(-\lambda t)} = \exp(-\lambda s) = P\{T > s\}.$$

9.10. Using integration by parts

$$\begin{array}{ll} u(x) = x^t & v(x) = -e^{-x} \\ u'(z) = tx^{t-1} & v'(x) = e^{-x} \end{array}$$

$$\Gamma(t+1) = \int_0^\infty x^t e^{-x}\, dx = -x^t e^{-x}\Big|_0^\infty + t\int_0^\infty x^{t-1}e^{-x}\, dx = t\Gamma(t).$$

The first term is 0 because $x^t e^{-x} \to 0$ as $x \to \infty$.

For the case $n = 1$, $\Gamma(1) = \int_0^\infty e^{-s}\, ds = 1 = (1-1)!$. Now assume that $\Gamma(n) = (n-1)!$ We have just shown this identity in the case $n = 1$. Now,

$$\Gamma(n+1) = n\Gamma(n) = n \cdot (n-1)! = n!.$$

Thus, by induction we have the formula for all integer values.

9.11. In order to be a probability density, we have that

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx.$$

We use this identity and (9.4) to compute the first two moments

$$\begin{aligned} E_{(\alpha,\beta)}X &= \int_0^1 x f_X(x|\alpha,\beta)\, dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^\alpha (1-x)^{\beta-1}\, dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\ &= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+1)}{\Gamma(\alpha+\beta+1)\Gamma(\alpha)} = \frac{\Gamma(\alpha+\beta)\alpha\Gamma(\alpha)}{(\alpha+\beta)\Gamma(\alpha+\beta)\Gamma(\alpha)} = \frac{\alpha}{\alpha+\beta}. \end{aligned}$$

and

$$\begin{aligned} E_{(\alpha,\beta)}X^2 &= \int_0^1 x^2 f_X(x|\alpha,\beta)\, dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^{\alpha+1}(1-x)^{\beta-1}\, dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)} \\ &= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+2)}{\Gamma(\alpha+\beta+2)\Gamma(\alpha)} = \frac{\Gamma(\alpha+\beta)(\alpha+1)\alpha\Gamma(\alpha)}{(\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta)\Gamma(\alpha)} = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}. \end{aligned}$$

Thus,

$$\begin{aligned} \text{Var}_{(\alpha,\beta)}(X) &= E_{(\alpha,\beta)}X^2 - (E_{(\alpha,\beta)}X)^2 = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\ &= \frac{(\alpha+1)\alpha(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta+1)(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \end{aligned}$$

9.12. We first consider the case of g increasing on the range of the random variable X. In this case, g^{-1} is also an increasing function.

To compute the cumulative distribution of $Y = g(X)$ in terms of the cumulative distribution of X, note that

$$F_Y(y) = P\{Y \le y\} = P\{g(X) \le y\} = P\{X \le g^{-1}(y)\} = F_X(g^{-1}(y)).$$

Now use the chain rule to compute the density of Y

$$f_Y(y) = F_Y'(y) = \frac{d}{dy}F_X(g^{-1}(y)) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y).$$

For g decreasing on the range of X,

$$F_Y(y) = P\{Y \le y\} = P\{g(X) \le y\} = P\{X \ge g^{-1}(y)\} = 1 - F_X(g^{-1}(y)),$$

and the density

$$f_Y(y) = F_Y'(y) = -\frac{d}{dy}F_X(g^{-1}(y)) = -f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y).$$

For g decreasing, we also have g^{-1} decreasing and consequently the density of Y is indeed positive,

We can combine these two cases to obtain

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|.$$

9.13. Let X be a normal random variable, then $Y = \exp X$ is log-normal. Thus $y = g(x) = e^x$, $g^{-1}(y) = \ln y$, and $\frac{d}{dy}g^{-1}(y) = \frac{1}{y}$. Note that y must be positive. Thus,

$$f_Y(y|\mu,\sigma) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right)\frac{1}{y}.$$

9.14. Let X be $Exp(\lambda)$ and $y = g(x) = e^x$. As in the previous exercise,

$$f_Y(y|\lambda) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| = \exp(-\lambda\ln(y))\frac{1}{y} = \lambda\exp(\ln(y^{-\lambda}))\frac{1}{y} = \lambda y^{-\lambda}\frac{1}{y} = \frac{\lambda}{y^{\lambda+1}}$$

which is the density of a $Pareto(1,\lambda)$ random variable.

9.15. Let X be a standard normal random variable, then $Y = X^2$ is χ^2_1. From the hint, the distribution function of Y,

$$F_Y(y) = P\{Y \le y\} = P\{-\sqrt{y} \le X \le \sqrt{y}\} = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

Now take a derivative with respect to y.

$$\begin{aligned} f_Y(y) = P\{Y \le y\} &= f_X(\sqrt{y})\left(\frac{1}{2\sqrt{y}}\right) - f_X(-\sqrt{y})\left(-\frac{1}{2\sqrt{y}}\right) \\ &= (f_X(\sqrt{y}) + f_X(-\sqrt{y}))\frac{1}{2\sqrt{y}} \\ &= \left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y}{2}\right) + \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y}{2}\right)\right)\frac{1}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y}{2}\right)\frac{1}{\sqrt{y}} \end{aligned}$$

Finally, $\Gamma(1/2) = \sqrt{\pi}$.

9.16. For both parts, we use the identity in Exercise 9.11.

(a) Let $y = g(s) = s\sqrt{m/kT}$, then $g^{-1}(y) = y\sqrt{kT/m}$ and

$$f_Y(y) = f_S(y\sqrt{kT/m})\sqrt{kT/m} = \frac{1}{\sqrt{(2\pi)^3}}4\pi y^2 e^{-y^2/2} = \frac{1}{\sqrt{2\pi}}y^2 e^{-y^2/2}.$$

(b) Let $x = g(y) = y^2$, then $g^{-1}(x) = \sqrt{x}$ and

$$f_X(x) = \frac{1}{\sqrt{2\pi}}xe^{-x/2}\frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi}}x^{1/2}e^{-x/2}.$$

This is χ^2_3. Notice that $\Gamma(3/2) = \sqrt{\pi}/2$.

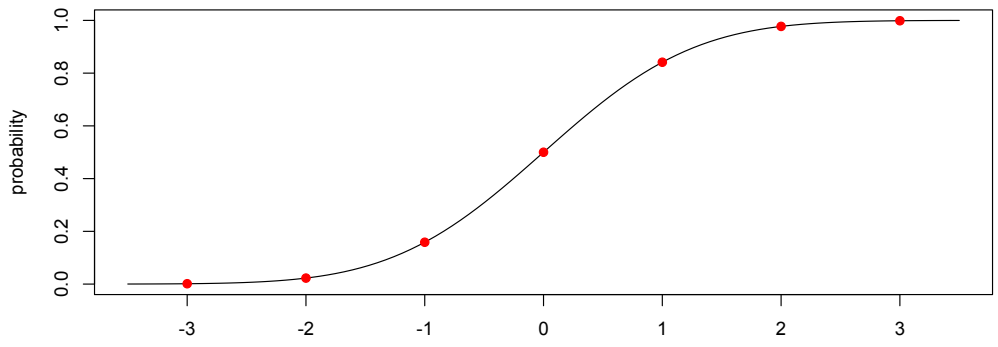


Figure 9.11: Distiribution function to the standard normal Z. Values for $P\{Z \leq z\}$ for $z = -3, -2, -1, 0, 1, 2, 3$ indicated in red.

9.19. First we plot the distribution function for the normal, then we add the points.

```
> curve(pnorm(x),-3.5,3.5,xlab=c("z"),ylab=c("probability"))
> z<- -3:3
> points(z,pnorm(z),pch=19,col="red")
```

.

Topic 10

The Law of Large Numbers

Individuals vary, but percentages remain constant. So says the statistician – Sir Arthur Conan Doyle

10.1 Introduction

A public health official want to ascertain the mean weight of healthy newborn babies in a given region under study. If we randomly choose babies and weigh them, keeping a running average, then, because individual weights vary, at the beginning we might see some larger fluctuations in our average. However, as we continue to make measurements, we expect to see this running average settle and converge to the true mean weight of newborn babies. This phenomena is informally known as the **law of averages**. In probability theory, we call this the **law of large numbers**.

Example 10.1. *We can simulate babies' weights with independent normal random variables, mean 3 kg and standard deviation 0.5 kg. The following* R *commands perform this simulation and computes a running average of the heights. The results are displayed in Figure 10.1.*

```
> n<-c(1:100)
> x<-rnorm(100,3,0.5)
> s<-cumsum(x)
> plot(s/n,xlab="n",ylim=c(2,4),type="l")
```

Here, we begin with a sequence $X_1, X_2, \ldots$ of random variables having a common distribution. Their average, the **sample mean**,

$$\bar{X} = \frac{1}{n}S_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n),$$

is itself a random variable.

If the common mean for the X_i's is μ, then by the linearity property of expectation, the mean of the average,

$$E\left[\frac{1}{n}S_n\right] = \frac{1}{n}ES_n = \frac{1}{n}(EX_1 + EX_2 + \cdots + EX_n) = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{1}{n}n\mu = \mu. \tag{10.1}$$

is also μ.

If, in addition, the X_i's are independent with common variance σ^2, then first by the quadratic identity and then the Pythagorean identity for the variance of independent random variables, we find that the variance of $\bar{X}$,

$$\sigma^2_{\bar{X}} = \text{Var}(\frac{1}{n}S_n) = \frac{1}{n^2}(\text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)) = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2. \tag{10.2}$$

So the mean of these running averages remains at μ but the variance is decreasing to 0 at a rate inversely proportional to the number of terms in the sum. For example, the mean of the average weight of 100 newborn babies is 3

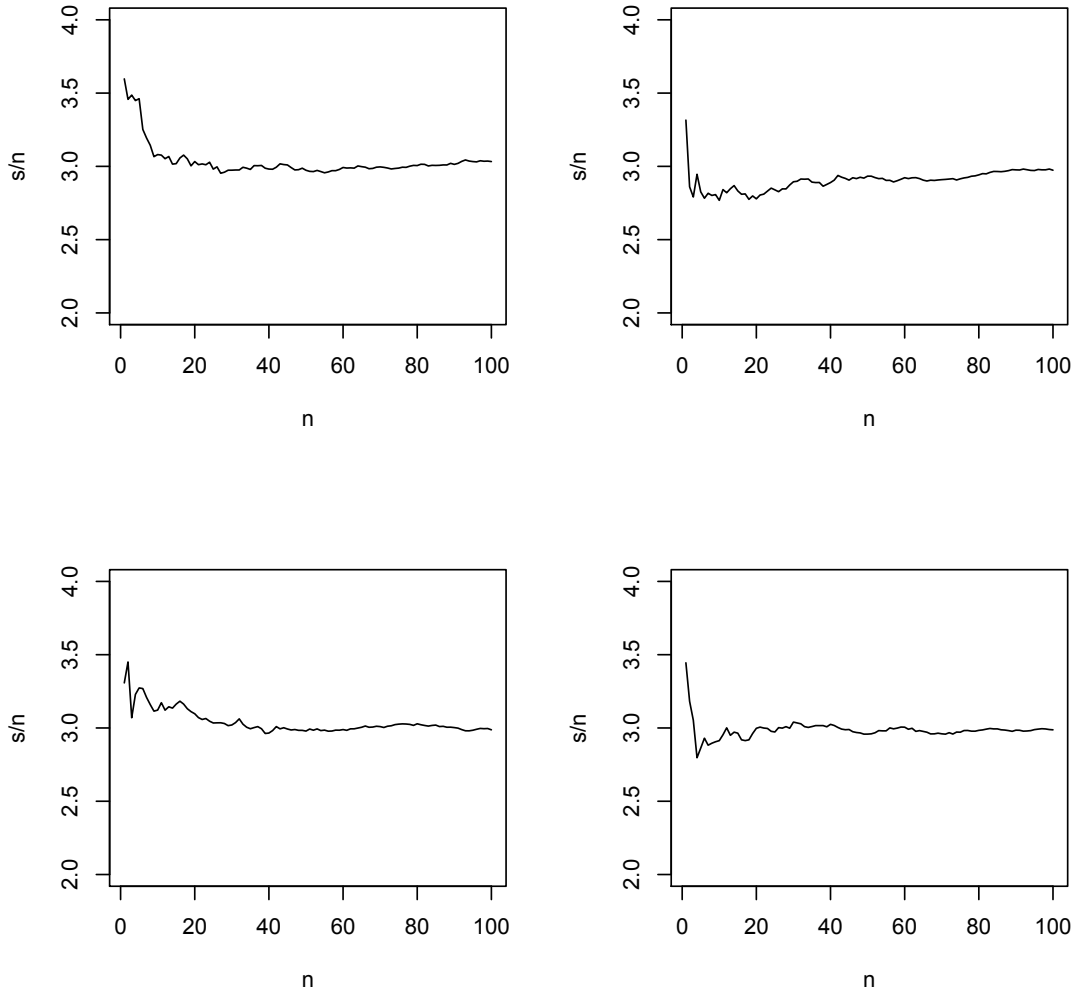


Figure 10.1: Four simulations of the running average S_n/n, $n = 1, 2, \ldots, 100$ for independent normal random variables, mean 3 kg and standard deviation 0.5 kg. Notice that the running averages have large fluctuations for small values of n but settle down converging to the mean value μ = 3 kilograms for newborn birth weight. This behavior could have been predicted using the law of large numbers. The size of the fluctuations, as measured by the standard deviation of S_n/n, is $\sigma/\sqrt{n}$ where σ is the standard deviation of newborn birthweight.

kilograms, the standard deviation is $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 0.5/\sqrt{100} = 0.05$ kilograms = 50 grams. For 10,000 males, the mean remains 3 kilograms, the standard deviation is $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 0.5/\sqrt{10000} = 0.005$ kilograms = 5 grams. Notice that

- as we increase n by a factor of **100**,
- we decrease $\sigma_{\bar{X}}$ by a factor of **10**.

The mathematical result, the **law of large numbers**, tells us that the results of these simulation could have been anticipated.

Theorem 10.2. *For a sequence of independent random variables $X_1, X_2, \ldots$ having a common distribution, their*

running average

$$\frac{1}{n}S_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

has a limit as $n \to \infty$ if and only if this sequence of random variables has a common mean μ. In this case the limit is μ.

The theorem also states that if the random variables do not have a mean, then as the next example shows, the limit will fail to exist. We shall show with the following example. When we look at methods for estimation, one approach, the method of moments, will be based on using the law of large numbers to estimate the mean μ or a function of μ.

Care needs to be taken to ensure that the simulated random variables indeed have a mean. For example, use the `runif` command to simulate uniform transform variables, and choose a transformation $Y = g(U)$ that results in an integral

$$\int_0^1 g(u)\, du$$

that does not converge. Then, if we simulate independent uniform random variables, the running average

$$\frac{1}{n}(g(U_1) + \cdots + g(U_n))$$

will not converge. This issue is the topic of the next exercise and example.

Exercise 10.3. *Let U be a uniform random variable on the interval $[0, 1]$. Give the value for p for which the mean is finite and the values for which it is infinite. Simulate the situation for a value of p for which the integral converges and a second value of p for which the integral does not converge and cheek has in Example 10.1 a plot of S_n/n versus n.*

Example 10.4. *The standard Cauchy random variable X has density function*

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2} \quad x \in \mathbb{R}.$$

Let $Y = |X|$. In an attempt to compute the improper integral for $EY = E|X|$, note that

$$\int_{-b}^{b} |x| f_X(x)\, dx = 2\int_0^b \frac{1}{\pi}\frac{x}{1+x^2}\, dx = \frac{1}{\pi}\ln(1+x^2)\Big|_0^b = \frac{1}{\pi}\ln(1+b^2) \to \infty$$

as $b \to \infty$. Thus, Y has infinite mean. We now simulate 1000 independent Cauchy random variables.

```
> n<-c(1:1000)
> y<-abs(rcauchy(1000))
> s<-cumsum(y)
> plot(s/n,xlab="n",ylim=c(-6,6),type="l")
```

These random variables do not have a finite mean. As you can see in Figure 10.2 that their running averages do not seem to be converging. Thus, if we are using a simulation strategy that depends on the law of large numbers, we need to check that the random variables have a mean.

Exercise 10.5. *Using simulations, check the failure of the law of large numbers of Cauchy random variables. In the plot of running averages, note that the shocks can jump either up or down.*

10.2 Monte Carlo Integration

Monte Carlo methods use stochastic simulations to approximate solutions to questions that are very difficult to solve analytically. This approach has seen widespread use in fields as diverse as statistical physics, astronomy, population genetics, protein chemistry, and finance. Our introduction will focus on examples having relatively rapid computations. However, many research groups routinely use Monte Carlo simulations that can take weeks of computer time to perform.

For example, let $X_1, X_2, \ldots$ be independent random variables uniformly distributed on the interval $[a, b]$ and write

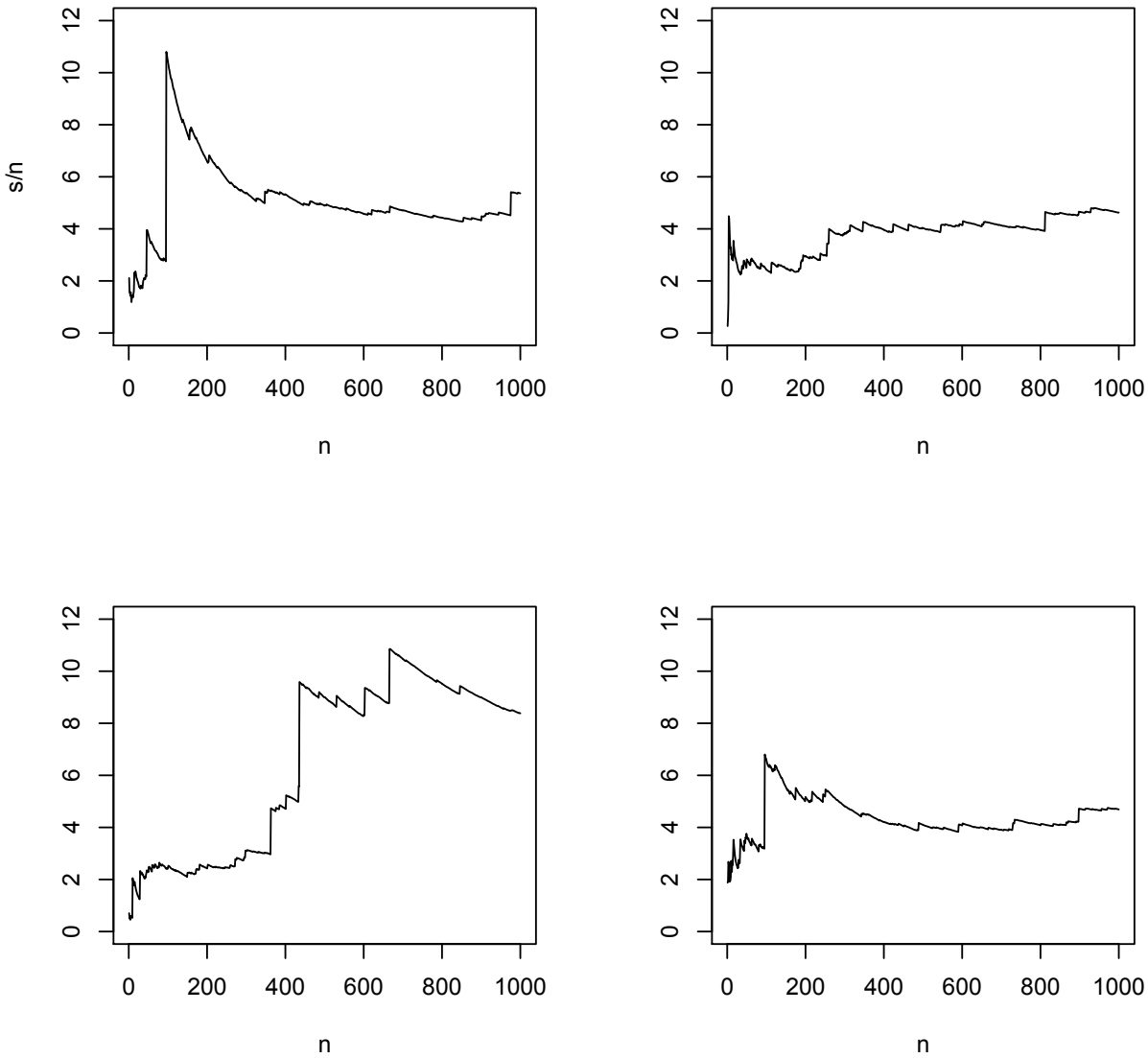


Figure 10.2: Four simulations of the running average S_n/n, $n = 1, 2, \ldots, 1000$ for the absolute value of independent Cauchy random variables. Note that the running average does not seem to be settling down and is subject to "shocks". Because Cauchy random variables do not have a mean, we know, from the law of large numbers, that the running averages do not converge.

f_X for their common density..

Then, by the law of large numbers, for n large we have that

$$\overline{g(X)}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \approx Eg(X_1) = \int_a^b g(x) f_X(x)\, dx = \frac{1}{b-a}\int_a^b g(x)\, dx.$$

Thus,

$$\int_a^b g(x)\, dx. \approx (b-a)\overline{g(X)}_n.$$

Recall that in calculus, we defined the average of g to be

$$\frac{1}{b-a}\int_a^b g(x)\, dx.$$

We can also interpret this integral as the expected value of $g(X_1)$.

Thus, Monte Carlo integration leads to a procedure for estimating integrals.

- Simulate uniform random variables $X_1, X_2, \ldots, X_n$ on the interval $[a, b]$.
- Evaluate $g(X_1), g(X_2), \ldots, g(X_n)$.
- Average this values and multiply by $b - a$ to estimate the integral.

Example 10.6. *Let* $g(x) = \sqrt{1+\cos^3(x)}$ *for* $x \in [0, \pi]$*, to find* $\int_0^\pi g(x)\, dx$*. The three steps above become the following* R *code.*

```
> x<-runif(1000,0,pi)
> g<-sqrt(1+cos(x)^3)
> pi*mean(g)
[1] 2.991057
```

Example 10.7. *To find the integral of* $g(x) = \cos^2(\sqrt{x^3+1})$ *on the interval* $[-1, 2]$*, we simulate* n *random variables uniformly using* `runif(n,-1,2)` *and then compute* `mean(cos(sqrt(x^3+1))^2)`*. The choices* $n = 25$ *and* $n = 250$ *are shown in Figure 10.3*

The variation in estimates for the integral can be described by the variance as given in equation (10.2).

$$\text{Var}(\overline{g(X)}_n) = \frac{1}{n}\text{Var}(g(X_1)).$$

where $\sigma^2 = \text{Var}(g(X_1)) = E(g(X_1) - \mu_{g(X_1)})^2 = \int_a^b (g(x) - \mu_{g(X_1)})^2 f_X(x)\, dx$. Typically this integral is more difficult to estimate than $\int_a^b g(x)\, dx$, our original integral of interest. However, we can see that the variance of the estimator is inversely proportional to n, the number of random numbers in the simulation. Thus, the standard deviation is inversely proportional to $\sqrt{n}$.

Monte Carlo techniques are rarely the best strategy for estimating one or even very low dimensional integrals. R does integration numerically using the `function` and the `integrate` commands. For example,

```
> g<-function(x){sqrt(1+cos(x)^3)}
> integrate(g,0,pi)
2.949644 with absolute error < 3.8e-06
```

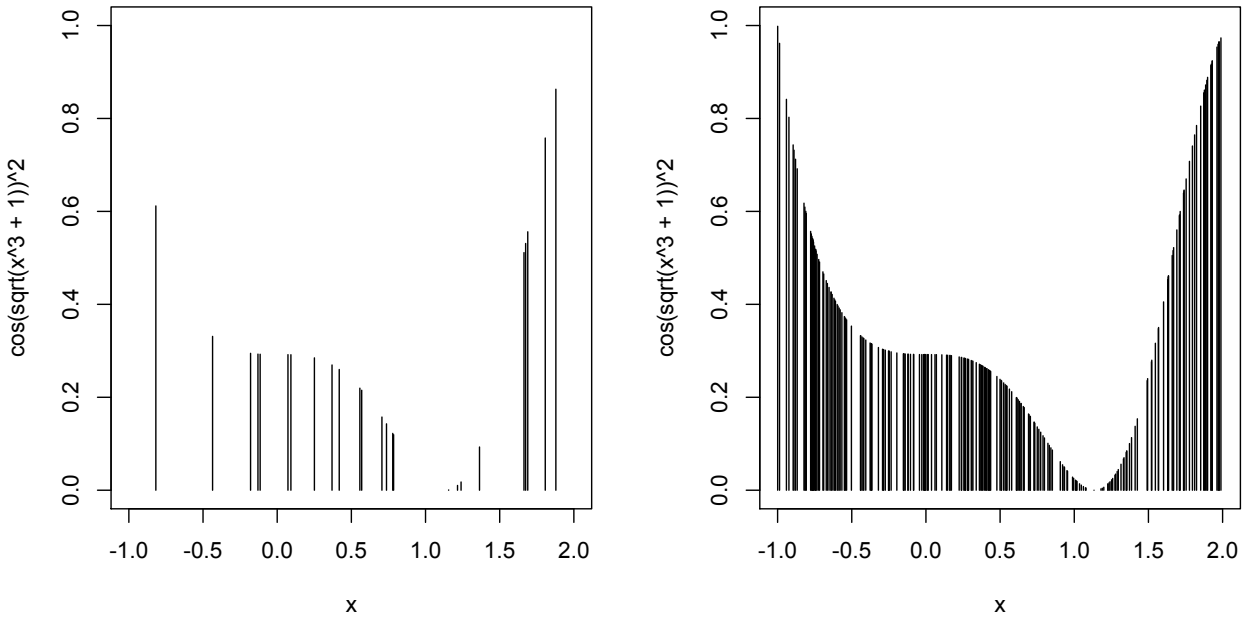


Figure 10.3: Monte Carlo integration of $g(x) = \cos^2(\sqrt{x^3+1})$ on the interval $[-1, 2]$, with (left) $n = 25$ and (right) $n = 250$. $\overline{g(X)}_n$ is the average heights of the n lines whose x values are uniformly chosen on the interval. By the law of large numbers, this estimates the average value of g. This estimate is multiplied by 3, the length of the interval to give $\int_{-1}^{2} g(x)\,dx$. In this example, the estimate os the integral is 0.905 for $n = 25$ and 1.028 for $n = 250$. Using the `integrate` command, a more precise numerical estimate of the integral gives the value 1.000194.

With only a small change in the algorithm, we can also use this to evaluate high dimensional multivariate integrals. For example, in three dimensions, the integral

$$I(g) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} g(x, y, z)\, dz\, dy\, dx$$

can be estimated using Monte Carlo integration by generating three sequences of uniform random variables,

$$X_1, X_2, \ldots, X_n, \quad Y_1, Y_2, \ldots, Y_n, \quad \text{and} \quad Z_1, Z_2, \ldots Z_n$$

Then,

$$I(g) \approx (b_1 - a_1)(b_2 - a_2)(b_3 - a_3)\frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i, Z_i). \tag{10.3}$$

Example 10.8. *Consider the function*

$$g(x, y, z) = \frac{32x^3}{3(y + z^4 + 1)}$$

with x, y and z all between 0 and 1.

To obtain a sense of the distribution of the approximations to the integral $I(g)$, we perform 1000 simulations using 100 uniform random variable for each of the three coordinates to perform the Monte Carlo integration. The command `Ig<-numeric(1000)` *creates a space for a vector of length 1000. This is added so that* R *creates a place ahead of the simulations to store the results.*

```
> Ig<-numeric(1000)
> for(i in 1:1000){x<-runif(100);y<-runif(100);z<-runif(100);
  g<-32*x^3/(3*(y+z^4+1)); Ig[i]<-mean(g)}
> hist(Ig)
```

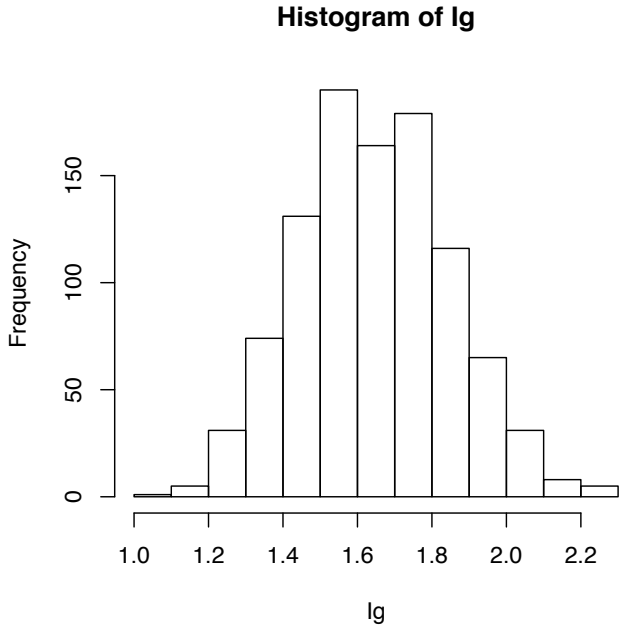


Figure 10.4: Histogram of 1000 Monte Carlo estimates for the integral $\int_0^1 \int_0^1 \int_0^1 32x^3/(y+z^4+1)\,dx\,dy\,dz$. The sample standard deviation is 0.188.

```
> summary(Ig)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.045   1.507   1.644   1.650   1.788   2.284
> var(Ig)
[1] 0.03524665
> sd(Ig)
[1] 0.1877409
```

Thus, our Monte Carlo estimate the standard deviation of the estimated integral is 0.188.

Exercise 10.9. *Estimate the variance and standard deviation of the Monte Carlo estimator for the integral in the example above based on $n = 500$ and 1000 random numbers.*

Exercise 10.10. *How many observations are needed in estimating the integral in the example above so that the standard deviation of the average is 0.05?*

To modify this technique for a region $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ use indepenent uniform random variables X_i, Y_i, and Z_i on the respective intervals, then

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i, Z_i) \approx Eg(X_1, Y_1, Z_1) = \frac{1}{b_1 - a_1}\frac{1}{b_2 - a_2}\frac{1}{b_3 - a_3}\int_{a_1}^{b_1}\int_{a_2}^{b_2}\int_{a_3}^{b_3} g(x, y, z)\,dz\,dy\,dx.$$

Thus, the estimate for the integral is

$$\frac{(b_1 - a_1)(b_2 - a_2)(b_3 - a_3)}{n}\sum_{i=1}^{n} g(X_i, Y_i, Z_i).$$

Exercise 10.11. *Use Monte Carlo integration to estimate*

$$\int_0^3 \int_{-2}^{2} \frac{\cos(\pi(x+y))}{\sqrt[4]{1+xy^2}}\,dydx.$$

10.3 Importance Sampling

In many of the large simulations, the dimension of the integral can be in the hundreds and the function g can be very close to zero for large regions in the domain of g. Simple Monte Carlo simulation will then frequently choose values for g that are close to zero. These values contribute very little to the average. Due to this inefficiency, a more sophisticated strategy is employed. **Importance sampling methods** begin with the observation that a better strategy may be to concentrate the random points where the integrand g is large in absolute value.

For example, for the integral

$$\int_0^1 \frac{e^{-x/2}}{\sqrt{x(1-x)}} dx, \tag{10.4}$$

the integrand is much bigger for values near $x = 0$ or $x = 1$. (See Figure 10.4) Thus, we can hope to have a more accurate estimate by concentrating our sample points in these places.

With this in mind, we perform the Monte Carlo integration beginning with $Y_1, Y_2, \ldots$ independent random variables with common densityf_Y. The goal is to find a density f_Y that is large when $|g|$ is large and small when $|g|$ is small. The density f_Y is called the **importance sampling function** or the **proposal density**. With this choice of density, we define the **importance sampling weights** so that

$$g(y) = w(y) f_Y(y). \tag{10.5}$$

To justify this choice, note that, the sample mean

$$\overline{w(Y)}_n = \frac{1}{n}\sum_{i=1}^{n} w(Y_i) \approx \int_{-\infty}^{\infty} w(y) f_Y(y)\, dy = \int_{-\infty}^{\infty} g(y) dy = I(g).$$

Thus, the average of the importance sampling weights, by the strong law of large numbers, still approximates the integral of g. This is an improvement over simple Monte Carlo integration if the variance decreases, i.e.,

$$\text{Var}(w(Y_1)) = \int_{-\infty}^{\infty} (w(y) - I(g))^2 f_Y(y)\, dy = \sigma_f^2 << \sigma^2.$$

As the formula shows, this can be best achieved by having the weight $w(y)$ be close to the integral $I(g)$. Referring to equation (10.5), we can now see that we should endeavor to have f_Y proportional to g.

Importance leads to the following procedure for estimating integrals.

- Write the integrand $g(x) = w(x) f_Y(x)$. Here f_Y is the density function for a random variable Y that is chosen to capture the changes in g.
- Simulate variables $Y_1, Y_2 \ldots, Y_n$ with density f_Y. This will sometimes require integrating the density function to obtain the distribution function $F_Y(x)$, and then finding its inverse function $F_Y^{-1}(u)$. This sets up the use of the probability transform to obtain $Y_i = F_Y^{-1}(U_i)$ where $U_1, U_2 \ldots, U_n$, independent random variables uniformly distributed on the interval $[0, 1]$.
- Compute the average of $w(Y_1), w(Y_2) \ldots, w(Y_n)$ to estimate the integral of g.

Note that the use of the probability transform removes the need to multiply $b - a$, the length of the interval.

Example 10.12. *For the integral (10.4) we can use Monte Carlo simulation based on uniform random variables.*

```
> Ig<-numeric(1000)
> for(i in 1:1000){x<-runif(100);g<-exp(-x/2)*1/sqrt(x*(1-x));Ig[i]<-mean(g)}
> summary(Ig)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.970   2.277   2.425   2.484   2.583   8.586
> sd(Ig)
[1] 0.3938047
```

Based on a 1000 simulations, we find a sample mean value of 2.484 and a sample standard deviation of 0.394. Because the integrand is very large near $x = 0$ and $x = 1$, we choose look for a density f_Y to concentrate the random samples near the ends of the intervals.

Our choice for the proposal density is a $Beta(1/2, 1/2)$, then

$$f_Y(y) = \frac{1}{\pi} y^{1/2-1}(1-y)^{1/2-1}$$

on the interval $[0, 1]$. Thus the weight

$$w(y) = \pi e^{-y/2}$$

is the ratio $g(x)/f_Y(x)$

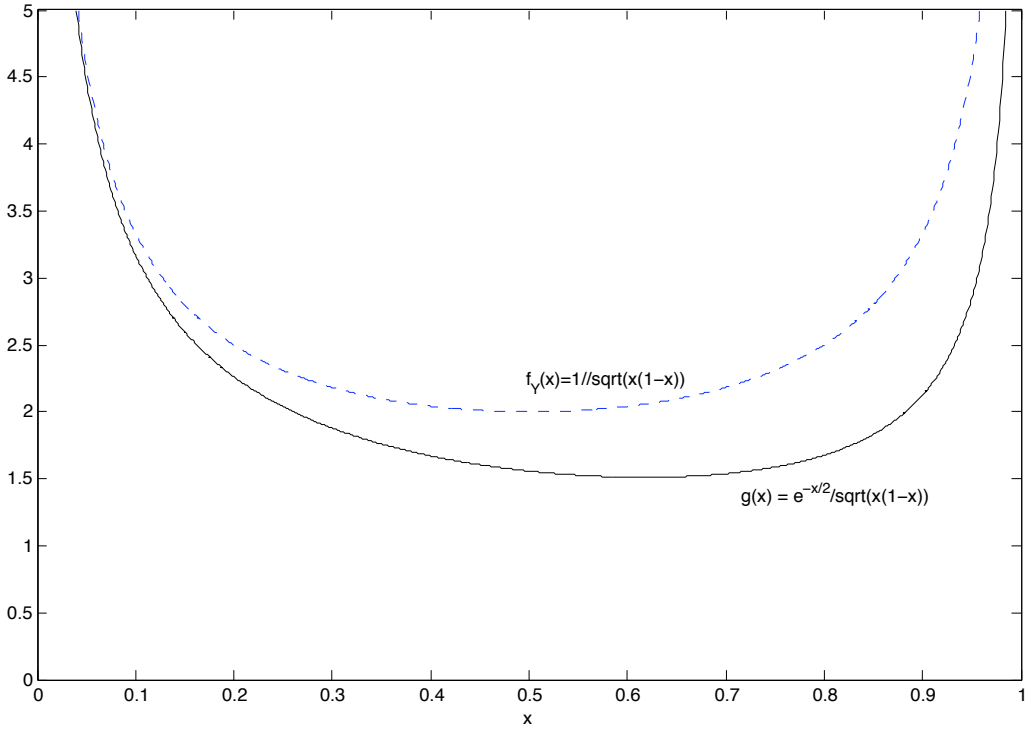


Figure 10.5: (left) Showing converges of the running average S_n/n to its limit 2 for $p = 1/2$. (right) Showing lack of convergence for the case $p = 2$.

Again, we perform the simulation multiple times.

```
> IS<-numeric(1000)
> for(i in 1:1000){y<-rbeta(100,1/2,1/2);w<-pi*exp(-y/2);IS[i]<-mean(w)}
> summary(IS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.321   2.455   2.483   2.484   2.515   2.609
> var(IS)
[1] 0.0002105915
> sd(IS)
[1] 0.04377021
```

Based on 1000 simulations, we find a sample mean value again of 2.484 and a sample standard deviation of 0.044, about 1/9th the size of the Monte Carlo sample standard deviation. Part of the gain is illusory. Beta random variables take longer to simulate. If they require a factor more than 81 times longer to simulate, then the extra work needed

to create a good importance sample is not helpful in producing a more accurate estimate for the integral. Numerical integration gives

```
> g<-function(x){exp(-x/2)*1/sqrt(x*(1-x))}
> integrate(g,0,1)
2.485054 with absolute error < 2e-06
```

Exercise 10.13. *Evaluate the integral*

$$\int_0^1 \frac{e^{-x}}{\sqrt[3]{x}}\,dx$$

1000 times using $n = 200$ *sample points using directly Monte Carlo integration and using importance sampling with random variables having density*

$$f_X(x) = \frac{2}{3\sqrt[3]{x}}$$

on the interval $[0, 1]$. *For the second part, you will need to use the probability transform. Compare the means and standard deviations of the 1000 estimates for the integral. The integral is approximately 1.04969.*

10.4 Answers to Selected Exercises

10.3. For $p \neq 1$, the expected value

$$EU^{-p} = \int_0^1 u^{-p}\,dp = \frac{1}{1-p}u^{1-p}\Big|_0^1 = \frac{1}{1-p} < \infty$$

provided that $1 - p > 0$ or $p < 1$. For $p > 1$, we evaluate the integral in the interval $[b, 1]$ and take a limit as $b \to 0$,

$$\int_b^1 u^{-p}\,dp = \frac{1}{1-p}u^{1-p}\Big|_b^1 = \frac{1}{1-p}(1 - b^{1-p}) \to \infty.$$

For $p = 1$,

$$\int_b^1 u^{-1}\,dp = \ln u\Big|_b^1 = -\ln b \to \infty.$$

We use the case $p = 1/2$ for which the integral converges. and $p = 2$ in which the integral does not. Indeed,

$$\int_)^1 u^{1/2}\,du = 2u^{3/2}\Big|_0^1 = 2$$

```
> par(mfrow=c(1,2))
> u<-runif(1000)
> x<-1/u^(1/2)
> s<-cumsum(x)
> plot(s/n,n,type="l")
> x<-1/u^2
> s<-cumsum(x)
> plot(n,s/n,type="l")
```

10.5. Here are the R commands:

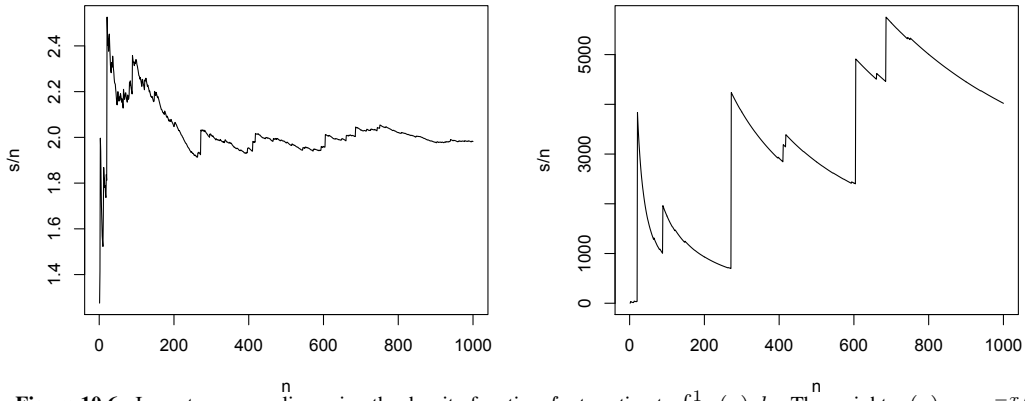


Figure 10.6: Importance sampling using the density function f_Y to estimate $\int_0^1 g(x)\,dx$. The weight $w(x) = \pi e^{-x/2}$.

```
> par(mfrow=c(2,2))
> x<-rcauchy(1000)
> s<-cumsum(x)
> plot (n,s/n,type="l")
> x<-rcauchy(1000)
> s<-cumsum(x)
> plot (n,s/n,type="l")
> x<-rcauchy(1000)
> s<-cumsum(x)
> plot (n,s/n,type="l")
> x<-rcauchy(1000)
> s<-cumsum(x)
> plot (n,s/n,type="l")
```

This produces in Figure 10.5. Notice the differences for the values on the x-axis

10.9. The standard deviation for the average of n observations is $\sigma/\sqrt{n}$ where σ is the standard deviation for a single observation. From the output

```
> sd(Ig)
[1] 0.1877409
```

We have that $0.1877409 \approx \sigma/\sqrt{100} = \sigma/10$. Thus, $\sigma \approx 1.877409$. Consequently, for 500 observations, $\sigma/\sqrt{500} \approx 0.08396028$. For 1000 observations $\sigma/\sqrt{1000} \approx 0.05936889$

10.10. For $\sigma/\sqrt{n} = 0.05$, we have that $n = (\sigma/0.05)^2 \approx 1409.866$. So we need approximately 1410 observations.

10.11. To view the surface for $\frac{\cos(\pi(x+y))}{\sqrt[4]{1+xy^2}}$, $0 \le x \le 3$, $-2 \le y \le 2$, we type

```
> x <- seq(0,3, len=30)
> y <- seq(-2,2, len=30)
> f <- outer(x, y, function(x, y)
(cos(pi*(x+y)))/(1+x*y^2)^(1/4))
> persp(x,y,f,col="orange",phi=45,theta=30)
```

Using 1000 random numbers uniformly distributed for both x and y, we have

```
> x<-runif(1000,0,3)
> y<-runif(1000,-2,2)
> g<-(cos(pi*(x+y)))/(1+x*y^2)^(1/4)
```

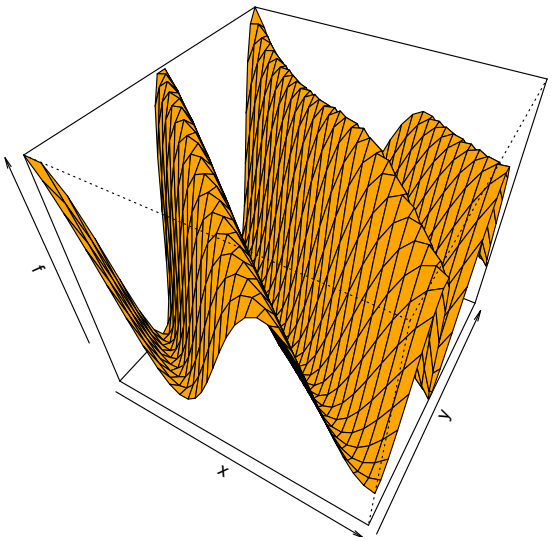


Figure 10.8: Surface plot of function used in Exercise 10.10

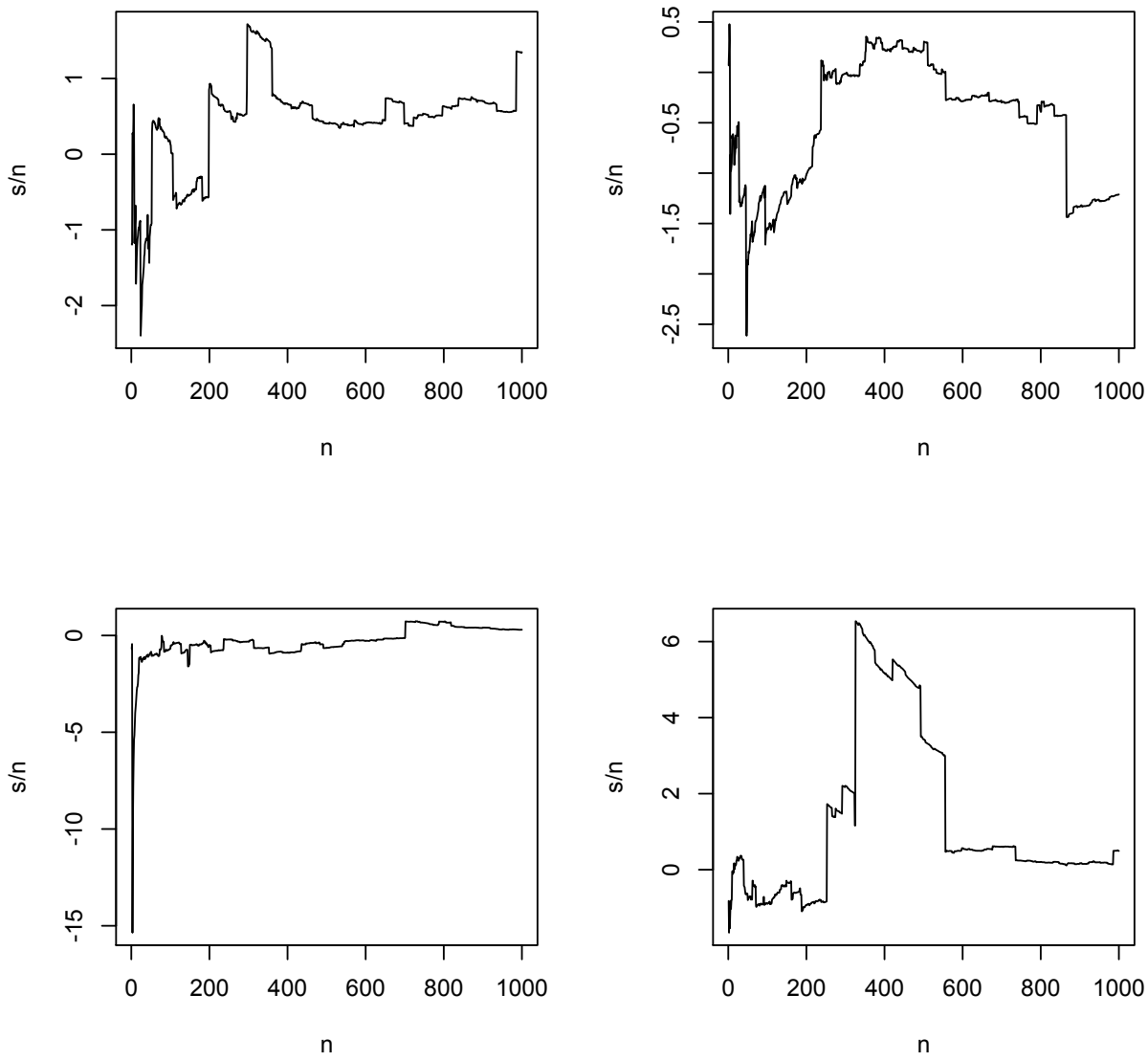


Figure 10.7: Plots of running averages of Cauchy random variables.

```
> 3*4*mean(g)
[1] 0.2452035
```

To finish, we need to multiply the average of g as estimated by `mean(g)` by the area associated to the integral $(3-0) \times (2-(-2)) = 12$.

10.13. For the direct Monte Carlo simulation, we have

```
> Ig<-numeric(1000)
> for (i in 1:1000){x<-runif(200);g<-exp(-x)/x^(1/3);Ig[i]<-mean(g)}
```

```
> mean(Ig)
[1] 1.048734
> sd(Ig)
[1] 0.07062628
```

For the importance sampler, the integral is

$$\frac{3}{2}\int_0^1 e^{-x} f_X(x)\,dx.$$

To simulate independent random variables with density f_X, we first need the cumulative distribution function for X,

$$F_X(x) = \int_0^x \frac{2}{3\sqrt[3]{t}}\,dt = t^{2/3}\Big|_0^x = x^{2/3}.$$

Then, to find the probability transform, note that

$$u = F_X(x) = x^{2/3} \quad \text{and} \quad x = F_X^{-1}(u) = u^{3/2}.$$

Thus, to simulate X, we simulate a uniform random variable U on the interval $[0, 1]$ and evaluate $U^{3/2}$. This leads to the following R commands for the importance sample simulation:

```
> ISg<-numeric(1000)
> for (i in 1:1000){u<-runif(200);x<-u^(3/2); w<-3*exp(-x)/2;ISg[i]<-mean(w)}
> mean(ISg)
[1] 1.048415
> sd(ISg)
[1] 0.02010032
```

Thus, the standard deviation using importance sampling is about 2/7-ths the standard deviation using simple Monte Carlo simulation. Consequently, we can decrease the number of samples using importance sampling by a factor of $(2/7)^2 \approx 0.08$.

Topic 11

The Central Limit Theorem

The occurrence of the Gaussian probability density $1 = e^{-x^2}$ *in repeated experiments, in errors of measurements, which result in the combination of very many and very small elementary errors, in diffusion processes etc., can be explained, as is well-known, by the very same limit theorem, which plays a central role in the calculus of probability.* - George Polya

11.1 Introduction

In the discussion leading to the law of large numbers, we saw visually that the sample means from a sequence of independent random variables converge to their common distributional mean as the number of random variables increases. In symbols,

$$\frac{1}{n}S_n = \bar{X}_n \to \mu \text{ as } n \to \infty.$$

Using the Pythagorean identity for independent random variables, we obtained the more precise statement that the standard deviation of of the sample mean, $\bar{X}_n$, is inversely proportional to $\sqrt{n}$, the square root of the number of observations. For example, for simulations based on observations of independent random variables, uniformly distributed on the interval $[0, 1]$, we see, as anticipated, the running averages converging to

$$\mu = \int_0^1 x f_X(x)\, dx = \int_0^1 x\, dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2},$$

the distributional mean.

Now, we zoom around the mean value of $\mu = 1/2$. Because the standard deviation $\sigma_{\bar{X}_n} \propto 1/\sqrt{n}$, we magnify the difference between the running average and the mean by a factor of $\sqrt{n}$ and investigate the graph of

$$\sqrt{n}\left(\frac{1}{n}S_n - \mu\right) = \frac{S_n - n\mu}{\sqrt{n}} \tag{11.1}$$

versus n. The results of a simulation are displayed in Figure 11.1.

As we see in Figure 11.2, even if we extend this simulation for larger values of n, we continue to see fluctuations about the center of roughly the same size and the size of the fluctuations for a single realization of a simulation cannot be predicted in advance.

Thus, we focus on addressing a broader question: *Does the distribution of the size of these fluctuations have any regular and predictable structure?* This question and the investigation that led to led to its answer, the **central limit theorem**, constitute one of the most important episodes in mathematics.

Exercise 11.1. *Repeat the exercise above times. looking at the centered and magnified running averages, (11.1) for 2000 steps for* $U(0, 1)$ *random variables. Give 1000 simulations of the final value for the simulations in (11.1), and describe the histogram.*

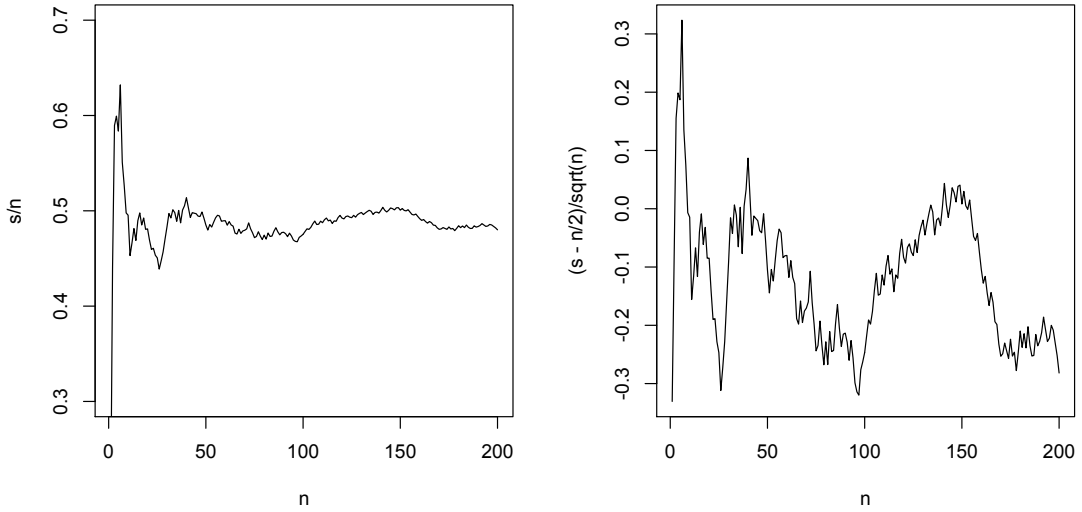


Figure 11.1: a. Running average of independent random variables, uniform on $[0, 1]$. b. Running average centered at the mean value of 1/2 and magnified by $\sqrt{n}$.

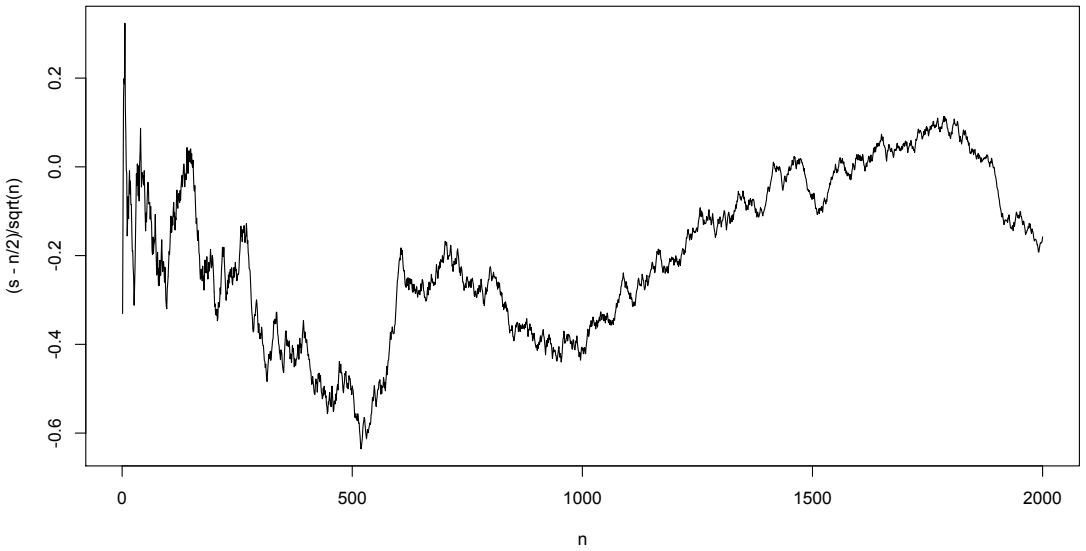


Figure 11.2: Plot of (11.1, the running average centered at the mean value of 1/2, and magnified by $\sqrt{n}$ extended to 2000 steps.

11.2 The Classical Central Limit Theorem

Let's begin by examining the distribution for the sum of $X_1, X_2 \ldots X_n$, independent and identically distributed random variables

$$S_n = X_1 + X_2 + \cdots + X_n,$$

what distribution do we see? Let's look first to the simplest case, X_i Bernoulli random variables. In this case, the sum S_n is a binomial random variable. We examine two cases - in the first we keep the number of trials the same at $n = 100$ and vary the success probability p. In the second case, we keep the success probability the same at $p = 1/2$, but vary the number of trials.

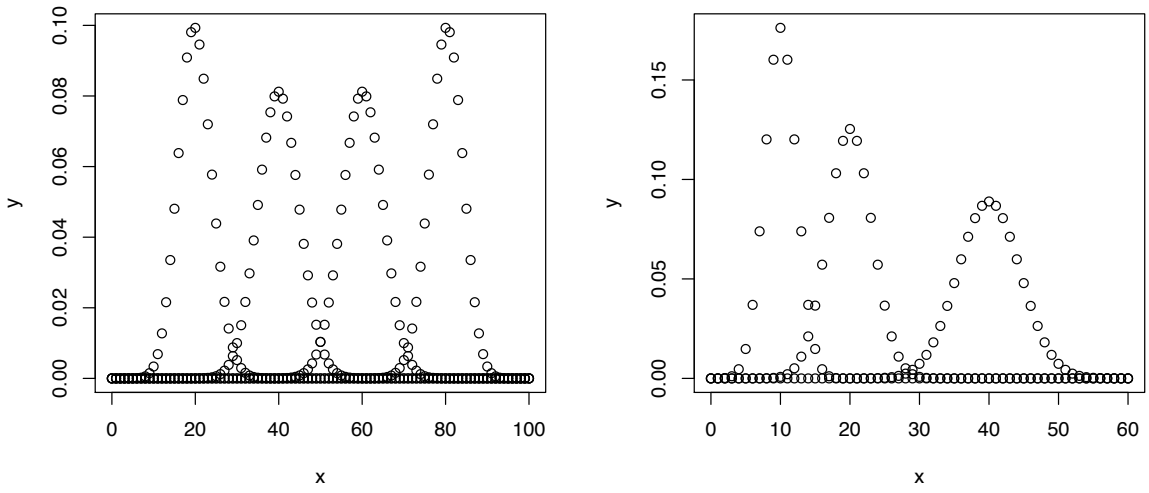


Figure 11.3: a. Successes in 100 Bernoulli trials with $p = 0.2, 0.4, 0.6$ and 0.8. b. Successes in Bernoulli trials with $p = 1/2$ and $n = 20, 40$ and 80.

The curves in Figure 11.3 look like bell curves. Their center and spread vary in ways that are predictable. The binomial random variable S_n has

$$\text{mean } np \text{ and standard deviation } \sqrt{np(1-p)}.$$

Thus, if we take the standardized version of these sums of Bernoulli random variables

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}},$$

then these bell curve graphs would lie on top of each other.

For our next example, we look at the density of the sum of standardized exponential random variables. The exponential density is strongly skewed and so we have have to wait for larger values of n before we see the bell curve emerge. In order to make comparisons, we examine standardized versions of the random variables with mean μ and variance σ^2.

To accomplish this,

- we can either standardize using the **sum** S_n having mean $n\mu$ and standard deviation $\sigma\sqrt{n}$, *or*
- we can standardize using the **sample mean** $\bar{X}_n$ having mean μ and standard deviation $\sigma/\sqrt{n}$.

This yields two equivalent versions of the z-score.

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu). \tag{11.2}$$

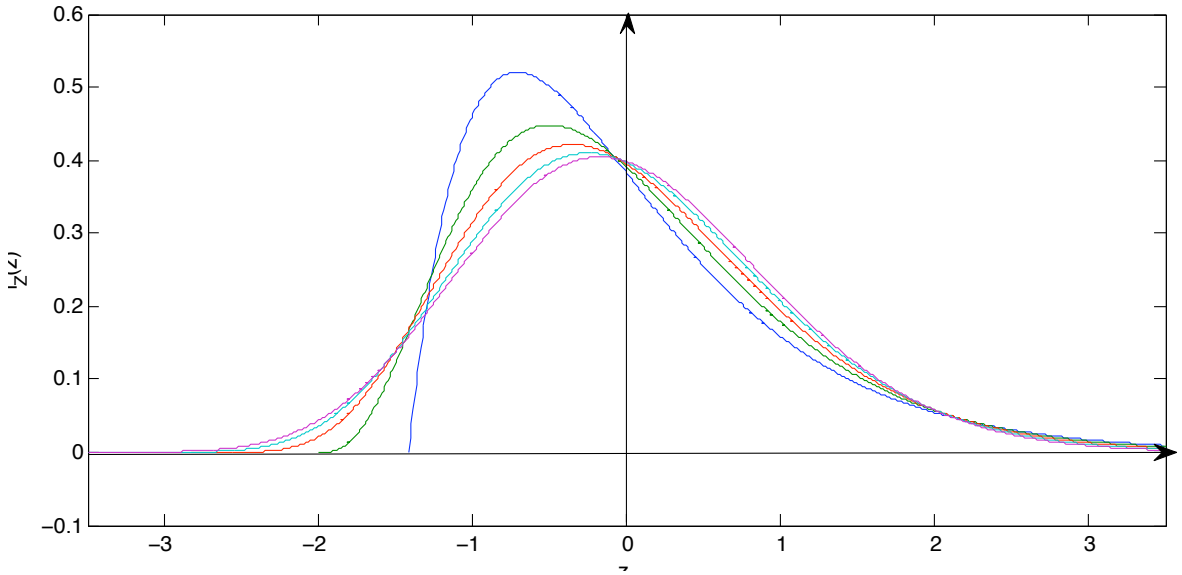


Figure 11.4: Displaying the central limit theorem graphically. Density of the standardized version of the sum of n independent exponential random variables for $n = 2$ (dark blue), 4 (green), 8 (red), 16 (light blue), and 32 (magenta). Note how the skewness of the exponential distribution slowly gives way to the bell curve shape of the normal distribution.

In Figure 11.4, we see the densities approaching that of the bell curve for a standard normal random variables. Even for the case of $n = 32$, we see a small amount of skewness that is a remnant of the skewness in the exponential density.

Exercise 11.2. *Show that the skewness for the sum of n independent $Exp(\lambda)$ random variables is $2/\sqrt{n}$. Thus the skewness of the normalized sums converges to 0 as $n \to \infty$.* Hint*: The sum S_n is a $\Gamma(n, \lambda)$ random variable.*

Exercise 11.3. *More generally, show the skewness for the sum of n independent random variables having a common distribution is $\gamma_1/\sqrt{n}$ where γ_1 is the skewness of any one of the random variables in the sum. Consequently, the skewness of S_n a $Bin(n, p)$ random variable is $(1 - 2p)/\sqrt{np(1 - p)}$.*

The theoretical result behind these numerical explorations is called the **classical central limit theorem**:

Theorem 11.4. *Let $\{X_i; i \geq 1\}$ be independent random variables having a common distribution. Let μ be their mean and σ^2 be their variance. Then Z_n, the standardized scores defined by equation (11.2), converges* **in distribution** *to Z a standard normal random variable. This statement is shorthand for the more precise statement that the distribution function F_{Z_n} converges to Φ, the distribution function of the standard normal.*

$$\lim_{n\to\infty} F_{Z_n}(z) = \lim_{n\to\infty} P\{Z_n \leq z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx = \Phi(z).$$

In practical terms the central limit theorem states that $P\{a < Z_n \leq b\} \approx P\{a < Z \leq b\} = \Phi(b) - \Phi(a)$.

This theorem is an enormously useful tool in providing good estimates for probabilities of events depending on either S_n or $\bar{X}_n$. We shall begin to show this in the following examples.

Example 11.5. *For Bernoulli random variables, $\mu = p$ and $\sigma = \sqrt{p(1 - p)}$. S_n is the number of successes in n Bernoulli trials. In this situation, the sample mean is the fraction of trials that result in a success. This is generally denoted by $\hat{p}$ to indicate the fact that it is a* **sample proportion**.

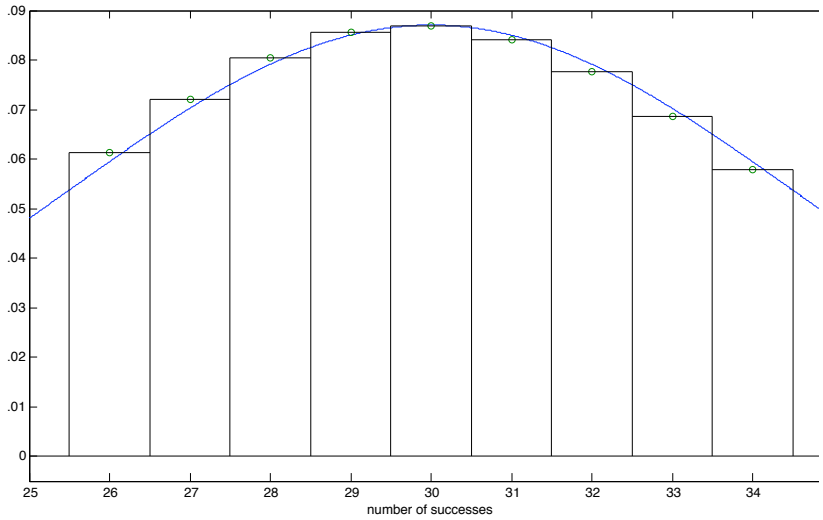


Figure 11.5: Mass function for a $Bin(100, 0.3)$ random variable (black) and approximating normal density $N(100 \cdot 0.3, \sqrt{100 \cdot 0.3 \cdot 0.7})$.

The normalized versions of S_n and $\hat{p}$ are equal to

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}} = \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}},$$

For example, in 100 tosses of a fair coin, $\mu = 1/2$ and $\sigma = \sqrt{1/2(1-1/2)} = 1/2$. Thus,

$$Z_{100} = \frac{S_{100} - 50}{5}.$$

To find $P\{S_{100} > 65\}$, convert the event $\{S_{100} > 65\}$ to an event concerning Z_{100}.

$$P\{S_{100} > 65\} = P\{S_{100} - 50 > 15\} = P\left\{\frac{S_{100} - 50}{5} > 3\right\} = P\{Z_{100} > 3\} \approx P\{Z > 3\} = 0.0013.$$

```
> 1-pnorm(3)
[1] 0.001349898
```

We could also write,

$$Z_{100} = \frac{\hat{p} - 1/2}{1/20} = 20(\hat{p} - 1/2).$$

and

$$P\{\hat{p} \leq 0.40\} = P\{\hat{p} - 1/2 \leq 0.40 - 1/2\} = P\{20(\hat{p} - 1/2) \leq 20(0.4 - 1/2)\} = P\{Z_{100} \leq -2\} \approx P\{Z \leq -2\} = 0.023.$$

```
> pnorm(-2)
[1] 0.02275013
```

Remark 11.6. *We can improve the normal approximation to the binomial random variable by employing the* **continuity correction**. *For a binomial random variable X, the distribution function*

$$P\{X \leq x\} = P\{X < x + 1\} = \sum_{y=0}^{x} P\{X = y\}$$

can be realized as the area of $x+1$ rectangles, height $P\{X = y\}$, $y = 0, 1, \ldots, x$ and width 1. These rectangles look like a Riemann sum for the integral up to the value $x + 1/2$. For the example in Figure 11.5, $P\{X \leq 32\} = P\{X < 33\}$ is the area of 33 rectangles. This right side of rectangles is at the value 32.5. Thus, for the approximating normal random variable Y, this suggests computing $P\{Y \leq 32.5\}$. In this example the exact value

```
> pbinom(32,100,0.3)
[1] 0.7107186
```

Comparing this to possible choices for the normal approximations

```
> n<-100
> p<-0.3
> mu<-n*p
> sigma<-sqrt(p*(1-p))
> prob<-pnorm((x-mu)/(sigma*sqrt(n)))
> x<-c(32,32.5,33)
> data.frame(x,prob)
     x      prob
1 32.0 0.6687397
2 32.5 0.7073105
3 33.0 0.7436546
```

This shows a difference of 0.0034 for the choice $x = 32.5$ and larger differences for the choices $x = 32$ or $x = 33$.

Example 11.7. *Opinion polls are generally designed to be modeled as Bernoulli trials. The number of trials n is set to give a prescribed value m of two times the standard deviation of $\hat{p}$. This value of m is an example of a* **margin of error**. *The standard deviation*

$$\sqrt{p(1-p)/n}$$

takes on its maximum value for $p = 1/2$. For this case,

$$m = 2\sqrt{\frac{1}{2}\left(1 - \frac{1}{2}\right)/n} = \frac{1}{\sqrt{n}}$$

Thus,

$$n = \frac{1}{m^2}$$

We display the results in R *for typical values of m.*

```
> m<-seq(0.01,0.05,0.01)
> n<-1/m^2
> data.frame(m,n)
     m         n
1 0.01 10000.000
2 0.02  2500.000
3 0.03  1111.111
4 0.04   625.000
5 0.05   400.000
```

So, a 5% margin of error can be achieved with a modest sample size of $n = 400$, whereas a 1% margin of error requires 10,000 samples.

Exercise 11.8. *We have two approximation methods for a large number n of Bernoulli trials - Poisson, which applies then p is small and their product $\lambda = np$ is moderate and normal when the mean number of successes np or the mean number of failures $n(1-p)$ is sufficiently large. Investigate the approximation of the distribution, X, a Poisson random variable, by the distribution of a normal random variable, Y, for the case $\lambda = 16$. Use the continuity correction to compare*

$$P\{X \le x\} \quad to \quad P\{Y \le x + \frac{1}{2}\}.$$

Example 11.9. *For exponential random variables $\mu = 1/\lambda$ and $\sigma = 1/\lambda$ and therefore*

$$Z_n = \frac{S_n - n/\lambda}{\sqrt{n}/\lambda} = \frac{\lambda S_n - n}{\sqrt{n}}.$$

Let T_{64} be the sum of 64 independent with parameter $\lambda = 1$. Then, $\mu = 1$ and $\sigma = 1$. So,

$$P\{T_{64} < 60\} = P\left\{\frac{T_{64} - 64}{8} < \frac{60 - 64}{8}\right\} = P\left\{\frac{T_{64} - 64}{8} < -\frac{1}{2}\right\} = P\{Z_{64} < -0.5\} \approx 0.309.$$

Example 11.10. *Pictures on your smartphone have a mean size of 400 kilobytes (KB) and a standard deviation of 50 KB. You want to store 100 pictures on your cell phone. If we assume that the size of the pictures $X_1, X_2, \cdots, X_{100}$ are independent, then $\bar{X}$ has mean $\mu_{\bar{X}} = 400$ KB and standard deviation $\sigma_{\bar{X}} = 50/\sqrt{100} = 5$ KB. So, the probability that the average picture size is between 394 and 406 kilobytes is*

$$P\{394 \le \bar{X} \le 406\} = P\left\{\frac{394 - 400}{5} \le \frac{\bar{X} - 400}{5} \le \frac{406 - 400}{5}\right\} = P\{-1.2 \le Z_{100} \le 1.2\} \approx 0.230.$$

S_{100} be the total storage space needed for the 100 pictures has mean $100 \times 400 = 40,000$ KB and standard deviation $\sigma_{S_{100}} = 50\sqrt{100} = 500$ KB. To estimate the space required to be 99% certain that the pictures will have storage space on the phone, note that

```
> qnorm(0.99)
[1] 2.326348
```

Thus,

$$Z_{100} = \frac{S_{100} - 40000}{500} \ge 2.326348, \quad S_{100} - 40000 \ge 1163.174, \quad S_{100} \ge 41163.174$$

kilobyres.

Exercise 11.11. *If your smartphone has 42000 KB of storage space for pictures, Use the central limit theorem to estimate the number of pictures you can have necessary to have a 1% chance of running out of space.*

Exercise 11.12. *Simulate 1000 times, $\bar{x}$, the sample mean of 100 random variables, uniformly distributed on $[0, 1]$. Show a histogram for these simulations to see the approximation to a normal distribution. Find the mean and standard deviations for the simulations and compare them to their distributional values. Use both the simulation and the central limit theorem to estimate the 35th percentile of $\bar{X}$.*

11.3 Propagation of Error

Propagation of error or **propagation of uncertainty** is a strategy to estimate the impact on the standard deviation of the consequences of a nonlinear transformation of a measured quantity whose measurement is subject to some uncertainty.

For any random variable Y with mean μ_Y and standard deviation σ_Y, we will be looking at linear functions $aY + b$ for Y. Using the linearity of expectation and the quadratic identity for variance, we have that

$$E[a + bY] = a + b\mu_Y, \qquad \text{Var}(a + bY) = b^2\text{Var}(Y). \tag{11.3}$$

Exercise 11.13. *Show that*

$$E[a + b(Y - \mu_Y)] = a, \qquad \text{Var}(a + b(Y - \mu_Y)) = b^2\text{Var}(Y).$$

We will apply this to the linear approximation of $g(Y)$ about the point μ_Y.

$$g(Y) \approx g(\mu_Y) + g'(\mu_Y)(Y - \mu_Y). \tag{11.4}$$

If we take expected values, then

$$Eg(Y) \approx E[g(\mu_Y) + g'(\mu_Y)(Y - \mu_Y)] = g(\mu_Y).$$

The variance

$$\text{Var}(g(Y)) \approx \text{Var}(g(\mu_Y) + g'(\mu_Y)(Y - \mu_Y)) = g'(\mu_Y)^2\sigma_Y^2.$$

Thus, the standard deviation

$$\sigma_{g(Y)} \approx |g'(\mu_Y)|\sigma_Y \tag{11.5}$$

gives what is known as the **propagation of error**.

If Y is meant to be some measurement of a quantity q with a measurement subject to error, then saying that

$$q = \mu_Y = EY$$

is stating that Y is an **unbiased estimator** of q. In other words, Y does not systematically overestimate or underestimate q. The standard deviation σ_Y gives a sense of the variability in the measurement apparatus. However, if we measure Y but want to give not an estimate for q, but an estimate for a function of q, namely $g(q)$, its standard deviation is approximation by formula (11.5).

Example 11.14. *Let Y be the measurement of a side of a cube with length ℓ. Then Y^3 is an estimate of the volume of the cube. If the measurement error has standard deviation σ_Y, then, taking $g(y) = y^3$, we see that the standard deviation of the error in the measurement of the volume*

$$\sigma_{Y^3} \approx 3q^2\sigma_Y.$$

If we estimate q with Y, then

$$\sigma_{Y^3} \approx 3Y^2\sigma_Y.$$

To estimate the coefficient volume expansion α_3 of a material, we begin with a material of known length ℓ_0 at temperature T_0 and measure the length ℓ_1 at temperature T_1. Then, the coefficient of linear expansion

$$\alpha_1 = \frac{\ell_1 - \ell_0}{\ell_0(T_1 - T_0)}.$$

If the measure length of ℓ_1 is Y. We estimate this by

$$\hat{\alpha}_1 = \frac{Y - \ell_0}{\ell_0(T_1 - T_0)}.$$

Then, if a measurement Y of ℓ_1 has variance σ_Y^2, then

$$\text{Var}(\hat{\alpha}_1) = \frac{\sigma_Y^2}{\ell_0^2(T_1 - T_0)^2} \quad \sigma_{\hat{\alpha}_1} = \frac{\sigma_Y}{\ell_0|T_1 - T_0|}.$$

Now, we estimate

$$\alpha_3 = \frac{\ell_1^3 - \ell_0^3}{\ell_0^3(T_1 - T_0)} \quad \text{by} \quad \hat{\alpha}_3 = \frac{Y^3 - \ell_0^3}{\ell_0^3(T_1 - T_0)}$$

and

$$\sigma_{\hat{\alpha}_3} \approx 3Y^2\frac{\sigma_Y}{\ell_0^3|T_1 - T_0|}.$$

Exercise 11.15. *In a effort to estimate the angle θ of the sun, the length ℓ of a shadow from a 10 meter flag pole is measured. If $\sigma_{\hat{\ell}}$ is the standard deviation for the length measurement, use propagation of error to estimate $\sigma_{\hat{\theta}}$, the standard deviation in the estimate of the angle.*

Often, the function g is a function of several variables. We will show the multivariate propagation of error in the two dimensional case noting that extension to the higher dimensional case is straightforward. Now, for random variables Y_1 and Y_2 with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, the linear approximation about the point (μ_1, μ_2) is

$$g(Y_1, Y_2) \approx g(\mu_1, \mu_2) + \frac{\partial g}{\partial y_1}(\mu_1, \mu_2)(Y_1 - \mu_1) + \frac{\partial g}{\partial y_2}(\mu_1, \mu_2)(Y_2 - \mu_2).$$

As before,

$$Eg(Y_1, Y_2) \approx g(\mu_1, \mu_2).$$

For Y_1 and Y_2 **independent**, we also have that the random variables

$$\frac{\partial g}{\partial y_1}(\mu_1, \mu_2)(Y_1 - \mu_1) \quad \text{and} \quad \frac{\partial g}{\partial y_2}(\mu_1, \mu_2)(Y_2 - \mu_2)$$

are independent. Because the variance of the sum of independent random variables is the sum of their variances, we have the approximation

$$\begin{aligned}\sigma^2_{g(Y_1,Y_2)} = \text{Var}(g(Y_1, Y_2)) &\approx \text{Var}(\frac{\partial g}{\partial y_1}(\mu_1, \mu_2)(Y_1 - \mu_1)) + \text{Var}(\frac{\partial g}{\partial y_2}(\mu_1, \mu_2)(Y_2 - \mu_2)) \\ &= \left(\frac{\partial g}{\partial y_1}(\mu_1, \mu_2)\right)^2 \sigma_1^2 + \left(\frac{\partial g}{\partial y_2}(\mu_1, \mu_2)\right)^2 \sigma_2^2. \end{aligned} \tag{11.6}$$

and consequently, the standard deviation,

$$\sigma_{g(Y_1,Y_2)} \approx \sqrt{\left(\frac{\partial g}{\partial y_1}(\mu_1, \mu_2)\right)^2 \sigma_1^2 + \left(\frac{\partial g}{\partial y_2}(\mu_1, \mu_2)\right)^2 \sigma_2^2}.$$

Exercise 11.16. *Repeat the exercise in the case that the height h if the flag poll is also unknown and is measured independently of the shadow length with standard deviation $\sigma_{\hat{h}}$. Comment on the case in which the two standard deviations are equal.*

Exercise 11.17. *Generalize the formula for the variance to the case of $g(Y_1, Y_2, \ldots, Y_d)$ for independent random variables $Y_1, Y_2, \ldots, Y_d$.*

Example 11.18. *In the previous example, we now estimate the volume of an $\ell_0 \times w_0 \times h_0$ rectangular solid with the measurements Y_1, Y_2, and Y_3 for, respectively, the length ℓ_0, width w_0, and height h_0 with respective standard deviations σ_ℓ, σ_w, and σ_h. Here, we take $g(\ell, w, h) = \ell wh$, then*

$$\frac{\partial g}{\partial \ell}(\ell, w, h) = wh, \quad \frac{\partial g}{\partial w}(\ell, w, h) = \ell h, \quad \frac{\partial g}{\partial h}(\ell, w, h) = \ell w,$$

and

$$\begin{aligned}&\sigma_{g(Y_1,Y_2,Y_3)} \\ &\approx \sqrt{\left(\frac{\partial g}{\partial \ell}(\ell_0, w_0, h_0)\right)^2 \sigma_\ell^2 + \left(\frac{\partial g}{\partial w}(\ell_0, w_0, h_0)\right)^2 \sigma_w^2 + \left(\frac{\partial g}{\partial h}(\ell_0, w_0, h_0)\right)^2 \sigma_h^2} \\ &= \sqrt{(wh)^2\sigma_\ell^2 + (\ell h)^2\sigma_w^2 + (\ell w)^2\sigma_h^2}.\end{aligned}$$

11.4 Delta Method

Let's use repeated independent measurements, $Y_1, Y_2, \dots Y_n$ to estimate a quantity q by its sample mean $\bar{Y}$. If each measurement has mean μ_Y and variance σ_Y^2, then $\bar{Y}$ has mean $q = \mu_Y$ and variance σ_Y^2/n. By the central limit theorem,

$$Z_n = \frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{n}} \tag{11.7}$$

has a distribution that can be approximated by a standard normal. We can apply the propagation of error analysis based on a linear approximation of $g(\bar{Y})$ to obtain

$$g(\bar{Y}) \approx g(\mu_Y), \quad \text{and} \quad \text{Var}(g(\bar{Y})) \approx g'(\mu_Y)^2 \frac{\sigma_Y^2}{n}.$$

Thus, the reduction in the variance in the estimate of q "propagates" to a reduction in variance in the estimate of $g(q)$.

However, the central limit theorem gives us some additional information. Returning to the linear approximation (11.4)

$$g(\bar{Y}) \approx g(\mu_Y) + g'(\mu_Y)(\bar{Y} - \mu_Y). \tag{11.8}$$

The central limit theorem tells us that $\bar{Y}$ has a nearly normal distribution. Thus, the linear approximation to $g(\bar{Y})$ also has nearly a normal distribution. Moreover, with repeated measurements, the variance of $\bar{Y}$ is the variance of a single measurement divided by n. As a consequence, the linear approximation under repeated measurements yields a better approximation because the reduction in variance implies that the difference $\bar{Y} - \mu_Y$ is more likely to be small.

The **delta method** combines the central limit theorem and the propagation of error. To see use (11.8) to write,

$$\frac{g(\bar{Y}) - g(\mu_y)}{\sigma_{g(\bar{Y})}} \approx \frac{g'(\mu_Y)(\bar{Y} - \mu_Y)}{|g'(\mu_Y)|\sigma_Y/\sqrt{n}} = \pm Z_n.$$

The last equality uses (11.7). The $\pm$ sign depends on the sign of the derivative $g'(\mu_Y)$. Because the negative of a standard normal is also a standard normal, we have the desired approximation to the standard normal.

Then, Z_n converges in distribution to a standard normal random variable. In this way, the delta method greatly extends the applicability of the central limit theorem.

Let's return to our previous example on thermal expansion.

Example 11.19. *Let $Y_1, Y_2, \dots, Y_n$ be repeated unbiased measurement of a side of a cube with length ℓ_1 and temperature T_1. We use the sample mean $\bar{Y}$ of these measurements to estimate the length at temperature T_1 for the coefficient of linear expansion.*

$$\hat{\alpha}_1 = \frac{\bar{Y} - \ell_0}{\ell_0(T_1 - T_0)}.$$

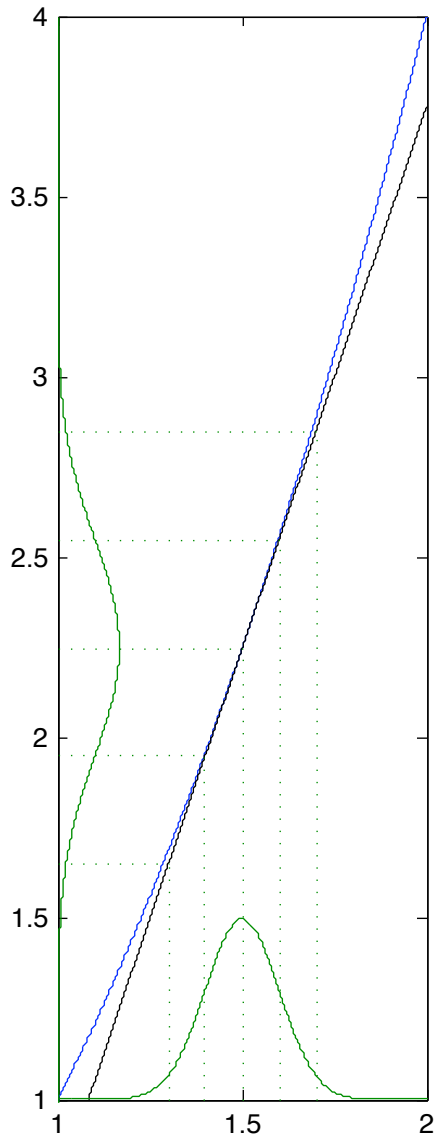


Figure 11.6: Illustrating the delta method. Here $\mu = 1.5$ and the blue curve $g(x) = x^2$. Thus, $g(\bar{X})$ is approximately normal with approximate mean 2.25 and $\sigma_{g(\bar{X})} \approx 3\sigma_{\bar{X}}$. The bell curve on the y-axis is the reflection of the bell curve on the x-axis about the (black) tangent line $y = g(\mu) + g'(\mu)(x - \mu)$.

Then, if each measurement Y_i has variance σ_Y^2,

$$\text{Var}(\hat{\alpha}_1) = \frac{\sigma_Y^2}{\ell_0^2(T_1 - T_0)^2 n} \quad \sigma_{\hat{\alpha}_1} = \frac{\sigma_Y}{\ell_0 |T_1 - T_0| \sqrt{n}}.$$

Now, we estimate the coefficient of volume expansion by

$$\hat{\alpha}_3 = \frac{\bar{Y}^3 - \ell_0^3}{\ell_0^3 (T_1 - T_0)}$$

and

$$\sigma_{\hat{\alpha}_3} \approx \frac{3\bar{Y}^2 \sigma_Y}{\ell_0^3 |T_1 - T_0| \sqrt{n}}.$$

By the delta method,

$$Z_n = \frac{\hat{\alpha}_3 - \alpha_3}{\sigma_{\hat{\alpha}_3}}$$

has a distribution that can be well approximated by a standard normal random variable.

Material	Coefficient of linear expansion
alumium	23.1
bradd	19
concrete	12
diamond	1
gasoline	317
glass	8.5
water	69

Table I: Coefficient of linear expansion at 20°C in units $10^{-6}/^{\circ}\text{C}$.

The next natural step is to take the approach used for the propagation of error in a multidimensional setting and extend the delta method. Focusing on the three dimensional case, we have three **independent** sequences $(Y_{1,1}, \ldots, Y_{1,n_1})$, $(Y_{2,1}, \ldots, Y_{2,n_2})$ and $(Y_{3,1}, \ldots, Y_{3,n_3})$ of independent random variables. The observations in the i-th sequence have mean μ_i and variance σ_i^2 for $i = 1, 2$ and 3. We shall use $\bar{Y}_1$, $\bar{Y}_2$ and $\bar{Y}_3$ to denote the sample means for the three sets of observations. Then, $\bar{Y}_i$ has

$$\text{mean} \quad \mu_i \quad \text{and} \quad \text{variance} \quad \tfrac{\sigma_i^2}{n_i} \quad \text{for } i = 1, 2, 3.$$

From the propagation of error linear approximation, we obtain

$$Eg(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3) \approx g(\mu_1, \mu_2, \mu_3).$$

For the variance, look to the multidimensional propagation of error variance formula (11.6) replacing the measurements Y_i by the sample mean $\bar{Y}_i$.

$$\begin{aligned} \sigma^2_{g(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)} &= \text{Var}(g(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)) \\ &\approx \frac{\partial g}{\partial y_1}(\mu_1, \mu_2, \mu_3)^2 \frac{\sigma_1^2}{n_1} + \frac{\partial g}{\partial y_2}(\mu_1, \mu_2, \mu_3)^2 \frac{\sigma_2^2}{n_2} + \frac{\partial g}{\partial y_3}(\mu_1, \mu_2, \mu_3)^2 \frac{\sigma_3^2}{n_3}. \end{aligned} \tag{11.9}$$

To obtain the normal approximation associated with the delta method, we need to have the additional fact that **the sum of independent normal random variables is also a normal random variable**. Thus, we have that, for n large,

$$Z_n = \frac{g(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3) - g(\mu_1, \mu_2, \mu_3)}{\sigma_{g(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)}}$$

is approximately a standard normal random variable.

Example 11.20. Fecundity *is the reproductive rate of a community or of a population. Fecundity can change over time due to both genetic and environmental circumstances. In avian biology, the fecundity B is defined as the number of female fledglings per female per year. $B > 1$ indicates a growing population and $B < 1$, a declining population. B is a product of three quantities,*

$$B = F \cdot p \cdot N,$$

where

- *F equals the mean number of female fledglings per successful nest,*
- *p equals nest survival probability, and*
- *N equals the mean number of nests built per female per year.*

Let's

- *collect measurement $F_1, \ldots, F_{n_F}$ on n_F nests to count female fledglings in a successful nest, and determine the sample mean $\bar{F}$,*
- *check n_p nests for survival probability, and determine the sample proportion $\hat{p}$, and*
- *follow n_N females to count the number $N_1, \ldots, N_{n_N}$ of successful nests per year and determine the sample mean $\bar{N}$.*

Our experimental design is structured so that measurements are independent. Then, taking the appropriate partial derivatives in (11.9) to $B = g(F, p, N) = F \cdot p \cdot N$, we obtain an estimate for the variance of $\hat{B} = g(\bar{F}, \hat{p}, \bar{N})$,

$$\begin{aligned}\sigma^2_{\hat{B}} &\approx \left(\frac{\partial B}{\partial F}(\mu_F, p, \mu_N)\right)^2 \frac{\sigma^2_F}{n_F} + \left(\frac{\partial B}{\partial p}(\mu_F, p, \mu_N)\right)^2 \frac{\sigma^2_p}{n_p} + \left(\frac{\partial B}{\partial N}(\mu_F, p, \mu_N)\right)^2 \frac{\sigma^2_N}{n_N}. \qquad (11.10)\\ &= (\mu_p\mu_N)^2\frac{\sigma^2_F}{n_F} + (\mu_F\mu_N)^2\frac{\sigma^2_p}{n_p} + (\mu_p\mu_F)^2\frac{\sigma^2_N}{n_N}.\end{aligned}$$

The checks of nest survival form a sequence of Bernoulli trials. Thus, $\mu_p = p$ and $\sigma^2_p = p(1-p)$ for a Bernoulli random variable, we can write the expression above upon dividing by B^2 as

$$\begin{aligned}\left(\frac{\sigma_{\hat{B}}}{B}\right)^2 &\approx \frac{1}{n_F}\left(\frac{\sigma_F}{\mu_F}\right)^2 + \frac{1}{n_p}\left(\frac{\sigma_p}{\mu_p}\right)^2 + \frac{1}{n_N}\left(\frac{\sigma_N}{\mu_N}\right)^2\\ &= \frac{1}{n_F}\left(\frac{\sigma_F}{\mu_F}\right)^2 + \frac{1}{n_p}\left(\frac{1-p}{p}\right) + \frac{1}{n_N}\left(\frac{\sigma_N}{\mu_N}\right)^2.\end{aligned}$$

This gives the individual contributions to the variance of B from each of the three data collecting activities - female fledglings, nest survivability, nest building. The values of n_F, n_p, and n_N can be adjusted in the collection of data to adjust the variance of $\hat{B}$ under a variety of experimental designs.

Estimates for $\sigma^2_{\hat{B}}$ can be found from the field data. Compute sample means

$$\bar{F}, \quad \hat{p}, \quad \textit{and} \quad \bar{N},$$

and sample variance

$$s^2_F, \quad \hat{p}(1-\hat{p}) \quad \textit{and} \quad s^2_N.$$

Using (11.10), we estimate the variance in fecundity

$$s^2_{\hat{B}} \approx \frac{1}{n_F}(\hat{p}\bar{N}s_F)^2 + \frac{1}{n_p}(\bar{F}\bar{N})^2\hat{p}(1-\hat{p}) + \frac{1}{n_N}(\hat{p}\bar{F}s_N)^2$$

If we make multiple measurements on the collection of female adult birds. The three observations (female offspring, net survivability, and number of nests) may be correlated, but the observations made on different female adults are independent. This leads to the following extension of the delta method.

Exercise 11.21. *In the case in which the n observations $Y_i = (Y_{i,1}, Y_{i,2}, \ldots Y_{i,n})$ are independent vector, but the entries in this vector may not be independent, show that the formula for the delta method is*

$$\text{Var}(g(\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_n)) \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial}{\partial y_i}g(\mu_1, \mu_2 \ldots, \mu_n)\frac{\partial}{\partial y_j}g(\mu_1, \mu_2 \ldots, \mu_n)\rho_{i,j}\sigma_i\sigma_j. \qquad (11.11)$$

Here the distributional means μ_i, sample means $\bar{Y}_i$, and distributional standard deviations σ_i for the i-th set of observations are as above and where $\rho_{i,j}$ is the correlation of i-th and j-th set of observations.

Exercise 11.22. *Use this formula for $\sigma^2_{B,n}$ in the case that measurements for F, p, and N for a given female adult are not independent.*

We will now move to a fundamental issue in statistic - estimation. The analysis of the properties an estimator, namely, its accuracy and its precision, are based to a large extent on the tools in probability theory that we have developed here - the law of large numbers, the central limit theorem and their extensions.

We finish the discussion on the central limit theorem with a summary of some of its applications.

11.5 Summary of Normal Approximations

The standardized score or z-score of some random quantity is

$$Z = \frac{\text{random quantity} - \text{mean}}{\text{standard deviation}}.$$

The **central limit theorem** and extensions like the **delta method** tell us when the z-score has an approximately standard normal distribution. Thus, using R, we can find good approximations by computing the probabilities of $P\{Z < z\}$, `pnorm(z)` and $P\{Z > z\}$ using `1-pnorm(z)` or $P\{z_1 < Z < z_2\}$ using the difference `pnorm(z2) - pnorm(z1)`.

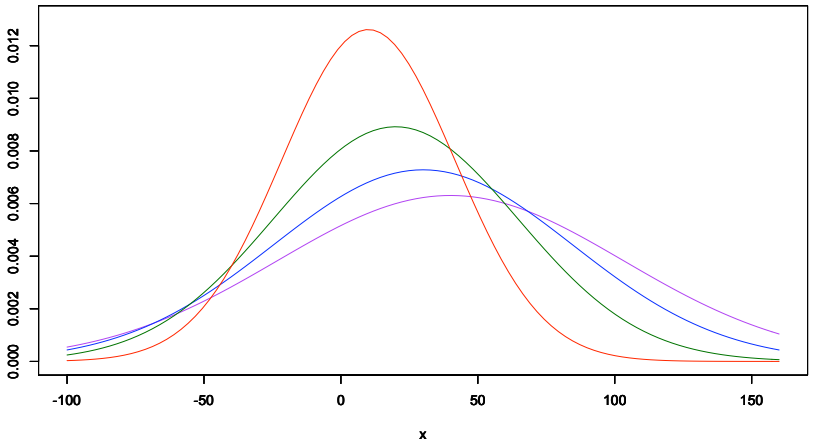


Figure 11.7: The density function for S_n for a random sample of size $n = 10$ (red), 20 (green), 30 (blue), and 40 (purple). In this example, the observations are normally distributed with mean $\mu = 1$ and standard deviation $\sigma = 10$.

11.5.1 Sample Sum

If we have a sum S_n of n independent random variables, $X_1, X_2, \ldots X_n$ whose common distribution has mean μ and variance σ^2, then

- the mean $ES_n = n\mu$,
- the variance $\text{Var}(S_n) = n\sigma^2$,
- the standard deviation is $\sigma\sqrt{n}$.

Thus, S_n is approximately normal with mean $n\mu$ and variance $n\sigma^2$. The z-score in this case is

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

We can approximate $P\{S_n < x\}$ by noting that this is the same as computing the probability

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} < \frac{x - n\mu}{\sigma\sqrt{n}} = z$$

and finding $P\{Z_n < z\}$ using the standard normal distribution.

11.5.2 Sample Mean

For a sample mean

$$\bar{X} = (X_1 + X_2 + \cdots + X_n)/n,$$

- the mean $E\bar{X} = \mu$,
- the variance $\text{Var}(\bar{X}) = \sigma^2/n$,
- the standard deviation is $\sigma/\sqrt{n}$.

Thus, $\bar{X}$ is approximately normal with mean μ and variance σ^2/n. The z-score in this case is

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Thus,

$$\bar{X} < x \quad \text{is equivalent to} \quad Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{x - \mu}{\sigma/\sqrt{n}}.$$

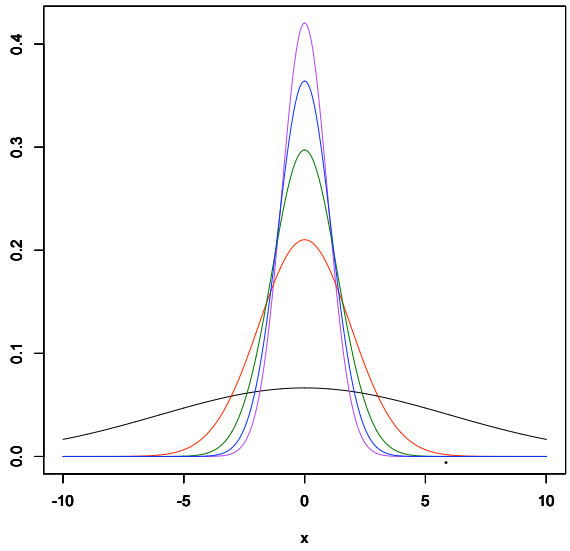


Figure 11.8: The density function for $\bar{X} - \mu$ for a random sample of size $n = 1$ (black), 10 (red), 20 (green), 30 (blue), and 40 (purple). In this example, the observations are normally distributed with standard deviation $\sigma = 10$.

11.5.3 Sample Proportion

For Bernoulli trials $X_1, X_2, \ldots X_n$ with success probability p, let $\hat{p} = (X_1 + X_2 + \cdots + X_n)/n$ be the sample proportion. Then

- the mean $E\hat{p} = p$,
- the variance $\text{Var}(\hat{p}) = p(1-p)/n$,
- the standard deviation is $\sqrt{p(1-p)/n}$.

Thus, $\hat{p}$ is approximately normal with mean p and variance $p(1-p)/n$. The z-score in this case is

$$Z_n = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}.$$

For the special case of Bernoulli trials, normal approximations often use a continuity correction.

11.5.4 Delta Method

For the delta method in one variable using $\bar{X}$ and a function g, for a sample mean $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$, we have

- the mean $Eg(\bar{X}) \approx g(\mu)$,
- the variance $\text{Var}(g(\bar{X})) \approx g'(\mu)^2\sigma^2/n$,
- the standard deviation is $|g'(\mu)|\sigma/\sqrt{n}$.

Thus, $g(\bar{X})$ is approximately normal with mean $g(\mu)$ and variance $g'(\mu)^2\sigma^2/n$. The z-score is

$$Z_n = \frac{g(\bar{X}) - g(\mu)}{|g'(\mu)|\sigma/\sqrt{n}}.$$

For the two variable delta method, we now have two independent sequences of independent random variables, $X_{1,1}, X_{1,2}, \ldots X_{1,n_1}$ whose common distribution has mean μ_1 and variance σ_1^2 and $X_{2,1}, X_{2,2}, \ldots X_{2,n_2}$ whose common distribution has mean μ_2 and variance σ_2^2. For a function g of the sample means, we have that

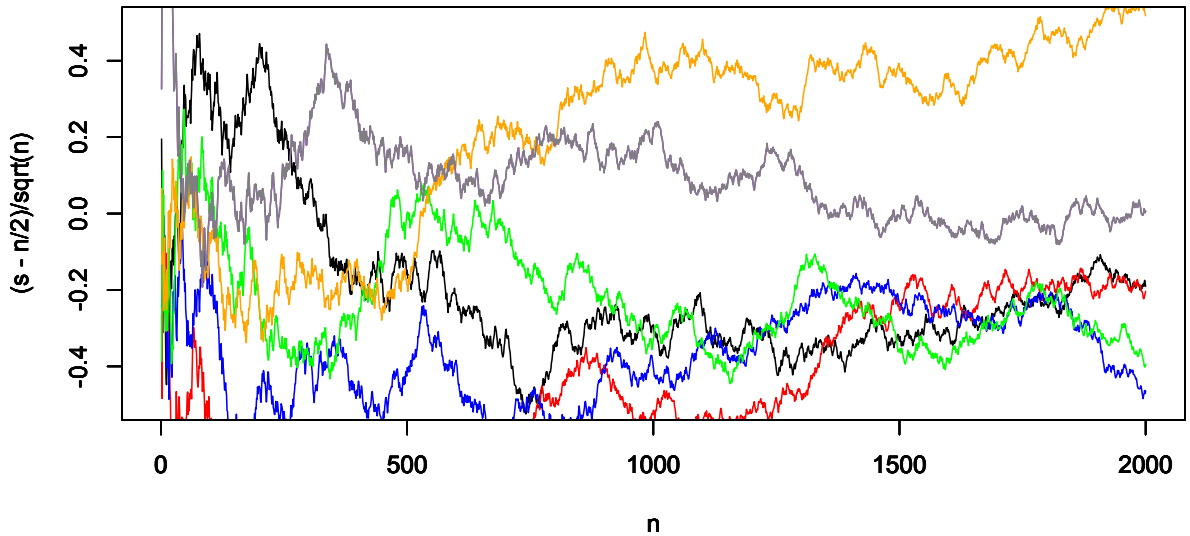


Figure 11.9: Six plots of of $(S_n - n/2)/\sqrt{n}$ for $U(0,1)$ random variables.

- the mean $Eg(\bar{X}_1, \bar{X}_2) \approx g(\mu_1, \mu_2)$,
- the variance
$$\text{Var}(g(\bar{X}_1, \bar{X}_2)) = \sigma^2_{g,n} \approx \left(\frac{\partial}{\partial x} g(\mu_1, \mu_2)\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{\partial}{\partial y} g(\mu_1, \mu_2)\right)^2 \frac{\sigma_2^2}{n_2},$$
- the standard deviation is $\sigma_{g,n}$.

Thus, $g(\bar{X}_1, \bar{X}_2)$ is approximately normal with mean $g(\mu_1, \mu_2)$ and variance $\sigma^2_{g,n}$. The z-score is

$$Z_n = \frac{g(\bar{X}_1, \bar{X}_2) - g(\mu_1, \mu_2)}{\sigma_{g,n}}.$$

The generalization of the delta method to higher dimensional data will add terms to the variance formula. Fuse or multiple observations on n individuals, (**??** to determine the standard deviation in the z-score.

11.6 Answers to Selected Exercises

11.1 Here is the code for one of the plots of of $(S_n - n/2)/\sqrt{n}$ in Figure 11.9.

```
> n<-1:2000
> x<-runif(2000)
> s<-cumsum(x)
> plot(n,(s-n/2)/sqrt(n),type="l",
  ylim=c(-0.5,0.5),col="orange")
```

For the 1000 simulations of $(S_{2000} - 1000)/\sqrt{2000}$ with a summary and a histogram, we have

```
> score<-numeric(1000)
> for (i in 1:1000){x<-runif(2000);
  s<-sum(x);score[i]<-(s-1000)/sqrt(2000)}
```

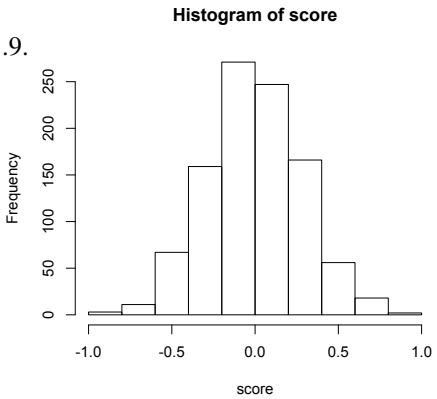


Figure 11.10: 1000 simulations of $(S_{2000} - 1000)/\sqrt{2000}$, the final values for the simulations above with $n = 2000$.

```
> mean(score)
[1] -0.001879942
> sd(score)
[1] 0.28578
> hist(score)
```

This histogram has a bell shape with the mean at 0 and a standard deviation of 0.2857.

11.2. The sum S_n of exponential random variables is $\Gamma(n, \lambda)$ and thus has mean n/λ and standard deviation $\sqrt{n}/\lambda$. The skewness is the third moment of the standardized sum,

$$\frac{S_n - n/\lambda}{\sqrt{n}\lambda} = \frac{\lambda S_n - n}{\sqrt{n}} = \frac{T_n - n}{\sqrt{n}}$$

where $T_n = \lambda S_n$ is $\Gamma(n, 1)$. (Check this!) Thus, the skewness

$$E\left[\left(\frac{T_n - n}{\sqrt{n}}\right)^3\right] = \frac{1}{n^{3/2}}(ET_n^3 - 3nET_n^2 + 3n^2ET_n - n^3).$$

Here we use the linearity properties of expectation. Now, the first moment, $ET_n = n$. For the second moment,

$$ET_n^2 = \int_0^\infty \frac{1}{\Gamma(n)}x^2x^{n-1}e^{-x}dx = \frac{\Gamma(n+2)}{\Gamma(n)}\int_0^\infty \frac{1}{\Gamma(n+2)}x^{(n+2)-1}e^{-x}dx = (n+1)n.$$

Notice that we are integrating the density function of a $\Gamma(n+2, 1)$ random variable. Similarly,

$$ET_n^3 = \int_0^\infty \frac{1}{\Gamma(n)}x^3x^{n-1}e^{-x}dx = \frac{\Gamma(n+3)}{\Gamma(n)}\int_0^\infty \frac{1}{\Gamma(n+3)}x^{(n+3)-1}e^{-x}dx = (n+2)(n+1)n.$$

Returning to the skewness, we substitute for each of the first three moments of T_n

$$\begin{aligned} E\left[\left(\frac{T_n - n}{\sqrt{n}}\right)^3\right] &= \frac{1}{n^{3/2}}((n+2)(n+1)n - 3n(n+1)n + 3n^2n - n^3) \\ &= \frac{1}{\sqrt{n}}((n+2)(n+1) - 3(n+1)n + 2n^3) \\ &= \frac{1}{\sqrt{n}}(n^2 + 3n + 2 - 3n^2 - 3n + 2n^2) = \frac{2}{\sqrt{n}}. \end{aligned}$$

11.3. Let μ and σ denote the common mean and standard deviation the standardized random variable

$$\left(\frac{S_n - n\mu}{\sigma\sqrt{n}}\right)^3 = \frac{1}{n^{3/2}}\left(\frac{S_n - n\mu}{n\sigma}\right)^3 = \frac{1}{n^{3/2}}\left(\sum_{i=1}^n \frac{X_i - \mu}{\sigma}\right)^3 = \frac{1}{n^{3/2}}\left(\sum_{i=1}^n X_i^*\right)^3, \tag{11.12}$$

where $X_i^* = (X_i - \mu)/\sigma$ is the standardized version of X_i. In particular, $EX_i^* = 0$.

We expand the cube of the sum in (11.12) and use the property of the expectation of the product of independent random variables. In so doing, we see that the expected value of the cube of a sum involves terms $\gamma_1 = EX_i^{*3}$, the skewness of a single random variable in the sum, terms $EX_i^{*2}X_j^* = EX_i^{*2}EX_j^*$ with $i \neq j$, and terms $EX_i^*X_j^*X_k^* = EX_i^*EX_j^*EX_k^*$ where i, j, and k all differ. Each of the last two types of terms equals 0 and consequently does not contribute to the expected value. Thus,

$$E\left[\left(\frac{S_n - n\mu}{\sigma\sqrt{n}}\right)^3\right] = \frac{1}{n^{3/2}}\sum_{i=1}^n E[X_i^{*3}] = \frac{1}{n^{3/2}}n\gamma_1 = \frac{1}{\sqrt{n}}\gamma_1.$$

The binomial random variable S_n can be realized as the sum of n independent Bernoulli random variables with skewness $\gamma_1 = (1-2p)/\sqrt{p(1-p)}$.

11.7. A Poisson random variable with parameter $\lambda = 16$ has mean 16 and standard deviation $4 = \sqrt{16}$. Thus, we first look at the maximum difference in the distribution function of a $Pois(4)$ random variable, X, and a $N(16, 4)$ random variable, Y, by comparing $P\{X \le x\}$ to $P\{Y \le x + \frac{1}{2}\}$ in a range around the mean value of 16.

```
> x<-c(4:28)
> max(abs(pnorm(x+0.5,16,4)-ppois(x,16)))
[1] 0.01648312
```

The maximum difference between the distribution function is approximately 1.6%. To compare the density functions, we have the R commands. (See Figure 11.11.)

```
> poismass<-dpois(x,16)
> plot(x,poismass,ylim=c(0,0.1),
   ylab="probability")
> par(new=TRUE)
> x<-seq(4,28,0.01)
> normd<-dnorm(x,16,4)
> plot(x,normd,ylim=c(0,0.1),
   ylab="probability",type="l",col="red")
```

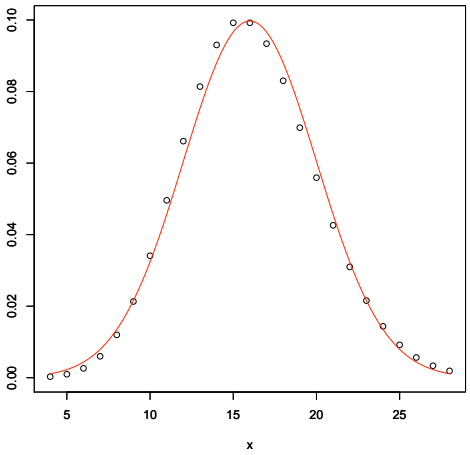


Figure 11.11: Circles indicate the mass function for a $Pois(16)$ random variable. The red curve is the density function of a $N(16, 4)$ random variable. The plots show that the Poisson random variable is slightly more skewed to the right that the normal.

11.11. For Z a standard normal random variable to determine $z_{0.01}$ that satisfies $P\{Z > z_{0.01}\} = 0.01$, we use the R command

```
> qnorm(0.99)
[1] 2.326348
```

Thus, we look for the value n that gives a standardized score of $z_{0.01}$.

$$\begin{aligned} 2.326348 = z_{0.01} = \frac{400n - 42000}{50 \cdot \sqrt{n}} = \frac{8n - 840}{\cdot\sqrt{n}} \\ 2.326348\sqrt{n} = 8n - 840 = 8(n - 105) \\ 0.08456085n = n^2 - 210n + 11025 \\ 0 = n^2 - 209.9154n + 11025 \end{aligned}$$

By the quadratic formula, we solve for n.

$$n = \frac{209.9154 + \sqrt{(209.9154)^2 - 4 \cdot 1 \cdot 11025}}{2 \cdot 1} = 388.70152$$

So, take $n = 388$.

11.12. The R code for the simulations is

```
> xbar<-numeric(1000)
> for (i in 1:1000)
   {x<-runif(100);xbar[i]<-mean(x)}
> hist(xbar)
> mean(xbar)
[1] 0.498483
```

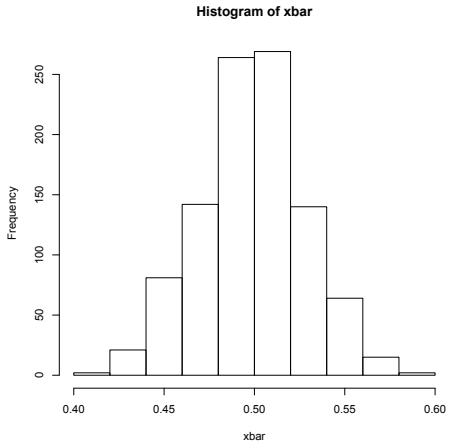


Figure 11.12: Histogram of the sample means of 100 random variables, uniformly distributed on $[0, 1]$.

```
> sd(xbar)
[1] 0.02901234
> quantile(xbar,0.35)
     35%
0.488918
> qnorm(0.35)
[1] -0.3853205
```

The mean of a $U[0,1]$ random variable is $\mu = 1/2$ and its variance is $\sigma^2 = 1/12$. Thus the mean of $\bar{X}$ is 1/2, its standard deviation is $\sqrt{1/(12 \cdot 100)} = 0.0289$, close to the simulated values.

Use `qnorm(0.35)` to see that the 35th percentile corresponds to a z-score of -0.3853205. Thus, the 35th percentile for $\bar{X}$ is approximately

$$\mu + z_{0.35}\frac{\sigma}{\sqrt{n}} = 0.5 - 0.3853205\frac{1}{\sqrt{1200}} = 0.4888768,$$

agreeing to four decimal places the value given by the simulations of `xbar`.

11.12. Using (11.3)

$$E[a + b(Y - \mu_Y)] = E[a - b\mu_y + bY] = a - b\mu_Y + b\mu_Y = b.$$

and

$$\text{Var}(a + b(Y - \mu_Y)) = \text{Var}(a - b\mu_y + bY) = b^2\text{Var}(Y).$$

11.15. Using right triangle trigonometry, we have that

$$\theta = g(\ell) = \tan^{-1}\left(\frac{\ell}{10}\right). \quad \text{Thus,} \quad g'(\ell) = \frac{1/10}{1 + (\ell/10)^2} = \frac{10}{100 + \ell^2}.$$

So, $\sigma_{\hat{\theta}} \approx 10/(100 + \ell^2) \cdot \sigma_\ell$. For example, set $\sigma_\ell = 0.1$ meter and $\ell = 5$. Then, $\sigma_{\hat{\theta}} \approx 10/125 \cdot 0.1 = 1/125$ radians = $0.49°$.

11.16. In this case,

$$\theta = g(\ell, h) = \tan^1\left(\frac{\ell}{h}\right).$$

For the partial derivatives, we use the chain rule

$$\frac{\partial g}{\partial \ell}(\ell, h) = \frac{1}{1 + (\ell/h)^2}\left(\frac{1}{h}\right) = \frac{h}{h^2 + \ell^2} \quad \frac{\partial g}{\partial h}(\ell, h) = \frac{1}{1 + (\ell/h)^2}\left(\frac{-\ell}{h^2}\right) = -\frac{\ell}{h^2 + \ell^2}$$

Thus,

$$\sigma_{\hat{\theta}} \approx \sqrt{\left(\frac{h}{h^2 + \ell^2}\right)^2 \sigma_\ell^2 + \left(\frac{\ell}{h^2 + \ell^2}\right)^2 \sigma_h^2} = \frac{1}{h^2 + \ell^2}\sqrt{h^2\sigma_\ell^2 + \ell^2\sigma_h^2}.$$

If $\sigma_h = \sigma_\ell$, let σ denote their common value. Then

$$\sigma_{\hat{\theta}} \approx \frac{1}{h^2 + \ell^2}\sqrt{h^2\sigma^2 + \ell^2\sigma^2} = \frac{\sigma}{\sqrt{h^2 + \ell^2}}.$$

In other words, $\sigma_{\hat{\theta}}$ is inversely proportional to the length of the hypotenuse.

11.17. Let μ_i be the mean of the i-th measurement. Then

$$\sigma_{g(Y_1,Y_2,\cdot,Y_d t)} \approx \sqrt{\left(\frac{\partial g}{\partial y_1}(\mu_1, \ldots, \mu_d)\right)^2 \sigma_1^2 + \left(\frac{\partial g}{\partial y_2}(\mu_1, \ldots, \mu_d)\right)^2 \sigma_2^2 + \cdots + \left(\frac{\partial g}{\partial y_d}(\mu_1, \ldots, \mu_d)\right)^2 \sigma_d^2}.$$

11.21. Recall that for random variables $X_1, X_2, \dots X_n$ and constants $c_1, c_2, \dots c_n$,

$$\text{Var}(c_0 + c_1X_1 + c_2X_2 + \cdots + c_nX_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_j\text{Cov}(X_i, X_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_j\rho_{i,j}\sigma_{X_i}\sigma_{X_j}.$$

where $\rho_{i,j}$ is the correlation of X_i and X_j. Note that the correlation of a random variable with itself, $\rho_{i,i} = 1$. For the delta method,

$$c_i = \frac{\partial}{\partial y_i}g(\mu_1, \mu_2 \dots, \mu_n), \quad X_i = \bar{Y}_i, \quad \text{and} \quad \sigma_{X_i} = \frac{\sigma_i}{\sqrt{n_i}}.$$

. Then

$$\text{Var}(g(\bar{Y}_1, \bar{Y}_2, \dots, Y_n)) \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial}{\partial y_i}g(\mu_1, \mu_2 \dots, \mu_n)\frac{\partial}{\partial y_j}g(\mu_1, \mu_2 \dots, \mu_n)\rho_{i,j}\sigma_i\sigma_j.$$

11.22. Let μ_F, p, μ_N be the means of the variables under consideration. Then we have the linear approximation,

$$\begin{aligned} g(\bar{F}, \hat{p}, \bar{N}) &\approx g(F, p, N) + \frac{\partial g}{\partial F}(F, p, N)(\bar{F} - F) + \frac{\partial g}{\partial p}(F, p, N)(\hat{p} - p) + \frac{\partial g}{\partial N}(F, p, N)(\bar{N} - \mu_N). \\ &= g(F, p, N) + pN(\bar{F} - F) + FN(\hat{p} - p) + Fp(\bar{N} - \mu_N) \end{aligned}$$

Matching this to the covariance formula, we have

$$c_0 = g(F, p, N), \quad c_1 = pN, \quad c_2 = FN, \quad c_3 = Fp,$$

$$X_1 = \bar{F}, \quad X_2 = \hat{p}, \quad X_3 = \bar{N}.$$

Thus,

$$\begin{aligned} \sigma^2_{B,n} = &\frac{1}{n_F}(pN\sigma_F)^2 + \frac{1}{n_p}(FN\sigma_p)^2 + \frac{1}{n_N}(pF\sigma_N)^2 \\ &+2FpN^2\rho_{F,p}\frac{\sigma_F\sigma_p}{\sqrt{n_Fn_p}} + 2Fp^2N\rho_{F,N}\frac{\sigma_F\sigma_N}{\sqrt{n_Fn_N}} + 2F^2pN\rho_{p,N}\frac{\sigma_p\sigma_N}{\sqrt{n_Fn_N}}. \end{aligned}$$

The subscripts on the correlation coefficients ρ have the obvious meaning.

Part III

Estimation

Topic 12

Overview of Estimation

> *Inference is the problem of turning data into knowledge, where knowledge often is expressed in terms of entities that are not present in the data per se but are present in models that one uses to interpret the data. Statistical rigor is necessary to justify the inferential leap from data to knowledge, and many difficulties arise in attempting to bring statistical principles to bear on massive data. Overlooking this foundation may yield results that are, at best, not useful, or harmful at worst. In any discussion of massive data and inference, it is essential to be aware that it is quite possible to turn data into something resembling knowledge when actually it is not. Moreover, it can be quite difficult to know that this has happened.* - page 2, Frontiers in Massive Data Analysis by the National Research Council, 2013.

The balance of this book is devoted to developing formal procedures of statistical inference. In this introduction to inference, we will be basing our analysis on the premise that the data have been collected according to well-designed procedures. We will focus our presentation on parametric estimation and hypothesis testing based on a given family of probability models chosen in line with the science under investigation and with the data collection procedures.

12.1 Introduction

In the simplest possible terms, the goal of **estimation theory** is to answer the question:

What is that number?

What is the length, the reaction rate, the fraction displaying a particular behavior, the temperature, the kinetic energy, the Michaelis constant, the speed of light, mutation rate, the melting point, the probability that the dominant allele is expressed, the elasticity, the force, the mass, the free energy, the mean number of offspring, the focal length, mean lifetime, the slope and intercept of a line?

The next step is to perform an experiment that is well designed to estimate one (or more) numbers. However, before we can embark on such a design, we must be informed by some principles of estimation in order to have some understanding of the properties of a good estimator and to present our uncertainties concerning the estimate. Statistics has provided two distinct approaches - typically called **classical** or frequentist and **Bayesian**. We shall give an overview of both approaches. However, this text will emphasize the classical approach.

Let's begin with a definition:

Definition 12.1. *A* **statistic** *is a function of the data that does not depend on any unknown parameter.*

We have to this point, seen a variety of statistics.

Example 12.2.

- *sample mean,* $\bar{x}$

- *sample variance,* s^2
- *sample standard deviation,* s
- *sample median, sample quartiles* Q_1, Q_3*, percentiles and other quantiles*
- *standardized scores* $(x_i - \bar{x})/s$
- *order statistics* $x_{(1)}, x_{(2)}, \dots x_{(n)}$*, including sample maximum and minimum*
- *sample moments*

$$\overline{x^m} = \frac{1}{n}\sum_{k=1}^{n} x_k^m, \quad m = 1, 2, 3, \dots.$$

Here, we will look at a particular type of **parameter estimation**, in which we consider $X = (X_1, \dots, X_n)$, independent random variables chosen according to one of a family of probabilities P_θ where θ is element from the **parameter space** Θ. Based on our analysis, we choose an **estimator** $\hat{\theta}(X)$. If the data $\mathbf{x}$ takes on the values $x_1, x_2, \dots, x_n$, then

$$\hat{\theta}(x_1, x_2, \dots, x_n)$$

is called the **estimate** of θ. Thus we have three closely related objects,

1. θ - the parameter, an element of the parameter space Θ. This is a number or a vector.
2. $\hat{\theta}(x_1, x_2, \dots, x_n)$ - the estimate. This again is a number or a vector obtained by evaluating the estimator on the data $\mathbf{x} = (x_1, x_2, \dots, x_n)$.
3. $\hat{\theta}(X_1, \dots, X_n)$ - the estimator. This is a random variable. We will analyze the distribution of this random variable to decide how well it performs in estimating θ.

The first of these three objects is a number. The second is a statistic. The third can be analyzed and its properties described using the theory of probability. Keeping the relationship among these three objects in mind is essential in understanding the fundamental issues in statistical estimation.

Example 12.3. *For Bernoulli trials* $X = (X_1, \dots, X_n)$*, we have*

1. *p, a single parameter, the probability of success, with parameter space* $[0, 1]$.
2. $\hat{p}(x_1, \dots, x_n)$ *is the* **sample proportion** *of successes in the data set.*
3. $\hat{p}(X_1, \dots, X_n)$*, the sample mean of the random variables*

$$\hat{p}(X_1, \dots, X_n) = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}S_n$$

is an estimator of p. We can give the distribution of this estimator because S_n *is a binomial random variable.*

Example 12.4. *Given pairs of observations* $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ *that display a general linear pattern, we use* **ordinary least squares regressn** *for*

1. *parameters - the slope* β *and intercept* α *of the regression line. So, the parameter space is* $\mathbb{R}^2$*, pairs of real numbers.*
2. *They are estimated using the statistics* $\hat{\beta}$ *and* $\hat{\alpha}$ *in the equations*

$$\hat{\beta}(\mathbf{x}, \mathbf{y}) = \frac{\operatorname{cov}(\mathbf{x}, \mathbf{y})}{\operatorname{var}(\mathbf{x})}, \quad \bar{y} = \hat{\alpha}(\mathbf{x}, \mathbf{y}) + \hat{\beta}(\mathbf{x}, \mathbf{y})\bar{x}.$$

3. *Later, when we consider statistical inference for linear regression, we will analyze the distribution of the estimators.*

Exercise 12.5. *Let* $X = (X_1, \dots, X_n)$ *be independent uniform random variables on the interval* $[0, \theta]$ *with* θ *unknown. Give some estimators of* θ *from the statistics above.*

12.2 Classical Statistics

In classical statistics, the **state of nature** is assumed to be fixed, but unknown to us. Thus, one goal of estimation is to determine which of the P_θ is the source of the data. The **estimate** is a statistic

$$\hat{\theta} : \text{data} \to \Theta.$$

Introduction to estimation in the classical approach to statistics is based on two fundamental questions:

- How do we determine estimators?
- How do we evaluate estimators?

We can ask if this estimator in any way systematically under or over estimate the parameter, if it has large or small variance, and how does it compare to a notion of best possible estimator. How easy is it to determine and to compute and how does the procedure improve with increased sample size?

The raw material for our analysis of any estimator is the **distribution of the random variables** that underlie the data under any possible value θ of the parameter. To simplify language, we shall use the term **density function** to refer to both continuous and discrete random variables. Thus, to each parameter value $\theta \in \Theta$, there exists a density function which we denote

$$\mathbf{f}_X(\mathbf{x}|\theta).$$

We focus on experimental designs based on a **simple random sample**. To be more precise, the observations are based on an experimental design that yields a sequence of random variables

$$X_1, \dots, X_n,$$

drawn from a family of distributions having common density $f_X(x|\theta)$ where the parameter value θ is unknown and must be estimated. Because the random variables are independent, the **joint density** is the product of the **marginal densities**.

$$\mathbf{f}_X(\mathbf{x}|\theta) = \prod_{k=i}^{n} f_X(x_k|\theta) = f_X(x_1|\theta) f_X(x_2|\theta) \cdots f_X(x_n|\theta).$$

In this circumstance, the data $\mathbf{x}$ are known and the parameter θ is unknown. Thus, we write the density function as

$$L(\theta|\mathbf{x}) = \mathbf{f}_X(\mathbf{x}|\theta)$$

and call L the **likelihood function**.

Because the algebra and calculus of $\mathbf{f}_X(\mathbf{x}|\theta)$ are a bit unfamiliar, we will look at several examples.

Example 12.6 (Parametric families of densities)**.**

1. *For Bernoulli trials with a known number of trials n but unknown success probability parameter p has joint density*

$$\begin{aligned}\mathbf{f}_X(\mathbf{x}|p) &= p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} \cdots p^{x_n}(1-p)^{1-x_n} = p^{\sum_{k=1}^n x_k}(1-p)^{\sum_{k=1}^n (1-x_k)} \\ &= p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k} = p^{n\bar{x}}(1-p)^{n(1-\bar{x})}\end{aligned}$$

2. *Normal random variables with known variance σ_0 but unknown mean μ has joint density*

$$\begin{aligned}\mathbf{f}_X(\mathbf{x}|\mu) &= \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(x_1-\mu)^2}{2\sigma_0^2}\right) \cdot \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(x_2-\mu)^2}{2\sigma_0^2}\right) \cdots \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma_0^2}\right) \\ &= \frac{1}{(\sigma_0\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{k=1}^{n}(x_k-\mu)^2\right)\end{aligned}$$

3. Normal random variables with unknown mean μ and variance σ has density

$$\mathbf{f}_X(\mathbf{x}|\mu,\sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\right).$$

4. Beta random variables with parameters α and β has joint desity

$$\begin{aligned}\mathbf{f}_X(x|\alpha,\beta) &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^n (x_1\cdot x_2\cdots x_n)^{\alpha-1}((1-x_1)\cdot(1-x_2)\cdots(1-x_n))^{\beta-1}\\ &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^n \left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\left(\prod_{i=1}^{n}(1-x_i)\right)^{\beta-1}\end{aligned}$$

Exercise 12.7. *Give the likelihood function for n observations of independent $\Gamma(\alpha,\beta)$ random variables .*

The choice of a **point estimator** $\hat{\theta}$ is often the first step. The next two topics will be devoted to consider two approaches for determining estimators - method of moments and maximum likelihood. We next move to analyze the quality of the estimator. With this in view, we will give methods for approximating the bias and the variance of the estimators. Typically, this information is, in part, summarized though what is know as an **interval estimator**. This is a procedure that determines a subset of the parameter space with high probability that it contains the real state of nature. We see this most frequently in the use of **confidence intervals**.

12.3 Bayesian Statistics

For a few tosses of a coin always that always turn up tails, the estimate $\hat{p} = 0$ for the probability of heads did not seem reasonable to Thomas Bayes. He wanted a way to place our uncertainly of the value for p into the procedure for estimation.

Today, the **Bayesian approach to statistics** takes into account not only the density

$$\mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)$$

for the data collected for any given experiment but also external information to determine a **prior density** π on the parameter space Θ. Thus, in this approach, both the parameter and the data are modeled as random. Estimation is based on Bayes formula. We now want to take Bayes theorem, previously derived for a finite partition and obtain a formula useful for Bayesian estimation. The first step will yield a formula based on a discrete mixture. We will then need to introduce the notion of a continuous mixture to give us the final formula, (12.4).

Let $\tilde{\Theta}$ be a random variable having the given prior density π. In the case in which both $\tilde{\Theta}$ and the data take on only a finite set of values, $\tilde{\Theta}$ is a discrete random variable and π is a mass function

$$\pi\{\psi\} = P\{\tilde{\Theta}=\psi\}.$$

Let $C_\psi = \{\tilde{\Theta}=\psi\}$ be the event that $\tilde{\Theta}$ takes on the value ψ and $A = \{X = \mathbf{x}\}$ be the values taken on by the data. Then $\{C_\psi, \psi\in\Theta\}$ from a partition of the probability space. Bayes formula is

$$P(C_\theta|A) = \frac{P(A|C_\theta)P(C_\theta)}{\sum_\psi P(A|C_\psi)P(C_\psi)} \quad \text{or} \tag{12.1}$$

$$f_{\Theta|X}(\theta|\mathbf{x}) = P\{\tilde{\Theta}=\theta|X=\mathbf{x}\} = \frac{P\{X=\mathbf{x}|\tilde{\Theta}=\theta\}P\{\tilde{\Theta}=\theta\}}{\sum_\psi P\{X=\mathbf{x}|\tilde{\Theta}=\psi\}P\{\tilde{\Theta}=\psi\}} = \frac{\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi\{\theta\}}{\sum_\psi \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi\{\psi\}}.$$

Given data $\mathbf{x}$, the function of θ, $f_{\Theta|X}(\theta|\mathbf{x}) = P\{\tilde{\Theta}=\theta|X=\mathbf{x}\}$ is called the **posterior density**.

r

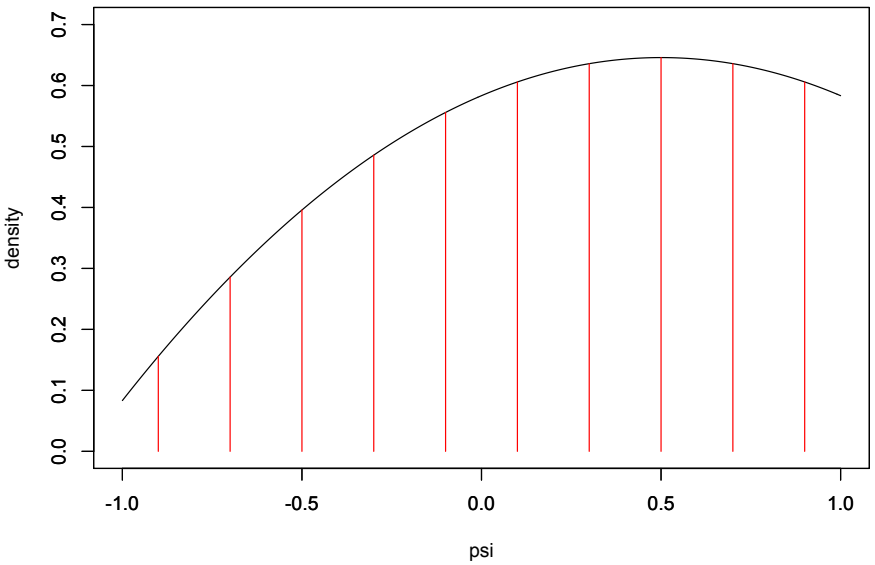


Figure 12.1: Plot of density $\pi(\psi)$ in black for continuous mixture and the approximating discrete mixture with weights $\tilde{\pi}(\tilde{\psi})$ proportional to the heights of the red vertical lines.

Remark 12.8. *As we learned in tte section on Random Variables and Distribution functions, the expression*

$$\sum_{\psi} \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi\{\psi\}$$

is the **mixture** *of the densities $\mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)$ for ψ in the finite set with weights $\pi(\psi)$. Typically the parameter space Θ is continuous and so we want to use the density π of a continuous random variable. To determine an expression for a continuous mixture, we will be guided by the ideas used deriving the formula for the expected value for a continuous random variable based on the formula for a discrete random variable. Beginning with the property*

$$\int_{\theta} \pi(\psi)d\psi = 1 \quad \textit{we have, for a Riemann sum,} \quad \sum_{\tilde{\psi}} \pi(\tilde{\psi})\Delta\psi \approx 1.$$

Now, write

$$\tilde{\pi}\{\tilde{\psi}\} = \pi(\tilde{\psi})\Delta\psi. \tag{12.2}$$

Then $\tilde{\pi}$ is (approximately) the density function for a discrete random variable. If we take a mixture of $\mathbf{f}_{X|\Theta}(\mathbf{x}|\tilde{\psi})$ with weights $\tilde{\pi}(\tilde{\psi})$, we have the mixture density

$$\sum_{\tilde{\psi}} \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\tilde{\pi}\{\tilde{\psi}\} = \sum_{\tilde{\psi}} \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\tilde{\psi})\Delta\psi.$$

This last sum is a Riemann sum and so taking limits as $\Delta\psi \to 0$, we have that the Riemann sum converges to the definite integral. This gives us the continuous mixture

$$\mathbf{f}_X(\mathbf{x}) = \int_{\Theta} \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\, d\psi. \tag{12.3}$$

Exercise 12.9. *Show that the expression (12.3) for $\mathbf{f}_X(\mathbf{x})$ is a valid density function*

Returning to the expression (12.1), substituting (12.2), we have in the interval from θ to $\theta + \Delta\theta$,

$$f_{\Theta|X}(\theta|\mathbf{x})\Delta\theta = \frac{\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)\Delta\theta}{\sum_{\psi} \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\Delta\psi}.$$

After dividing by $\Delta\theta$ and taking a limit as $\Delta\psi \to 0$, we have, for π, a density for a continuous random variable, that the sum in Bayes formula becomes an integral for a continuous mixture,

$$f_{\Theta|X}(\theta|\mathbf{x}) = \frac{\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)}{\int \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\, d\psi} \tag{12.4}$$

Sometimes we shall write (12.4) as

$$f_{\Theta|X}(\theta|\mathbf{x}) = c(\mathbf{x})\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)$$

where $c(\mathbf{x})$, the reciprocal of continuous mixture (12.3) in the denominator of (12.4), is the value necessary so that the integral of the posterior density $f_{\Theta|X}(\theta|\mathbf{x})$ with respect to θ equals 1. We might also write

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto \mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta) \tag{12.5}$$

where $c(\mathbf{x})$ is the constant of proportionality.

Estimation, e.g., point and interval estimates, in the Bayesian approach is based on the data and an analysis using the posterior density. For example, one way to estimate θ is to use the mean of the posterior distribution, or more briefly, the **posterior mean**,

$$\hat{\theta}(\mathbf{x}) = E[\theta|\mathbf{x}] = \int \theta f_{\Theta|X}(\theta|\mathbf{x})\, d\theta.$$

Example 12.10. *As suggested in the original question of Thomas Bayes, we will make independent flips of a biased coin and use a Bayesian approach to make some inference for the probability of heads. We first need to set a prior distribution for $\tilde{P}$. The beta family $Beta(\alpha, \beta)$ of distributions takes values in the interval $[0, 1]$ and provides a convenient prior density π. Thus,*

$$\pi(p) = c_{\alpha,\beta}p^{(\alpha-1)}(1-p)^{(\beta-1)}, \quad 0 < p < 1.$$

Any density on the interval $[0, 1]$ that can be written a a power of p times a power of $1 - p$ times a constant chosen so that

$$1 = \int_0^1 \pi(p)\, dp$$

is a member of the beta family. This distribution has

$$\text{mean } \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \text{variance } \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{12.6}$$

Thus, the mean is the ratio of α and $\alpha + \beta$. If the two parameters are each multiplied by a factor of k, then the mean does not change. However, the variance is reduced by a factor close to k. The prior gives a sense of our prior knowledge of the mean through the ratio of α to $\alpha + \beta$ and our uncertainly through the size of α and β

If we perform n Bernoulli trials, $\mathbf{x} = (x_1, \ldots, x_n)$, then the joint density

$$\mathbf{f}_X(\mathbf{x}|p) = p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k}.$$

Thus the posterior distribution of the parameter $\tilde{P}$ given the data $\mathbf{x}$, using (12.5), we have.

$$\begin{aligned} f_{\tilde{P}|X}(p|\mathbf{x}) \propto \mathbf{f}_{X|\tilde{P}}(\mathbf{x}|p)\pi(p) &= p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k} \cdot c_{\alpha,\beta}p^{(\alpha-1)}(1-p)^{(\beta-1)}. \\ &= c_{\alpha,\beta}p^{\alpha+\sum_{k=1}^n x_k-1}(1-p)^{\beta+n-\sum_{k=1}^n x_k-1}. \end{aligned}$$

Consequently, the posterior distribution is also from the beta family with parameters

$$\alpha + \sum_{k=1}^n x_k \quad \text{and} \quad \beta + n - \sum_{k=1}^n x_k = \beta + \sum_{k=1}^n (1 - x_k).$$

$$\alpha + \#\text{ successes} \quad \text{and} \quad \beta + \#\text{ failures}.$$

Notice that the posterior mean can be written as

$$\begin{aligned}\frac{\alpha + \sum_{k=1}^n x_k}{\alpha+\beta+n} &= \frac{\alpha}{\alpha+\beta+n} + \frac{\sum_{k=1}^n x_k}{\alpha+\beta+n} \\ &= \frac{\alpha}{\alpha+\beta}\cdot\frac{\alpha+\beta}{\alpha+\beta+n} + \frac{1}{n}\sum_{k=1}^n x_k \cdot \frac{n}{\alpha+\beta+n} \\ &= \frac{\alpha}{\alpha+\beta}\cdot\frac{\alpha+\beta}{\alpha+\beta+n} + \bar{x}\cdot\frac{n}{\alpha+\beta+n}.\end{aligned}$$

This expression allow us to see that the posterior mean can be expresses as a weighted average $\alpha/(\alpha+\beta)$ from the prior mean and $\bar{x}$, the sample mean from the data. The relative weights are

$$\alpha+\beta \text{ from the prior} \quad \text{and} \quad n, \text{ the number of observations.}$$

Thus, if the number of observations n is small compared to $\alpha+\beta$, then most of the weight is placed on the prior mean $\alpha/(\alpha+\beta)$. As the number of observations n increase, then

$$\frac{n}{\alpha+\beta+n}$$

increases towards 1. The weight result in a shift the posterior mean away from the prior mean and towards the sample mean $\bar{x}$.

This brings forward two central issues in the use of the Bayesian approach to estimation.

- *If the number of observations is small, then the estimate relies heavily on the quality of the choice of the prior distribution π. Thus, an unreliable choice for π leads to an unreliable estimate.*
- *As the number of observations increases, the estimate relies less and less on the prior distribution. In this circumstance, the prior may simply be playing the roll of a catalyst that allows the machinery of the Bayesian methodology to proceed.*

Exercise 12.11. *Show that this answer is equivalent to having α heads and β tails in the data set before actually flipping coins.*

Example 12.12. *If we flip a coin $n = 14$ times with 8 heads, then the classical estimate of the success probability p is 8/14=4/7. For a Bayesian analysis with a beta prior distribution, using (12.6) we have a beta posterior distribution with the following parameters.*

prior				data		posterior			
α	β	mean	variance	heads	tails	α	β	mean	variance
6	6	1/2	1/52=0.0192	8	6	14	12	14/(12+14)=7/13	168/18542=0.0092
9	3	3/4	3/208=0.0144	8	6	17	9	17/(17+9) =17/26	153/18252=0.0083
3	9	1/4	3/208=0.0144	8	6	11	15	11/(15+11)=11/26	165/18542=0.0090

In his original example, Bayes chose was the uniform distribution ($\alpha = \beta = 1$) for his prior. In this case the posterior mean is

$$\frac{1}{2+n}\left(1 + \sum_{k=1}^n x_k\right).$$

For the example above

prior				data		posterior			
α	β	mean	variance	heads	tails	α	β	mean	variance
1	1	1/2	1/12=0.813	8	6	9	7	9/(9+7)=9/16	63/4352=0.0144

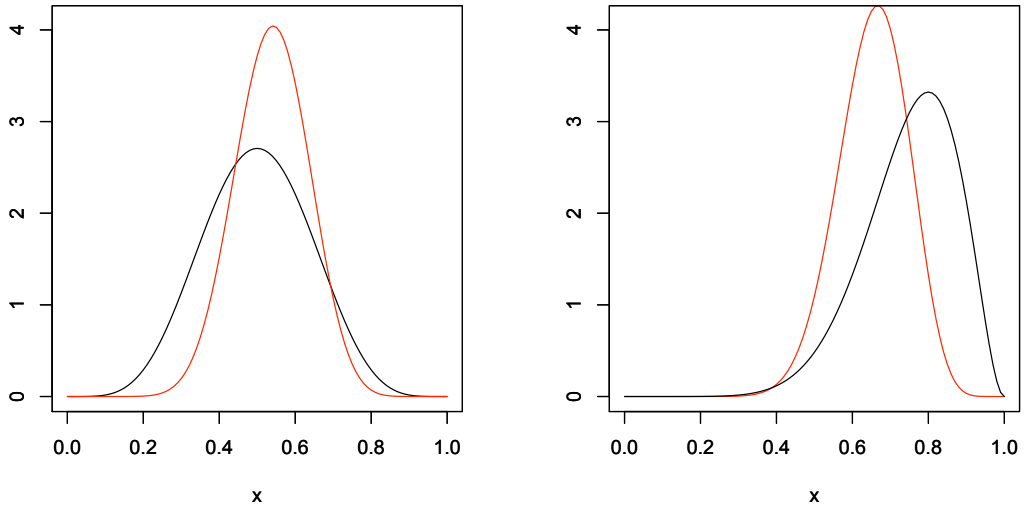


Figure 12.2: Example of prior (black) and posterior (red) densities based on 14 coin flips, 8 heads and 6 tails. Left panel: Prior is $Beta(6,6)$, Right panel: Prior is $Beta(9,3)$. Note how the peak is narrowed. This shows that the posterior variance is smaller than the prior variance. In addition, the peal moves from the prior towards $\hat{p} = 4/7$, the sample proportion of the number of heads.

Reliability engineering emphasizes dependability of a product by assessing the ability of a system or component to function. We introduce the Bayesian perspective to reliability through a simple devise. Our analysis is based on an extension the ideas of Bernoulli trials example above.

Example 12.13. *A devise consists of two independent units. Let A_i, $i = 1, 2$ be the event that the i-th unit is operating appropriately and define the probability $p_i = P(A_i)$. Then, the devise works with probability $p_1 p_2 = P(A_1 \cap A_2)$. For each unit, we place independent $Beta(\alpha_i, \beta_i)$ prior distributions. Next we test n_i units of type i. Repeating the steps in a previous exercise, we find that y_i units are functioning, then the posterior distribution are also in the beta family,*

$$Beta(\alpha_1 + y_1, \beta_1 + n_1 - y_1) \quad \textit{and} \quad Beta(\alpha_2 + y_2, \beta_2 + n_2 - y_2),$$

respectively. This results in a joint posterior density

$$f_{P_1,P_2|Y_1,Y_2}(p_1, p_2|y_1, y_2) = c(\alpha_1, \beta_1, n_1)c(\alpha_2, \beta_2, n_2)\, p_1^{\alpha_1+x_1}(1-p_1)^{\beta_1+n_1-y_1} \cdot p_2^{\alpha_2+x_2}(1-p_2)^{\beta_1+n_2-y_2}.$$

To find the posterior distribution of $p = p_1 p_2$ that the devise functions, we integrate to find the cumulative distribution function.

$$F_{P|Y_1,Y_2}(p|y_1, y_2) = \int\int_{\{p_1 p_2 \le p\}} f_{P_1,P_2|Y_1,Y_2}(p_1, p_2|y_1, y_2)\, dp_2 dp_1.$$

We can also simulate using the `rbeta` *command to estimate values for this distribution functions.*

To provide a concrete example, assume a uniform prior ($\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1$) and test twenty units of each type ($n_1 = n_2 = 20$). If 15 and 17 of the devises work ($y_1 = 15, y_2 = 17$), then the posteriors distributions are

$$Beta(16, 6) \quad \textit{and} \quad Beta(18, 4),$$

We simulate this in R *to find the distribution of the posterior probability of $p = p_1 p_2$.*

```
> p1<-rbeta(10000,16,6);p2<-rbeta(10000,18,4)
> p<-p1*p2
```

We then give a table of deciles for the posterior distribution function and present a histogram.

```
> data.frame(quantile(p,d))
     quantile.p..d.
0%        0.2825593
10%       0.4660896
20%       0.5094321
30%       0.5422747
40%       0.5712765
50%       0.5968341
60%       0.6209610
70%       0.6477835
80%       0.6776208
90%       0.7187307
100%      0.9234834
> hist(p)
```

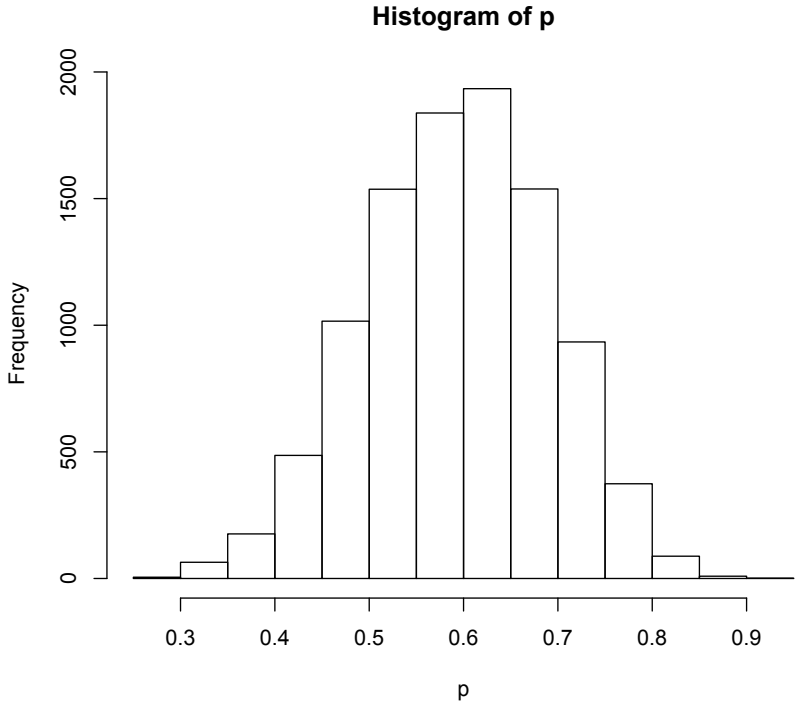


Figure 12.3: Histogram of simulated posterior distribution of the reliability $p = p_1 p_2$ where p_1 and p_2 have independent beta distributions based on the prior distributions and the data.

The posterior density $f_{P|Y_1,Y_2}(p|y_1, y_2)$ is non-negative throughout the interval from 0 to 1, but is very samll for values near 0 and 1. Indeed, none of the 10,000 simulations give a posterior probability below 0.282 or above 0.923. We could take the mean of the simulated sample as a point estimate $\hat{p}$ for p.

```
> mean(p);sd(p)
[1] 0.5935356
[1] 0.09661065
```

This is very close to the means from the beta distributions.

$$E\hat{p} = Ep_1p_2 = Ep_1Ep_2 = \frac{16}{22} \cdot \frac{18}{22} = \frac{72}{121} = 0.595.$$

Exercise 12.14. *The simulation variance is also indicated. Compare this answer with the answer given by the delta method.*

Example 12.15. *Suppose that the prior density is a normal random variable with mean θ_0 and variance $1/\lambda$. This way of giving the variance may seem unusual, but we will see that λ is a measure of* **information**. *Thus, low variance means high information. Our data $\mathbf{x}$ are a realization of independent normal random variables with unknown mean θ. We shall choose the variance to be 1 to set a scale for the size of the variation in the measurements that yield the data $\mathbf{x}$. We will present this example omitting some of the algebraic steps to focus on the central ideas.*

The prior density is

$$\pi(\theta) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(\theta - \theta_0)^2\right).$$

We rewrite the density for the data to empathize the difference between the parameter θ for the mean and the $\bar{x}$, the sample mean.

$$\begin{aligned} f_{X|\Theta}(\mathbf{x}|\theta) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{n}{2}(\theta - \bar{x})^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right). \end{aligned}$$

The posterior density is proportional to the product $f_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)$, Becsuse the posterior is a function of θ, we need only keep track of the terms which involve θ. Consequently, we write the posterior density as

$$\begin{aligned} f_{\Theta|X}(\theta|\mathbf{x}) &= c(\mathbf{x}) \exp\left(-\frac{1}{2}(n(\theta - \bar{x})^2 + \lambda(\theta - \theta_0)^2)\right) \\ &= \tilde{c}(\mathbf{x}) \exp(-\frac{n+\lambda}{2}(\theta - \theta_1(\mathbf{x}))^2). \end{aligned}$$

where

$$\theta_1(\mathbf{x}) = \frac{\lambda}{\lambda + n}\theta_0 + \frac{n}{\lambda + n}\bar{x}. \tag{12.7}$$

Notice that the posterior distribution is normal with mean $\theta_1(\mathbf{x})$ that results from the weighted average with relative weights

$$\lambda \quad \textit{from the information from the prior} \quad \textit{and} \quad n \quad \textit{from the data.}$$

The variance in inversely proportional to the total information $\lambda + n$. Thus, if n is small compared to λ, then $\theta_1(\mathbf{x})$ is near θ_0. If n is large compared to λ, $\theta_1(\mathbf{x})$ is near $\bar{x}$.

Exercise 12.16. *Fill in the steps in the derivation of the posterior density in the example above.*

For these two examples, we see that the prior distribution and the posterior distribution are members of the same parameterized family of distributions, namely the beta family and the normal family. In these two cases, we say that the prior density and the density of the data form a **conjugate pair**. In the case of coin tosses, we find that the beta and the Bernoulli families form a conjugate pair. In Example 12.11, we learn that the normal density is conjugate to itself.

Typically, the computation of the posterior density is much more computationally intensive that what was shown in the two examples above. The choice of conjugate pairs is enticing because the posterior density is a determined from a simple algebraic computation.

Bayesian statistics is seeing increasing use in the sciences, including the life sciences, as we see the explosive increase in the amount of data. For example, using a classical approach, mutation rates estimated from genetic sequence data are, due to the paucity of mutation events, often not very precise. However, we now have many data sets that can be synthesized to create a prior distribution for mutation rates and will lead to estimates for this and other parameters of interest that will have much smaller variance than under the classical approach.

Exercise 12.17. *Show that the gamma family of distributions is a conjugate prior for the Poisson family of distributions. Give the posterior mean based on n observations.*

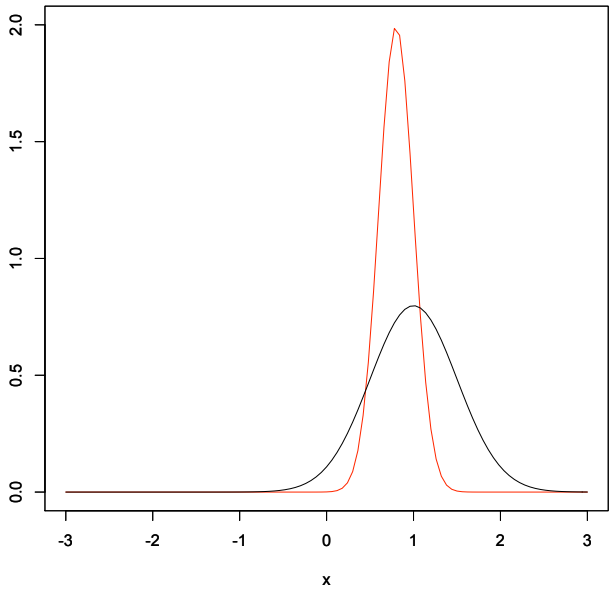


Figure 12.4: Example of prior (black) and posterior (red) densities for a normal prior distribution and normally distributed data. In this figure the prior density is $N(1, 1/2)$. Thus, $\theta_0 = 1$ and $\lambda = 2$. Here the data consist of 3 observations having sample mean $\bar{x} = 2/3$. Thus, the posterior mean from equation (12.7) is $\theta_1(\mathbf{x}) = 4/5$ and the variance is 1/(2+3) = 1/5.

12.4 Answers to Selected Exercises

12.5. Double the average, $2\bar{X}$. Take the maximum value of the data, $\max_{1\le i\le n} x_i$. Double the difference of the maximum and the minimum, $2(\max_{1\le i\le n} x_i - \min_{1\le i\le n} x_i)$.

12.7. The density of a gamma random variable

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Thus, for n observations

$$\begin{aligned} L(\theta|\mathbf{x}) &= f(x_1|\alpha,\beta) f(x_2|\alpha,\beta) \cdots f(x_n|\alpha,\beta) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} x_1^{\alpha-1} e^{-\beta x_1} \frac{\beta^\alpha}{\Gamma(\alpha)} x_2^{\alpha-1} e^{-\beta x_2} \cdots \frac{\beta^\alpha}{\Gamma(\alpha)} x_n^{\alpha-1} e^{-\beta x_n} \\ &= \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} (x_1 x_2 \cdots x_n)^{\alpha-1} e^{-\beta(x_1+x_2+\cdots+x_n)} \end{aligned}$$

12.9. We need to verify that the density $\mathbf{f}$ is a non-negative function and that the integral over the sample space is 1. Note that both $\mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)$ and $\pi(\psi)$ and thus their product is positive. Consequently,

$$\mathbf{f}_X(x) = \int_\Theta \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\, d\psi \ge 0 \quad \text{for all } x.$$

Next, we reverse the order of the double integral,

$$\int_{\mathbb{R}^n} \mathbf{f}_X(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^n} \left(\int_\Theta \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\, d\psi \right) dx = \int_\Theta \left(\int_{\mathbb{R}^n} \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\, dx \right) \pi(\psi) d\psi.$$

Because $\mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)$, is a density function, the integral inside the parentheses is 1. Now use the fact that π is a probability density,

$$\int_{\mathbb{R}^n} \mathbf{f}_X(\mathbf{x})d\mathbf{x} = \int_\Theta \pi(\psi)d\psi = 1.$$

12.10. In this case the total number of observations is $\alpha + \beta + n$ and the total number of successes is $\alpha + \sum_{i=1}^n x_i$. Their ratio is the posterior mean.

12.13. To include some of the details in the computation, we first add and subtract $\bar{x}$ in the sum for the joint density,

$$f_{X|\Theta}(\mathbf{x}|\theta) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n ((x_i - \bar{x}) + (\bar{x} - \theta))^2\right)$$

Then we expand the square in the sum to obtain

$$\begin{aligned}\sum_{i=1}^n ((x_i - \bar{x}) + (\bar{x} - \theta))^2 &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2\left(\sum_{i=1}^n (x_i - \bar{x})\right)(\bar{x} - \theta) + \sum_{i=1}^n (\bar{x} - \theta)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 0 + n(\bar{x} - \theta)^2\end{aligned}$$

This gives the joint density

$$f_{X|\Theta}(\mathbf{x}|\theta) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{n}{2}(\theta - \bar{x})^2 - \frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2\right).$$

The posterior density is

$$\begin{aligned}f_{\Theta|X}(\theta|\mathbf{x}) &= c(\mathbf{x}) f_{X|\Theta}(\mathbf{x}|\theta) \cdot f_\Theta(\theta) \\ &= c(\mathbf{x}) \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{n}{2}(\theta - \bar{x})^2 - \frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2\right) \cdot \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(\theta - \theta_0)^2\right) \\ &= \left(c(\mathbf{x}) \frac{1}{(2\pi)^{n/2}} \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2\right)\right) \exp\left(-\frac{1}{2}(n(\theta - \bar{x})^2 + \lambda(\theta - \theta_0)^2)\right) \\ &= c_1(\mathbf{x}) \exp\left(-\frac{1}{2}(n(\theta - \bar{x})^2 + \lambda(\theta - \theta_0)^2)\right).\end{aligned}$$

Here $c_1(\mathbf{x})$ is the function of $\mathbf{x}$ in parenthesis. We now expand the expressions in the exponent,

$$\begin{aligned}n(\theta - \bar{x})^2 + \lambda(\theta - \theta_0)^2 &= (n\theta^2 - 2n\bar{x}\theta + n\bar{x}^2) + (\lambda\theta^2 - 2\lambda\theta_0\theta + \lambda\theta_0^2) \\ &= (n + \lambda)\theta^2 - 2(n\bar{x} + \lambda\theta_0)\theta + (n\bar{x}^2 + \lambda\theta_0^2) \\ &= (n + \lambda)\left(\theta^2 - 2\frac{n\bar{x} + \lambda\theta_0}{n + \lambda}\theta\right) + (n\bar{x}^2 + \lambda\theta_0^2) \\ &= (n + \lambda)\left(\theta^2 - 2\theta_1(\mathbf{x})\theta + \theta_1(\mathbf{x})^2\right) - (n + \lambda)\theta_1(\mathbf{x})^2 + (n\bar{x}^2 + \lambda\theta_0^2) \\ &= (n + \lambda)(\theta - \theta_1(\mathbf{x}))^2 - (n + \lambda)\theta_1(\mathbf{x})^2 + (n\bar{x}^2 + \lambda\theta_0^2)\end{aligned}$$

using the definition of $\theta_1(\mathbf{x})$ in (12.7) and completing the square.

$$\begin{aligned}f_{\Theta|X}(\theta|\mathbf{x}) &= c_1(\mathbf{x}) \exp\left(-\frac{1}{2}((n\bar{x}^2 + \lambda\theta_0^2) - (n + \lambda)\theta_1(\mathbf{x})^2 + (n + \lambda)(\theta - \theta_1(\mathbf{x}))^2)\right) \\ &= \left(c_1(\mathbf{x}) \exp\left(-\frac{1}{2}((n\bar{x}^2 + \lambda\theta_0^2) - (n + \lambda)\theta(\mathbf{x})^2)\right)\right) \exp(-\frac{n + \lambda}{2}(\theta - \theta_1(\mathbf{x}))^2) \\ &= c_2(\mathbf{x}) \exp(-\frac{n + \lambda}{2}(\theta - \theta_1(\mathbf{x}))^2)\end{aligned}$$

where $c_2(\mathbf{x})$ is the function of $\mathbf{x}$ in parenthesis. This give a posterior density that is normal, mean $\theta_1(\mathbf{x})$ and variance $n + \lambda$.

12.15. Our goal is to estimate the variance of $p = g(p_1, p_2) = p_1 p_2$ using the delta method. Let μ_i, σ_i^2, be the posterior mean and variance for unit $i = 1, 2$, respectively. For this we write

$$\begin{aligned}\sigma^2_{g(\mu_1,\mu_2)} &\approx \frac{\partial g}{\partial p_1}(\mu_1,\mu_2)\sigma_1^2 + \frac{\partial g}{\partial p_2}(\mu_1,\mu_2)\sigma_2^2 = \mu_2\sigma_1^2 + \mu_1\sigma_2^2 \\ &= \frac{16}{22}\cdot\frac{18\cdot 4}{22^2\cdot 23} + \frac{18}{22}\cdot\frac{16\cdot 6}{22^2\cdot 23} = \frac{16\cdot 18}{22^3\cdot 23}(4+6) = \frac{8\cdot 9\cdot 5}{11^3\cdot 23} = 0.01176.\end{aligned}$$

The estimated standard deviation $\sigma^2_{g(\mu_1,\mu_2)} \approx 0.1084$ os about 12% higher than the estimate from the simulation.

12.16. For n observations $x_1, x_2, \ldots, x_n$ of independent Poisson random variables having parameter λ, the joint density is the product of the n marginal densities.

$$\mathbf{f}_X(\mathbf{x}|\lambda) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\cdot\frac{\lambda^{x_2}}{x_2!}e^{-\lambda}\cdots\frac{\lambda^{x_n}}{x_n!}e^{-\lambda} = \frac{1}{x_1!x_2!\cdots x_n!}\lambda^{x_1+x_2+\cdots+x_n}e^{-n\lambda} = \frac{1}{x_1!x_2!\cdots x_n!}\lambda^{n\bar{x}}e^{-n\lambda}.$$

The prior density on λ has a $\Gamma(\alpha,\beta)$ density

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}.$$

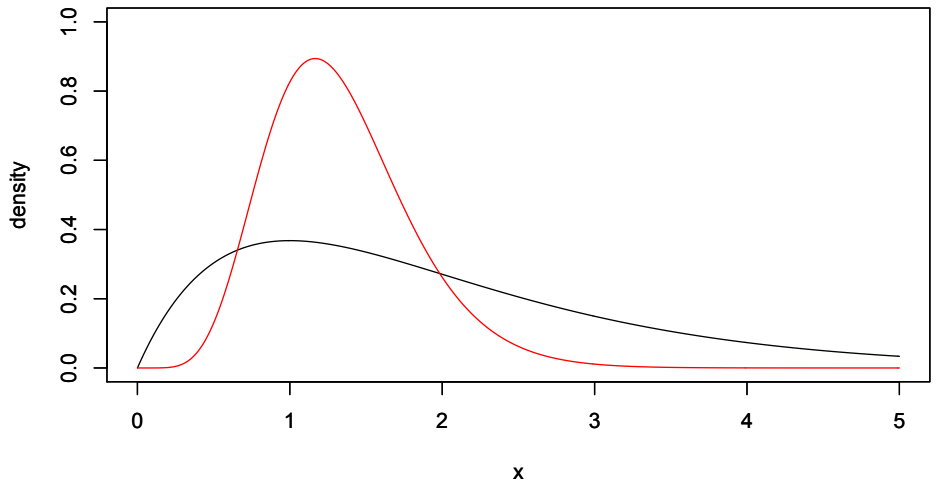
Thus, the posterior density

$$f_{\Lambda|X}(\lambda|\mathbf{x}) = c(\mathbf{x})\lambda^{\alpha-1}e^{-\beta\lambda}\cdot\lambda^{n\bar{x}}e^{-n\lambda} = c(\mathbf{x})\lambda^{\alpha+n\bar{x}-1}e^{-(\beta+n)\lambda}$$

is the density of a $\Gamma(\alpha+n\bar{x}, \beta+n)$ random variable. Its mean can be written as the weighted average

$$\frac{\alpha+n\bar{x}}{\beta+n} = \frac{\alpha}{\beta}\cdot\frac{\beta}{\beta+n} + \bar{x}\cdot\frac{n}{\beta+n}$$

of the prior mean α/β and the sample mean $\bar{x}$. The weights are, respectively, proportional to β and the number of observations n.



The figure above demonstrate the case with a $\Gamma(2,1)$ prior density on λ and a sum $x_1+x_2+x_3+x_4+x_5 = 6$ for 5 values for independent observations of a Poisson random random variable. Thus the posterior has a $\Gamma(2+6, 1+5) = \Gamma(8,6)$ distribution.

Topic 13

Method of Moments

13.1 Introduction

Method of moments estimation is based solely on the law of large numbers, which we repeat here:

Let $M_1, M_2, \ldots$ be independent random variables having a common distribution possessing a mean μ_M. Then the sample means converge to the distributional mean as the number of observations increase.

$$\bar{M}_n = \frac{1}{n}\sum_{i=1}^{n} M_i \to \mu_M \quad \text{as } n \to \infty.$$

To show how the method of moments determines an estimator, we first consider the case of one parameter. We start with independent random variables $X_1, X_2, \ldots$ chosen according to the probability density $f_X(x|\theta)$ associated to an unknown parameter value θ. The common mean of the X_i, μ_X, is a function $k(\theta)$ of θ. For example, if the X_i are continuous random variables, then

$$\mu_X = \int_{-\infty}^{\infty} x f_X(x|\theta)\, dx = k(\theta).$$

The law of large numbers states that

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu_X \quad \text{as } n \to \infty.$$

Thus, if the number of observations n is large, the distributional mean, $\mu = k(\theta)$, should be well approximated by the sample mean, i.e.,

$$\bar{X} \approx k(\theta).$$

This can be turned into an estimator $\hat{\theta}$ by setting

$$\bar{X} = k(\hat{\theta}).$$

and solving for $\hat{\theta}$.

We shall next describe the procedure in the case of a vector of parameters and then give several examples. We shall see that the delta method can be used to estimate the variance of method of moment estimators.

13.2 The Procedure

More generally, for independent random variables $X_1, X_2, \ldots$ chosen according to the probability distribution derived from the parameter value θ and m a real valued function, if $k(\theta) = E_\theta m(X_1)$, then

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i) \to k(\theta) \quad \text{as } n \to \infty.$$

The **method of moments** results from the choices $m(x) = x^m$. Write

$$\mu_m = EX^m = k_m(\theta). \tag{13.1}$$

for the m-th moment.

Our estimation procedure follows from these 4 steps to link the sample moments to parameter estimates.

- **Step 1**. If the model has d parameters, we compute the functions k_m in equation (13.1) for the first d moments,

  $$\mu_1 = k_1(\theta_1, \theta_2 \ldots, \theta_d), \quad \mu_2 = k_2(\theta_1, \theta_2 \ldots, \theta_d), \quad \ldots, \quad \mu_d = k_d(\theta_1, \theta_2 \ldots, \theta_d),$$

  obtaining d equations in d unknowns.

- **Step 2**. We then solve for the d parameters as a function of the moments.

  $$\theta_1 = g_1(\mu_1, \mu_2, \cdots, \mu_d), \quad \theta_2 = g_2(\mu_1, \mu_2, \cdots, \mu_d), \quad \ldots, \quad \theta_d = g_d(\mu_1, \mu_2, \cdots, \mu_d). \tag{13.2}$$

- **Step 3.** Now, based on the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, we compute the first d **sample moments**,

  $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \ldots, \quad \overline{x^d} = \frac{1}{n}\sum_{i=1}^{n} x_i^d.$$

  Using the law of large numbers, we have, for each moment, $m = 1, \ldots, d$, that $\mu_m \approx \overline{x^m}$.

- **Step 4**. We replace the distributional moments μ_m by the sample moments $\overline{x^m}$, then the solutions in (13.2) give us formulas for the **method of moment estimators** $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d)$. For the data $\mathbf{x}$, these estimates are

  $$\hat{\theta}_1(\mathbf{x}) = g_1(\bar{x}, \overline{x^2}, \cdots, \overline{x^d}), \quad \hat{\theta}_2(\mathbf{x}) = g_2(\bar{x}, \overline{x^2}, \cdots, \overline{x^d}), \quad \ldots, \quad \hat{\theta}_d(\mathbf{x}) = g_d(\bar{x}, \overline{x^2}, \cdots, \overline{x^d}).$$

How this abstract description works in practice can be best seen through examples.

13.3 Examples

Example 13.1. *Let $X_1, X_2, \ldots, X_n$ be a simple random sample of Pareto random variables with density*

$$f_X(x|\beta) = \frac{\beta}{x^{\beta+1}}, \quad x > 1.$$

The cumulative distribution function is

$$F_X(x) = 1 - x^{-\beta}, \quad x > 1.$$

The mean and the variance are, respectively,

$$\mu = \frac{\beta}{\beta - 1}, \qquad \sigma^2 = \frac{\beta}{(\beta-1)^2(\beta-2)}.$$

In this situation, we have one parameter, namely β. Thus, in step 1, we will only need to determine the first moment

$$\mu_1 = \mu = k_1(\beta) = \frac{\beta}{\beta - 1}$$

to find the method of moments estimator $\hat{\beta}$ for β.

For step 2, we solve for β as a function of the mean μ.

$$\beta = g_1(\mu) = \frac{\mu}{\mu - 1}.$$

Consequently, a method of moments estimate for β is obtained by replacing the distributional mean μ by the sample mean $\bar{X}$.

$$\hat{\beta} = \frac{\bar{X}}{\bar{X} - 1}.$$

A good estimator should have a small variance . To use the delta method to estimate the variance of $\hat{\beta}$,

$$\sigma^2_{\hat{\beta}} \approx g_1'(\mu)^2 \frac{\sigma^2}{n}.$$

we compute

$$g_1'(\mu) = -\frac{1}{(\mu - 1)^2}, \quad \textit{giving} \quad g_1'\left(\frac{\beta}{\beta - 1}\right) = -\frac{1}{(\frac{\beta}{\beta-1} - 1)^2} = -\frac{(\beta - 1)^2}{(\beta - (\beta - 1))^2} = -(\beta - 1)^2$$

and find that $\hat{\beta}$ has mean approximately equal to β and variance

$$\sigma^2_{\hat{\beta}} \approx g_1'(\mu)^2 \frac{\sigma^2}{n} = (\beta - 1)^4 \frac{\beta}{n(\beta - 1)^2(\beta - 2)} = \frac{\beta(\beta - 1)^2}{n(\beta - 2)}$$

As a example, let's consider the case with $\beta = 3$ and $n = 100$. Then,

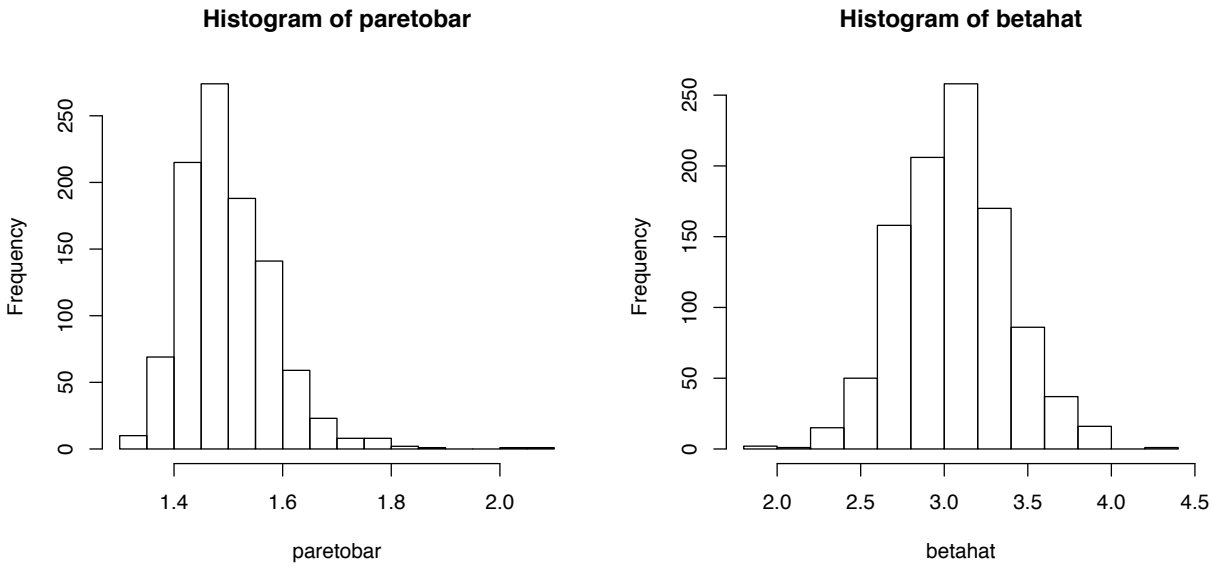
$$\sigma^2_{\hat{\beta}} \approx \frac{3 \cdot 2^2}{100 \cdot 1} = \frac{12}{100} = \frac{3}{25}, \quad \sigma_{\hat{\beta}} \approx \frac{\sqrt{3}}{5} = 0.346.$$

To simulate this, we first need to simulate Pareto random variables. Recall that the probability transform states that if the X_i are independent Pareto random variables, then $U_i = F_X(X_i)$ are independent uniform random variables on the interval $[0, 1]$. Thus, we can simulate X_i with $F_X^{-1}(U_i)$. If

$$u = F_X(x) = 1 - x^{-3}, \quad \textit{then} \quad x = (1 - u)^{-1/3} = v^{-1/3}, \quad \textit{where } v = 1 - u.$$

Note that if U_i are uniform random variables on the interval $[0, 1]$ then so are $V_i = 1 - U_i$. Consequently, $1/\sqrt[\beta]{V_1}, 1/\sqrt[\beta]{V_2}, \cdots$ have the appropriate Pareto distribution. ..

```
> paretobar<-numeric(1000)
> for (i in 1:1000){v<-runif(100);pareto<-1/v^(1/3);paretobar[i]<-mean(pareto)}
> hist(paretobar)
> betahat<-paretobar/(paretobar-1)
> hist(betahat)
> mean(betahat)
[1] 3.053254
> sd(betahat)
[1] 0.3200865
```



The sample mean for the estimate for β at 3.053 is close to the simulated value of 3. In this example, the estimator $\hat{\beta}$ is **biased upward**, *In other words, on average the estimate is greater than the parameter, i. e., $E_\beta\hat{\beta} > \beta$. The sample standard deviation value of 0.320 is close to the value 0.346 estimated by the delta method. When we examine unbiased estimators, we will learn that this bias could have been anticipated.*

Exercise 13.2. *The* **muon** *is an elementary particle with an electric charge of -1 and a* **spin** *(an intrinsic angular momentum) of 1/2. It is an unstable subatomic particle with a mean lifetime of 2.2 μs. Muons have a mass of about 200 times the mass of an electron. Since the muon's charge and spin are the same as the electron, a muon can be viewed as a much heavier version of the electron. The collision of an accelerated proton (p) beam having energy 600 MeV (million electron volts) with the nuclei of a production target produces positive pions (π^+) under one of two possible reactions.*

$$p + p \to p + n + \pi^+ \quad \textit{or} \quad p + n \to n + n + \pi^+$$

From the subsequent decay of the pions (mean lifetime 26.03 ns), positive muons (μ^+), are formed via the two body decay

$$\pi^+ \to \mu^+ + \nu_\mu$$

where ν_μ is the symbol of a **muon neutrino**. *The decay of a muon into a* **positron** *(e^+), an* **electron neutrino** *(ν_e), and a* **muon antineutrino** *($\bar{\nu}_\mu$)*

$$\mu^+ \to e^+ + \nu_e + \bar{\nu}_\mu$$

has a distribution angle t with density given by

$$f(t|\alpha) = \frac{1}{2\pi}(1 + \alpha\cos t), \quad 0 \le t \le 2\pi,$$

with t the angle between the positron trajectory and the μ^+-spin and **anisometry parameter** *$\alpha \in [-1/3, 1/3]$ depends the polarization of the muon beam and positron energy. Based on the measurement $t_1, \dots t_n$, give the method of moments estimate $\hat{\alpha}$ for α. (*Note: *In this case the mean is 0 for all values of α, so we will have to compute the second moment to obtain an estimator.)*

Example 13.3 (Lincoln-Peterson method of mark and recapture)**.** *The size of an animal population in a habitat of interest is an important question in conservation biology. However, because individuals are often too difficult to find,*

a census is not feasible. One estimation technique is to capture some of the animals, mark them and release them back into the wild to mix randomly with the population.

Some time later, a second capture from the population is made. In this case, some of the animals were not in the first capture and some, which are tagged, are recaptured. Let

- t *be the number captured and tagged,*
- k *be the number in the second capture,*
- r *be the number in the second capture that are tagged, and let*
- N *be the total population size.*

Thus, t *and* k *is under the control of the experimenter. The value of* r *is random and the populations size* N *is the parameter to be estimated. We will use a method of moments strategy to estimate* N*. First, note that we can guess the the estimate of* N *by considering two proportions.*

$$\text{the proportion of the tagged fish in the second capture} \approx \text{the proportion of tagged fish in the population}$$
$$\frac{r}{k} \approx \frac{t}{N}$$

This can be solved for N *to find* $N \approx kt/r$*. The advantage of obtaining this as a method of moments estimator is that we evaluate the precision of this estimator by determining, for example, its variance. To begin, let*

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th individual in the second capture has a tag.} \\ 0 & \text{if the } i\text{-th individual in the second capture does not have a tag.} \end{cases}$$

The X_i *are Bernoulli random variables with success probability*

$$P\{X_i = 1\} = \frac{t}{N}.$$

They are not Bernoulli trials because the outcomes are not independent. We are sampling **without replacement**. *For example,*

$$P\{\text{the second individual is tagged}|\text{first individual is tagged}\} = \frac{t-1}{N-1}.$$

In words, we are saying that the probability model behind mark and recapture is one where the number recaptured is random and follows a **hypergeometric distribution**. *The number of tagged individuals is* $X = X_1 + X_2 + \cdots + X_k$ *and the expected number of tagged individuals is*

$$\mu = EX = EX_1 + EX_2 + \cdots + EX_k = \frac{t}{N} + \frac{t}{N} + \cdots + \frac{t}{N} = \frac{kt}{N}.$$

The proportion of tagged individuals, $\bar{X} = (X_1 + \cdots + X_k)/k$*, has expected value*

$$E\bar{X} = \frac{\mu}{k} = \frac{t}{N}.$$

Thus,

$$N = \frac{kt}{\mu}.$$

Now in this case, we are estimating μ*, the mean number recaptured with* r*, the actual number recaptured. So, to obtain the estimate* $\hat{N}$*. we replace* μ *with the previous equation by* r*.*

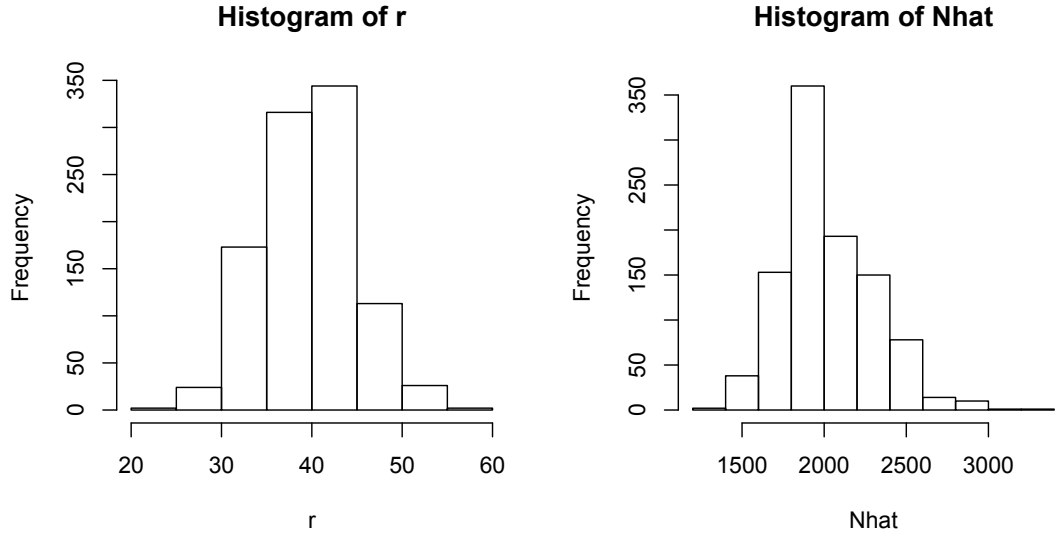
$$\hat{N} = \frac{kt}{r}$$

To simulate mark and capture, consider a population of 2000 fish, tag 200, and capture 400. We perform 1000 simulations of this experimental design. (The R *command* `rep(x,n)` *repeats* `n` *times the value* `x`*.)....*

```
> r<-numeric1000)
> fish<-c(rep(1,200),rep(0,1800))
> for (j in 1:1000){r[j]<-sum(sample(fish,400))}
> Nhat<-200*400/r
```

The command `sample(fish,400)` *creates a vector of length 400 of zeros and ones for, respectively, untagged and tagged fish. Thus, the* `sum` *command gives the number of tagged fish in the simulation. This is repeated 1000 times and stored in the vector* `r`*. Let's look a summaries of* `r` *and the estimates* $\hat{N}$ *of the population.*

```
> mean(r)
[1] 40.09
> sd(r)
[1] 5.245705
> mean(Nhat)
[1] 2031.031
> sd(Nhat)
[1] 276.6233
```



To estimate the population of pink salmon in Deep Cove Creek in southeastern Alaska, 1709 fish were tagged. Of the 6375 carcasses that were examined, 138 were tagged. The estimate for the population size

$$\hat{N} = \frac{6375 \times 1709}{138} \approx 78948.$$

Exercise 13.4. *Use the delta method to estimate Var*$(\hat{N})$ *and* $\sigma_{\hat{N}}$*. Apply this to the simulated sample and to the Deep Cove Creek data.*

Example 13.5. *Fitness is a central concept in the theory of evolution. Relative fitness is quantified as the average number of surviving progeny of a particular genotype compared with average number of surviving progeny of competing genotypes after a single generation. Consequently, the distribution of fitness effects, that is, the distribution of fitness for newly arising mutations is a basic question in evolution. A basic understanding of the distribution of fitness effects is still in its early stages. Eyre-Walker (2006) examined one particular distribution of fitness effects, namely, deleterious amino acid changing mutations in humans. His approach used a gamma-family of random variables and gave the estimate of* $\hat{\alpha} = 0.23$ *and* $\hat{\beta} = 5.35$*.*

A $\Gamma(\alpha, \beta)$ random variable has mean α/β and variance α/β^2. Because we have two parameters, the method of moments methodology requires us, in step 1, to determine the first two moments.

$$E_{(\alpha,\beta)}X_1 = \frac{\alpha}{\beta} \quad \textit{and} \quad E_{(\alpha,\beta)}X_1^2 = \textit{Var}_{(\alpha,\beta)}(X_1) + E_{(\alpha,\beta)}[X_1]^2 = \frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha(1+\alpha)}{\beta^2} = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2}.$$

Thus, for step 1, we find that

$$\mu_1 = k_1(\alpha, \beta) = \frac{\alpha}{\beta}, \quad \mu_2 = k_2(\alpha, \beta) = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2}.$$

For step 2, we solve for α and β. Note that

$$\mu_2 - \mu_1^2 = \frac{\alpha}{\beta^2},$$

$$\frac{\mu_1}{\mu_2 - \mu_1^2} = \frac{\alpha/\beta}{\alpha/\beta^2} = \beta,$$

and

$$\mu_1 \cdot \frac{\mu_1}{\mu_2 - \mu_1^2} = \frac{\alpha}{\beta} \cdot \beta = \alpha, \quad \textit{or} \quad \alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

So set

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{and} \quad \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

to obtain estimators

$$\hat{\beta} = \frac{\bar{X}}{\overline{X^2} - (\bar{X})^2} \quad \textit{and} \quad \hat{\alpha} = \hat{\beta}\bar{X} = \frac{(\bar{X})^2}{\overline{X^2} - (\bar{X})^2}.$$

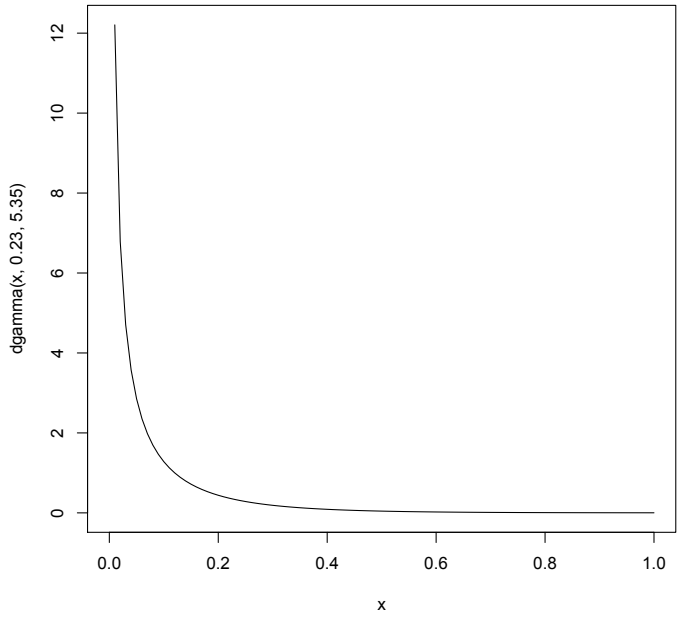


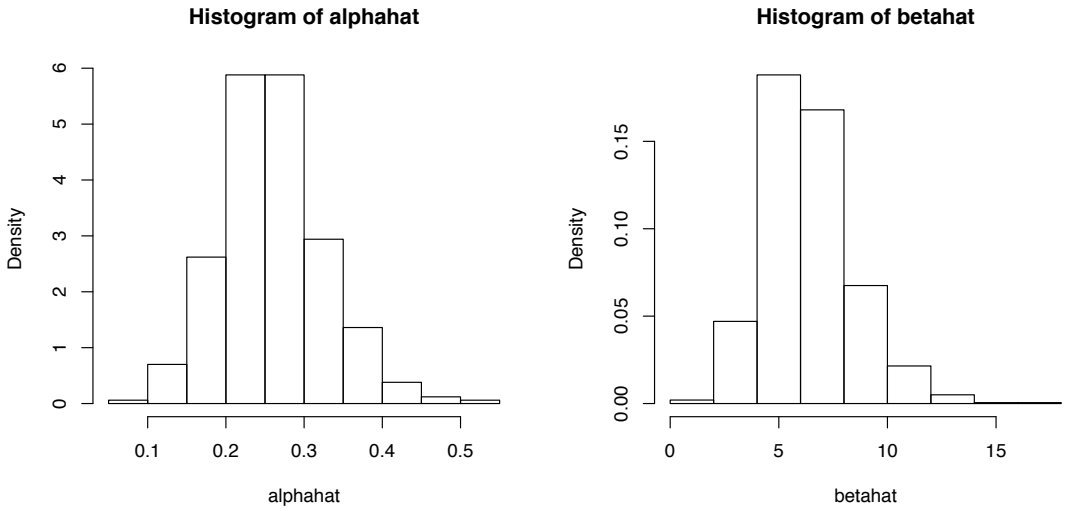
Figure 13.1: The density of a $\Gamma(0.23, 5.35)$ random variable.

To investigate the method of moments on simulated data using R*, we consider 1000 repetitions of 100 independent observations of a* $\Gamma(0.23, 5.35)$ *random variable.*

```
> xbar <- numeric(1000)
> x2bar <- numeric(1000)
> for (i in 1:1000){x<-rgamma(100,0.23,5.35);xbar[i]<-mean(x);x2bar[i]<-mean(x^2)}
> betahat <- xbar/(x2bar-(xbar)^2)
> alphahat <- betahat*xbar
> mean(alphahat)
[1] 0.2599894
> sd(alphahat)
[1] 0.06672909
> mean(betahat)
[1] 6.315644
> sd(betahat)
[1] 2.203887
```

To obtain a sense of the distribution of the estimators $\hat{\alpha}$ *and* $\hat{\beta}$*, we give histograms.*

```
> hist(alphahat,probability=TRUE)
> hist(betahat,probability=TRUE)
```



As we see, the variance in the estimate of β *is quite large. We will revisit this example using maximum likelihood estimation in the hopes of reducing this variance. The use of the delta method is more difficult in this case because it must take into account the correlation between* $\bar{X}$ *and* $\overline{X^2}$ *for independent gamma random variables. Indeed, from the simulation, we have an estimate..*

```
> cor(xbar,x2bar)
[1] 0.8120864
```

Moreover, the two estimators $\hat{\alpha}$ *and* $\hat{\beta}$ *are fairly strongly positively correlated. Again, we can estimate this from the simulation.*

```
> cor(alphahat,betahat)
[1] 0.7606326
```

In particular, an estimate of $\hat{\alpha}$ *and* $\hat{\beta}$ *are likely to be overestimates or underestimates in tandem.*

13.4 Answers to Selected Exercises

13.2. Let T be the random variable that is the angle between the positron trajectory and the μ^+-spin

$$\mu_2 = E_\alpha T^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} t^2(1+\alpha\cos t)dt = \frac{\pi^2}{3} - 2\alpha$$

Thus, $\alpha = (\mu_2 - \pi^2/3)/2$. This leads to the method of moments estimate

$$\hat{\alpha} = \frac{1}{2}\left(\overline{t^2} - \frac{\pi^2}{3}\right)$$

where $\overline{t^2}$ is the sample mean of the square of the observations.

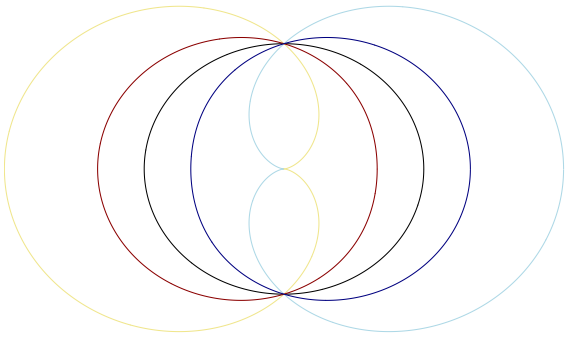


Figure 13.2: Densities $f(t|\alpha)$ for the values of $\alpha = -1$ (yellow). $-1/3$ (red), 0 (black), 1/3 (blue), 1 (light blue).

13.4. Let X be the random variable for the number of tagged fish. Then, X is a hypergeometric random variable with

$$\text{mean } \mu_X = \frac{kt}{N} \quad \text{and variance} \quad \sigma_X^2 = k\frac{t}{N}\frac{N-t}{N}\frac{N-k}{N-1}$$

$$N = g(\mu_X) = \frac{kt}{\mu_X}. \quad \text{Thus, } g'(\mu_X) = -\frac{kt}{\mu_X^2}.$$

The variance of $\hat{N}$

$$\begin{aligned}\text{Var}(\hat{N}) \approx g'(\mu)^2\sigma_X^2 &= \left(\frac{kt}{\mu_X^2}\right)^2 k\frac{t}{N}\frac{N-t}{N}\frac{N-k}{N-1} = \left(\frac{kt}{\mu_X^2}\right)^2 k\frac{t}{kt/\mu_X}\frac{kt/\mu_X - t}{kt/\mu_X}\frac{kt/\mu_X - k}{kt/\mu_X - 1}\\ &= \left(\frac{kt}{\mu_X^2}\right)^2 k\frac{\mu_X t}{kt}\frac{kt-\mu_X t}{kt}\frac{kt-k\mu_X}{kt-\mu_X} = \left(\frac{kt}{\mu_X^2}\right)^2 k\frac{\mu_X}{k}\frac{k-\mu_X}{k}\frac{k(t-\mu_X)}{kt-\mu_X}\\ &= \frac{k^2t^2}{\mu_X^3}\frac{(k-\mu_X)(t-\mu_X)}{kt-\mu_X}\end{aligned}$$

Now if we replace μ_X by its estimate r we obtain

$$\sigma_{\hat{N}}^2 \approx \frac{k^2t^2}{r^3}\frac{(k-r)(t-r)}{kt-r}.$$

For $t = 200, k = 400$ and $r = 40$, we have the estimate $\sigma_{\hat{N}} = 268.4$. This compares to the estimate of 276.6 from simulation.

For $t = 1709, k = 6375$ and $r = 138$, we have the estimate $\sigma_{\hat{N}} = 6373.4$.

Topic 14

Unbiased Estimation

14.1 Introduction

In creating a parameter estimator, a fundamental question is whether or not the estimator differs from the parameter in a systematic manner. Let's examine this by looking a the computation of the mean and the variance of 16 flips of a fair coin.

Give this task to 10 individuals and ask them report the number of heads. We can simulate this in R as follows .

```
> (x<-rbinom(10,16,0.5))
 [1]   8   5   9   7   7   9   7   8   8 10
```

Our estimate is obtained by taking these 10 answers and averaging them. Intuitively we anticipate an answer around 8. For these 10 observations, we find, in this case, that

```
> sum(x)/10
[1] 7.8
```

The result is a bit below 8. Is this systematic? To assess this, we appeal to the ideas behind Monte Carlo to twice perform a 1000 simulations of the example above.

```
> meanx<-replicate(1000,mean(rbinom(10,16,0.5)))
> mean(meanx)
[1] 7.9799
> meanx<-replicate(1000,mean(rbinom(10,16,0.5)))
> mean(meanx)
[1] 8.0049
```

From this, we surmise that we the estimate of the sample mean $\bar{x}$ neither systematically overestimates or underestimates the distributional mean. From our knowledge of the binomial distribution, we know that the mean $\mu = np = 16 \cdot 0.5 = 8$. In addition, the sample mean $\bar{X}$ also has mean

$$E\bar{X} = \frac{1}{10}(8+8+8+8+8+8+8+8+8+8) = \frac{80}{10} = 8$$

verifying that we have no systematic error.

The phrase that we use is that the sample mean $\bar{X}$ is an **unbiased** estimator of the distributional mean μ. Here is the precise definition.

Definition 14.1. *For observations $X = (X_1, X_2, \ldots, X_n)$ based on a distribution having parameter value θ, and for $d(X)$ an estimator for $k(\theta)$, the* **bias** *is the mean of the difference $d(X) - k(\theta)$, i.e.,*

$$b_d(\theta) = E_\theta d(X) - k(\theta). \tag{14.1}$$

If $b_d(\theta) = 0$ for all values of the parameter, then $d(X)$ is called an **unbiased estimator**. *Any estimator that is not unbiased is called* **biased**.

Example 14.2. *Let $X_1, X_2, \ldots, X_n$ be Bernoulli trials with success parameter p and set the estimator for p to be $d(X) = \bar{X}$, the sample mean. Then,*

$$E_p\bar{X} = \frac{1}{n}(EX_1 + EX_2 + \cdots + EX_n) = \frac{1}{n}(p + p + \cdots + p) = p$$

Thus, $\bar{X}$ is an unbiased estimator for p. In this circumstance, we generally write $\hat{p}$ instead of $\bar{X}$. In addition, we can use the fact that for independent random variables, the variance of the sum is the sum of the variances to see that

$$\begin{aligned}\text{Var}(\hat{p}) &= \frac{1}{n^2}(\text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)) \\ &= \frac{1}{n^2}(p(1-p) + p(1-p) + \cdots + p(1-p)) = \frac{1}{n}p(1-p).\end{aligned}$$

Example 14.3. *If $X_1, \ldots, X_n$ form a simple random sample with unknown finite mean μ, then $\bar{X}$ is an unbiased estimator of μ. If the X_i have variance σ^2, then*

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}. \tag{14.2}$$

We can assess the quality of an estimator by computing its **mean square error**, defined by

$$E_\theta[(d(X) - k(\theta))^2]. \tag{14.3}$$

Estimators with smaller mean square error are generally preferred to those with larger. Next we derive a simple relationship between mean square error and variance. We begin by substituting (14.1) into (14.4), rearranging terms, and expanding the square.

$$\begin{aligned}E_\theta[(d(X) - k(\theta))^2] &= E_\theta[(d(X) - (E_\theta d(X) - b_d(\theta)))^2] = E_\theta[((d(X) - E_\theta d(X)) + b_d(\theta))^2] \\ &= E_\theta[(d(X) - E_\theta d(X))^2] + 2b_d(\theta)E_\theta[d(X) - E_\theta d(X)] + b_d(\theta)^2 \\ &= \text{Var}_\theta(d(X)) + b_d(\theta)^2\end{aligned} \tag{14.4}$$

Thus, the representation of the mean square error as equal to the variance of the estimator plus the square of the bias is called the **bias-variance decomposition**. In particular:

- The mean square error for an unbiased estimator is its variance.
- Bias always increases the mean square error.

14.2 Computing Bias

For the variance σ^2, we have been presented with two choices:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{14.5}$$

Using bias as our criterion, we can now resolve between the two choices for the estimators for the variance σ^2. Again, we use simulations to make a conjecture, we then follow up with a computation to verify our guess. For 16 tosses of a fair coin, we know that the variance is $np(1-p) = 16 \cdot 1/2 \cdot 1/2 = 4$

For the example above, we begin by simulating the coin tosses and compute the sum of squares $\sum_{i=1}^{10}(x_i - \bar{x})^2$,

```
> ssx<-numeric(1000)
> for (i in 1:1000){x<-rbinom(10,16,0.5);ssx[i]<-sum((x-mean(x))^2)}
> mean(ssx)
[1] 35.8511
```

The choice is to divide either by 10, for the first choice, or 9, for the second.

```
> mean(ssx)/10;mean(ssx)/9
[1] 3.58511
[1] 3.983456
```

Exercise 14.4. *Repeat the simulation above, compute the sum of squares $\sum_{i=1}^{10}(x_i-8)^2$. Show that these simulations support dividing by 10 rather than 9. verify that $\sum_{i=1}^{n}(X_i-\mu)^2/n$ is an unbiased estimator for σ^2 for independent random variable $X_1,\dots,X_n$ whose common distribution has mean μ and variance σ^2.*

In this case, because we know all the aspects of the simulation, and thus we know that the answer ought to be near 4. Consequently, division by 9 appears to be the appropriate choice. Let's check this out, beginning with what seems to be the *inappropriate choice* to see what goes wrong..

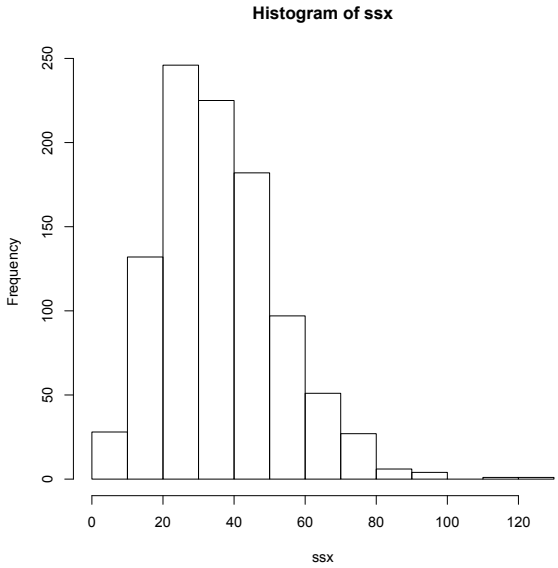


Figure 14.1: Sum of squares about $\bar{x}$ for 1000 simulations.

Example 14.5. *If a simple random sample $X_1, X_2, \dots,$ has unknown finite variance σ^2, then, we can consider the sample variance*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

To find the mean of S^2, we divide the difference between an observation X_i and the distributional mean into two steps - the first from X_i to the sample mean $\bar{x}$ and and then from the sample mean to the distributional mean, i.e.,

$$X_i - \mu = (X_i - \bar{X}) + (\bar{X} - \mu).$$

We shall soon see that the lack of knowledge of μ is the source of the bias. Make this substitution and expand the square to obtain

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}((X_i-\bar{X})+(\bar{X}-\mu))^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu) + \sum_{i=1}^{n}(\bar{X}-\mu)^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X}) + n(\bar{X}-\mu)^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2\end{aligned}$$

(Check for yourself that the middle term in the third line equals 0.) Subtract the term $n(\bar{X}-\mu)^2$ from both sides and divide by n to obtain the identity

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\bar{X}-\mu)^2.$$

Using the identity above and the linearity property of expectation we find that

$$\begin{aligned} ES^2 &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\bar{X}-\mu)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}E[(X_i-\mu)^2]-E[(\bar{X}-\mu)^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}\text{Var}(X_i)-\text{Var}(\bar{X}) \\ &= \frac{1}{n}n\sigma^2-\frac{1}{n}\sigma^2=\frac{n-1}{n}\sigma^2\neq\sigma^2. \end{aligned}$$

The last line uses (14.2). This shows that S^2 is a biased estimator for σ^2. Using the definition in (14.1), we can see that it is biased downwards.

$$b(\sigma^2)=\frac{n-1}{n}\sigma^2-\sigma^2=-\frac{1}{n}\sigma^2.$$

Note that the bias is equal to $-\text{Var}(\bar{X})$. In addition, because

$$E\left[\frac{n}{n-1}S^2\right]=\frac{n}{n-1}E\left[S^2\right]=\frac{n}{n-1}\left(\frac{n-1}{n}\sigma^2\right)=\sigma^2$$

and

$$S_u^2=\frac{n}{n-1}S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

is an unbiased estimator for σ^2. As we shall learn in the next section, because the square root is concave downward, $S_u=\sqrt{S_u^2}$ as an estimator for σ is **downwardly biased**.

Example 14.6. *We have seen, in the case of n Bernoulli trials having x successes, that $\hat{p}=x/n$ is an unbiased estimator for the parameter p. This is the case, for example, in taking a simple random sample of genetic markers at a particular biallelic locus. Let one allele denote the wildtype and the second a variant. If the circumstances in which variant is recessive, then an individual expresses the variant phenotype only in the case that both chromosomes contain this marker. In the case of independent alleles from each parent, the probability of the variant phenotype is p^2. Naïvely, we could use the estimator $\hat{p}^2$. (Later, we will see that this is the maximum likelihood estimator.) To determine the bias of this estimator, note that*

$$E\hat{p}^2=(E\hat{p})^2+\text{Var}(\hat{p})=p^2+\frac{1}{n}p(1-p). \tag{14.6}$$

Thus, the bias $b(p)=p(1-p)/n$ and the estimator $\hat{p}^2$ is biased upward.

Exercise 14.7. *For Bernoulli trials $X_1,\ldots,X_n$,*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{p})^2=\hat{p}(1-\hat{p}).$$

Based on this exercise, and the computation above yielding an unbiased estimator, S_u^2, for the variance,

$$E\left[\frac{1}{n-1}\hat{p}(1-\hat{p})\right]=\frac{1}{n}E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{p})^2\right]=\frac{1}{n}E[S_u^2]=\frac{1}{n}Var(X_1)=\frac{1}{n}p(1-p).$$

In other words,

$$\frac{1}{n-1}\hat{p}(1-\hat{p})$$

is an unbiased estimator of $p(1-p)/n$. *Returning to (14.6),*

$$E\left[\hat{p}^2 - \frac{1}{n-1}\hat{p}(1-\hat{p})\right] = \left(p^2 + \frac{1}{n}p(1-p)\right) - \frac{1}{n}p(1-p) = p^2.$$

Thus,

$$\widehat{p^2}_u = \hat{p}^2 - \frac{1}{n-1}\hat{p}(1-\hat{p})$$

is an unbiased estimator of p^2.

To compare the two estimators for p^2, *assume that we find 13 variant alleles in a sample of 30, then* $\hat{p} = 13/30 = 0.4333$,

$$\hat{p}^2 = \left(\frac{13}{30}\right)^2 = 0.1878, \quad \textit{and} \quad \widehat{p^2}_u = \left(\frac{13}{30}\right)^2 - \frac{1}{29}\left(\frac{13}{30}\right)\left(\frac{17}{30}\right) = 0.1878 - 0.0085 = 0.1793.$$

The bias for the estimate $\hat{p}^2$, *in this case 0.0085, is subtracted to give the unbiased estimate* $\widehat{p^2}_u$.

The **heterozygosity** *of a biallelic locus is* $h = 2p(1-p)$. *From the discussion above, we see that* h *has the unbiased estimator*

$$\hat{h} = \frac{2n}{n-1}\hat{p}(1-\hat{p}) = \frac{2n}{n-1}\left(\frac{x}{n}\right)\left(\frac{n-x}{n}\right) = \frac{2x(n-x)}{n(n-1)}.$$

14.3 Compensating for Bias

In the methods of moments estimation, we have used $g(\bar{X})$ as an estimator for $g(\mu)$. If g is a **convex function**, we can say something about the bias of this estimator. In Figure 14.2, we see the method of moments estimator for the estimator $g(\bar{X})$ for a parameter β in the Pareto distribution. The choice of $\beta = 3$ corresponds to a mean of $\mu = 3/2$ for the Pareto random variables. The central limit theorem states that the sample mean $\bar{X}$ is nearly normally distributed with mean 3/2. Thus, the distribution of $\bar{X}$ is nearly symmetric around 3/2. From the figure, we can see that the interval from 1.4 to 1.5 under the function g maps into a longer interval above $\beta = 3$ than the interval from 1.5 to 1.6 maps below $\beta = 3$. Thus, the function g spreads the values of $\bar{X}$ above $\beta = 3$ more than below. Consequently, we anticipate that the estimator $\hat{\beta}$ will be **upwardly biased**.

To address this phenomena in more general terms, we use the characterization of a convex function as a differentiable function whose graph lies above any tangent line. If we look at the value μ for the convex function g, then this statement becomes

$$g(x) - g(\mu) \geq g'(\mu)(x-\mu).$$

Now replace x with the random variable $\bar{X}$ and take expectations.

$$E_\mu[g(\bar{X}) - g(\mu)] \geq E_\mu[g'(\mu)(\bar{X}-\mu)] = g'(\mu)E_\mu[\bar{X}-\mu] = 0.$$

Consequently,

$$E_\mu g(\bar{X}) \geq g(\mu) \tag{14.7}$$

and $g(\bar{X})$ is **biased upwards**. The expression in (14.7) is known as **Jensen's inequality**.

Exercise 14.8. *Show that the estimator* S_u *is a downwardly biased estimator for* σ.

To estimate the size of the bias, we look at a quadratic approximation for g centered at the value μ

$$g(x) - g(\mu) \approx g'(\mu)(x-\mu) + \frac{1}{2}g''(\mu)(x-\mu)^2.$$

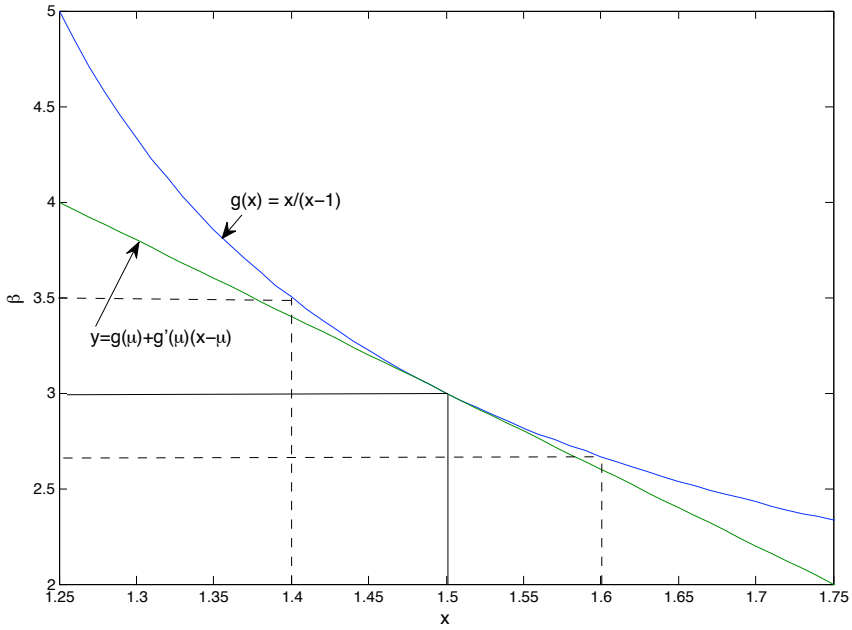


Figure 14.2: Graph of a convex function. Note that the tangent line is below the graph of g. Here we show the case in which $\mu = 1.5$ and $\beta = g(\mu) = 3$. Notice that the interval from $x = 1.4$ to $x = 1.5$ has a longer range than the interval from $x = 1.5$ to $x = 1.6$ Because g spreads the values of $\bar{X}$ above $\beta = 3$ more than below, the estimator $\hat{\beta}$ for β is biased upward. We can use a second order Taylor series expansion to correct most of this bias.

Again, replace x in this expression with the random variable $\bar{X}$ and then take expectations. Then, the bias

$$b_g(\mu) = E_\mu[g(\bar{X})] - g(\mu) \approx E_\mu[g'(\mu)(\bar{X} - \mu)] + \frac{1}{2}E[g''(\mu)(\bar{X} - \mu)^2] = \frac{1}{2}g''(\mu)\text{Var}(\bar{X}) = \frac{1}{2}g''(\mu)\frac{\sigma^2}{n}. \quad (14.8)$$

(Remember that $E_\mu[g'(\mu)(\bar{X} - \mu)] = 0$.) Thus, the bias has the intuitive properties of being

- large for strongly convex functions, i.e., ones with a large value for the second derivative evaluated at the mean μ,
- large for observations having high variance σ^2, and
- small when the number of observations n is large.

In addition, this provides an estimate of the **mean square error** for a method of moment estimator $\hat{\theta}(X) = g(\bar{X})$ based on the relationship $\theta = g(\mu)$. Based on the variance-bias identity in (14.4), we use the delta method to approximate $\text{Var}_\theta(g(\bar{X}))$ and (14.8) to approximate the bias. Consequently,

$$E_\theta[(\hat{\theta}(X) - \theta)^2] = E_\theta[(g(\bar{X}) - g(\mu))^2] = \text{Var}_\theta(g(\bar{X})) + b_g(\mu)^2 \approx g'(\mu)^2\frac{\sigma^2}{n} + \frac{1}{4}\left(g''(\mu)\frac{\sigma^2}{n}\right)^2.$$

Finally, we make the substitution $\mu = k(\theta)$ to give an expression for the approximation of the mean square error for $\hat{\theta}$ as a function of θ. Notice that the contribution to mean square error from the variance of the estimator is inversely proportional to n, the sample size. The contribution from the bias decreases more rapidly, inversely proportional to n^2,

Exercise 14.9. *If a method of moments estimator $\hat{\theta}$ is a linear function of the sample mean $\bar{x}$, then it is unbiased,*

Exercise 14.10. *Use (14.8) to estimate the bias in using $\hat{p}^2$ as an estimate of p^2 is a sequence of n Bernoulli trials and note that it matches the value (14.6).*

Example 14.11. *For the method of moments estimator for the Pareto random variable, we determined that*

$$g(\mu) = \frac{\mu}{\mu - 1}.$$

and that $\bar{X}$ has

$$\textit{mean} \quad \mu = \tfrac{\beta}{\beta-1} \quad \textit{and} \quad \textit{variance} \quad \tfrac{\sigma^2}{n} = \tfrac{\beta}{n(\beta-1)^2(\beta-2)}$$

By taking the second derivative, we see that $g''(\mu) = 2(\mu-1)^{-3} > 0$ and, because $\mu > 1$, g is a convex function. Next, we have

$$g''\left(\frac{\beta}{\beta-1}\right) = \frac{2}{\left(\frac{\beta}{\beta-1} - 1\right)^3} = 2(\beta-1)^3.$$

Thus, the bias

$$b_g(\beta) \approx \frac{1}{2} g''(\mu)\frac{\sigma^2}{n} = \frac{1}{2} 2(\beta-1)^3 \frac{\beta}{n(\beta-1)^2(\beta-2)} = \frac{\beta(\beta-1)}{n(\beta-2)}.$$

So, for $\beta = 3$ and $n = 100$, the bias is approximately 0.06. Compare this to the estimated value of 0.053 from the simulation in the previous section.

Example 14.12. *For estimating the population in mark and recapture, we used the estimate*

$$N = g(\mu) = \frac{kt}{\mu}$$

for the total population. Here μ is the mean number recaptured, k is the number captured in the second capture event and t is the number tagged. The second derivative

$$g''(\mu) = \frac{2kt}{\mu^3} > 0$$

and hence the method of moments estimate is biased upwards. In this siutation, $n = 1$ and the number recaptured is a hypergeometric random variable. Hence its variance

$$\sigma^2 = \frac{kt}{N}\frac{(N-t)(N-k)}{N(N-1)}.$$

Thus, the bias

$$b_g(N) = \frac{1}{2}\frac{2kt}{\mu^3}\frac{kt}{N}\frac{(N-t)(N-k)}{N(N-1)} = \frac{(N-t)(N-k)}{\mu(N-1)} = \frac{(kt/\mu - t)(kt/\mu - k)}{\mu(kt/\mu - 1)} = \frac{kt(k-\mu)(t-\mu)}{\mu^2(kt-\mu)}.$$

In the simulation example, $N = 2000, t = 200, k = 400$ and $\mu = 40$. This gives an estimate for the bias of 36.02. We can compare this to the bias of 2031.03-2000 = 31.03 based on the simulation in Example 13.2.

This suggests a new estimator by taking the method of moments estimator and subtracting the approximation of the bias.

$$\hat{N} = \frac{kt}{r} - \frac{kt(k-r)(t-r)}{r^2(kt-r)} = \frac{kt}{r}\left(1 - \frac{(k-r)(t-r)}{r(kt-r)}\right).$$

The delta method gives us that the standard deviation of the estimator is $|g'(\mu)|\sigma/\sqrt{n}$. Thus the ratio of the bias of an estimator to its standard deviation as determined by the delta method is approximately

$$\frac{g''(\mu)\sigma^2/(2n)}{|g'(\mu)|\sigma/\sqrt{n}} = \frac{1}{2}\frac{g''(\mu)}{|g'(\mu)|}\frac{\sigma}{\sqrt{n}}.$$

If this ratio is $\ll 1$, then the bias correction is not very important. In the case of the example above, this ratio is

$$\frac{36.02}{268.40} = 0.134$$

and its usefulness in correcting bias is small.

Example 14.13. *As noted earlier, because S^2 is an unbiased estimator of σ^2 and the square root function is concave downward. Thus, S is biased downwards as an estimator of σ. To illustrate this, we first simulate both ES_n and ES_n^2 for values n from 2 to 25 using normal random variables with mean $\mu = 0$ and standard deviation $\sigma = 1$.*

```
> s<-numeric(24); s2<-numeric(24)
> for (i in 1:24){s[i]<-mean(replicate(10000,sd(rnorm(i+1))))}
> for (i in 1:24){s2[i]<-mean(replicate(10000,var(rnorm(i+1))))}
```

We can also find an analytical expression for ES_n.

Exercise 14.14. *Note that f $Z_1, Z_2, \ldots, Z_n$ independent standard normal variables, then*

$$V_n = \sum_{i=1}^{n} (Z_i - \bar{Z})^2$$

is χ^2_{n-1}. Let

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})^2. \quad \textit{Show that} \quad ES_n = \sqrt{\frac{2}{n-1}}\, \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}. \tag{14.9}$$

Then, we add the values of ES to form a table that has ES along with the simulated values of ES and ES^2.

```
> df<-1:24
> means<-sqrt(2/df)*gamma((df+1)/2)/gamma(df/2)
> head(data.frame(df,means,s,s2))
  df     means         s        s2
1  1 0.7978846 0.7913008 1.0008486
2  2 0.8862269 0.8920162 0.9811749
3  3 0.9213177 0.9190756 1.0145019
4  4 0.9399856 0.9419472 1.0006425
5  5 0.9515329 0.9498657 1.0055365
6  6 0.9593688 0.9563602 0.9985865
```

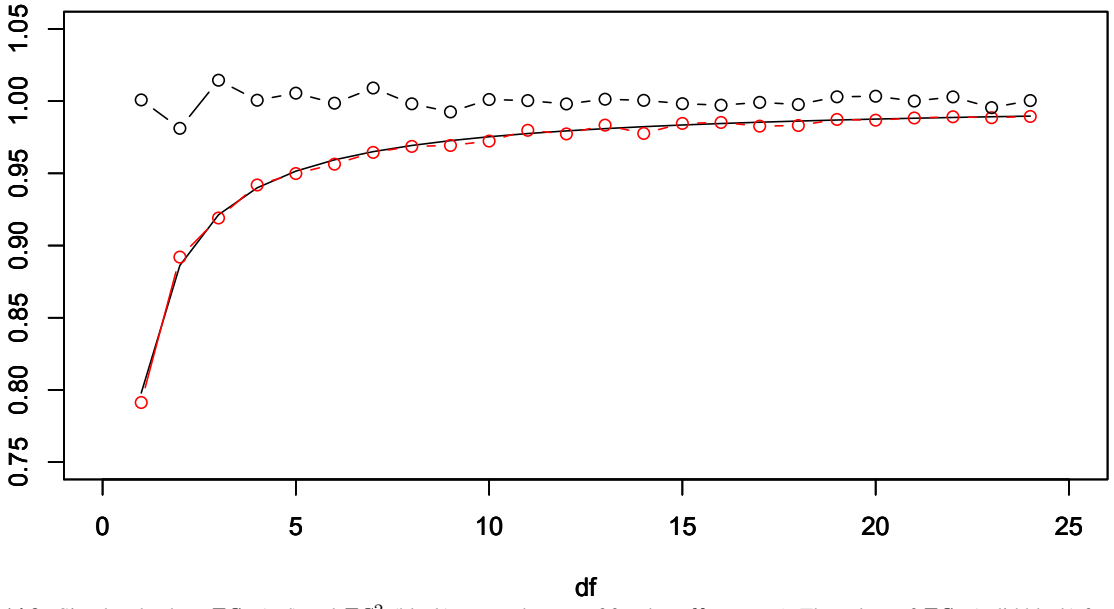


Figure 14.3: Simulated values ES_n (red) and ES_n^2 (black) versus degrees of freedom $df = n - 1$. The values of ES_n (solid black) from (14.9).

As we can see from both the analytical expression and the simulations, $ES < \sigma = 1$ and approaches 1 as the degrees of freedom $df = n - 1$ increase. Notice that the simulated values for the variance is close to $1 = ES^2$.

Exercise 14.15. *The Stirling approximation states that*

$$\Gamma(t) \approx \sqrt{2\pi(t-1)} \left(\frac{t-1}{e} \right)^{t-1}, \quad t = 1.2\ldots.$$

Use this in (14.9) to show that

$$\lim_{n\to\infty} ES_n = 1.$$

14.4 Consistency

Despite the desirability of using an unbiased estimator, sometimes such an estimator is hard to find and at other times impossible. However, note that in the examples above both the size of the bias and the variance in the estimator decrease inversely proportional to n, the number of observations. Thus, these estimators improve, under both of these criteria, with more observations. A concept that describes properties such as these is called **consistency**.

Definition 14.16. *Given data $X_1, X_2, \ldots$ and a real valued function h of the parameter space, a sequence of estimators d_n, based on the first n observations, is called* **consistent** *if for every choice of θ*

$$\lim_{n\to\infty} d_n(X_1, X_2, \ldots, X_n) = k(\theta)$$

whenever θ is the true state of nature.

Thus, the bias of the estimator disappears in the limit of a large number of observations. In addition, the distribution of the estimators $d_n(X_1, X_2, \ldots, X_n)$ become more and more concentrated near $k(\theta)$.

For the next example, we need to recall the sequence definition of continuity: A function g is continuous at a real number x provided that for every sequence $\{x_n; n \geq 1\}$ with

$$x_n \to x, \text{ then, we have that } g(x_n) \to g(x).$$

A function is called continuous if it is continuous at every value of x in the domain of g. Thus, we can write the expression above more succinctly by saying that for every convergent sequence $\{x_n; n \geq 1\}$,

$$\lim_{n\to\infty} g(x_n) = g(\lim_{n\to\infty} x_n).$$

Example 14.17. *For a method of moment estimator, let's focus on the case of a single parameter ($d = 1$). For independent observations, $X_1, X_2, \ldots$, having mean $\mu = k(\theta)$, we have that*

$$E\bar{X}_n = \mu,$$

i. e. $\bar{X}_n$, the sample mean for the first n observations, is an unbiased estimator for $\mu = k(\theta)$. Also, by the law of large numbers, we have that

$$\lim_{n\to\infty} \bar{X}_n = \mu.$$

Assume that k has a continuous inverse $g = k^{-1}$. In particular, because $\mu = k(\theta)$, we have that $g(\mu) = \theta$. Next, using the methods of moments procedure, define, for n observations, the estimators

$$\hat{\theta}_n(X_1, X_2, \ldots, X_n) = g\left(\frac{1}{n}(X_1 + \cdots + X_n) \right) = g(\bar{X}_n).$$

for the parameter θ. Using the continuity of g, we find that

$$\lim_{n\to\infty} \hat{\theta}_n(X_1, X_2, \ldots, X_n) = \lim_{n\to\infty} g(\bar{X}_n) = g(\lim_{n\to\infty} \bar{X}_n) = g(\mu) = \theta$$

and so we have that $g(\bar{X}_n)$ is a consistent sequence of estimators for θ.

14.5 Cramér-Rao Bound

This topic is somewhat more advanced and can be skipped for the first reading. This section gives us an introduction to the log-likelihood and its derivative, the **score** functions. We shall encounter these functions again when we introduce maximum likelihood estimation. In addition, the Cramér Rao bound, which is based on the variance of the score function, known as the Fisher information, gives a lower bound for the variance of an unbiased estimator. These concepts will be necessary to describe the variance for maximum likelihood estimators.

Among unbiased estimators, one important goal is to find an estimator that has as small a variance as possible, A more precise goal would be to find an unbiased estimator d that has **uniform minimum variance**. In other words, $d(X)$ has has a smaller variance than for any other unbiased estimator $\tilde{d}$ for every value θ of the parameter.

$$\text{Var}_\theta d(X) \le \text{Var}_\theta \tilde{d}(X) \quad \text{for all } \theta \in \Theta.$$

The **efficiency** $e(\tilde{d})$ of unbiased estimator $\tilde{d}$ is the minimum value of the ratio

$$\frac{\text{Var}_\theta d(X)}{\text{Var}_\theta \tilde{d}(X)}$$

over all values of θ. Thus, the efficiency is between 0 and 1 with a goal of finding estimators with efficiency as near to one as possible.

For unbiased estimators, the Cramér-Rao bound tells us how small a variance is ever possible. The formula is a bit mysterious at first. However, we shall soon learn that this bound is a consequence of the bound on correlation that we have previously learned

Recall that for two random variables Y and Z, the correlation

$$\rho(Y,Z) = \frac{\text{Cov}(Y,Z)}{\sqrt{\text{Var}(Y)\text{Var}(Z)}}. \tag{14.10}$$

takes values between -1 and 1. Thus, $\rho(Y,Z)^2 \le 1$ and so

$$\text{Cov}(Y,Z)^2 \le \text{Var}(Y)\text{Var}(Z). \tag{14.11}$$

Exercise 14.18. *If* $EZ = 0$*, the* $\text{Cov}(Y,Z) = EYZ$

We begin with data $X = (X_1, \dots, X_n)$ drawn from an unknown probability P_θ. The parameter space $\Theta \subset \mathbb{R}$. Denote the joint density of these random variables

$$\mathbf{f}(\mathbf{x}|\theta), \quad \text{where } \mathbf{x} = (x_1 \dots, x_n).$$

In the case that the data comes from a simple random sample then the joint density is the product of the marginal densities.

$$\mathbf{f}(\mathbf{x}|\theta) = f(x_1|\theta) \cdots f(x_n|\theta) \tag{14.12}$$

For continuous random variables, the two basic properties of the density are that $\mathbf{f}(\mathbf{x}|\theta) \ge 0$ for all $\mathbf{x}$ and that

$$1 = \int_{\mathbb{R}^n} \mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x}. \tag{14.13}$$

Now, let d be the unbiased estimator of $k(\theta)$, then by the basic formula for computing expectation, we have for continuous random variables

$$k(\theta) = E_\theta d(X) = \int_{\mathbb{R}^n} d(\mathbf{x})\mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x}. \tag{14.14}$$

If the functions in (14.13) and (14.14) are differentiable with respect to the parameter θ and we can pass the derivative through the integral, then we first differentiate both sides of equation (14.13), and then use the logarithm function to write this derivate as the expectation of a random variable,

$$0 = \int_{\mathbb{R}^n} \frac{\partial \mathbf{f}(\mathbf{x}|\theta)}{\partial \theta}\, d\mathbf{x} = \int_{\mathbb{R}^n} \frac{\partial \mathbf{f}(\mathbf{x}|\theta)/\partial\theta}{\mathbf{f}(\mathbf{x}|\theta)} \mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x} = \int_{\mathbb{R}^n} \frac{\partial \ln \mathbf{f}(\mathbf{x}|\theta)}{\partial \theta} \mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x} = E_\theta \left[\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right]. \quad (14.15)$$

From a similar calculation using (14.14),

$$k'(\theta) = E_\theta \left[d(X) \frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right]. \quad (14.16)$$

Now, return to the review on correlation with $Y = d(X)$, the unbiased estimator for $k(\theta)$ and the **score function** $Z = \partial \ln \mathbf{f}(X|\theta)/\partial\theta$. From equations (14.16) and then (14.11), we find that

$$k'(\theta)^2 = E_\theta \left[d(X) \frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right]^2 = \text{Cov}_\theta \left(d(X), \frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right) \le \text{Var}_\theta(d(X)) \text{Var}_\theta \left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right),$$

or,

$$\text{Var}_\theta(d(X)) \ge \frac{k'(\theta)^2}{I(\theta)}. \quad (14.17)$$

where

$$I(\theta) = \text{Var}_\theta \left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right) = E_\theta \left[\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right)^2 \right]$$

is called the **Fisher information**. For the equality, recall that the variance $\text{Var}(Z) = EZ^2 - (EZ)^2$ and recall from equation (14.15) that the random variable $Z = \partial \ln \mathbf{f}(X|\theta)/\partial\theta$ has mean $EZ = 0$.

Equation (14.17), called the **Cramér-Rao lower bound** or the **information inequality**, states that the lower bound for the variance of an unbiased estimator is the reciprocal of the Fisher information. In other words, the *higher* the information, the *lower* is the possible value of the variance of an unbiased estimator.

If we return to the case of a simple random sample, then take the logarithm of both sides of equation (14.12)

$$\ln \mathbf{f}(\mathbf{x}|\theta) = \ln f(x_1|\theta) + \cdots + \ln f(x_n|\theta)$$

and then differentiate with respect to the parameter θ,

$$\frac{\partial \ln \mathbf{f}(\mathbf{x}|\theta)}{\partial \theta} = \frac{\partial \ln f(x_1|\theta)}{\partial \theta} + \cdots + \frac{\partial \ln f(x_n|\theta)}{\partial \theta}.$$

The random variables $\{\partial \ln f(X_k|\theta)/\partial\theta; 1 \le k \le n\}$ are independent and have the same distribution. Using the fact that the variance of the sum is the sum of the variances for independent random variables, we see that I_n, the Fisher information for n observations is n times the Fisher information of a single observation.

$$I_n(\theta) = \text{Var} \left(\frac{\partial \ln f(X_1|\theta)}{\partial \theta} + \cdots + \frac{\partial \ln f(X_n|\theta)}{\partial \theta} \right) = n\text{Var}(\frac{\partial \ln f(X_1|\theta)}{\partial \theta}) = nE[(\frac{\partial \ln f(X_1|\theta)}{\partial \theta})^2].$$

Notice the correspondence. *Information* is *linearly* proportional to the number of observations. If our estimator is a sample mean or a function of the sample mean, then the *variance* is *inversely* proportional to the number of observations.

Example 14.19. *For independent Bernoulli random variables with unknown success probability θ, the density is*

$$f(x|\theta) = \theta^x (1-\theta)^{(1-x)}.$$

The mean is θ and the variance is $\theta(1-\theta)$. Taking logarithms, we find that

$$\ln f(x|\theta) = x\ln\theta + (1-x)\ln(1-\theta),$$

$$\frac{\partial}{\partial\theta}\ln f(x|\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}.$$

The Fisher information associated to a single observation

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right)^2\right] = \frac{1}{\theta^2(1-\theta)^2}E[(X-\theta)^2] = \frac{1}{\theta^2(1-\theta)^2}\text{Var}(X)$$
$$= \frac{1}{\theta^2(1-\theta)^2}\theta(1-\theta) = \frac{1}{\theta(1-\theta)}.$$

Thus, the information for n observations $I_n(\theta) = n/(\theta(1-\theta))$. Thus, by the Cramér-Rao lower bound, any unbiased estimator of θ based on n observations must have variance al least $\theta(1-\theta)/n$. Now, notice that if we take $d(\mathbf{x}) = \bar{x}$, then

$$E_\theta\bar{X} = \theta, \quad \textit{and} \quad \text{Var}_\theta d(X) = \text{Var}(\bar{X}) = \frac{\theta(1-\theta)}{n}.$$

These two equations show that $\bar{X}$ is a unbiased estimator having uniformly minimum variance.

Exercise 14.20. *For independent normal random variables with known variance σ_0^2 and unknown mean μ, $\bar{X}$ is a uniformly minimum variance unbiased estimator.*

Exercise 14.21. *Take two derivatives of $\ln \mathbf{f}(x|\theta)$ to show that*

$$I(\theta) = E_\theta\left[\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right)^2\right] = -E_\theta\left[\frac{\partial^2 \ln \mathbf{f}(X|\theta)}{\partial\theta^2}\right]. \tag{14.18}$$

This identity is often a useful alternative to compute the Fisher Information.

Example 14.22. *For an exponential random variable,*

$$\ln f(x|\lambda) = \ln\lambda - \lambda x, \quad \frac{\partial^2 f(x|\lambda)}{\partial\lambda^2} = -\frac{1}{\lambda^2}.$$

Thus, by (14.18),

$$I(\lambda) = \frac{1}{\lambda^2}.$$

Now, $\bar{X}$ is an unbiased estimator for $h(\lambda) = 1/\lambda$ with variance

$$\frac{1}{n\lambda^2}.$$

By the Cramér-Rao lower bound, we have that

$$\frac{g'(\lambda)^2}{nI(\lambda)} = \frac{1/\lambda^4}{n\lambda^2} = \frac{1}{n\lambda^2}.$$

Because $\bar{X}$ has this variance, it is a uniformly minimum variance unbiased estimator.

Example 14.23. *To give an estimator that does not achieve the Cramér-Rao bound, let $X_1, X_2, \ldots, X_n$ be a simple random sample of Pareto random variables with density*

$$f_X(x|\beta) = \frac{\beta}{x^{\beta+1}}, \quad x > 1.$$

The mean and the variance

$$\mu = \frac{\beta}{\beta - 1}, \qquad \sigma^2 = \frac{\beta}{(\beta-1)^2(\beta-2)}.$$

Thus, $\bar{X}$ is an unbiased estimator of $\mu = \beta/(\beta-1)$

$$\text{Var}(\bar{X}) = \frac{\beta}{n(\beta-1)^2(\beta-2)}.$$

To compute the Fisher information, note that

$$\ln f(x|\beta) = \ln \beta - (\beta+1)\ln x \quad \textit{and thus} \quad \frac{\partial^2 \ln f(x|\beta)}{\partial \beta^2} = -\frac{1}{\beta^2}.$$

Using (14.18), we have that

$$I(\beta) = \frac{1}{\beta^2}.$$

Next, for

$$\mu = g(\beta) = \frac{\beta}{\beta-1}, \quad g'(\beta) = -\frac{1}{(\beta-1)^2}, \quad \textit{and} \quad g'(\beta)^2 = \frac{1}{(\beta-1)^4}.$$

Thus, the Cramér-Rao bound for the estimator is

$$\frac{g'(\beta)^2}{I_n(\beta)} = \frac{\beta^2}{n(\beta-1)^4}.$$

and the efficiency compared to the Cramér-Rao bound is

$$\frac{g'(\beta)^2/I_n(\beta)}{\text{Var}(\bar{X})} = \frac{\beta^2}{n(\beta-1)^4} \cdot \frac{n(\beta-1)^2(\beta-2)}{\beta} = \frac{\beta(\beta-2)}{(\beta-1)^2} = 1 - \frac{1}{(\beta-1)^2}.$$

The Pareto distribution does not have a variance unless $\beta > 2$. For β just above 2, the efficiency compared to its Cramér-Rao bound is low but improves with larger β.

14.6 A Note on Efficient Estimators

For an efficient estimator, we need find the cases that lead to equality in the correlation inequality (14.10). Recall that equality occurs precisely when the correlation is ± 1. This occurs when the estimator $d(X)$ and the score function $\partial \ln f_X(X|\theta)/\partial\theta$ are linearly related with probability 1.

$$\frac{\partial}{\partial\theta} \ln f_X(X|\theta) = a(\theta)d(X) + b(\theta). \tag{14.19}$$

After integrating, we obtain,

$$\ln f_X(X|\theta) = \int a(\theta)d\theta \; d(X) + \int b(\theta)d\theta + j(X) = \pi(\theta)d(X) + B(\theta) + j(X)$$

Note that the constant of integration of integration is a function of X. Now exponentiate both sides of this equation

$$f_X(X|\theta) = c(\theta)h(X)\exp(\pi(\theta)d(X)). \tag{14.20}$$

Here $c(\theta) = \exp B(\theta)$ and $h(X) = \exp j(X)$.

We shall call density functions satisfying equation (14.20) an **exponential family** with **natural parameter** $\pi(\theta)$. Thus, if we have independent random variables $X_1, X_2, \ldots X_n$, then the joint density is the product of the densities, namely,

$$\mathbf{f}(X|\theta) = c(\theta)^n h(X_1)\cdots h(X_n)\exp(\pi(\theta)(d(X_1) + \cdots + d(X_n)). \tag{14.21}$$

In addition, as a consequence of this linear relation in (14.19),

$$\overline{d(X)} = \frac{1}{n}(d(X_1) + \cdots + d(X_n))$$

is an efficient estimator for $k(\theta)$.

Example 14.24 (Poisson random variables)**.**

$$f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\frac{1}{x!}\exp(x\ln\lambda).$$

Thus, Poisson random variables are an exponential family with $c(\lambda) = \exp(-\lambda)$, $h(x) = 1/x!$, and natural parameter $\pi(\lambda) = \ln\lambda$. Because

$$\lambda = E_\lambda \bar{X},$$

$\bar{X}$ is an unbiased estimator of the parameter λ.

The score function

$$\frac{\partial}{\partial\lambda}\ln f(x|\lambda) = \frac{\partial}{\partial\lambda}(x\ln\lambda - \ln x! - \lambda) = \frac{x}{\lambda} - 1.$$

The Fisher information for one observation is

$$I(\lambda) = E_\lambda\left[\left(\frac{X}{\lambda} - 1\right)^2\right] = \frac{1}{\lambda^2}E_\lambda[(X-\lambda)^2] = \frac{1}{\lambda}.$$

Thus, $I_n(\lambda) = n/\lambda$ is the Fisher information for n observations. In addition,

$$Var_\lambda(\bar{X}) = \frac{\lambda}{n}$$

and $d(x) = \bar{x}$ has efficiency

$$\frac{Var(\bar{X})}{1/I_n(\lambda)} = 1.$$

This could have been predicted. The density of n independent observations is

$$\mathbf{f}(\mathbf{x}|\lambda) = \frac{e^{-\lambda}}{x_1!}\lambda^{x_1}\cdots\frac{e^{-\lambda}}{x_n!}\lambda^{x_n} = \frac{e^{-n\lambda}\lambda^{x_1\cdots+x_n}}{x_1!\cdots x_n!} = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{x_1!\cdots x_n!}$$

and so the score function

$$\frac{\partial}{\partial\lambda}\ln\mathbf{f}(\mathbf{x}|\lambda) = \frac{\partial}{\partial\lambda}(-n\lambda + n\bar{x}\ln\lambda) = -n + \frac{n\bar{x}}{\lambda}$$

showing that the estimate $\bar{x}$ and the score function are linearly related.

Exercise 14.25. *Show that a Bernoulli random variable with parameter p is an exponential family.*

Exercise 14.26. *Show that a normal random variable with known variance σ_0^2 and unknown mean μ is an exponential family.*

14.7 Answers to Selected Exercises

14.4. Repeat the simulation, replacing `mean(x)` by 8.

```
> ssx<-numeric(1000)
> for (i in 1:1000){x<-rbinom(10,16,0.5);ssx[i]<-sum((x-8)^2)}
> mean(ssx)/10;mean(ssx)/9
[1] 3.9918
[1] 4.435333
```

Note that division by 10 gives an answer very close to the correct value of 4. To verify that the estimator is unbiased, we write

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right]=\frac{1}{n}\sum_{i=1}^{n}E[(X_i-\mu)^2]=\frac{1}{n}\sum_{i=1}^{n}\text{Var}(X_i)=\frac{1}{n}\sum_{i=1}^{n}\sigma^2=\sigma^2.$$

14.7. For a Bernoulli trial note that $X_i^2=X_i$. Expand the square to obtain

$$\sum_{i=1}^{n}(X_i-\hat{p})^2=\sum_{i=1}^{n}X_i^2-\hat{p}\sum_{i=1}^{n}X_i+n\hat{p}^2=n\hat{p}-2n\hat{p}^2+n\hat{p}^2=n(\hat{p}-\hat{p}^2)=n\hat{p}(1-\hat{p}).$$

Divide by n to obtain the result.

14.8. Recall that $ES_u^2=\sigma^2$. Check the second derivative to see that $g(t)=\sqrt{t}$ is concave down for all t. For concave down functions, the direction of the inequality in Jensen's inequality is reversed. Setting $t=S_u^2$, we have that

$$ES_u=Eg(S_u^2)\le g(ES_u^2)=g(\sigma^2)=\sigma$$

and S_u is a downwardly biased estimator of σ.

14.9. In oder to have a linear method of moments estimator

$$\hat{\theta}=a+b\bar{x},$$

we must have, for mean μ,

$$\theta=a+b\mu.$$

Thus,

$$E_\theta\hat{\theta}=E[a+b\bar{X}]=a+bE\bar{X}=a+b\mu=\theta$$

and $\hat{\theta}$ is unbiased.

14.10. Set $g(p)=p^2$. Then, $g''(p)=2$. Recall that the variance of a Bernoulli random variable $\sigma^2=p(1-p)$ and the bias

$$b_g(p)\approx\frac{1}{2}g''(p)\frac{\sigma^2}{n}=\frac{1}{2}2\frac{p(1-p)}{n}=\frac{p(1-p)}{n}.$$

14.14. V_n is χ^2_{n-1},

$$\begin{aligned}E\sqrt{V_n}&=\int_0^\infty\sqrt{v}f(v|n-1)\,dv=\frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)}\int_0^\infty\sqrt{v}v^{(n-1)/2-1}e^{-v/2}\,dv\\&=\frac{\sqrt{2}\Gamma(n/2)}{\Gamma((n-1)/2)}\frac{1}{2^{n/2}\Gamma(n/2)}\int_0^\infty v^{n/2-1}e^{-v/2}\,dv=\frac{\sqrt{2}\Gamma(n/2)}{\Gamma((n-1)/2)}\int_0^\infty f(v|n)\,dv\\&=\frac{\sqrt{2}\Gamma(n/2)}{\Gamma((n-1)/2)}\end{aligned}$$

Now, $S_n^2=\frac{1}{n-1}V$. Thus

$$ES_n=\frac{1}{\sqrt{n-1}}E\sqrt{V_n}=\sqrt{\frac{2}{n-1}}\,\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}.$$

14.15. We substitute Stirling's approximation and drop terms from the expression that have limit 1.

$$\begin{aligned}
\lim_{n\to\infty} ES_n &= \lim_{n\to\infty} \sqrt{\frac{2}{n-1}}\, \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \\
&= \lim_{n\to\infty} \sqrt{\frac{2}{n-1}}\, \sqrt{2\pi(n/2-1)} \left(\frac{n/2-1}{e}\right)^{n/2-1} \frac{1}{\sqrt{2\pi((n-1)/2-1)}} \left(\frac{e}{(n-1)/2-1}\right)^{(n-1)/2-1} \\
&= \lim_{n\to\infty} \sqrt{\frac{2}{n-1}}\, \sqrt{\frac{n/2-1}{(n-1)/2-1}} \left(\frac{n/2-1}{e}\right)^{n/2-1} \left(\frac{e}{(n-1)/2-1}\right)^{(n-1)/2-1} \\
&= \lim_{n\to\infty} \sqrt{\frac{2}{e(n-1)}}\, \frac{(n/2-1)^{n/2-1/2}}{((n-1)/2-1)^{(n-1)/2-1/2}} = \lim_{n\to\infty} \sqrt{\frac{1}{e}\frac{(n-1)/2-1}{(n-1)/2}}\, \frac{(n/2-1)^{n/2-1/2}}{((n-1)/2-1)^{n/2-1/2}} \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{e}} \left(\frac{n/2-1}{(n-1)/2-1}\right)^{n/2-1/2} = \lim_{n\to\infty} \frac{1}{\sqrt{e}} \left(\frac{n-2}{n-3}\right)^{n/2} \left(\frac{n-2}{n-3}\right)^{-1/2} \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{e}}\, \frac{(1-2/n)^{n/2}}{(1-3/n)^{n/2}} = \frac{1}{e^{1/2}} \frac{e^{-1}}{e^{-3/2}} = 1,
\end{aligned}$$

recalling that

$$\lim_{x\to\infty} \left(1-\frac{\alpha}{x}\right)^x = e^{-\alpha}.$$

14.18. $\mathrm{Cov}(Y,Z) = EYZ - EY \cdot EZ = EYZ$ whenever $EZ = 0$.

14.20. For independent normal random variables with known variance σ_0^2 and unknown mean μ, the density

$$f(x|\mu) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma_0^2},$$

$$\ln f(x|\mu) = -\ln(\sigma_0\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma_0^2}.$$

Thus, the score function

$$\frac{\partial}{\partial\mu} \ln f(x|\mu) = \frac{1}{\sigma_0^2}(x-\mu).$$

and the Fisher information associated to a single observation

$$I(\mu) = E\left[\left(\frac{\partial}{\partial\mu} \ln f(X|\mu)\right)^2\right] = \frac{1}{\sigma_0^4} E[(X-\mu)^2] = \frac{1}{\sigma_0^4}\mathrm{Var}(X) = \frac{1}{\sigma_0^2}.$$

Again, the information is the reciprocal of the variance. Thus, by the Cramér-Rao lower bound, any unbiased estimator based on n observations must have variance al least σ_0^2/n. However, if we take $d(\mathbf{x}) = \bar{x}$, then

$$\mathrm{Var}_\mu d(X) = \frac{\sigma_0^2}{n}.$$

and $\bar{x}$ is a uniformly minimum variance unbiased estimator.

14.21. First, we take two derivatives of $\ln \mathbf{f}(x|\theta)$.

$$\frac{\partial \ln \mathbf{f}(x|\theta)}{\partial\theta} = \frac{\partial \mathbf{f}(x|\theta)/\partial\theta}{\mathbf{f}(x|\theta)} \tag{14.22}$$

and

$$\frac{\partial^2 \ln \mathbf{f}(x|\theta)}{\partial \theta^2} = \frac{\partial^2 \mathbf{f}(x|\theta)/\partial\theta^2}{\mathbf{f}(x|\theta)} - \frac{(\partial \mathbf{f}(x|\theta)/\partial\theta)^2}{\mathbf{f}(x|\theta)^2} = \frac{\partial^2 \mathbf{f}(x|\theta)/\partial\theta^2}{\mathbf{f}(x|\theta)} - \left(\frac{\partial \mathbf{f}(x|\theta)/\partial\theta)}{\mathbf{f}(x|\theta)}\right)^2$$
$$= \frac{\partial^2 \mathbf{f}(x|\theta)/\partial\theta^2}{\mathbf{f}(x|\theta)} - \left(\frac{\partial \ln \mathbf{f}(x|\theta)}{\partial\theta}\right)^2$$

upon substitution from identity (14.22). Thus, the expected values satisfy

$$E_\theta\left[\frac{\partial^2 \ln \mathbf{f}(X|\theta)}{\partial\theta^2}\right] = E_\theta\left[\frac{\partial^2 \mathbf{f}(X|\theta)/\partial\theta^2}{\mathbf{f}(X|\theta)}\right] - E_\theta\left[\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right)^2\right].$$

Consquently, the exercise is complete if we show that $E_\theta\left[\frac{\partial^2 \mathbf{f}(X|\theta)/\partial\theta^2}{\mathbf{f}(X|\theta)}\right] = 0$. However, for a continuous random variable,

$$E_\theta\left[\frac{\partial^2 \mathbf{f}(X|\theta)/\partial\theta^2}{\mathbf{f}(X|\theta)}\right] = \int_{\mathbb{R}^n} \frac{\partial^2 \mathbf{f}(x|\theta)/\partial\theta^2}{\mathbf{f}(x|\theta)}\, \mathbf{f}(x|\theta)\, dx = \int_{\mathbb{R}^n} \frac{\partial^2 \mathbf{f}(x|\theta)}{\partial\theta^2}\, dx = \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}^n} \mathbf{f}(x|\theta)\, dx = \frac{\partial^2}{\partial\theta^2}\, 1 = 0.$$

Note that the computation require that we be able to pass two derivatives with respect to θ through the integral sign.

14.25. The Bernoulli density

$$f(x|p) = p^x(1-p)^{1-x} = (1-p)\left(\frac{p}{1-p}\right)^x = (1-p)\exp\left(x\ln\left(\frac{p}{1-p}\right)\right).$$

Thus, $c(p) = 1-p, h(x) = 1$ and the natural parameter $\pi(p) = \ln\left(\frac{p}{1-p}\right)$, the **log-odds**.

14.27. The normal density

$$f(x|\mu) = \frac{1}{\sigma_0\sqrt{2\pi}}\exp -\frac{(x-\mu)^2}{2\sigma_0^2} = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\mu^2/2\sigma_0} e^{-x^2/2\sigma_0}\exp\frac{x\mu}{\sigma_0^2}$$

Thus, $c(\mu) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\mu^2/2\sigma_0}, h(x) = e^{-x^2/2\sigma_0}$ and the natural parameter $\pi(\mu) = \mu/\sigma_0^2$.

Topic 15

Maximum Likelihood Estimation

The solution of the problems of calculating from a sample the parameters of the hypothetical population, which we have put forward in the method of maximum likelihood, consists, then, simply of choosing such values of these parameters as have the maximum likelihood. - R. A. Fisher, *Phil. Trans. Royal Soc. Ser.* A. 222, (1922),

15.1 Introduction

The **principle of maximum likelihood** is relatively straightforward to state. As before, we begin with observations $X = (X_1, \ldots, X_n)$ of random variables chosen according to one of a family of probabilities P_θ. In addition, $\mathbf{f}(\mathbf{x}|\theta)$, $\mathbf{x} = (x_1, \ldots, x_n)$ will be used to denote the density function for the data when θ is the true state of nature.

Then, the principle of maximum likelihood yields a choice of the estimator $\hat{\theta}$ as the value for the parameter that makes the observed data most probable.

Definition 15.1. *The* **likelihood function** *is the density function regarded as a function of* θ*.*

$$\mathbf{L}(\theta|\mathbf{x}) = \mathbf{f}(\mathbf{x}|\theta),\ \theta \in \Theta. \tag{15.1}$$

The **maximum likelihood estimate (MLE)***,*

$$\hat{\theta}(\mathbf{x}) = \arg\max_{\theta} \mathbf{L}(\theta|\mathbf{x}). \tag{15.2}$$

Thus, we are presuming that a unique global maximum exists.

We will learn that especially for large samples, the maximum likelihood estimators have many desirable properties. However, especially for high dimensional data, the likelihood can have many local maxima. Thus, finding the *global maximum* can be a major computational challenge.

This class of estimators has an important **invariance property**. If $\hat{\theta}(\mathbf{x})$ is a maximum likelihood estimate for θ, then $g(\hat{\theta}(\mathbf{x}))$ is a maximum likelihood estimate for $g(\theta)$. For example, if θ is a parameter for the variance and $\hat{\theta}$ is the maximum likelihood estimate for the variance, then $\sqrt{\hat{\theta}}$ is the maximum likelihood estimate for the standard deviation. This flexibility in estimation criterion seen here is not available in the case of unbiased estimators.

For independent observations, the likelihood is the product of density functions. Because the logarithm of a product is the sum of the logarithms, finding zeroes of the **score function**, $\partial \ln \mathbf{L}(\theta|\mathbf{x})/\partial\theta$, the derivative of the logarithm of the likelihood, will be easier. Having the parameter values be the variable of interest is somewhat unusual, so we will next look at several examples of the likelihood function.

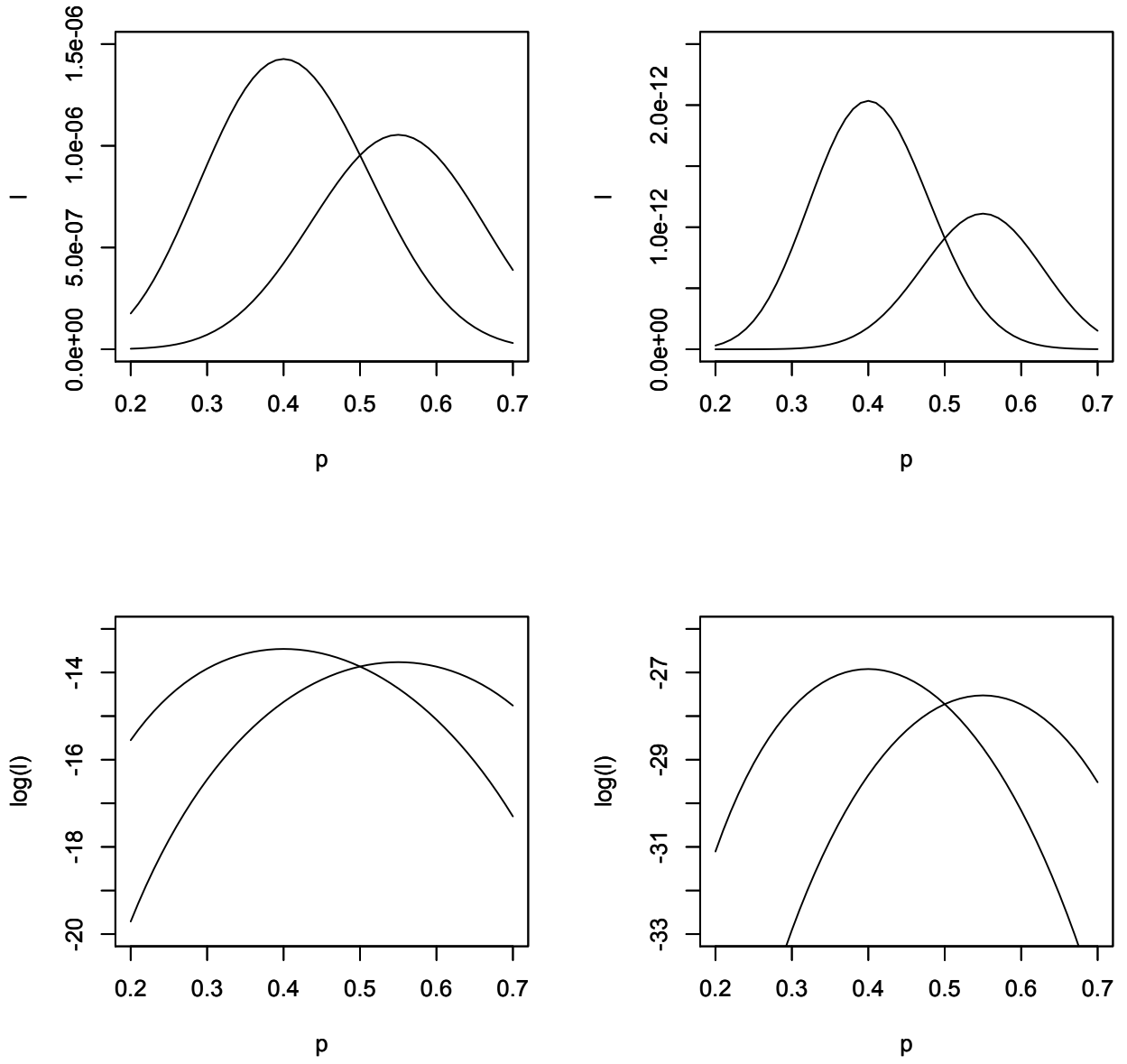


Figure 15.1: Likelihood function (top row) and its logarithm (bottom row) for Bernouli trials. The left column is based on 20 trials having 8 and 11 successes. The right column is based on 40 trials having 16 and 22 successes. Notice that the maximum likelihood is approximately 10^{-6} for 20 trials and 10^{-12} for 40. In addition, note that the peaks are more narrow for 40 trials rather than 20. We shall later be able to associate this property to the variance of the maximum likelihood estimator.

15.2 Examples

Example 15.2 (Bernoulli trials)**.** *If the experiment consists of n Bernoulli trials with success probability p, then*

$$\mathbf{L}(p|\mathbf{x}) = p^{x_1}(1-p)^{(1-x_1)}\cdots p^{x_n}(1-p)^{(1-x_n)} = p^{(x_1+\cdots+x_n)}(1-p)^{n-(x_1+\cdots+x_n)}.$$

$$\ln \mathbf{L}(p|\mathbf{x}) = \ln p(\sum_{i=1}^{n} x_i) + \ln(1-p)(n - \sum_{i=1}^{n} x_i) = n(\bar{x}\ln p + (1-\bar{x})\ln(1-p)).$$

$$\frac{\partial}{\partial p}\ln \mathbf{L}(p|\mathbf{x}) = n\left(\frac{\bar{x}}{p} - \frac{1-\bar{x}}{1-p}\right) = n\frac{\bar{x}-p}{p(1-p)}$$

This equals zero when $p = \bar{x}$.

Exercise 15.3. *Check that this is a maximum.*

Thus,

$$\hat{p}(\mathbf{x}) = \bar{x}.$$

In this case the maximum likelihood estimator is also unbiased.

Example 15.4 (Normal data)**.** *Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n normal random variables, we can use the properties of the exponential function to simplify the likelihood function.*

$$\mathbf{L}(\mu,\sigma^2|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_1-\mu)^2}{2\sigma^2}\right)\cdots\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_n-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

The log-likelihood

$$\ln \mathbf{L}(\mu,\sigma^2|\mathbf{x}) = -\frac{n}{2}(\ln 2\pi + \ln\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

The score function is now a vector. $\left(\frac{\partial}{\partial\mu}\ln \mathbf{L}(\mu,\sigma^2|\mathbf{x}), \frac{\partial}{\partial\sigma^2}\ln \mathbf{L}(\mu,\sigma^2|\mathbf{x})\right)$. Next we find the zeros to determine the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$

$$\frac{\partial}{\partial\mu}\ln \mathbf{L}(\hat{\mu},\hat{\sigma}^2|\mathbf{x}) = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu}) = \frac{1}{\hat{\sigma}^2}n(\bar{x}-\hat{\mu}) = 0$$

Because the second partial derivative with respect to μ is negative,

$$\hat{\mu}(\mathbf{x}) = \bar{x}$$

is the maximum likelihood estimator. For the derivative of the log-likelihood with respect to the parameter σ^2,

$$\frac{\partial}{\partial\sigma^2}\ln \mathbf{L}(\mu,\sigma^2|\mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = -\frac{n}{2(\sigma^2)^2}\left(\sigma^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right) = 0.$$

Recalling that $\hat{\mu}(\mathbf{x}) = \bar{x}$, we obtain

$$\hat{\sigma}^2(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Note that the maximum likelihood estimator is a biased estimator.

Example 15.5 (Lincoln-Peterson method of mark and recapture)**.** *Let's recall the variables in mark and recapture:*

- t *be the number captured and tagged,*
- k *be the number in the second capture,*
- r *the the number in the second capture that are tagged, and let*
- N *be the total population.*

Here t and k is set by the experimental design; r is an observation that may vary. The total population N is unknown. The likelihood function for N is the hypergeometric distribution.

$$L(N|r) = \frac{\binom{t}{r}\binom{N-t}{k-r}}{\binom{N}{k}}$$

Exercise 15.6. *Show that the maximum likelihood estimator*

$$\hat{N} = \left[\frac{tk}{r}\right].$$

where $[\cdot]$ mean the greater integer less than*.*

Thus, the maximum likelihood estimator is, in this case, obtained from the method of moments estimator by rounding down to the next integer.

Let look at the example of mark and capture from the previous topic. There $N = 2000$, the number of fish in the population, is unknown to us. We tag $t = 200$ fish in the first capture event, and obtain $k = 400$ fish in the second capture.

```
> N<-2000
> t<-200
> fish<-c(rep(1,t),rep(0,N-t))
```

This creates a vector of length N with t ones representing tagged fish and and $N - t$ zeroes representing the untagged fish.

```
> k<-400
> r<-sum(sample(fish,k))
> r
[1] 42
```

This samples k for the recaptured and adds up the ones to obtained, in this simulation, the number $r = 42$ of recaptured fish. For the likelihood function, we look at a range of values for N that is symmetric about 2000. Here, the maximum likelihood estimate $\hat{N} = [200 \cdot 400/42] = 1904$. ..

```
> N<-c(1800:2200)
> L<-dhyper(r,t,N-t,k)
> plot(N,L,type="l",ylab="L(N|42)",col="green")
```

The likelihood function for this example is shown in Figure 15.2.

Example 15.7 (Linear regression)**.** *Our data are n observations with one explanatory variable and one response variable. The model is that the responses y_i are linearly related to the explanatory variable x_i with an "error" ϵ_i, i.e.,*

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Here we take the ϵ_i to be independent mean 0 normal random variables. The (unknown) variance is σ^2. Consequently, our model has three parameters, the intercept α, the slope β, and the variance of the error, σ^2.

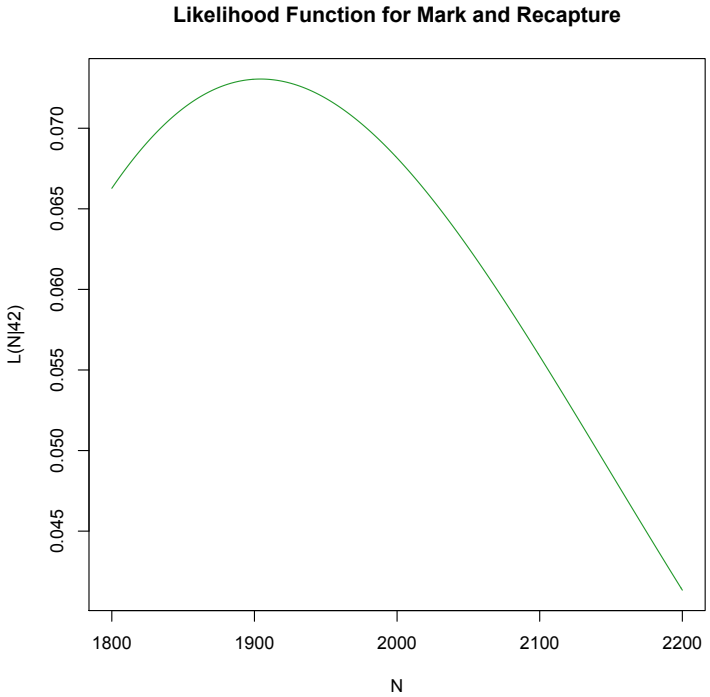


Figure 15.2: Likelihood function $L(N|42)$ for mark and recapture with $t = 200$ tagged fish, $k = 400$ in the second capture with $r = 42$ having tags and thus recapture. Note that the maximum likelihood estimator for the total fish population is $\hat{N} = 1904$.

Thus, the joint density for the ϵ_i is

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{\epsilon_1^2}{2\sigma^2}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{\epsilon_2^2}{2\sigma^2}\cdots\frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{\epsilon_n^2}{2\sigma^2}=\frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2$$

Since $\epsilon_i = y_i - (\alpha + \beta x_i)$, the likelihood function

$$L(\alpha,\beta,\sigma^2|\mathbf{y},\mathbf{x})=\frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2.$$

The logarithm

$$\ln L(\alpha,\beta,\sigma^2|\mathbf{y},\mathbf{x})=-\frac{n}{2}(\ln 2\pi+\ln\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2. \tag{15.3}$$

Consequently, maximizing the likelihood function for the parameters α and β is equivalent to minimizing

$$SS(\alpha.\beta)=\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2.$$

Thus, the principle of maximum likelihood is equivalent to the **least squares criterion** *for ordinary linear regression. The maximum likelihood estimators α and β give the regression line*

$$\hat{y}_i=\hat{\alpha}+\hat{\beta}x_i.$$

with

$$\hat{\beta}=\frac{\text{cov}(x,y)}{\text{var}(x)},\quad \text{and } \hat{\alpha} \text{ determined by solving}\quad \bar{y}=\hat{\alpha}+\hat{\beta}\bar{x}.$$

Exercise 15.8. *Show that the* **maximum likelihood estimator** *for* σ^2 *is*

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{k=1}^{n}(y_i - \hat{y}_i)^2. \tag{15.4}$$

Frequently, software will report the **unbiased estimator**. For ordinary least square procedures, this is

$$\hat{\sigma}^2_U = \frac{1}{n-2}\sum_{k=1}^{n}(y_i - \hat{y}_i)^2.$$

For the measurements on the lengths in centimeters of the femur and humerus for the five specimens of *Archeopteryx*, we have the following R output for linear regression. ..

```
> femur<-c(38,56,59,64,74)
> humerus<-c(41,63,70,72,84)
> summary(lm(humerus~femur))

Call:
lm(formula = humerus ~ femur)

Residuals:
      1       2       3       4       5
-0.8226 -0.3668  3.0425 -0.9420 -0.9110

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.65959    4.45896  -0.821 0.471944
femur        1.19690    0.07509  15.941 0.000537 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.982 on 3 degrees of freedom
Multiple R-squared: 0.9883,Adjusted R-squared: 0.9844
F-statistic: 254.1 on 1 and 3 DF,  p-value: 0.0005368
```

The residual standard error of 1.982 centimeters is obtained by squaring the 5 residuals, dividing by $3 = 5 - 2$ and taking a square root.

Example 15.9 (weighted least squares)**.** *If we know the relative size of the variances of the* ϵ_i*, then we have the model*

$$y_i = \alpha + \beta x_i + \gamma(x_i)\epsilon_i$$

where the ϵ_i *are, again, independent mean 0 normal random variable with unknown variance* σ^2*. In this case,*

$$\epsilon_i = \frac{1}{\gamma(x_i)}(y_i - \alpha + \beta x_i)$$

are independent normal random variables, mean 0 and (unknown) variance σ^2*. the likelihood function*

$$\mathbf{L}(\alpha, \beta, \sigma^2|\mathbf{y}, \mathbf{x}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp -\frac{1}{2\sigma^2}\sum_{i=1}^{n} w(x_i)(y_i - (\alpha + \beta x_i))^2$$

where $w(x) = 1/\gamma(x)^2$*. In other words, the weights are inversely proportional to the variances. The log-likelihood is*

$$\ln \mathbf{L}(\alpha, \beta, \sigma^2|\mathbf{y}, \mathbf{x}) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} w(x_i)(y_i - (\alpha + \beta x_i))^2.$$

Exercise 15.10. *Show that the maximum likelihood estimators $\hat{\alpha}_w$ and $\hat{\beta}_w$ have formulas*

$$\hat{\beta}_w = \frac{\text{cov}_w(x,y)}{\text{var}_w(x)}, \quad \bar{y}_w = \hat{\alpha}_w + \hat{\beta}_w \bar{x}_w$$

where $\bar{x}_w$ and $\bar{y}_w$ are the weighted means

$$\bar{x}_w = \frac{\sum_{i=1}^n w(x_i)x_i}{\sum_{i=1}^n w(x_i)}, \quad \bar{y}_w = \frac{\sum_{i=1}^n w(x_i)y_i}{\sum_{i=1}^n w(x_i)}.$$

The weighted covariance and variance are, respectively,

$$\text{cov}_w(x,y) = \frac{\sum_{i=1}^n w(x_i)(x_i-\bar{x}_w)(y_i-\bar{y}_w)}{\sum_{i=1}^n w(x_i)}, \quad \text{var}_w(x) = \frac{\sum_{i=1}^n w(x_i)(x_i-\bar{x}_w)^2}{\sum_{i=1}^n w(x_i)},$$

The maximum likelihood estimator for σ^2 is

$$\hat{\sigma}^2_{MLE} = \frac{\sum_{k=1}^n w(x_i)(y_i-\hat{y}_i)^2}{\sum_{i=1}^n w(x_i)}.$$

In the case of weighted least squares, the predicted value for the response variable is

$$\hat{y}_i = \hat{\alpha}_w + \hat{\beta}_w x_i.$$

Exercise 15.11. *Show that $\hat{\alpha}_w$ and $\hat{\beta}_w$ are unbiased estimators of α and β. In particular, ordinary (unweighted) least square estimators are unbiased.*

In computing the optimal values using introductory differential calculus, the maximum can occur at either critical points or at the endpoints. The next example show that the maximum value for the likelihood can occur at the end point of an interval.

Example 15.12 (Uniform random variables)**.** *If our data $X = (X_1, \ldots, X_n)$ are a simple random sample drawn from uniformly distributed random variable whose maximum value θ is unknown, then each random variable has density*

$$f(x|\theta) = \begin{cases} 1/\theta & \text{if } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the joint density or the likelihood

$$\mathbf{f}(x|\theta) = \mathbf{L}(\theta|\mathbf{x}) = \begin{cases} 1/\theta^n & \text{if } 0 \le x_i \le \theta \text{ for all } i, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, the joint density is 0 whenever any *of the $x_i > \theta$. Restating this in terms of likelihood, no value of θ is possible that is less than any of the x_i. Conseuently, any value of θ less than any of the x_i has likelihood 0. Symbolically,*

$$\mathbf{L}(\theta|\mathbf{x}) = \begin{cases} 0 & \text{for } \theta < \max_i x_i = x_{(n)}, \\ 1/\theta^n & \text{for } \theta \ge \max_i x_i = x_{(n)}. \end{cases}$$

Recall the notation $x_{(n)}$ for the top **order statistic** *based on n observations.*

The likelihood is 0 on the interval $(0, x_{(n)})$ and is positive and decreasing on the interval $[x_{(n)}, \infty)$. Thus, to maximize $\mathbf{L}(\theta|\mathbf{x})$, we should take the minimum value of θ on this interval. In other words,

$$\hat{\theta}(\mathbf{x}) = x_{(n)}.$$

Because the estimator is always less than the parameter value it is meant to estimate, the estimator

$$\hat{\theta}(X) = X_{(n)} < \theta,$$

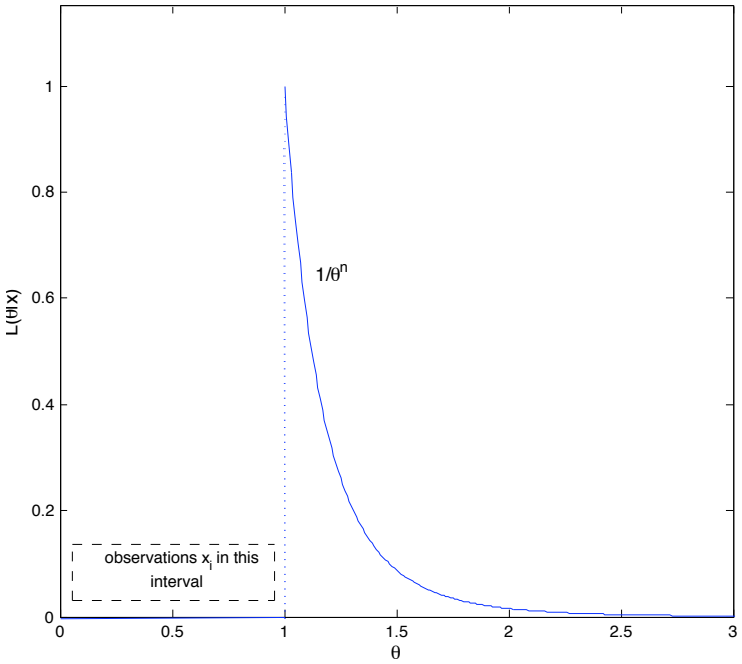


Figure 15.3: Likelihood function for uniform random variables on the interval $[0, \theta]$. The likelihood is 0 up to $\max_{1\le i\le n} x_i$ and $1/\theta^n$ afterwards.

Thus, we suspect it is biased downwards, i. e..

$$E_\theta X_{(n)} < \theta. \tag{15.5}$$

In order to compute the expected value in (15.5), note that $X_{(n)} = \max_{1\le i\le n} X_i \le x$ *if and only if each of the* $X_i \le x$*. Thus, for* $0 \le x \le \theta$*, the distribution function for* $X_{(n)}$ *is*

$$\begin{aligned} F_{X_{(n)}}(x) &= P\{\max_{1\le i\le n} X_i \le x\} = P\{X_1 \le x, X_2 \le x, \dots, X_n \le x\} \\ &= P\{X_1 \le x\}P\{X_2 \le x\}\cdots P\{X_n < x\} \end{aligned}$$

each of these random variables have the same distribution function

$$F_{X_i}(x) = P\{X_i \le x\} = \begin{cases} 0 & \text{for } x \le 0, \\ \frac{x}{\theta} & \text{for } 0 < x \le \theta, \\ 1 & \text{for } \theta < x. \end{cases}$$

Thus, the distribution function for $X_{(n)}$ *is the product* $F_{X_1}(x)F_{X_2}(x)\cdots F_{X_n}(x)$*, i.e.,*

$$F_{X_{(n)}}(x) = \begin{cases} 0 & \text{for } x \le 0, \\ \left(\frac{x}{\theta}\right)^n & \text{for } 0 < x \le \theta, \\ 1 & \text{for } \theta < x. \end{cases}$$

Take the derivative to find the density,

$$f_{X_{(n)}}(x) = \begin{cases} 0 & \text{for } x \le 0, \\ \frac{nx^{n-1}}{\theta^n} & \text{for } 0 < x \le \theta, \\ 0 & \text{for } \theta < x. \end{cases}$$

The mean

$$E_\theta X_{(n)} = \int_0^\theta x f_{X_{(n)}}(x)\, dx = \int_0^\theta x \frac{n x^{n-1}}{\theta^n}\, dx$$
$$= \frac{n}{\theta^n} \int_0^\theta x^n\, dx = \frac{n}{(n+1)\theta^n} x^{n+1}\Big|_0^\theta = \frac{n}{n+1}\theta.$$

This confirms the bias of the estimator $X_{(n)}$ and gives us a strategy to find an unbiased estimator. Note that the choice

$$d(X) = \frac{n+1}{n} X_{(n)}$$

yields an unbiased estimator of θ.

15.3 Summary of Estimators

Look to the text above for the definition of variables.

parameter	estimate	
Bernoulli trials		
p	$\hat{p} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$	unbiased
mark recapture		
N	$\hat{N} = \left[\frac{kt}{r}\right]$	biased upward
normal observations		
μ	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$	unbiased
σ^2	$\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$	biased downward
	$\hat{\sigma}^2_u = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$	unbiased
σ	$\hat{\sigma}_{mle} = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$	biased downward
linear regression		
β	$\hat{\beta} = \frac{\text{cov}(x,y)}{\text{var}(x)}$	unbiased
α	$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$	unbiased
σ^2	$\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x))^2$	biased downward
	$\hat{\sigma}^2_u = \frac{1}{n-2}\sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x))^2$	unbiased
σ	$\hat{\sigma}_{mle} = \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x))^2}$	biased downward
	$\hat{\sigma}_u = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x))^2}$	biased downward
uniform $[0, \theta]$		
θ	$\hat{\theta} = \max_i x_i$	biased downward
	$\hat{\theta} = \frac{n+1}{n}\max_i x_i$	unbiased

15.4 Asymptotic Properties

Much of the attraction of maximum likelihood estimators is based on their properties for large sample sizes. We summarizes some the important properties below, saving a more technical discussion of these properties for later.

1. **Consistency**. If θ_0 is the state of nature and $\hat{\theta}_n(X)$ is the maximum likelihood estimator based on n observations from a simple random sample, then
$$\hat{\theta}_n(X) \to \theta_0 \quad \text{as } n \to \infty.$$
In words, as the number of observations increase, the distribution of the maximum likelihood estimator becomes more and more concentrated about the true state of nature.

2. **Asymptotic normality and efficiency.** Under some assumptions that allows, among several analytical properties, the use of a central limit theorem holds. Here we have
$$\sqrt{n}(\hat{\theta}_n(X) - \theta_0)$$
converges in distribution as $n \to \infty$ to a normal random variable with mean 0 and variance $1/I(\theta_0)$, the Fisher information for one observation. Thus,
$$\text{Var}_{\theta_0}(\hat{\theta}_n(X)) \approx \frac{1}{nI(\theta_0)},$$
the lowest variance possible under the Crámer-Rao lower bound. This property is called **asymptotic efficiency**. We can write this in terms of the z-score. Let
$$Z_n = \frac{\hat{\theta}(X) - \theta_0}{1/\sqrt{nI(\theta_0)}}.$$
Then, as with the central limit theorem, Z_n converges in distribution to a standard normal random variable.

3. **Properties of the log likelihood surface**. For large sample sizes, the variance of a maximum likelihood estimator of a single parameter is approximately the reciprocal of the the Fisher information
$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} \ln L(\theta|X)\right].$$
The Fisher information can be approximated by the **observed information** based on the data $\mathbf{x}$,
$$J(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2} \ln L(\hat{\theta}(\mathbf{x})|\mathbf{x}),$$
giving the negative of the curvature of the log-likelihood surface at the maximum likelihood estimate $\hat{\theta}(\mathbf{x})$. If the curvature is small near the maximum likelihood estimator, then the likelihood surface is nearty flat and the variance is large. If the curvature is large, the likelihood decreases quickly at the maximum and thus the variance is small.

We now look at these properties in some detail by revisiting the example of the distribution of fitness effects. For this example, we have two parameters - α and β for the gamma distribution and so, we will want to extend the properties above to circumstances in which we are looking to estimate more than one parameter.

15.5 Multidimensional Estimation

For a multidimensional parameter space $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, the Fisher information $I(\theta)$ is now a matrix . As with one-dimensional case, the ij-th entry has two alternative expressions, namely,
$$I(\theta)_{ij} = E_\theta\left[\frac{\partial}{\partial\theta_i} \ln L(\theta|X) \frac{\partial}{\partial\theta_j} \ln L(\theta|X)\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln L(\theta|X)\right].$$

Rather than taking reciprocals to obtain an estimate of the variance, we find the matrix inverse $I(\theta)^{-1}$. This inverse will provide estimates of both variances and covariances. To be precise, for n observations, let $\hat{\theta}_{i,n}(X)$ be the maximum likelihood estimator of the i-th parameter. Then

$$\mathrm{Var}_\theta(\hat{\theta}_{i,n}(X)) \approx \frac{1}{n} I(\theta)^{-1}_{ii} \qquad \mathrm{Cov}_\theta(\hat{\theta}_{i,n}(X), \hat{\theta}_{j,n}(X)) \approx \frac{1}{n} I(\theta)^{-1}_{ij}.$$

When the i-th parameter is θ_i, the asymptotic normality and efficiency can be expressed by noting that the z-score

$$Z_{i,n} = \frac{\hat{\theta}_i(X) - \theta_i}{I(\theta)^{-1}_{ii}/\sqrt{n}}.$$

is approximately a standard normal. As we saw in one dimension, we can replace the information matrix with the observed information matrix,

$$J(\hat{\theta})_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln L(\hat{\theta}(\mathbf{x})|\mathbf{x}).$$

Example 15.13. *To obtain the maximum likelihood estimate for the gamma family of random variables, write the likelihood*

$$\mathbf{L}(\alpha, \beta|\mathbf{x}) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)} x_1^{\alpha-1} e^{-\beta x_1}\right) \cdots \left(\frac{\beta^\alpha}{\Gamma(\alpha)} x_n^{\alpha-1} e^{-\beta x_n}\right) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n (x_1 x_2 \cdots x_n)^{\alpha-1} e^{-\beta(x_1+x_2+\cdots+x_n)}.$$

and its logarithm

$$\ln \mathbf{L}(\alpha, \beta|\mathbf{x}) = n(\alpha \ln \beta - \ln \Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^n \ln x_i - Whendoyouwanttocomeover?\beta \sum_{i=1}^n x_i.$$

To determine the parameters that maximize the likelihood, we solve the equations

$$\frac{\partial}{\partial\alpha} \ln \mathbf{L}(\hat{\alpha}, \hat{\beta}|\mathbf{x}) = n(\ln \hat{\beta} - \frac{d}{d\alpha} \ln \Gamma(\hat{\alpha})) + \sum_{i=1}^n \ln x_i = 0$$

and

$$\frac{\partial}{\partial\beta} \ln \mathbf{L}(\hat{\alpha}, \hat{\beta}|\mathbf{x}) = n\frac{\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^n x_i = 0, \quad or \quad \bar{x} = \frac{\hat{\alpha}}{\hat{\beta}}.$$

Recall that the mean μ of a gamma distribution is α/β. Thus. by the invariance property of maximum likelihood estimators

$$\hat{\mu} = \frac{\hat{\alpha}}{\hat{\beta}} = \bar{x},$$

and the sample mean is the maximum likelihood estimate for the distributional mean.

Substituting $\hat{\beta} = \hat{\alpha}/\bar{x}$ into the first equation results the following relationship for $\hat{\alpha}$

$$n(\ln \hat{\alpha} - \ln \bar{x} - \frac{d}{d\alpha} \ln \Gamma(\hat{\alpha})) + \sum_{i=1}^n \ln x_i = 0$$

which can be solved numerically. The derivative of the logarithm of the gamma function

$$\psi(\alpha) = \frac{d}{d\alpha} \ln \Gamma(\alpha)$$

is know as the **digamma function** *and is called in* R *with* `digamma`.

For the example for the distribution of fitness effects $\alpha = 0.23$ and $\beta = 5.35$ with $n = 100$, a simulated data set yields $\hat{\alpha} = 0.2376$ and $\hat{\beta} = 5.690$ for maximum likelihood estimator. (See Figure 15.4.)

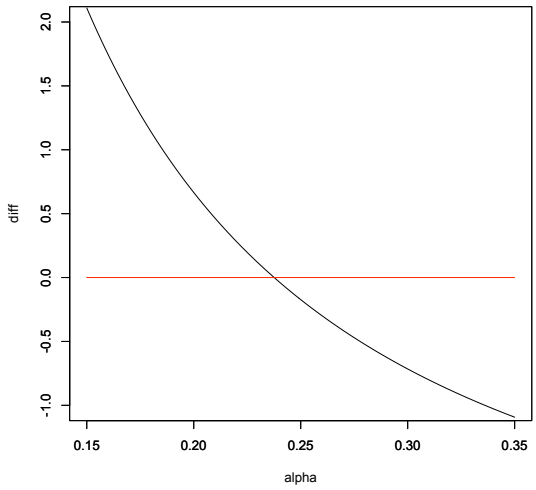


Figure 15.4: The graph of $n(\ln\hat{\alpha} - \ln\bar{x} - \frac{d}{d\alpha}\ln\Gamma(\hat{\alpha})) + \sum_{i=1}^{n}\ln x_i$ crosses the horizontal axis at $\hat{\alpha} = 0.2376$. The fact that the graph of the derivative is decreasing states that the score function moves from increasing to decreasing with α and confirming that $\hat{\alpha}$ is a maximum.

To determine the variance of these estimators, we first compute the Fisher information matrix. Taking the appropriate derivatives, we find that each of the second order derivatives are constant and thus the expected values used to determine the entries for Fisher information matrix are the negative of these constants.

$$I(\alpha,\beta)_{11} = -\frac{\partial^2}{\partial\alpha^2}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}) = n\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha), \quad I(\alpha,\beta)_{22} = -\frac{\partial^2}{\partial\beta^2}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}) = n\frac{\alpha}{\beta^2},$$

$$I(\alpha,\beta)_{12} = -\frac{\partial^2}{\partial\alpha\partial\beta}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}) = -n\frac{1}{\beta}.$$

This give a Fisher information matrix

$$I(\alpha,\beta) = n\begin{pmatrix} \frac{d^2}{d\alpha^2}\ln\Gamma(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}.$$

The second derivative of the logarithm of the gamma function

$$\psi_1(\alpha) = \frac{d^2}{d\alpha^2}\ln\Gamma(\alpha)$$

is known as the **trigamma function** *and is called in* R *with* `trigamma`.

The inverse

$$I(\alpha,\beta)^{-1} = \frac{1}{n\alpha(\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha) - 1)}\begin{pmatrix} \alpha & \beta \\ \beta & \beta^2\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha) \end{pmatrix}.$$

For the example for the distribution of fitness effects $\alpha = 0.23$ *and* $\beta = 5.35$ *and* $n = 100$*, and*

$$I(0.23, 5.35)^{-1} = \frac{1}{100(0.23)(19.12804)}\begin{pmatrix} 0.23 & 5.35 \\ 5.35 & 5.35^2(20.12804) \end{pmatrix} = \begin{pmatrix} 0.0001202 & 0.01216 \\ 0.01216 & 1.3095 \end{pmatrix}.$$

$$\text{Var}_{(0.23,5.35)}(\hat{\alpha}) \approx 0.0001202, \quad \text{Var}_{(0.23,5.35)}(\hat{\beta}) \approx 1.3095.$$

$$\sigma_{(0.23,5.35)}(\hat{\alpha}) \approx 0.0110, \quad \sigma_{(0.23,5.35)}(\hat{\beta}) \approx 1.1443.$$

Compare this to the empirical values of 0.0662 *and* 2.046 *for the method of moments. This gives the following table of standard deviations for* $n = 100$ *observation*

method	$\hat{\alpha}$	$\hat{\beta}$
maximum likelihood	0.0110	1.1443
method of moments	0.0662	2.046
ratio	0.166	0.559

Thus, the standard deviation for the maximum likelihood estimator is respectively 17% and 56% that of method of moments estimator. We will look at the impact as we move on to our next topic - interval estimation and the confidence intervals.

Exercise 15.14. *If the data are a simple random sample of 100 observations of a* $\Gamma(0.23, 5.35)$ *random variable. Use the approximate normality of maximum likelihood estimators to estimate*

$$P\{\hat{\alpha} \geq 0.2376\} \qquad P\{\hat{\beta} \geq 5.690\}.$$

15.6 Comparison of Estimation Procedures

For n observations, $x_1, x_2, \ldots x_n$ and a single parameter θ:

	method of moments	maximum likelihood
estimate	If $\mu = k(\theta)$, then $\hat{\theta} = g(\bar{x})$, where $g = k^{-1}$.	$\hat{\theta} = \arg\max_\theta L(\theta\|\mathbf{x})$
variance	$\approx g'(\mu)^2 \frac{\sigma^2}{n}$, delta method	$\approx \frac{1}{nI(\theta)}$, Fisher information
bias	$\approx g''(\mu)\frac{\sigma^2}{n}$	*

* If g is continuous at μ, then both estimators are consistent. For a vector of parameters, we will need to perform a higher dimenstional delta method or invert the Fisher information matrix to estimate variance.

15.7 Choice of Estimators

With all of the desirable properties of the maximum likelihood estimator, the question arises as to why would one choose a method of moments estimator?

One answer is that the use maximum likelihood techniques relies on knowing the density function explicitly. Moreover, the form of the density must be amenable to the analysis necessary to maximize the likelihood and find the Fisher information.

However, much less about the experiment is need in order to compute moments. Thus far, we have computed moments using the density

$$E_\theta X^m = \int_{-\infty}^{\infty} x^m f_X(x|\theta)\, dx.$$

However, consider the case of determining parameters in the distribution in the number of proteins in a tissue. If the tissue has several cell types, then we would need

- the distribution of cell types, and
- a density function for the number of proteins in each cell type.

These two pieces of information can be used to calculate the mean and variance for the number of cells with some ease. However, giving an explicit expression for the density and hence the likelihood function is more difficult to obtain. This leads to quite intricate computations to carry out the desired analysis of the likelihood function.

15.8 Technical Aspects

We can use concepts previously introduced to obtain the properties for the maximum likelihood estimator. For example, θ_0 is more likely that a another parameter value θ

$$\mathbf{L}(\theta_0|X) > \mathbf{L}(\theta|X) \quad \text{if and only if} \quad \frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(X_i|\theta_0)}{f(X_i|\theta)} > 0.$$

By the strong law of large numbers, this sum converges to

$$E_{\theta_0}\left[\ln \frac{f(X_1|\theta_0)}{f(X_1|\theta)}\right].$$

which is greater than 0. thus, for a large number of observations and a given value of θ, then with a probability nearly one, $\mathbf{L}(\theta_0|X) > \mathbf{L}(\theta|X)$ and so the maximum likelihood estimator has a high probability of being very near θ_0. This is a statement of the **consistency** of the estimator.

For the asymptotic normality and efficiency, we write the linear approximation of the score function

$$\frac{d}{d\theta}\ln L(\theta|X) \approx \frac{d}{d\theta}\ln L(\theta_0|X) + (\theta-\theta_0)\frac{d^2}{d\theta^2}\ln L(\theta_0|X).$$

Now substitute $\theta = \hat{\theta}$ and note that $\frac{d}{d\theta}\ln L(\hat{\theta}|X) = 0$. Then

$$\sqrt{n}(\hat{\theta}_n(X) - \theta_0) \approx -\sqrt{n}\frac{\frac{d}{d\theta}\ln L(\theta_0|X)}{\frac{d^2}{d\theta^2}\ln L(\theta_0|X)} = \frac{\frac{1}{\sqrt{n}}\frac{d}{d\theta}\ln L(\theta_0|X)}{-\frac{1}{n}\frac{d^2}{d\theta^2}\ln L(\theta_0|X)}$$

Now assume that θ_0 is the true state of nature. Then, the random variables $d\ln f(X_i|\theta_0)/d\theta$ are independent with mean 0 and variance $I(\theta_0)$. Thus, the distribution of numerator

$$\frac{1}{\sqrt{n}}\frac{d}{d\theta}\ln L(\theta_0|X) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d}{d\theta}\ln f(X_i|\theta_0)$$

converges, by the central limit theorem, to a normal random variable with mean 0 and variance $I(\theta_0)$. For the denominator, $-d^2\ln f(X_i|\theta_0)/d\theta^2$ are independent with mean $I(\theta_0)$. Thus,

$$-\frac{1}{n}\frac{d^2}{d\theta^2}\ln L(\theta_0|X) = -\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\ln f(X_i|\theta_0)$$

converges, by the law of large numbers, to $I(\theta_0)$. Thus, the distribution of the ratio, $\sqrt{n}(\hat{\theta}_n(X) - \theta_0)$, converges to a normal random variable with variance $I(\theta_0)/I(\theta_0)^2 = 1/I(\theta_0)$.

15.9 Answers to Selected Exercises

15.3. We have found that the score function

$$\frac{\partial}{\partial p}\ln \mathbf{L}(p|\mathbf{x}) = n\frac{\bar{x}-p}{p(1-p)}$$

Thus

$$\frac{\partial}{\partial p}\ln \mathbf{L}(p|\mathbf{x}) > 0 \quad \text{if } p < \bar{x}, \quad \text{and} \quad \frac{\partial}{\partial p}\ln \mathbf{L}(p|\mathbf{x}) < 0 \quad \text{if } p > \bar{x}$$

In words, $\ln \mathbf{L}(p|\mathbf{x})$ is increasing for $p < \bar{x}$ and decreasing for $p > \bar{x}$. Thus, $\hat{p}(\mathbf{x}) = \bar{x}$ is a maximum.

15.7. The log-likelihood function

$$\ln L(\alpha, \beta, \sigma^2 | \mathbf{y}, \mathbf{x}) = -\frac{n}{2}(\ln(2\pi) + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

leads to the ordinary least squares equations for the maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$. Take the partial derivative with respect to σ^2,

$$\frac{\partial}{\partial \sigma^2} L(\alpha, \beta, \sigma^2 | \mathbf{y}, \mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

This partial derivative is 0 at the maximum likelihood estimates $\hat{\sigma}^2, \hat{\alpha}$ and $\hat{\beta}$.

$$0 = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2$$

or

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2.$$

We would like to maximize the likelihood given the number of recaptured individuals r. Because the domain for N is the nonnegative integers, we cannot use calculus. However, we can look at the ratio of the likelihood values for successive value of the total population.

$$\frac{L(N|r)}{L(N-1|r)}$$

N is more likely that $N-1$ precisely when this ratio is larger than one. The computation below will show that this ratio is greater than 1 for small values of N and less than one for large values. Thus, there is a place in the middle which has the maximum. We expand the binomial coefficients in the expression for $L(N|r)$ and simplify.

$$\begin{aligned}
\frac{L(N|r)}{L(N-1|r)} &= \frac{\binom{t}{r}\binom{N-t}{k-r} / \binom{N}{k}}{\binom{t}{r}\binom{N-t-1}{k-r} / \binom{N-1}{k}} = \frac{\binom{N-t}{k-r}\binom{N-1}{k}}{\binom{N-t-1}{k-r}\binom{N}{k}} = \frac{\frac{(N-t)!}{(k-r)!(N-t-k+r)!} \frac{(N-1)!}{k!(N-k-1)!}}{\frac{(N-t-1)!}{(k-r)!(N-t-k+r-1)!} \frac{N!}{k!(N-k)!}} \\
&= \frac{(N-t)!(N-1)!(N-t-k+r-1)!(N-k)!}{(N-t-1)!N!(N-t-k+r)!(N-k-1)!} = \frac{(N-t)(N-k)}{N(N-t-k+r)}.
\end{aligned}$$

Thus, the ratio

$$\frac{L(N|r)}{L(N-1|r)} = \frac{(N-t)(N-k)}{N(N-t-k+r)}$$

exceeds 1if and only if

$$\begin{aligned}
(N-t)(N-k) &> N(N-t-k+r) \\
N^2 - tN - kN + tk &> N^2 - tN - kN + rN \\
tk &> rN \\
\frac{tk}{r} &> N
\end{aligned}$$

Writing $[x]$ for the integer part of x, we see that $L(N|r) > L(N-1|r)$ for $N < [tk/r]$ and $L(N|r) \leq L(N-1|r)$ for $N \geq [tk/r]$. This give the maximum likelihood estimator

$$\hat{N} = \left[\frac{tk}{r} \right].$$

15.8 .Take the derivative with respect to σ^2 in (15.3)

$$\frac{\partial}{\partial \sigma^2} \ln L(\alpha, \beta, \sigma^2 | \mathbf{y}, \mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

Now set this equal to zero, substitute $\hat{\alpha}$ for α, $\hat{\beta}$ for β and solve for σ^2 to obtain (15.4).

15.9. The maximum likelihood principle leads to a minimization problem for

$$SS_w(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} w(x_i)(y_i - (\alpha + \beta x_i))^2.$$

Following the steps to derive the equations for ordinary least squares, take partial derivatives to find that

$$\frac{\partial}{\partial \beta} SS_w(\alpha, \beta) = -2 \sum_{i=1}^{n} w(x_i) x_i (y_i - \alpha - \beta x_i) \quad \frac{\partial}{\partial \alpha} SS_w(\alpha, \beta) = -2 \sum_{i=1}^{n} w(x_i)(y_i - \alpha - \beta x_i).$$

Set these two equations equal to 0 and call the solutions $\hat{\alpha}_w$ and $\hat{\beta}_w$.

$$0 = \sum_{i=1}^{n} w(x_i) x_i (y_i - \hat{\alpha}_w - \hat{\beta}_w x_i) = \sum_{i=1}^{n} w(x_i) x_i y_i - \hat{\alpha}_w \sum_{i=1}^{n} w(x_i) x_i - \hat{\beta}_w \sum_{i=1}^{n} w(x_i) x_i^2 \tag{15.6}$$

$$0 = \sum_{i=1}^{n} w(x_i)(y_i - \hat{\alpha}_w - \hat{\beta}_w x_i) = \sum_{i=1}^{n} w(x_i) y_i - \hat{\alpha}_w \sum_{i=1}^{n} w(x_i) - \hat{\beta}_w \sum_{i=1}^{n} w(x_i) x_i \tag{15.7}$$

Multiply these equations by the appropriate factors to obtain

$$\begin{aligned} 0 = &\left(\sum_{i=1}^{n} w(x_i) \right) \left(\sum_{i=1}^{n} w(x_i) x_i y_i \right) - \hat{\alpha}_w \left(\sum_{i=1}^{n} w(x_i) \right) \left(\sum_{i=1}^{n} w(x_i) x_i \right) \\ &- \hat{\beta}_w \left(\sum_{i=1}^{n} w(x_i) \right) \left(\sum_{i=1}^{n} w(x_i) x_i^2 \right) \end{aligned} \tag{15.8}$$

$$0 = \left(\sum_{i=1}^{n} w(x_i) x_i \right) \left(\sum_{i=1}^{n} w(x_i) y_i \right) - \hat{\alpha}_w \left(\sum_{i=1}^{n} w(x_i) \right) \left(\sum_{i=1}^{n} w(x_i) x_i \right) - \hat{\beta}_w \left(\sum_{i=1}^{n} w(x_i) x_i \right)^2 \tag{15.9}$$

Now subtract the equation (15.9) from equation (15.8) and solve for $\hat{\beta}$.

$$\begin{aligned} \hat{\beta} &= \frac{\left(\sum_{i=1}^{n} w(x_i)\right) \left(\sum_{i=1}^{n} w(x_i) x_i y_i\right) - \left(\sum_{i=1}^{n} w(x_i) x_i\right) \left(\sum_{i=1}^{n} w(x_i) y_i\right)}{n \sum_{i=1}^{n} w(x_i) x_i^2 - \left(\sum_{i=1}^{n} w(x_i) x_i\right)^2} \\ &= \frac{\sum_{i=1}^{n} w(x_i)(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^{n} w(x_i)(x_i - \bar{x}_w)^2} = \frac{\text{cov}_w(x, y)}{\text{var}_w(x)}. \end{aligned}$$

Next, divide equation (15.9) by $\sum_{i=1}^{n} w(x_i)$ to obtain

$$\bar{y}_w = \hat{\alpha}_w + \hat{\beta}_w \bar{x}_w. \tag{15.10}$$

15.10. Because the ϵ_i have mean zero,

$$E_{(\alpha,\beta)}y_i = E_{(\alpha,\beta)}[\alpha + \beta x_i + \gamma(x_i)\epsilon_i] = \alpha + \beta x_i + \gamma(x_i)E_{(\alpha,\beta)}[\epsilon_i] = \alpha + \beta x_i.$$

Next, use the linearity property of expectation to find the mean of $\bar{y}_w$.

$$E_{(\alpha,\beta)}\bar{y}_w = \frac{\sum_{i=1}^n w(x_i)E_{(\alpha,\beta)}y_i}{\sum_{i=1}^n w(x_i)} = \frac{\sum_{i=1}^n w(x_i)(\alpha+\beta x_i)}{\sum_{i=1}^n w(x_i)} = \alpha + \beta\bar{x}_w. \tag{15.11}$$

Taken together, we have that $E_{(\alpha,\beta)}[y_i - \bar{y}_w] = (\alpha+\beta x_i.) - (\alpha+\beta x_i) = \beta(x_i - \bar{x}_w)$. To show that $\hat{\beta}_w$ is an unbiased estimator, we see that

$$\begin{aligned} E_{(\alpha,\beta)}\hat{\beta}_w &= E_{(\alpha,\beta)}\left[\frac{\text{cov}_w(x,y)}{\text{var}_w(x)}\right] = \frac{E_{(\alpha,\beta)}[\text{cov}_w(x,y)]}{\text{var}_w(x)} = \frac{1}{\text{var}_w(x)}E_{(\alpha,\beta)}\left[\frac{\sum_{i=1}^n w(x_i)(x_i-\bar{x}_w)(y_i-\bar{y}_w)}{\sum_{i=1}^n w(x_i)}\right] \\ &= \frac{1}{\text{var}_w(x)}\frac{\sum_{i=1}^n w(x_i)(x_i-\bar{x}_w)E_{(\alpha,\beta)}[y_i-\bar{y}_w]}{\sum_{i=1}^n w(x_i)} = \frac{\beta}{\text{var}_w(x)}\frac{\sum_{i=1}^n w(x_i)(x_i-\bar{x}_w)(x_i-\bar{x}_w)}{\sum_{i=1}^n w(x_i)} = \beta. \end{aligned}$$

To show that $\hat{\alpha}_w$ is an unbiased estimator, recall that $\bar{y}_w = \hat{\alpha}_w + \hat{\beta}_w\bar{x}_w$. Thus

$$E_{(\alpha,\beta)}\hat{\alpha}_w = E_{(\alpha,\beta)}[\bar{y}_w - \hat{\beta}_w\bar{x}_w] = E_{(\alpha,\beta)}\bar{y}_w - E_{(\alpha,\beta)}[\hat{\beta}_w]\bar{x}_w = \alpha + \beta\bar{x}_w - \beta\bar{x}_w = \alpha,$$

using (15.11) and the fact that $\hat{\beta}_w$ is an unbiased estimator of β

15.14. For $\hat{\alpha}$, we have the z-score

$$z_{\hat{\alpha}} = \frac{\hat{\alpha} - 0.23}{\sqrt{0.0001202}} \geq \frac{0.2376 - 0.23}{\sqrt{0.0001202}} = 0.6841.$$

Thus, using the normal approximation,

$$P\{\hat{\alpha} \geq 0.2367\} = P\{z_{\hat{\alpha}} \geq 0.6841\} = 0.2470.$$

For $\hat{\beta}$, we have the z-score

$$z_{\hat{\beta}} = \frac{\hat{\beta} - 5.35}{\sqrt{1.3095}} \geq \frac{5.690 - 5.35}{\sqrt{1.3095}} = 0.2971.$$

Here, the normal approximation gives

$$P\{\hat{\beta} \geq 5.690\} = P\{z_{\hat{\beta}} \geq 0.2971\} = 0.3832.$$

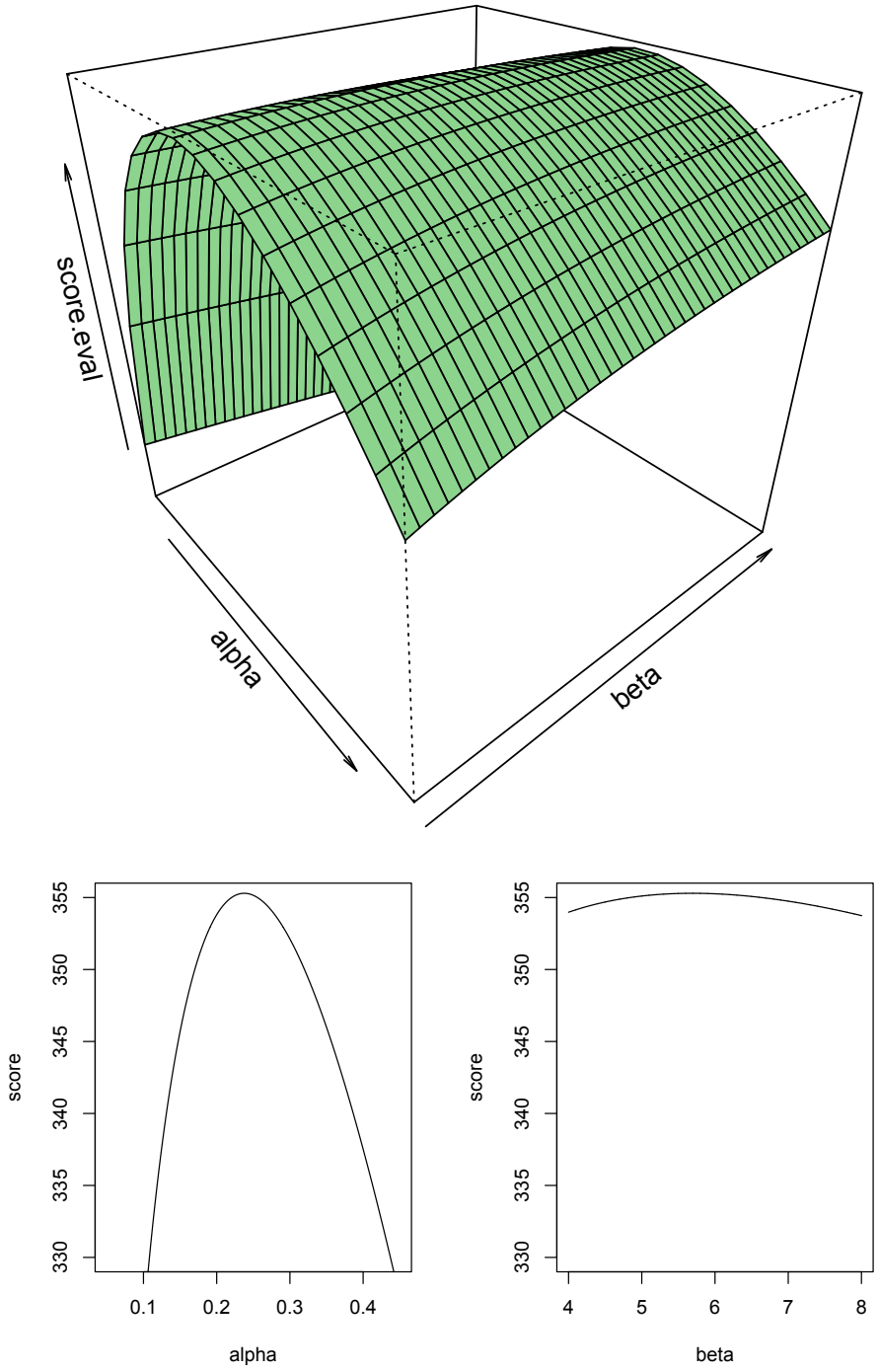


Figure 15.5: (**top**) The log-likelihood near the maximum likelihood estimators. The domain is $0.1 \leq \alpha \leq 0.4$ and $4 \leq \beta \leq 8$. (**bottom**) Graphs of vertical slices through the log-likelihood function surface. (left) $\hat{\alpha} = 0.2376$ and $4 \leq \beta \leq 8$ varies. (right) $\hat{\beta} = 5.690$ and $0.1 \leq \alpha \leq 0.4$. The variance of the estimator is approximately the negative reciprocal of the second derivative of the log-likelihood function at the maximum likelihood estimators (known as the observed information). Note that the log-likelihood function is nearly flat as β varies. This leads to the interpretation that a range of values for β are nearly equally likely and that the variance for the estimator for $\hat{\beta}$ will be high. On the other hand, the log-likelihood function has a much greater curvature for the α parameter and the estimator $\hat{\alpha}$ will have a much smaller variance than $\hat{\beta}$

Topic 16

Interval Estimation

> *The form of this solution consists in determining certain intervals, which I propose to call the confidence intervals..., in which we may assume are contained the values of the estimated characters of the population, the probability of an error is a statement of this sort being equal to or less than* $1 - \epsilon$*, where* ϵ *is any number* $0 < \epsilon < 1$*, chosen in advance. The number* ϵ *I call the confidence coefficient.* - Jerzy Neyman, 1934, On the Two Different Aspects of the Representative Method, *Journal of the Royal Statistical Society*

Our strategy to estimation thus far has been to use a method to find an estimator, e.g., method of moments, or maximum likelihood, and evaluate the quality of the estimator by evaluating the bias and the variance of the estimator. Often, we know more about the distribution of the estimator and this allows us to take a more comprehensive statement about the estimation procedure.

Interval estimation is an alternative to the variety of techniques we have examined. Given data $\mathbf{x}$, we replace the point estimate $\hat{\theta}(\mathbf{x})$ for the parameter θ by a statistic that is subset $\hat{C}(\mathbf{x})$ of the parameter space. We will consider both the classical and Bayesian approaches to choosing $\hat{C}(\mathbf{x})$. As we shall learn, the two approaches have very different interpretations.

16.1 Classical Statistics

In this case, the random set $\hat{C}(X)$ is chosen to have a prescribed high probability, γ, of containing the true parameter value θ. In symbols,

$$P_\theta\{\theta \in \hat{C}(X)\} = \gamma.$$

In this case, the set $\hat{C}(\mathbf{x})$ is called a γ-**level confidence set**. In the case of a one dimensional parameter set, the typical choice of confidence set is a **confidence interval**

$$\hat{C}(\mathbf{x}) = (\hat{\theta}_\ell(\mathbf{x}), \hat{\theta}_u(\mathbf{x})).$$

Often this interval takes the form

$$\hat{C}(\mathbf{x}) = (\hat{\theta}(\mathbf{x}) - m(\mathbf{x}), \hat{\theta}(\mathbf{x}) + m(\mathbf{x})) = \hat{\theta}(\mathbf{x}) \pm m(\mathbf{x})$$

where the two statistics,

- $\hat{\theta}(\mathbf{x})$ is a **point estimate**, and
- $m(\mathbf{x})$ is the **margin of error**.

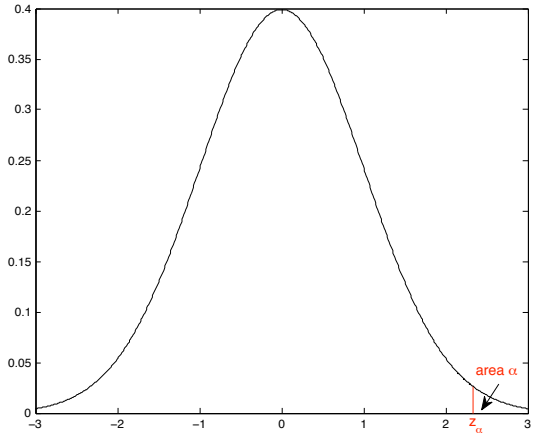


Figure 16.1: Upper tail critical values. α is the area under the standard normal density and to the right of the vertical line at critical value z_α

16.1.1 Means

Example 16.1 (1-sample z interval)**.** *If $X_1.X_2.\ldots X_n$ are normal random variables with unknown mean μ but known variance σ_0^2. Then,*

$$Z = \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}$$

is a standard normal random variable. For any α between 0 and 1, let z_α satisfy

$$P\{Z > z_\alpha\} = \alpha \quad \textit{or equivalently} \quad P\{Z \leq z_\alpha\} = 1 - \alpha.$$

The value is known as the **upper tail probability with critical value** z_α*. We can compute this in* R *using, for example*

```
> qnorm(0.975)
[1] 1.959964
```

for $\alpha = 0.025$.

If $\gamma = 1 - 2\alpha$, then $\alpha = (1-\gamma)/2$. In this case, we have that

$$P\{-z_\alpha < Z < z_\alpha\} = \gamma.$$

Let μ_0 is the state of nature. Taking in turn each the two inequalities in the line above and isolating μ_0, we find that

$$\begin{aligned} \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} = Z &< z_\alpha \\ \bar{X} - \mu_0 &< z_\alpha \frac{\sigma_0}{\sqrt{n}} \\ \bar{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}} &< \mu_0 \end{aligned}$$

Similarly,

$$\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} = Z > -z_\alpha$$

implies

$$\mu_0 < \bar{X} + z_\alpha \frac{\sigma_0}{\sqrt{n}}$$

Thus

$$\bar{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}} < \mu_0 < \bar{X} + z_\alpha \frac{\sigma_0}{\sqrt{n}}.$$

has probability γ. Thus, for data $\mathbf{x}$,

$$\bar{x} \pm z_{(1-\gamma)/2} \frac{\sigma_0}{\sqrt{n}}$$

is a confidence interval with confidence level γ. In this case,

$$\hat{\mu}(\mathbf{x}) = \bar{x} \textit{ is the estimate for the mean and } m(\mathbf{x}) = z_{(1-\gamma)/2}\sigma_0/\sqrt{n} \textit{ is the margin of error.}$$

*We can use the z-***interval** *above for the confidence interval for μ for data that is not necessarily normally distributed as long as the central limit theorem applies. For one population tests for means, $n > 30$ and data not strongly skewed is a good rule of thumb.*

Generally, the standard deviation is not known and must be estimated. So, let $X_1, X_2, \cdots, X_n$ be normal random variables with unknown mean and unknown standard deviation. Let S^2 be the **unbiased** sample variance. If we are forced to replace the unknown variance σ^2 with its unbiased estimate s^2, then the statistic is known as t:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

The term $s/\sqrt{n}$ which estimates the standard deviation of the sample mean is called the **standard error**. The remarkable discovery by William Gossett is that the distribution of the t statistic can be determined **exactly**. Write

$$T_{n-1} = \frac{\sqrt{n}(\bar{X} - \mu)}{S}.$$

Then, Gossett was able to establish the following three facts:

- The numerator is a standard normal random variable.
- The denominator is the square root of

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

  The sum has chi-square distribution with $n-1$ degrees of freedom.
- The numerator and denominator are **independent**.

With this, Gossett was able to compute the density of the t **distribution** with $n-1$ **degrees of freedom**. Gossett, who worked for the the brewery of Arthur Guinness in Dublin, was permitted to publish his results only if it appeared under a pseudonym. Gosset chose the name *Student*, thus the distribution is sometimes known as **Student's** t.

Again, for any α between 0 and 1, let **upper tail probability** $t_{n-1,\alpha}$ satisfy

$$P\{T_{n-1} > t_{n-1,\alpha}\} = \alpha \quad \text{or equivalently} \quad P\{T_{n-1} \le t_{n-1,\alpha}\} = 1 - \alpha.$$

We can compute this in R using, for example

```
> qt(0.975,12)
[1] 2.178813
```

for $\alpha = 0.025$ and $n - 1 = 12$.

Example 16.2. *For the data on the lengths of 200 Bacillus subtilis, we had a mean* $\bar{x} = 2.49$ *and standard deviation* $s = 0.674$. *For a 96% confidence interval* $\alpha = 0.02$ *and we type in* R,

```
> qt(0.98,199)
[1] 2.067298
```

Thus, the interval is

$$2.490 \pm 2.0674\frac{0.674}{\sqrt{200}} = 2.490 \pm 0.099 \quad \textit{or} \quad (2.391, 2.589)$$

Example 16.3. *We can obtain the data for the Michaelson-Morley experiment using* R *by typing*

```
> data(morley)
```

The data have 100 rows - 5 experiments (column 1) of 20 runs (column 2). The `Speed` *is in column 3. The values for speed are the amounts over 299,000 km/sec. Thus, a* t*-confidence interval will have 99 degrees of freedom. We can see a histogram by writing* `hist(morley$Speed)`. *To determine a 95% confidence interval, we find*

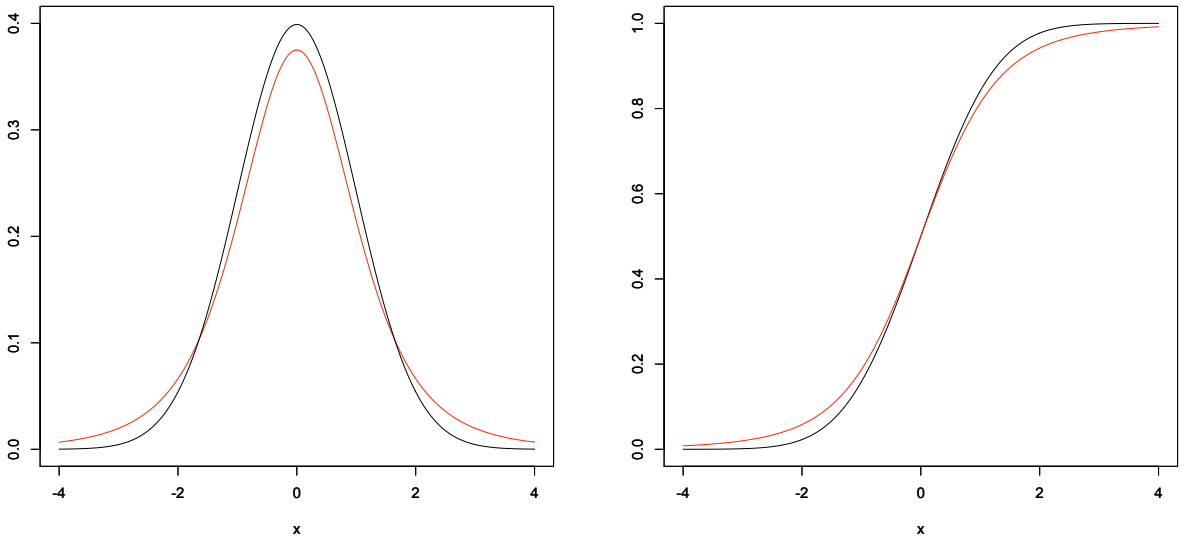


Figure 16.2: The density and distribution function for a standard normal random variable (black) and a t random variable with 4 degrees of freedom (red). The variance of the t distribution is $df/(df-2) = 4/(4-2) = 2$ is higher than the variance of a standard normal. This can be seen in the broader shoulders of the t density function or in the smaller increases in the t distribution function away from the mean of 0.

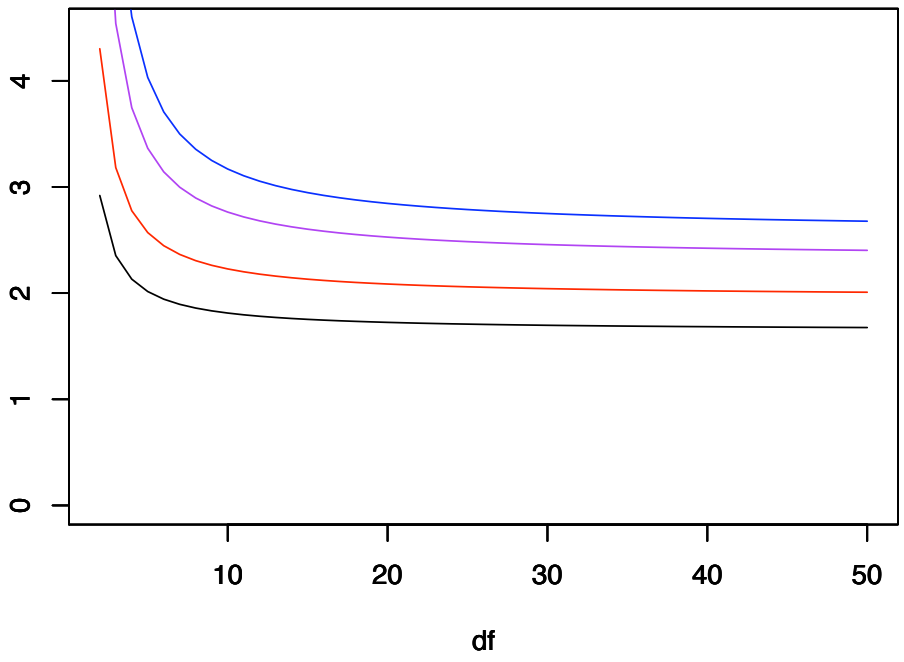


Figure 16.3: Upper critical values for the t confidence interval with $\gamma = 0.90$ (black), 0.95 (red), 0.98 (magenta) and 0.99 (blue) as a function of df, the number of degrees of freedom. Note that these critical values decrease to the critical value for the z confidence interval and increases with γ.

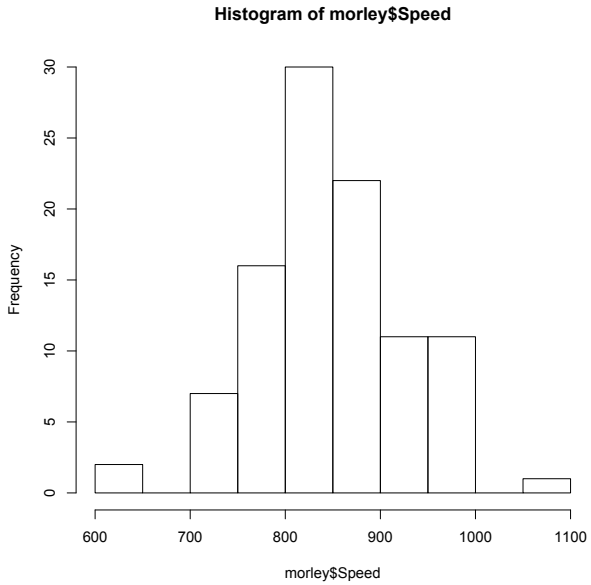


Figure 16.4: Measurements of the speed of light. Actual values are 299,000 kilometers per second plus the value shown.

```
> mean(morley$Speed)
[1] 852.4
> sd(morley$Speed)
[1] 79.01055
> qt(0.975,99)
[1] 1.984217
```

Thus, our confidence interval for the speed of light is

$$299,852.4 \pm 1.9842\frac{79.0}{\sqrt{100}} = 299,852.4 \pm 15.7 \quad \textit{or the interval } (299836.7, 299868.1)$$

This confidence interval does not *include the presently determined values of 299,792.458 km/sec for the speed of light. The confidence interval can also be found by tying* `t.test(morley$Speed)`*. We will study this command in more detail when we describe the* t*-test.*

Exercise 16.4. *Give a 90% and a 98% confidence interval for the example above.*

We often wish to determine a sample size that will guarantee a desired margin of error. For a γ-level t-interval, this is

$$m = t_{n-1,(1-\gamma)/2}\frac{s}{\sqrt{n}}.$$

Solving this for n yields

$$n = \left(\frac{t_{n-1,(1-\gamma)/2}\, s}{m}\right)^2.$$

Because the number of degrees of freedom, $n-1$, for the t distribution is unknown, the quantity n appears on both sides of the equation and the value of s is unknown. We search for a **conservative** value for n, i.e., a margin of error that will be no greater that the desired length. This can be achieved by overestimating $t_{n-1,(1-\gamma)/2}$ and s. For the

speed of light example above, if we desire a margin of error of $m = 10$ km/sec for a 95% confidence interval, then we set $t_{n-1,(1-\gamma)/2} = 2$ and $s = 80$ to obtain

$$n \approx \left(\frac{2 \cdot 80}{10}\right)^2 = 256$$

measurements are necessary to obtain the desired margin of error..

The next set of confidence intervals are determined, in the case in which the distributional variance in known, by finding the standardized score and using the normal approximation as given via the central limit theorem. In the cases in which the variance is unknown, we replace the distribution variance with a variance that is estimated from the observations. In this case, the procedure that is analogous to the standardized score is called the **studentized score**.

Example 16.5 (matched pair t interval)**.** *We begin with two quantitative measurements*

$$(X_{1,1}, \ldots, X_{1,n}) \quad \textit{and} \quad (X_{2,1}, \ldots, X_{2,n}),$$

on the same n individuals. Assume that the first set of measurements has mean μ_1 and the second set has mean μ_2.

If we want to determine a confidence interval for the difference $\mu_1 - \mu_2$, we can apply the t-procedure to the differences

$$(X_{1,1} - X_{2,1}, \ldots, X_{1,n} - X_{2,n})$$

to obtain the confidence interval

$$(\bar{X}_1 - \bar{X}_2) \pm t_{n-1,(1-\gamma)/2} \frac{S_d}{\sqrt{n}}$$

where S_d is the standard deviation of the difference.

Example 16.6 (2-sample z interval)**.** *If we have two* **independent** *samples of normal random variables*

$$(X_{1,1}, \ldots, X_{1,n_1}) \quad \textit{and} \quad (X_{2,1}, \ldots, X_{2,n_2}),$$

the first having mean μ_1 and variance σ_1^2 and the second having mean μ_2 and variance σ_2^2, then the difference in their sample means

$$\bar{X}_2 - \bar{X}_1$$

is also a normal random variable with

$$\textit{mean} \quad \mu_1 - \mu_2 \quad \textit{and} \quad \textit{variance} \quad \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Therefore,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is a standard normal random variable. In the case in which the variances σ_1^2 and σ_2^2 are known, this gives us a γ-level confidence interval for the difference in parameters $\mu_1 - \mu_2$.

$$(\bar{X}_1 - \bar{X}_2) \pm z_{(1-\gamma)/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Example 16.7 (2-sample t interval)**.** *If we know that $\sigma_1^2 = \sigma_2^2$, then we can pool the data to compute the standard deviation. Let S_1^2 and S_2^2 be the sample variances from the two samples. Then the* **pooled sample variance** *S_p is the weighted average of the sample variances with weights equal to their respective degrees of freedom.*

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

This give a statistic

$$T_{n_1+n_2-2} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

that has a t distribution with $n_1 + n_2 - 2$ degrees of freedom. Thus we have the γ level confidence interval

$$(\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2,(1-\gamma)/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

for $\mu_1 - \mu_2$.

If we do not know that $\sigma_1^2 = \sigma_2^2$, then the corresponding studentized random variable

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

no longer has a t-distribution.

Welch and Satterthwaite have provided an approximation to the t distribution with **effective degrees of freedom** *given by the Welch-Satterthwaite equation*

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 \cdot (n_1-1)} + \frac{s_2^4}{n_2^2 \cdot (n_2-1)}}. \tag{16.1}$$

This give a γ-level confidence interval

$$\bar{x}_1 - \bar{x}_2 \pm t_{\nu,(1-\gamma/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

For two sample tests, the number of observations per group may need to be at least 40 for a good approximation to the normal distribution.

Exercise 16.8. *Show that the effective degrees is between the worst case of the minimum choice from a one sample t-interval and the best case of equal variances.*

$$\min\{n_1, n_2\} - 1 \leq \nu \leq n_1 + n_2 - 2$$

For data on the life span in days of 88 wildtype and 99 transgenic mosquitoes, we have the summary

	observations	mean	standard deviation
wildtype	88	20.784	12.99
transgenic	99	16.546	10.78

Using the conservative 95% confidence interval based on $\min\{n_1, n_2\} - 1 = 87$ degrees of freedom, we use

```
> qt(0.975,87)
[1] 1.987608
```

to obtain the interval

$$(20.78 - 16.55) \pm 1.9876\sqrt{\frac{12.99^2}{88} + \frac{10.78^2}{99}} = (0.744, 7.733)$$

Using the the Welch-Satterthwaite equation, we obtain $\nu = 169.665$. The increase in the number of degrees of freedom gives a slightly narrower interval $(0.768, 7.710)$.

16.1.2 Linear Regression

For ordinary linear regression, we have given least squares estimates for the slope β and the intercept α. For data $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$, our model is

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where ϵ_i are independent $N(0, \sigma)$ random variables. Recall that the estimator for the slope

$$\hat{\beta}(x, y) = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

is unbiased.

Exercise 16.9. *Show that*

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{(n-1)\text{var}(x)}.$$

and

$$\text{Var}_{(\alpha,\beta)}(\hat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\text{var}(x)} \right).$$

Notice that $\text{Var}_{(\alpha,\beta)}(\hat{\alpha})$ *increases with the distance that the mean of the* x *values is from 0.*

If σ is known, this suggests a z-interval for a γ-level confidence interval

$$\hat{\beta} \pm z_{(1-\gamma)/2} \frac{\sigma}{s_x \sqrt{n-1}}.$$

Generally, σ is unknown. However, the variance of the residuals,

$$s_u^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - (\hat{\alpha} - \hat{\beta} x_i))^2 \tag{16.2}$$

is an unbiased estimator of σ^2 and s_u/σ has a t distribution with $n-2$ degrees of freedom. This gives the t-interval

$$\hat{\beta} \pm t_{n-2,(1-\gamma)/2} \frac{s_u}{s_x \sqrt{n-1}}.$$

As the formula shows, the margin of error is proportional to the standard deviation of the residuals. It is inversely proportional to the standard deviation of the x measurement. Thus, we can reduce the margin of error by taking a broader set of values for the explanatory variables.

For the data on the humerus and femur of the five specimens of *Archeopteryx*, we have $\hat{\beta} = 1.197$. $s_u = 1.982$, $s_x = 13.2$, and $t_{3,0.025} = 3.1824$, Thus, the 95% confidence interval is 1.197 ± 0.239 or $(0.958, 1.436)$.

16.1.3 Sample Proportions

Example 16.10 (proportions)**.** *For* n *Bernoulli trials with success parameter* p*, the sample proportion* $\hat{p}$ *has*

$$\textit{mean} \quad p \quad \textit{and} \quad \textit{variance} \quad \frac{p(1-p)}{n}.$$

The parameter p *appears in both the mean and in the variance. Thus, we need to make a choice* $\tilde{p}$ *to replace* p *in the confidence interval*

$$\hat{p} \pm z_{(1-\gamma)/2} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}. \tag{16.3}$$

One simple choice for $\tilde{p}$ *is* $\hat{p}$*. Based on extensive numerical experimentation, one more recent popular choice is*

$$\tilde{p} = \frac{x+2}{n+4}$$

where x is the number of successes.

For population proportions, we ask that the **mean number of successes** np *and the* **mean number of failures** $n(1-p)$ *each be at least 10. We have this requirement so that a normal random variable is a good approximation to the appropriate binomial random variable.*

Example 16.11. *For Mendel's data the F_2 generation consisted 428 for the dominant allele green pods and 152 for the recessive allele yellow pods. Thus, the sample proportion of green pod alleles is*

$$\hat{p} = \frac{428}{428+152} = 0.7379.$$

The confidence interval, using

$$\tilde{p} = \frac{428+2}{428+152+4} = 0.7363$$

is

$$0.7379 \pm z_{(1-\gamma)/2}\sqrt{\frac{0.7363 \cdot 0.2637}{580}} = 0.7379 \pm 0.0183 z_{(1-\gamma)/2}$$

For $\gamma = 0.98$, $z_{0.01} = 2.326$ and the confidence interval is $0.7379 \pm 0.0426 = (0.6953, 0.7805)$. Note that this interval contains the predicted value of $p = 3/4$.

A comparable formula gives confidence intervals based on more than two independent samples

Example 16.12. *For the difference in two proportions p_1 and p_2 based on n_1 and n_2 independent trials. We have, for the difference $p_1 - p_2$, the confidence interval*

$$\hat{p}_1 - \hat{p}_2 \pm \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Example 16.13 (transformation of a single parameter)**.** *If*

$$(\hat{\theta}_\ell, \hat{\theta}_u)$$

is a level γ confidence interval for θ and g is an increasing function, then

$$(g(\hat{\theta}_\ell), g(\hat{\theta}_u))$$

is a level γ confidence interval for $g(\theta)$

Exercise 16.14. *For the example above, find the confidence interval for the yellow pod genotype.*

16.1.4 Summary of Standard Confidence Intervals

The confidence interval is an extension of the idea of a point estimation of the parameter to an interval that is likely to contain the true parameter value. A level γ confidence interval for a population parameter θ is an interval computed from the sample data having probability γ of producing an interval containing θ.

For an estimate of a population mean or proportion, a level γ confidence interval often has the form

$$\text{estimate} \pm t^* \times \text{standard error}$$

where t^* is the upper $\frac{1-\gamma}{2}$ critical value for the t distribution with the appropriate number of degrees of freedom. If the number of degrees of freedom is **infinite**, we use the **standard normal** distribution to determine the critical value, usually denoted by z^*.

The margin of error $m = t^* \times$ standard error decreases if

- γ, the confidence level, decreases
- the standard deviation decreases
- n, the number of observations, increases

The procedures for finding the confidence interval are summarized in the table below.

procedure	parameter	estimate	standard error	degrees of freedom
one sample	μ	$\bar{x}$	$\frac{s}{\sqrt{n}}$	$n-1$
two sample	$\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	See (16.1)
pooled two sample	$\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2$	$s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	n_1+n_2-2
one proportion	p	$\hat{p}$	$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$, $\tilde{p}=\frac{x+2}{n+4}$	∞
two proportion	p_1-p_2	$\hat{p}_1-\hat{p}_2$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	∞
linear regression	β	$\hat{\beta}=\text{cov}(x,y)/\text{var}(x)$	$\frac{s_u}{s_x\sqrt{n-1}}$	$n-2$

The first confidence interval for $\mu_1-\mu_2$ is the two-sample t procedure. If we can assume that the two samples have a common standard deviation, then we pool the data to compute s_p, the pooled standard deviation. Matched pair procedures use a one sample procedure on the difference in the observed values.

For these tests, we need a sample size large enough so that the central limit theorem is a sufficiently good approximation. For one population tests for means, $n>30$ and data not strongly skewed is a good rule of thumb. For two population tests, $n>40$ may be necessary. For population proportions, we ask that the mean number of successes np and the mean number of failures $n(1-p)$ each be at least 10.

For the standard error for $\hat{\beta}$ in linear regression, s_u is defined in (16.2) and s_x is the standard deviation of the values of the explanatory variable.

16.1.5 Interpretation of the Confidence Interval

The confidence interval for a parameter θ is based on two statistics - $\hat{\theta}_\ell(\mathbf{x})$, the lower end of the confidence interval and $\hat{\theta}_u(\mathbf{x})$, the upper end of the confidence interval. As with all statistics, these two statistics cannot be based on the value of the parameter. In addition, these two statistics are determined in advance of having the actual data. The term confidence can be related to the production of confidence intervals. We can think of the situation in which we produce independent confidence intervals repeatedly. Each time, we may either succeed or fail to include the true parameter in the confidence interval. In other words, the inclusion of the parameter value in the confidence interval is a Bernoulli trial with success probability γ.

For example, after having seen these 100 intervals in Figure 5, we can conclude that the lowest and highest intervals are much less likely that 95% of containing the true parameter value. This phenomena can be seen in the presidential polls for the 2012 election. Three days before the election we see the following spread between Mr. Obama and Mr. Romney

0% -1% 0% 1% 5% 0% -5% -1% 1% 1%

with the 95% confidence interval having a margin of error $\sim 3\%$ based on a sample of size ~ 1000. Because these values are highly dependent, the values of $\pm 5\%$ is less likely to contain the true spread.

Exercise 16.15. *Perform the computations needed to determine the margin of error in the example above.*

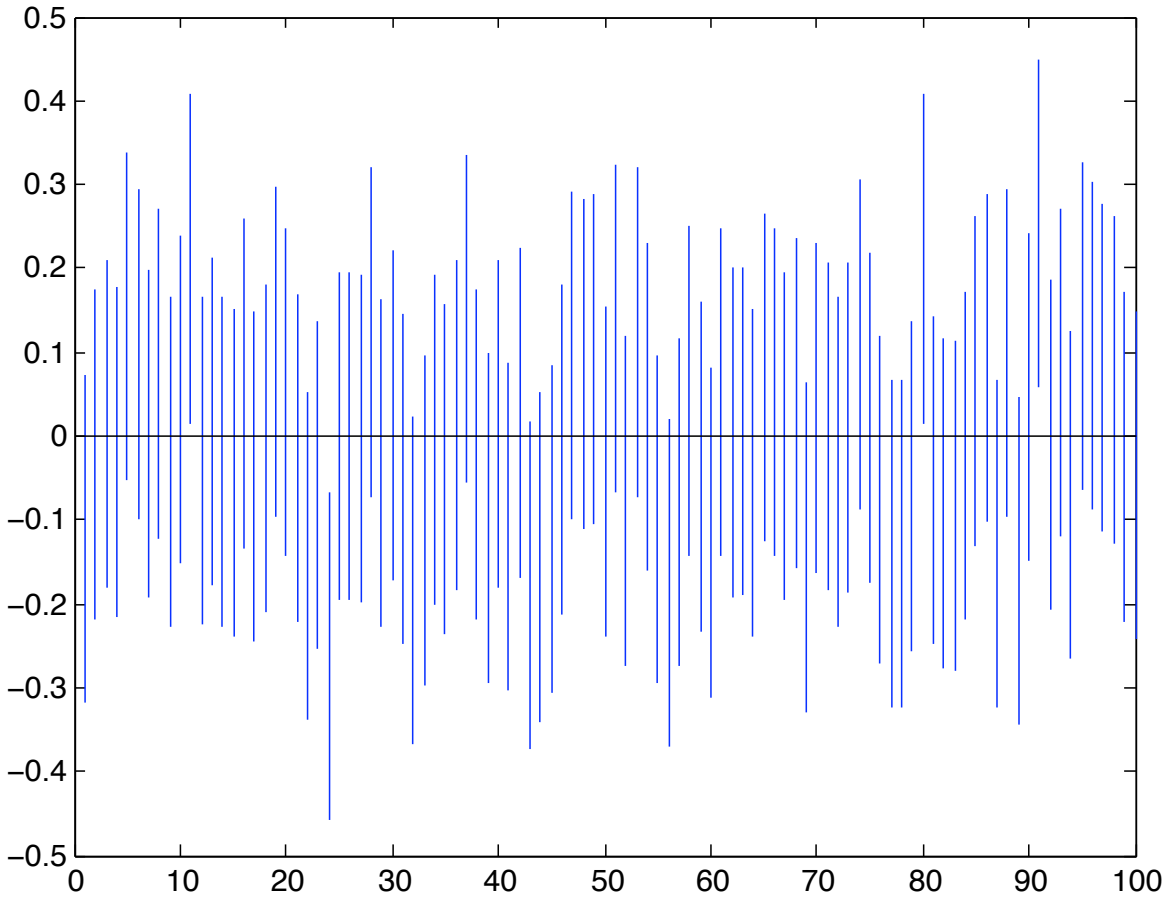


Figure 16.5: One hundred confidence build from repeatedly simulating 100 standard normal random variables and constructing 95% confidence intervals for the mean value - 0. Note that the 24th interval is entirely below 0 and so does not contain the actual parameter. The 11th, 80th and 91st intervals are entirely above 0 and again do not contain the parameter.

The following example, although never likely to be used in an actual problem, may shed some insight into the difference between confidence and probability.

Example 16.16. *Let X_1 and X_2 be two independent observations from a uniform distribution on the interval $[\theta - 1, \theta + 1]$ where θ is an unknown parameter. In this case, an observation is greater than θ with probability 1/2, and less than θ with probability 1/2. Thus,*

- *with probability 1/4, both observations are above θ,*
- *with probability 1/4, both observations are below θ, and*
- *with probability 1/2, one observation is below θ and the other is above.*

In the third case alone, the confidence interval contains the parameter value. As a consequence of these considerations, the interval

$$(\hat{\theta}_\ell(X_1, X_2), \hat{\theta}_u(X_1, X_2)) = (\min\{X_1, X_2\}, \max\{X_1, X_2\})$$

is a 50% confidence interval for the parameter.

Sometimes, $\max\{X_1, X_2\} - \min\{X_1, X_2\} > 1$. *Because any subinterval of the interval* $[\theta - 1, \theta + 1]$ *that has length at least 1 must contain* θ*, the midpoint of the interval, this confidence interval must contain the parameter value. In other words, sometimes the 50% confidence interval is certain to contain the parameter.*

Exercise 16.17. *For the example above, show that*

$$P\{\text{confidence interval has length } > 1\} = 1/4.$$

Hint: Draw the square $[\theta - 1, \theta + 1] \times [\theta - 1, \theta + 1]$ *and shade the region in which the confidence interval has length greater than 1.*

16.1.6 Extensions on the Use of Confidence Intervals

Example 16.18 (delta method)**.** *For estimating the distribution* μ *by the sample mean* $\bar{X}$*, the delta method provides an alternative for the example above. In this case, the standard deviation of* $g(\bar{X})$ *is approximately*

$$\frac{|g'(\mu)|\sigma}{\sqrt{n}}. \tag{16.4}$$

We replace μ *with* $\bar{X}$ *to obtain the confidence interval for* $g(\mu)$

$$g(\bar{X}) \pm z_{\alpha/2}\frac{|g'(\bar{X})|\sigma}{\sqrt{n}}.$$

Using the notation for the example of estimating α_3*, the coefficient of volume expansion based on independent length measurements,* $Y_1, Y_2, \ldots, Y_n$ *measured at temperature* T_1 *of an object having length* ℓ_0 *at temperature* T_0*.*

$$\frac{\bar{Y}^3 - \ell_0^3}{\ell_0^3|T_1 - T_0|} \pm z_{(1-\gamma)/2}\frac{3\bar{Y}^2\sigma_Y}{n}$$

Exercise 16.19. *Recall that the* **odds** *of an event having probability* p *is*

$$\psi = \frac{p}{1-p}. \tag{16.5}$$

Use the delta method to show that

$$\sigma_{\hat{\psi}}^2 \approx \frac{\psi(\psi+1)^2}{n}.$$

In the example above on green peas,

$$\hat{\psi} = \frac{\hat{p}}{1-\hat{p}} = \frac{0.7379}{1-0.7379} = 2.8153.$$

Using (16.4), we obtain a 98% confidence interval

$$\hat{\psi} \pm z_{0.01}\sqrt{\sigma_{\hat{\psi}}^2} \approx 2.8153 \pm 2.326\sqrt{\frac{2.8153(1+2.8153)^2}{580}} = (2.1970, 3.4337)$$

which includes the predicted value $\psi = 3$*.*

If we transform the 98% confidence interval $(0.6953, 0.7805)$ *for* p *to a confidence interval for the odds* ψ *using the transformation (16.5), we obtain an interval* $(2.2819, 3.5558)$ *that is slightly shifted upward from the confidence interval obtained by the delta method.*

For multiple independent samples, the simple idea using the transformation in the Example 12 no longer works. For example, to determine the confidence interval using $\bar{X}_1$ and $\bar{X}_2$ above, the confidence interval for $g(\mu_1, \mu_2)$, the delta methos gives the confidence interval

$$g(\bar{X}_1, \bar{X}_2) \pm z_{(1-\gamma)/2}\sqrt{\left(\frac{\partial}{\partial x}g(\bar{X}_1, \bar{X}_2)\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{\partial}{\partial y}g(\bar{X}_1, \bar{X}_2)\right)^2 \frac{\sigma_2^2}{n_2}}.$$

Example 16.20. *Let's return to the example of n_ℓ and n_h measurements y and y of, respectively, the length ℓ and the height h of a right triangle with the goal of giving the angle*

$$\theta = g(\ell, h) = \tan^{-1}\left(\frac{h}{\ell}\right)$$

between these two sides. Here are the measurements:

```
> x
 [1] 10.224484 10.061800  9.945213  9.982061  9.961353 10.173944  9.952279  9.855147
 [9]  9.737811  9.956345
> y
 [1] 4.989871 5.090002 5.021615 4.864633 5.024388 5.123419 5.033074 4.750892 4.985719
[10] 4.912719 5.027048 5.055755
> mean(x);sd(x)
[1] 9.985044
[1] 0.1417969
> mean(y);sd(y)
[1] 4.989928
[1] 0.1028745
```

The angle θ is the arctangent, here estimated using the mean and given in radians

```
>(thetahat<-atan(mean(y)/mean(x)))
[1] 0.4634398
```

Using the delta method, we have estimated the standard deviation of these measurements.

$$\sigma_{\hat{\theta}} \approx \frac{1}{h^2 + \ell^2}\sqrt{h^2\frac{\sigma_\ell^2}{n_\ell} + \ell^2\frac{\sigma_h^2}{n_h}}.$$

We estimate this with the sample means and standard deviations

$$s_{\hat{\theta}} \approx \frac{1}{\bar{y}^2 + \bar{x}^2}\sqrt{\bar{y}^2\frac{s_x^2}{n_\ell} + \bar{x}^2\frac{s_y^2}{n_h}} = 0.0030.$$

This gives a γ level z- confidence interval

$$\hat{\theta} \pm z_{(1-\gamma)/2}s_{\hat{\theta}}.$$

For a 95% confidence interval, this is $0.4634 \pm 0.0059 = (0.4575, 0.4693)$ radians or $(26.22°, 26.89°)$.

We can extend the Welch and Satterthwaite method to include the delta method to create a t-interval with effective degrees of freedom

$$\nu = \frac{\left(\frac{\partial g(\bar{x},\bar{y})^2}{\partial x}\frac{s_1^2}{n_1} + \frac{\partial g(\bar{x},\bar{y})^2}{\partial y}\frac{s_2^2}{n_2}\right)^2}{\frac{\partial g(\bar{x},\bar{y})^4}{\partial x}\frac{s_1^4}{n_1^2\cdot(n_1-1)} + \frac{\partial g(\bar{x},\bar{y})^4}{\partial y}\frac{s_2^4}{n_2^2\cdot(n_2-1)}}.$$

We compute to find that $\nu = 19.4$ and then use the t-interval

$$\hat{\theta} \pm t_{\nu,(1-\gamma)/2}s_{\hat{\theta}}.$$

For a 95% confidence, this is sightly larger interval $0.4634 \pm 0.0063 = (0.4571, 0.4697)$ radians or $(26.19°, 26.91°)$.

Example 16.21 (maximum likelihood estimation)**.** *The Fisher information is the main tool used to set an confidence interval for a maximum likelihood estimation. Two choices are typical. First, we can use the Fisher information $I_n(\theta)$. Of course, the estimator is for the unknown value of the parameter θ. This give an confidence interval*

$$\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{nI(\hat{\theta})}}.$$

More recently, the more popular method is to use the **observed information** *based on the observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.*

$$J(\hat{\theta}) = -\frac{\partial^2}{\partial \theta^2} \log L(\hat{\theta}|\mathbf{x}) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_X(x_i|\hat{\theta}).$$

This is the second derivative of the score function evaluated at the maximum likelihood estimator. Then, the confidence interval is

$$\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{J(\hat{\theta})}}.$$

Note that $E_\theta J(\theta) = nI(\theta)$, the Fisher information for n observations. Thus, by the law of large numbers

$$\frac{1}{n} J(\theta) \to I(\theta) \quad \text{as } n \to \infty.$$

If the estimator is consistent and I is continuous at θ, then

$$\frac{1}{n} J(\hat{\theta}) \to I(\theta) \quad \text{as } n \to \infty.$$

16.2 The Bootstrap

The confidence regions have been determined using aspects of the distribution of the data. In particular, these regions have often been specified by appealing to the central limit theorem and normal approximations. The notion behind **bootstrap** techniques begins with the concession that the information about the source of the data is insufficient to perform the analysis to produce the necessary description of the distribution of the estimator. This is particularly true for small data sets or highly skewed data.

The strategy is to take the data and treat it as if it were the distribution underlaying the data and to use a resampling protocol to describe the estimator. For the example above, we estimated the angle in a right triangle by estimating ℓ and h, the lengths two adjacent sides by taking the mean of our measurements and then computing the arctangent of the ratio of these means. Using the delta method, our confidence interval was based on a normal approximation of the estimator.

The bootstrap takes another approach. We take the data

$$x_1, x_2, \ldots, x_{n_1}, \quad y_1, y_2, \ldots, y_{n_2},$$

the **empirical distribution** of the measurements of ℓ and h and act as if it were the **actual distribution**. The next step is the use the data and randomly create the results of the experiment many times over. In the example, we choose, *with replacement* n_1 measurements from the x data and n_2 measurements from the y data. We then compute the bootstrap means

$$\bar{x}_b \quad \text{and} \quad \bar{y}_b$$

and the estimate

$$\hat{\theta}(\bar{x}_b, \bar{y}_b) = \tan^{-1}\left(\frac{\bar{y}_b}{\bar{x}_b}\right).$$

Repeating this many times gives an empirical distribution for the estimator $\hat{\theta}$. This can be accomplish in just a couple lines of R code.

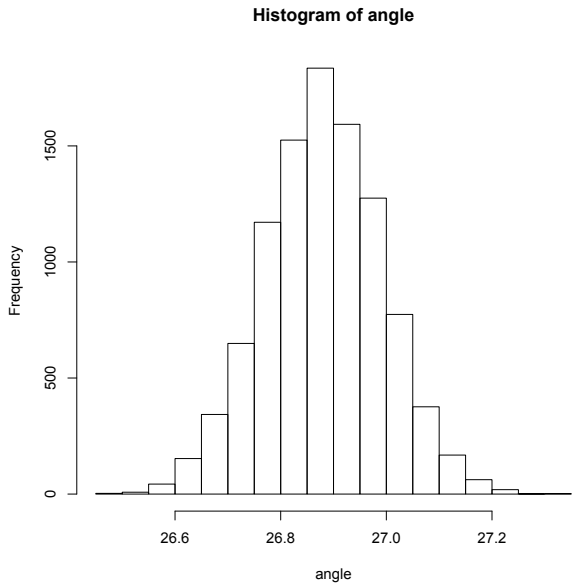


Figure 16.6: Bootstrap distribution of $\hat{\theta}$.

```
> angle<-numeric10000)
> for (i in 1:10000){xb<-sample(x,length(x),replace=TRUE);
  yb<-sample(y,length(y),replace=TRUE);angle[i]<-atan(mean(yb)/mean(xb))*180/pi}
> hist(angle)
```

We can use this bootstrap distribution of $\hat{\theta}$ to construct a confidence interval.

```
> q<-c(0.005,0.01,0.025,0.5,0.975,0.99,0.995)
> quantile(angle,q)
    0.5%       1%     2.5%      50%    97.5%      99%    99.5%
26.09837 26.14807 26.21860 26.55387 26.86203 26.91486 26.95065
```

A 95% confidence interval $(26.21°, 26.86°)$ can be accomplished using the 2.5th percentile as the lower end point and the 97.5th percentile as the upper end point. This confidence interval is very similar to the one obtained using the delta method.

Exercise 16.22. *Give the 98% bootstrap confidence interval for the angle in the example above.*

Exercise 16.23. *Bootstrap confidences are based on a simulation. So, the answer will vary with each simulation. Repeat the bootstrap above and compare.*

16.3 Bayesian Statistics

A Bayesian interval estimate is called a **credible interval**. Recall that for the Bayesian approach to statistics, both the data and the parameter are random Thus, the interval estimate is a statement about the posterior probability of the parameter θ.

$$P\{\tilde{\Theta} \in C(X)|X = x\} = \gamma. \tag{16.6}$$

Here $\tilde{\Theta}$ is the random variable having a distribution equal to the prior probability π. We have choices in defining this interval. For example, we can

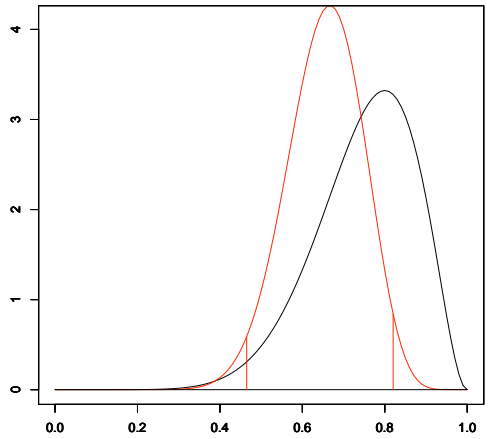


Figure 16.7: Bayesian credible interval. The 95% credible interval based on a $Beta(9, 3)$ prior distribution and data consisting of 8 heads and 6 tails. This interval has probability 0.025 both above and below the end of the interval. Because the density is higher for the upper value, an narrow credible interval can be obtained by shifting the values upward so that the densities are equal at the endpoints of the interval and (16.6) is satisfied.

- choose the narrowest interval, which involves choosing those values of highest posterior density.
- choosing the interval in which the probability of being below the interval is as likely as being above it.

We can look at this by examining the two examples given in the Introduction to Estimation.

Example 16.24 (coin tosses)**.** *In this example, we began with a beta prior distribution. Consequently, the posterior distribution will also be a member of the beta family of distributions. We then flipped a coin $n = 14$ times with 8 heads. Here, again, is the summary.*

prior				data		posterior		
α	β	mean	variance	heads	tails	α	β	mean
6	6	1/2	1/52	8	6	14	12	7/13
9	3	3/4	3/208	8	6	17	9	17/26
3	9	1/4	3/208	8	6	11	15	11/26

We use the R command `qbeta` to find the credible interval. For the second case in the table above and with $\gamma = 0.95$, we find values that give the lower and upper 2.5% of the posterior distribution.

```
> qbeta(0.025,17,9)
[1] 0.4649993
> qbeta(0.975,17,9)
[1] 0.8202832
```

This gives a 95% credible interval of $(0.4650, 0.8203)$. This is indicated in the figure above by the two vertical lines. Thus, the area under the density function from the vertical lines outward totals 5%.

The narrowest credible interval is $(0.4737, 0.8276)$. At these values, the density equals 0.695. The density is lower for more extreme values and higher between these values. The beta distribution has a probability 0.0306 below the lower value for the credible interval and 0.0194 above the upper value satisfying the criterion (16.6) with $\gamma = 0.95$.

Example 16.25. *For the example having both a normal prior distribution and normal data, we find that we also have a normal posterior distribution. In particular, if the prior is normal, mean θ_0, variance $1/\lambda$ and our data has sample mean $\bar{x}$ and each observation has variance 1.*

The classical statistics confidence interval

$$\bar{x} \pm z_{\alpha/2}\frac{1}{\sqrt{n}}.$$

For the Bayesian credible interval, note that the posterior distribution in normal with mean

$$\theta_1(\mathbf{x}) = \frac{\lambda}{\lambda+n}\theta_0 + \frac{n}{\lambda+n}\bar{x}.$$

and variance $1/(n+\lambda)$*. Thus the credible interval is*

$$\theta_1(\mathbf{x}) \pm z_{\alpha/2}\frac{1}{\sqrt{\lambda+n}}.$$

Thus, the center of the interval is influenced by the prior mean. The prior variance results in a narrower interval.

16.4 Answers to Selected Exercises

16.4. Using R to find upper tail probabilities, we find that

```
> qt(0.95,99)
[1] 1.660391
> qt(0.99,99)
[1] 2.364606
```

For the 90% confidence interval

$$299,852.4 \pm 1.6604\frac{79.0}{\sqrt{100}} = 299852.4 \pm 13.1 \quad \text{or the interval} \quad (299839.3, 299865.5).$$

For the 98% confidence interval

$$299,852.4 \pm 2.3646\frac{79.0}{\sqrt{100}} = 299852.4 \pm 18.7 \quad \text{or the interval} \quad (299833.7, 299871.1).$$

16.8. Let

$$c = \frac{s_1^2/n_1}{s_2^2/n_2}. \quad \text{Then, } \frac{s_2^2}{n_2} = c\frac{s_1^2}{n_1}.$$

Then, substitute for s_2^2/n_2 and divide by s_1^2/n_1 to obtain

$$\nu = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2\cdot(n_1-1)}+\frac{s_2^4}{n_2^2\cdot(n_2-1)}} = \frac{\left(\frac{s_1^2}{n_1}+\frac{cs_1^2}{n_1}\right)^2}{\frac{s_1^4}{n_1^2\cdot(n_1-1)}+\frac{c^2s_1^4}{n_1^2\cdot(n_2-1)}} = \frac{(1+c)^2}{\frac{1}{n_1-1}+\frac{c^2}{n_2-1}} = \frac{(n_1-1)(n_2-1)(1+c)^2}{(n_2-1)+(n_1-1)c^2}.$$

Take a derivative to see that

$$\begin{aligned}\frac{d\nu}{dc} &= (n_1-1)(n_2-1)\frac{((n_2-1)+(n_1-1)c^2)\cdot 2(1+c)-(1+c)^2\cdot 2(n_1-1)c}{((n_2-1)+(n_1-1)c^2)^2} \\ &= 2(n_1-1)(n_2-1)(1+c)\frac{((n_2-1)+(n_1-1)c^2)-(1+c)(n_1-1)c}{((n_2-1)+(n_1-1)c^2)^2} \\ &= 2(n_1-1)(n_2-1)(1+c)\frac{(n_2-1)-(n_1-1)c}{((n_2-1)+(n_1-1)c^2)^2}\end{aligned}$$

So the maximum takes place at $c = (n_2 - 1)/(n_1 - 1)$ with value of ν.

$$\begin{aligned}\nu &= \frac{(n_1-1)(n_2-1)(1+(n_2-1)/(n_1-1))^2}{(n_2-1)+(n_1-1)((n_2-1)/(n_1-1))^2}\\ &= \frac{(n_1-1)(n_2-1)((n_1-1)+(n_2-1))^2}{(n_1-1)^2(n_2-1)+(n_1-1)(n_2-1)^2}\\ &= \frac{((n_1-1)+(n_2-1))^2}{(n_1-1)+(n_2-1)} = n_1+n_2-2.\end{aligned}$$

Note that for this value

$$\frac{s_1^2}{s_2^2} = \frac{n_1}{n_2}c = \frac{n_1/(n_1-1)}{n_2/(n_2-1)}$$

and the variances are nearly equal. Notice that this is a global maximum with

$$\nu \to n_1 - 1 \text{ as } c \to 0 \text{ and } s_1 \ll s_2 \quad \text{and} \quad \nu \to n_2 - 1 \text{ as } c \to \infty \text{ and } s_2 \ll s_1.$$

The smaller of these two limits is the global minimum.

16.9. Recall that $\hat{\beta}$ is an unbiased estimator for β, thus $E_{(\alpha,\beta)}\hat{\beta} = \beta$, and $E_{(\alpha,\beta)}[(\hat{\beta}-\beta)^2]$ is the variance of $\hat{\beta}$.

$$\begin{aligned}\hat{\beta}(x,y) - \beta &= \frac{1}{(n-1)\text{var}(x)}\left(\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) - \beta\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})\right)\\ &= \frac{1}{(n-1)\text{var}(x)}\left(\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}-\beta(x_i-\bar{x}))\right)\\ &= \frac{1}{(n-1)\text{var}(x)}\left(\sum_{i=1}^n (x_i-\bar{x})((y_i-\beta x_i)-(\bar{y}-\beta\bar{x}))\right)\\ &= \frac{1}{(n-1)\text{var}(x)}\left(\sum_{i=1}^n (x_i-\bar{x})(y_i-\beta x_i) - \sum_{i=1}^n (x_i-\bar{x})(\bar{y}-\beta\bar{x})\right)\end{aligned}$$

The second sum is 0. For the first, we use the fact that $y_i - \beta x_i = \alpha + \epsilon_i$. Thus,

$$\begin{aligned}\text{Var}_{(\alpha,\beta)}(\hat{\beta}) &= \text{Var}_{(\alpha,\beta)}\left(\frac{1}{(n-1)\text{var}(x)}\sum_{i=1}^n (x_i-\bar{x})(\alpha+\epsilon_i)\right) = \frac{1}{(n-1)^2\text{var}(x)^2}\sum_{i=1}^n (x_i-\bar{x})^2\text{Var}_{(\alpha,\beta)}(\alpha+\epsilon_i)\\ &= \frac{1}{(n-1)^2\text{var}(x)^2}\sum_{i=1}^n (x_i-\bar{x})^2\sigma^2 = \frac{\sigma^2}{(n-1)\text{var}(x)}\end{aligned}$$

Because the ϵ_i are independent, we can use the Pythagorean identity that the variance of the sum is the sum of the variances.

Similarly, $\hat{\alpha}$ is an unbiased estimator for α. Because $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, we have by the law of cosines,

$$\begin{aligned}\text{Var}_{(\alpha,\beta)}(\hat{\alpha}) &= \text{Var}_{(\alpha,\beta)}(\bar{y}) + \text{Var}_{(\alpha,\beta)}(\hat{\beta}\bar{x}) - 2\text{Cov}_{(\alpha,\beta)}(\bar{y},\hat{\beta}\bar{x})\\ &= \text{Var}_{(\alpha,\beta)}(\bar{y}) + \bar{x}^2\text{Var}_{(\alpha,\beta)}(\hat{\beta}) - 2\bar{x}\text{Cov}_{(\alpha,\beta)}(\bar{y},\hat{\beta})\end{aligned}$$

For the first term, note that $\text{Var}(y_i) = \text{Var}(\epsilon_n)$.

$$\text{Var}_{(\alpha,\beta)}(\bar{y}) = \text{Var}_{(\alpha,\beta)}\left(\frac{1}{n}\sum_{i=1}^n y_i\right) = \frac{1}{n^2}\text{Var}_{(\alpha,\beta)}\left(\sum_{i=1}^n \epsilon_i\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

For the third term, note that

$$\text{Cov}_{(\alpha,\beta)}(y_i, y_j) = \begin{cases} 0 & \text{if } i \neq j, \\ \sigma^2 & \text{if } i = j. \end{cases}$$

and thus $\text{Cov}_{(\alpha,\beta)}(y_i, \bar{y}) = \sum_{j=1}^{n} \text{Cov}_{(\alpha,\beta)}(y_i, y_j)/n = \sigma^2/n$. Using the binlinear properties of covariance, we find that

$$\begin{aligned}
\text{Cov}_{(\alpha,\beta)}(\bar{y}, \hat{\beta}) &= \frac{1}{n\text{var(x)}} \sum_{i=1}^{n} \text{Cov}_{(\alpha,\beta)}(y_i, \text{cov}(x, y)) \\
&= \frac{1}{n\text{var(x)}} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}_{(\alpha,\beta)}(y_i, (x_j - \bar{x})(y_j - \bar{y})) \\
&= \frac{1}{n\text{var(x)}} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_j - \bar{x})(\text{Cov}_{(\alpha,\beta)}(y_i, y_j) - \text{Cov}_{(\alpha,\beta)}(y_i, \bar{y})) \\
&= \frac{1}{n\text{var(x)}} \sum_{j=1}^{n} (x_j - \bar{x}) \sum_{i=1}^{n} \sigma^2 \left(1 - \frac{1}{n}\right) = 0
\end{aligned}$$

because $\sum_{j=1}^{n} (x_j - \bar{x}) = 0$. Combining the results, we obtain

$$\text{Var}_{(\alpha,\beta)}(\hat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\text{var}(x)}\right).$$

16.12. For

$$\psi = g(p) = \frac{p}{1-p} \quad \text{we have that} \quad p = \frac{\psi}{\psi + 1} \quad \text{and} \quad g'(p) = \frac{1}{(1-p)^2}.$$

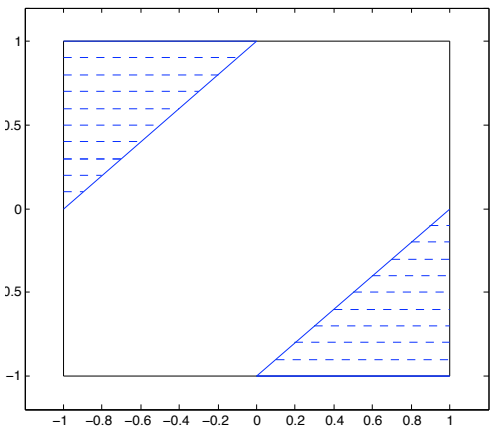
By the delta method,

$$\sigma^2_{\hat{\psi}} \approx g'(p)^2 \frac{\sigma_p^2}{n} = \frac{1}{(1-p)^4} \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3} = \frac{\psi(\psi+1)^2}{n}.$$

16.15. The confidence interval for the proportion yellow pod genes $1 - p$ is $(0.2195, 0.3047)$. The proportion of yellow pod phenotype is $(1-p)^2$ and a 95% confidence interval has as its endpoints the square of these numbers - $(0.0482, 0.0928)$.

16.16. The critical value $z_{0.025} = 1.96$. For $\hat{p} = 0.468$ and $n = 1500$, the number of successes is $x = 702$. The margin of error is

$$z_{0.025} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.025.$$



16.18. On the left is the square $[\theta - 1, \theta + 1] \times [\theta - 1, \theta + 1]$. For the random variables X_1, X_2, because they are independent and uniformly distributed over a square of area 4, their joint density is 1/4 on this square. The two diagonal line segments are the graph of $|x_1 - x_2| = 1$. In the shaded area, the region $|x_1 - x_2| > 1$, is precisely the region in which $\max\{x_1, x_2\} - \min\{x_1, x_2\} > 1$. Thus, for these values of the random variables, the confidence interval has length greater than 1. The area of each of the shaded triangles is $1/2 \cdot 1 \cdot 1 = 1/2$. Thus, the total area of the two triangles, 1, represents a probability of 1/4.

16.22. A 98% confidence interval $(26.14°, 26.91°)$ can be accomplished using the 1st percentile as the lower end point and the 99th percentile as the upper end point.

16.23. Here is the output of a second simulation.

```
> q<-c(0.005,0.01,0.025,0.5,0.975,0.99,0.995)
> quantile(angle,q)
    0.5%       1%     2.5%      50%    97.5%      99%    99.5%
26.12021 26.16463 26.22423 26.55800 26.85488 26.90665 26.94847
```

All of the values of within 0.04 of those from the first bootstrap.

Part IV

Hypothesis Testing

Topic 17

Simple Hypotheses

I can point to the particular moment when I understood how to formulate the undogmatic problem of the most powerful test of a simple statistical hypothesis against a fixed simple alternative. At the present time, the problem appears entirely trivial and within reach of a beginning undergraduate. But, with a degree of embarrassment, I must confess that it took something like half a decade of combined effort of E.S.P. and myself to put things straight. - Jerzy Neymann in the Festschrift in honor of Herman Wold, 1970, E.S.P is Egon Sharpe Pearson

17.1 Overview and Terminology

Statistical hypothesis testing is designed to address the question: *Do the data provide sufficient evidence to conclude that we must depart from our original assumption concerning the state of nature?*

The logic of hypothesis testing is similar to the one a juror faces in a criminal trial: *Is the evidence provided by the prosecutor sufficient for the jury to depart from its original assumption that the defendant is not guilty of the charges brought before the court?*

Two of the jury's possible actions are

- **Find the defendant guilty**.
- **Find the defendant not guilty**.

The weight of evidence that is necessary to find the defendant guilty depends on the type of trial. In a criminal trial the stated standard is that the prosecution must prove that *the defendant is guilty beyond any reasonable doubt*. In civil trials, the burden of proof may be the intermediate level of *clear and convincing evidence* or the lower level of the *preponderance of evidence*.

Given the level of evidence needed, a prosecutors task is to present the evidence in the most powerful and convincing manner possible. We shall see these notions reflected in the nature of hypothesis testing.

The simplest set-up for understanding the issues of **statistical hypothesis**, is the case of two values θ_0, and θ_1 in the parameter space. We write the test, known as a **simple hypothesis** as

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

H_0 is called the **null hypothesis**. H_1 is called the **alternative hypothesis**.

We now frame the issue of hypothesis testing using the classical approach. In this approach, the possible actions are:

- **Reject the hypothesis**. Rejecting the hypothesis when it is true is called a **type I error** or a **false positive**. Its probability α is called the **size of the test** or the **significance level**. Sometimes, $1 - \alpha$, the **true negatiive** is

called the **specificity**. In symbols, we write

$$\alpha = P_{\theta_0}\{\text{reject } H_0\}.$$

- **Fail to reject the hypothesis**. Failing to reject the hypothesis when it is false is called a **type II error** or a **false negative**, has probability β. The **power of the test**, $1-\beta$, the probability of rejecting the test when it is indeed false, is also called the **true positive fraction** or the the **sensitivity**. In symbols, we write

$$\beta = P_{\theta_1}\{\text{fail to reject } H_0\}.$$

hypothesis tests			criminal trials		
				the defendant is	
	H_0 is true	H_1 is true		innocent	guilty
reject H_0	type I error	OK	convict		OK
fail to reject H_0	OK	type II error	do not convict	OK	

Thus, the *higher* level necessary to secure conviction in a criminal trial corresponds to having *lower* significance levels. This analogy should not be taken too far. The nature of the data and the decision making process is quite dissimilar. For example, the prosecutor and the defense attorney are not always out to find the most honest manner to present information. In statistical inference for hypothesis testing, the goal is something that all participants in this endeavor ought to share.

The decision for the test is often based on first determining a **critical region** C. Data $\mathbf{x}$ in this region is determined to be too unlikely to have occurred when the null hypothesis is true. Thus, the decision is

$$\text{reject } H_0 \quad \text{if and only if} \quad \mathbf{x} \in C.$$

Given a choice α for the size of the test, the choice of a critical region C is called **best** or **most powerful** if for any other choice of critical region C^* for a size α test, i.e., both critical region lead to the same type I error probability,

$$\alpha = P_{\theta_0}\{X \in C\} = P_{\theta_0}\{X \in C^*\},$$

but perhaps different type II error probabiities

$$\beta = P_{\theta_1}\{X \notin C\}, \quad \beta^* = P_{\theta_1}\{X \notin C^*\},$$

we have the lowest probability of a type II error, ($\beta \leq \beta^*$) associated to the critical region C.

The two approaches to hypothesis testing, classical and Bayesian, begin with distinct starting points and end with different interpretations for implications of the data. Interestingly, both approaches result in a decision that is based on the values of a likelihood ratio. In the classical approach, we shall learn, based on the Neyman-Pearson lemma, that the decision is based on a level for this ratio based on setting the type I error probabilities. In the Bayesian approach, the decision on minimizing risk, a concept that we will soon define precisely.

17.2 The Neyman-Pearson Lemma

Many critical regions are either determined by the consequences of the **Neyman-Pearson lemma** or by using analogies of this fundamental lemma. Rather than presenting a proof of this lemma, we will provide some intuition for the choice of critical region through the following "game".

We will conduct a single observation X that can take values from -11 to 11 and based on that observation, decide whether or not to reject the null hypothesis. Basing a decision on a single observation, of course, is not the usual circumstance for hypothesis testing. We will first continue on this line of reasoning to articulate the logic behind the Neyman-Pearson lemma before examining more typical and reasonable data collection protocols.

To begin the game, corresponding to values for x running from -11 to 11, write a row of the number from 0 up to 10 and back down to 0 and add an additional 0 at each end. These numbers add to give 100. Now, scramble the numbers and write them under the first row. This can be created and displayed quickly in R using the commands:

```
> x<-c(-11:11)
> L0<-c(0,0:10,9:0,0)
> L1<-sample(L0,length(L0))
> data.frame(x,L0,L1)
```

The top row, giving the values of L_0, represents the likelihood for our one observation under the null hypothesis. The bottom row, giving the values of L_1, represents the likelihood under the alternative hypothesis. Note that the values for L_1 is a rearrangement of the values for L_0. Here is the output from one simulation.

x	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
$L_0(x)$	0	0	1	2	3	4	5	6	7	8	9	10	9	8	7	6	5	4	3	2	1	0	0
$L_1(x)$	3	8	7	5	7	1	3	10	6	0	6	4	2	5	0	1	0	4	0	8	2	9	9

The goal of this game is to pick values x so that your accumulated points increase as quickly as possible from your likelihood L_0 keeping your opponent's points from L_1 as low as possible. The natural start is to pick values of x so that $L_1(x) = 0$. Then, the points you collect begin to add up without your opponent gaining anything. We find 4 such values for x and record their values along with running totals for L_0 and L_1. 0

x	-2	3	5	7	
L_0 total	8	15	20	23	
L_1 total	0	0	0	0	

Being ahead by a score of 23 to 0 can be translated into a best critical region in the following way. If we take as our critical region C all the values for x except -2, 3, 5, and 7. Then, because the L_0-total is 23 points out of a possible 100, then we find

$$\text{the size of the test } \alpha = 0.77 \quad \text{and} \quad \text{the power of the test } 1 - \beta = 1.00$$

because there is *no* chance of type II error with this critical region. If the results of our one observation is one of -2, 3, 5, or 7, then we are never incorrect in failing to reject H_0.

Understanding the next choice is crucial. Candidates are

$$x = 4, \text{with } L_0(4) = 6 \text{ against } L_1(4) = 1 \quad \text{and} \quad x = 1, \text{with } L_0(1) = 9 \text{ against } L_1(1) = 2.$$

The choice 6 against 1 is better than 9 against 2. One way to see this is to note that choosing 6 against 1 twice will put us in a better place than the single choice of 9 against 2. Indeed, after choosing 6 against 1, a choice of 3 against 1 puts us in at least as good a position than the single choice of 9 against 2. The central point is that the best choice comes to picking the remaining value for x that has the *highest ratio* of $L_0(x)$ to $L_1(x)$

Now we can pick the next few candidates, keeping track of the size and the power of the test with the choice of critical region being the values of x not yet chosen.

x	-2	3	5	7	4	1	-6	0	-5	
$L_0(x)/L_1(x)$	∞	∞	∞	∞	6	9/2	4	5/2	5/3	
L_0 total	8	15	20	23	29	38	42	52	57	
L_1 total	0	0	0	0	1	3	4	8	11	
α	0.92	0.85	0.80	**0.77**	0.71	0.62	0.58	0.48	0.43	
$1 - \beta$	1.00	1.00	1.00	**1.00**	0.99	0.97	0.96	0.92	0.89	

From this exercise we see how the likelihood ratio test is the choice for a most powerful test. For example, for these likelihoods, the last column states that for a $\alpha = 0.43$ level test, the best region consists of those values of x so that

$$\frac{L_1(x)}{L_0(x)} \geq \frac{5}{3}.$$

The power is $1 - \beta = 0.89$ and thus the type II error probability is $\beta = 0.11$. In genuine examples, we will typically look for level much below 0.43 and we will make many observations. We now summarize carefully the insights from this game before examining more genuine examples. A proof of this theorem is provided in Section 17.4.

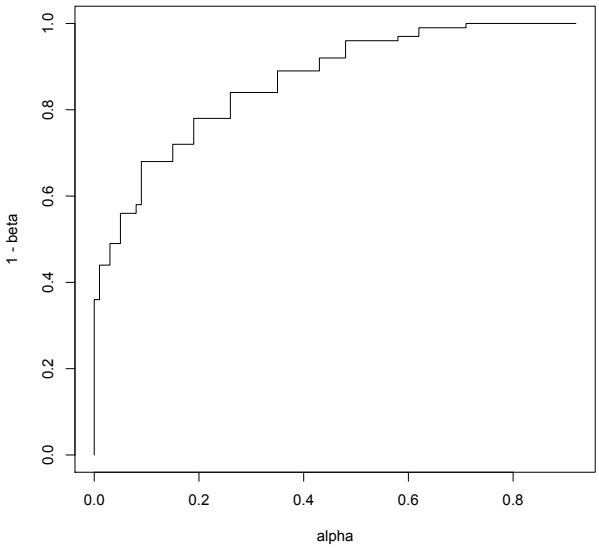


Figure 17.1: Receiver Operating Characteristic. The graph of $\alpha = P\{X \in C|H_0 \text{ is true}\}$ (significance) versus $1-\beta = P\{X \in C|H_1 \text{ is true}\}$ (power) in the example. The horizontal axis α is also called the **false positive fraction (FPF)**. The vertical axis $1-\beta$ is also called the **true positive fraction (TPF)**.

Theorem 17.1 (Neyman-Pearson Lemma)**.** *Let $L(\theta|\mathbf{x})$ denote the likelihood function for the random variable X corresponding to the probability P_θ. If there exists a critical region C of size α and a nonnegative constant k_α such that*

$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} \geq k_\alpha \quad \textit{for } \mathbf{x} \in C$$

and

$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} < k_\alpha \quad \textit{for } \mathbf{x} \notin C, \tag{17.1}$$

then C is the most powerful critical region of size α.

We, thus, reject the null hypothesis if and only if the likelihood ratio exceeds a value k_α with

$$\alpha = P_{\theta_0}\left\{\frac{L(\theta_1|X)}{L(\theta_0|X)} \geq k_\alpha\right\}.$$

We shall learn that many of the standard tests use critical values for the t-statistic, the chi-square statistic, or the F-statistic. These critical values are related to the critical value k_α in extensions of the ideas of likelihood ratios. In a few pages, we will take a glance at the Bayesian approach to hypothesis testing.

17.2.1 The Receiver operating Characteristic

Using R, we can complete the table for L_0 total and L_1 total.

```
> o<-order(L1/L0)
> sumL0<-cumsum(L0[o])
> sumL1<-cumsum(L1[o])
> alpha<-1-sumL0/100
> beta<-sumL1/100
> data.frame(x[o],L0[o],L1[o],sumL0,sumL1,alpha,1-beta)
```

Completing the curve, known as the **receiver operating characteristic (ROC)**, is shown in the figure above. The ROC shows the inevitable trade-offs between Type I and Type II errors. For example, by the mere fact that the graph is increasing, we can see that by setting a more rigorous test achieved by lowering α, the level of significance, (decreasing the value on the horizontal axis) necessarily reduces $1-\beta$, the power (decreasing the value on the vertical axis.). The unusual and slightly mystifying name is due to the fact that the ROC was first developed during World War II for detecting enemy objects in battlefields, Following the surprise military attack on Pearl Harbor in 1941, the United States saw the need to improve the prediction of the movement of aircraft from their radar signals.

Exercise 17.2. *Consider the following (ignorant) example. Flip a coin that gives heads with probability α. Ignore whatever data you have collected and reject if the coin turns up heads. This test has significance level α. Show that the receiver operating characteristic curve is the line through the origin having slope 1.*

This shows what a minimum acceptable ROC curve looks like - any hypothesis test ought be better than a coin toss that ignores the data. The ROC can be used as a test diagnostic. One commonly used is the area under the ROC, (AUC). For the example above, the AUC is 1/2. So any test should be improve on that value. The "nearly perfect test" would have have the power near to 1 for even very low significance level. In this case the AUC is very nearly equal to 1.



Figure 17.2: Heliconius butterflies

17.3 Examples

Example 17.3. *Mimicry is the similarity of one species to another in a manner that enhances the survivability of one or both species - the* **model** *and* **mimic** *. This similarity can be, for example, in appearance, behavior, sound, or scent. One method for producing a mimic species is* **hybridization***. This results in the transferring of adaptations from the model species to the mimic. The genetic signature of this has recently been discovered in* Heliconius *butterflies. Padro-Diaz et al sequenced chromosomal regions both linked and unlinked to the red color locus and found a region that displays an almost perfect genotype by phenotype association across four species in the genus* Heliconius

Let's consider a model butterfly species with mean wingspan $\mu_0 = 10$ cm and a mimic species with mean wingspan $\mu_1 = 7$ cm. Both species have standard deviation $\sigma_0 = 3$ cm. Collect 16 specimen to decide if the mimic species has migrated into a given region. If we assume, for the null hypothesis, that the habitat under study is populated by the model species, then

- *a type I error is falsely concluding that the species is the mimic when indeed the model species is resident and*
- *a type II error is falsely concluding that the species is the model when indeed the mimic species has invaded.*

If our action is to begin an eradication program if the mimic has invaded, then a type I error would result in the eradication of the resident model species and a type II error would result in the letting the invasion by the mimic take its course.

To begin, we set a significance level. The choice of an $\alpha = 0.05$ test means that we are accepting a 5% chance of having this error. If the goal is to design a test that has the lowest type II error probability, then the Neyman-Pearson lemma tells us that the critical region is determined by a threshold level k_α for the likelihood ratio.

$$C = \left\{ \mathbf{x}; \frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} \geq k_\alpha \right\}.$$

We next move to see how this critical region is determined.

Example 17.4. *Let $X = (X_1, \dots, X_n)$ be independent normal observations with unknown mean and known variance σ_0^2. The hypothesis is*

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_1 : \mu = \mu_1. \tag{17.2}$$

For the moment consider the case in which $\mu_1 < \mu_0$. We look to determine the critical region.

$$\begin{aligned}
\frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} &= \frac{\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp -\frac{(x_1-\mu_1)^2}{2\sigma_0^2} \cdots \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp -\frac{(x_n-\mu_1)^2}{2\sigma_0^2}}{\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp -\frac{(x_1-\mu_0)^2}{2\sigma_0^2} \cdots \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp -\frac{(x_n-\mu_1)^2}{2\sigma_0^2}} \\
&= \frac{\exp -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_1)^2}{\exp -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_0)^2} \\
&= \exp -\frac{1}{2\sigma_0^2} \sum_{i=1}^n \left((x_i - \mu_1)^2 - (x_i - \mu_0)^2 \right) \\
&= \exp -\frac{\mu_0 - \mu_1}{2\sigma_0^2} \sum_{i=1}^n (2x_i - \mu_1 - \mu_0)
\end{aligned}$$

Because the exponential function is increasing, the likelihood ratio test (17.1) is equivalent to

$$\frac{\mu_1 - \mu_0}{2\sigma_0^2} \sum_{i=1}^n (2x_i - \mu_1 - \mu_0), \tag{17.3}$$

exceeding some critical value. Continuing to simplify, this is equivalent to $\bar{x}$ bounded by some critical value,

$$\bar{x} \le \tilde{k}_\alpha,$$

where $\tilde{k}_\alpha$ is chosen to satisfy

$$P_{\mu_0}\{\bar{X} \le \tilde{k}_\alpha\} = \alpha.$$

(Note that division by the negative number $\mu_1 - \mu_0$ reverses the direction of the inequality.) Pay particular attention to the fact that the probability is computed under the null hypothesis specifying the mean to be μ_0. In this case, $\bar{X}$ is $N(\mu_0, \sigma_0/\sqrt{n})$ and consequently the standardized version of $\bar{X}$,

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}, \tag{17.4}$$

is a standard normal. Set z_α so that $P\{Z \le -z_\alpha\} = \alpha$. (This can be determined in R *using the* `qnorm` *command.) Then, by rearranging (17.4), we can determine $\tilde{k}_\alpha$.*

$$\bar{X} \le \mu_0 - z_\alpha \frac{\sigma_0}{\sqrt{n}} = \tilde{k}_\alpha.$$

.

Equivalently, we can use the standardized score Z as our test statistic and z_α as the critical value. Note that the only role played by μ_1, the value of the mean under the alternative, is that is less than μ_0. However, it will play a role in determining the power of the test.

Exercise 17.5. *In the example above, give the value of $\tilde{k}_\alpha$ explicitly in terms of $k_\alpha, \mu_0, \mu_1, \sigma_0^2$ and n.*

Returning to the example of the model and mimic bird species, we now see, by the Neyman-Person lemmaa that the critical region can be defined as

$$C = \left\{ \mathbf{x}; \bar{x} \le \tilde{k}_\alpha \right\} = \left\{ \mathbf{x}; \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \le -z_\alpha \right\}.$$

Under the null hypothesis, $\bar{X}$ has a normal distribution with mean $\mu_0 = 10$ and standard deviation $\sigma/\sqrt{n} = 3/4$. This using the distribution function of the normal we can find either $\tilde{k}_\alpha$

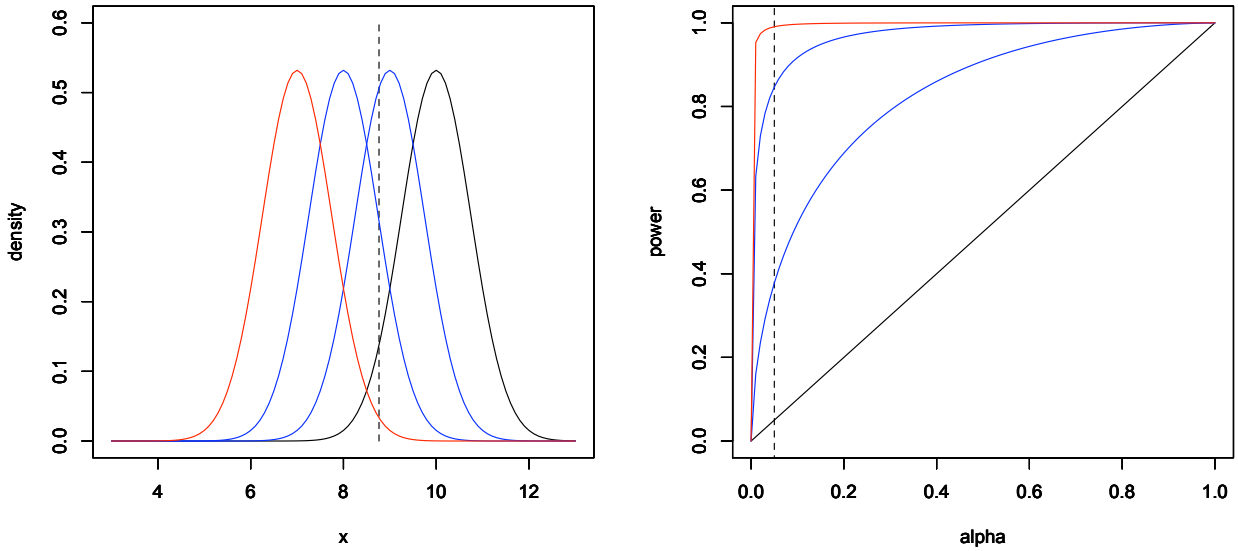


Figure 17.3: Left: (black) Density of $\bar{X}$ for normal data under the null hypothesis - $\mu_0 = 10$ and $\sigma_0/\sqrt{n} = 3/\sqrt{16} = 3/4$. With an $\alpha = 0.05$ level test, the critical value $\tilde{k}_\alpha = \mu_0 - z_\alpha \sigma_0/\sqrt{n} = 8.766$. Thus, the area to the left of the vertical dashed line and below the black density function is the significance level $\alpha = P_{\mu_0}\{\bar{X} \le k_\alpha\}$. The alternatives shown are $\mu_1 = 9$ and 8 (in blue) and $\mu_1 = 7$ (in red). The areas below these curves and to the left of the dashed line is the power $1 - \beta = P_{\mu_1}\{\bar{X} \le k_\alpha\}$. These values are 0.3777, 0.8466, and 0.9907 for respective alternatives $\mu_1 = 9, 8$ and 7. **Right**: The corresponding receiver operating characteristics curves of the power $1 - \beta$ versus the significance α using equation (17.6). The power for an $\alpha = 0.05$ test are indicated by the intersection of vertical dashed line and the receiver operating characteristics curves.

```
> qnorm(0.05,10,3/4)
[1] 8.76636
```

or $-z_\alpha$,

```
> qnorm(0.05)
[1] -1.644854
```

Thus, the critical value is $\tilde{k}_\alpha = 8.767$ for the test statistic $\bar{x}$ and $z_\alpha = -1.645$ for the test statistic z. Now let's look at data.

```
> x
 [1]  8.9  2.4 12.1 10.0  9.2  3.7 13.9  9.1  8.8  6.3 12.1 11.0 12.5  4.5  8.2 10.2
> mean(x)
[1] 8.93125
```

Then

$$\bar{x} = 8.931 \quad z = \frac{8.93124 - 10}{3/\sqrt{16}} = -1.425.$$

$\tilde{k}_\alpha = 8.766 < 8.931$ or $z_\alpha = -1.645 < -1.425$ and we fail to reject the null hypothesis.

Exercise 17.6. *Modify the calculations in the example above to show that for the case $\mu_0 < \mu_1$, using the same value of z_α as above, the we reject the null hypothesis precisely when*

$$\bar{X} \ge \mu_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}. \quad \textit{or} \quad Z \ge z_\alpha$$

Exercise 17.7. *Give an intuitive explanation why the power should*

- increase *as a function of* $|\mu_1 - \mu_0|$,
- decrease *as a function of* σ_0^2, *and*
- increase *as a function of* n.

Next we determine the type II error probability for the situation given by the previous exercise. We will be guided by the fact that

$$\frac{\bar{X} - \mu_1}{\sigma_0/\sqrt{n}}$$

is a *standard normal random variable* In the case that $H_1 : \mu = \mu_1$ is true.

For $\mu_1 > \mu_0$, we find that the type II error probability

$$\begin{aligned}\beta = P_{\mu_1}\{X \notin C\} &= P_{\mu_1}\{\bar{X} < \mu_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}\} \\ &= P_{\mu_1}\left\{\frac{\bar{X} - \mu_1}{\sigma_0/\sqrt{n}} < z_\alpha - \frac{|\mu_1 - \mu_0|}{\sigma_0/\sqrt{n}}\right\} = \Phi\left(z_\alpha - \frac{|\mu_1 - \mu_0|}{\sigma_0/\sqrt{n}}\right)\end{aligned}$$

and the power

$$1 - \beta = 1 - \Phi\left(z_\alpha + \frac{|\mu_1 - \mu_0|}{\sigma_0/\sqrt{n}}\right) \tag{17.5}$$

Exercise 17.8. *For* **sample size determination** *for the simple hypothesis (17.2) show that* n^**, the number of observations to obtain type I error probability* α *and type II error probability* β *must satisfy*

$$n^* \geq \frac{\sigma_0^2}{(\mu_1 - \mu_0)^2}(z_\alpha + z_\beta)^2.$$

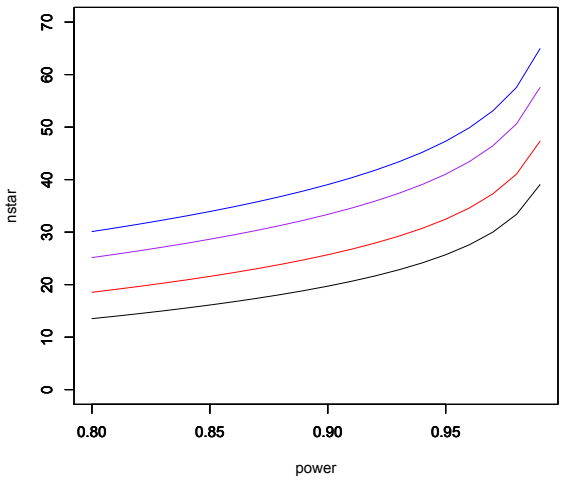


Figure 17.4: Sample size determination for the simple hypothesis (17.2) . Minimum sample sample size versus power for significance level $\alpha = 0.10$ (black), 0.05 (red), 0.02 (purple), and 0.01 (blue).

Notice that n^*

- *decreases* as a function of $|\mu_1 - \mu_0|$,
- *increases* as a function of σ_0^2, and
- *decreases* as a function of α and β. In other words, n^* *increases* as we decrease either type I or type II error.

Exercise 17.9. *Modify the calculations of power in (17.5) above to show that for the case* $\mu_1 < \mu_0$ *to show that*

$$1 - \beta = \Phi\left(-z_\alpha - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}}\right). \tag{17.6}$$

A type II error is falsely failing to conclude that the mimic species have inhabited the study area when indeed they have. To compute the probability of a type II error, note that for $\alpha = 0.05$, we substitute into (17.6),

$$-z_\alpha + \frac{\mu_0 - \mu_1}{\sigma_0/\sqrt{n}} = -1.645 + \frac{3}{3/\sqrt{16}} = 2.355$$

```
>  pnorm(2.355)
[1] 0.9907386
```

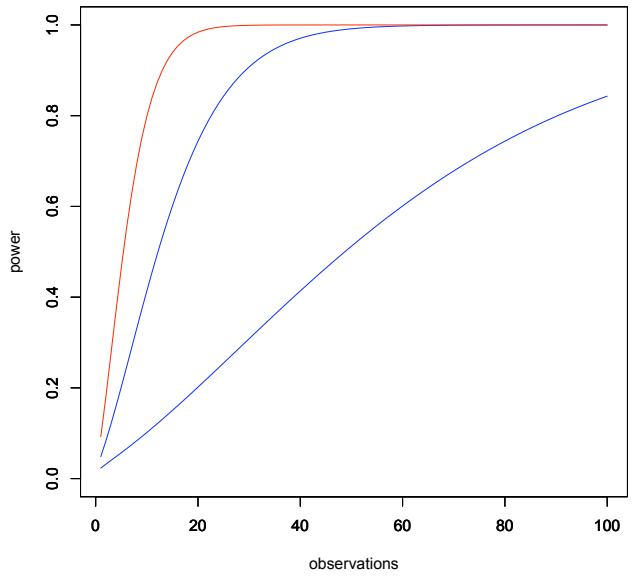


Figure 17.5: Power as a function of the number of observations for an $\alpha = 0.01$ level test. The null hypothesis - $\mu_0 = 10$. The alternatives shown are $\mu_1 = 9$ and 8 (in blue) and $\mu_1 = 7$ (in red). Here $\sigma_0 = 3$. The low level for α is chosen to reflect the desire to have a stringent criterion for rejecting the null hypothesis that the resident species is the model species.

and the type II error probability is $\beta = 1 - 0.9907 = 0.0093$, a bit under 1%.

Let's expand the examination of equation (17.6). As we move the alternative value μ_1 downward, the density of $\bar{X}$ moves leftward. The values for $\mu_1 = 9, 8$, and 7 are displayed on the left in Figure 17.2. This shift in the values is a way of saying that the alternative is becoming more and more distinct as μ_1 decreases. The mimic species becomes easier and easier to detect. We express this by showing that the test is more and more powerful with decreasing values of μ_1. This is displayed by the increasing area under the density curve to the left of the dashed line from 0.377 for the alternative $\mu_1 = 9$ to 0.9907 for $\mu_1 = 7$. We can also see this relationship in the receiver operating characteristic graphed, the graph of the power $1 - \beta$ versus the significance α. This is displayed for the significance level $\alpha = 0.05$ by the dashed line.

Exercise 17.10. *Determine the power of the test for $\mu_0 = 10$ cm and $\mu_1 = 9, 8$, and 7 cm with the significance level $\alpha = 0.01$. Does the power increase or decrease from its value when $\alpha = 0.01$? Explain your answer. How would the graphs in Figure 17.2 be altered to show this case?*

Often, we wish to know in advance the number of observations n needed to obtain a given power. In this case, we use (17.5) with a fixed value of α, the size of the test, and determine the power of the test as a function of n. We display this in Figure 17.4 with the value of $\alpha = 0.01$. Notice how the number of observations needed to achieve a desired power is high when the wingspan of the mimic species is close to that of the model species.

The example above is called the z-test. If n is sufficiently large, then even if the data are not normally distributed, $\bar{X}$ is well approximated by a normal distribution and, as long as the variance σ_0^2 is known, the z-test is used in this case. In addition, the z-test can be used when $g(\bar{X}_1, \ldots, \bar{X}_n)$ can be approximated by a normal distribution using the delta method.

Example 17.11 (Bernoulli trials)**.** *Here $X = (X_1, \ldots, X_n)$ is a sequence of Bernoulli trials with unknown success probability p, the likelihood*

$$\begin{aligned} L(p|\mathbf{x}) &= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} = p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \\ &= (1-p)^n \left(\frac{p}{1-p}\right)^{x_1+\cdots+x_n} \end{aligned}$$

For the test

$$H_0 : p = p_0 \quad \textit{versus} \quad H_1 : p = p_1$$

the likelihood ratio

$$\frac{L(p_1|\mathbf{x})}{L(p_0|\mathbf{x})} = \left(\frac{1-p_1}{1-p_0}\right)^n \left(\left(\frac{p_1}{1-p_1}\right) \Big/ \left(\frac{p_0}{1-p_0}\right)\right)^{x_1+\cdots+x_n}. \tag{17.7}$$

Exercise 17.12. *Show that the likelihood ratio (17.7) results in a test to reject H_0 whenever*

$$\sum_{i=1}^n x_i \geq \tilde{k}_\alpha \text{ when } p_0 < p_1 \quad \text{or} \quad \sum_{i=1}^n x_i \leq \tilde{k}_\alpha \text{ when } p_0 > p_1. \tag{17.8}$$

In words, if the alternative is a higher proportion than the null hypothesis, we reject H_0 when the data have too many successes. If the alternative is lower than the null, we eject H_0 when the data do not have enough successes .

In either situation, the number of successes $N = \sum_{i=1}^n X_i$ has a $Bin(n, p_0)$ distribution under the null hypothesis. Thus, in the case $p_0 < p_1$, we choose $\tilde{k}_\alpha$ so that

$$P_{p_0}\left\{\sum_{i=1}^n X_i \geq \tilde{k}_\alpha\right\} \leq \alpha. \tag{17.9}$$

In general, we cannot choose k_α to obtain exactly the value α. Thus, we take the minimum value of k_α to achieve the inequality in (17.9).

To give a concrete example take $p_0 = 0.6$ and $n = 20$ and look at a part of the cumulative distribution function.

x	$\cdots$	13	14	15	16	17	18	19	20
$F_N(x) = P\{N \leq x\}$	$\cdots$	0.7500	0.8744	0.9491	0.9840	0.9964	0.9994	0.99996	1

If we take $\alpha = 0.05$, then

$$P\{N \geq 16\} = 1 - P\{N \leq 15\} = 1 - 0.9491 = 0.0509 > 0.05$$
$$P\{N \geq 17\} = 1 - P\{N \leq 16\} = 1 - 0.9840 = 0.0160 < 0.05$$

Consequently, we need to have at least 17 successes in order to reject H_0.

Exercise 17.13. *Find the critical region in the example above for $\alpha = 0.10$ and $\alpha = 0.01$. For what values of α is $C = \{16, 17, 18, 19, 20\}$ a critical region for the likelihood ratio test.*

Example 17.14. *If np_0 and $n(1-p_0)$ are sufficiently large, then, by the central limit theorem, $\sum_{i=1}^n X_i$ has approximately a normal distribution. If we write the sample proportion*

$$\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i,$$

then, under the null hypothesis, we can apply the central limit theorem to see that

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

is approximately a standard normal random variable and we perform the z-test as in the previous exercise.

For example, if we take $p_0 = 1/2$ and $p_1 = 3/5$ and $\alpha = 0.05$, then with 110 heads in 200 coin tosses

$$Z = \frac{0.55 - 0.50}{0.05/\sqrt{2}} = \sqrt{2}.$$

```
> qnorm(0.95)
[1] 1.644854
```

Thus, $\sqrt{2} < 1.645 = z_{0.05}$ *and we fail to reject the null hypothesis.*

Example 17.15. *Honey bees store honey for the winter. This honey serves both as nourishment and insulation from the cold. Typically for a given region, the probability of survival of a feral bee hive over the winter is* $p_0 = 0.7$*. We are checking to see if, for a particularly mild winter, this probability moved up to* $p_1 = 0.8$*. This leads us to consider the hypotheses*

$$H_0 : p = p_0 \quad \textit{versus} \quad H_1 : p = p_1.$$

for a test of the probability that a feral bee hive survives a winter. If we use the central limit theorem, then, under the null hypothesis,

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

has a distribution approximately that of a standard normal random variable. For an α *level test, the critical value is* z_α *where* α *is the probability that a standard normal is at least* z_α *If the significance level is* $\alpha = 0.05$*, then we will reject* H_0 *for any value of* $z > z_\alpha = 1.645$

For this study, 112 colonies have been chosen and 88 survive. Thus $\hat{p} = 0.7875$ *and*

$$z = \frac{0.7875 - 0.7}{\sqrt{0.7(1-0.7)/112}} = 1.979.$$

Consequently, reject H_0*.*

For both of these previous examples, the usual method is to compute the z-score with the continuity correction. We shall soon see this with the use of `prop.test` in R.

17.4 Summary

For a simple hypothesis

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

we have two possible action, **reject** H_0 and **fail to reject** H_0, this leads to two possible types of errors

error	probability	alternative names
type I	$\alpha = P_{\theta_0}\{\text{reject } H_0\}$	level significance false positive
type II	$\beta = P_{\theta_1}\{\text{fail to reject } H_0\}$	false negative

The probability $1 - \beta = P_{\theta_1}\{\text{reject } H_0\}$ is called the true positive probability or **power** or **sensitivity**. The probability $1 - \alpha = P_{\theta_0}\{\text{fail to reject } H_0\}$ is called the **specificity**.

The procedure is to set a **significance level** α and find a critical region C so that the type II error probability is as small as possible. The Neyman-Pearson lemma lets us know that the critical region is determined by setting a level k_α for the likelihood ratio.

$$C = \left\{ \mathbf{x}; \frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} \geq k_\alpha \right\}$$

We continue, showing the procedure in the examples above.

	normal observations $\mu_1 \geq \mu_0$	Bernoulli trials $p_1 > p_0$
Simplify likelihood ratio to obtain a test statistic $T(\mathbf{x})$	$\bar{x}$ $z = \frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}$	$\sum_{i=1}^n x_i$
Use the distribution of $T(\mathbf{x})$ under H_0 to set a critical value $\tilde{k}_\alpha$ so that $P_{\theta_0}\{T(X) \geq \tilde{k}_\alpha\} = \alpha$	$\bar{X} \sim N(\mu_0, \sigma_0/\sqrt{n})$ $Z \sim N(0,1)$	$\sum_{i=1}^n X_i \sim Bin(n, p_0)$
Determine type II error probability $\beta = P_{\theta_1}\{T(X) \geq \tilde{k}_\alpha\}$	$P_{\mu_1}\{\bar{X} \geq \tilde{k}_\alpha\}$	$P_{p_1}\{\sum_{i=1}^n X_i \geq \tilde{k}_\alpha\}$

17.5 Proof of the Neyman-Pearson Lemma

For completeness in exposition, we include a proof of the Neyman-Pearson lemma.

Let C be the α critical region determined by the likelihood ratio test. In addition, let C^* be a critical region for a second test of size α. In symbols,

$$P_{\theta_0}\{X \in C^*\} = P_{\theta_0}\{X \in C\} = \alpha \tag{17.10}$$

As before, we use the symbols β and β^* denote, respectively, the probability of type II error for the critical regions C and C^* respectively. The Neyman-Pearson lemma is the statement that $\beta^* \geq \beta$.

Divide both critical regions C and C^* into two disjoint subsets, the subset that the critical regions share $S = C \cap C^*$ and the subsets $E = C \backslash C^*$ and $E^* = C^* \backslash C$ that are exclusive to one region. In symbols, we write this as the disjoint unions

$$C = S \cup E, \quad \text{and} \quad C^* = S \cup E^*.$$

Thus under either parameter value $\theta_i, i = 1, 2$,

$$P_{\theta_i}\{X \in C\} = P_{\theta_i}\{X \in S\} + P_{\theta_i}\{X \in E\} \quad \text{and} \quad P_{\theta_i}\{X \in C^*\} = P_{\theta_i}\{X \in S\} + P_{\theta_i}\{X \in E^*\}.$$

(See Figure 17.5)

First, we will describe the proof in words.

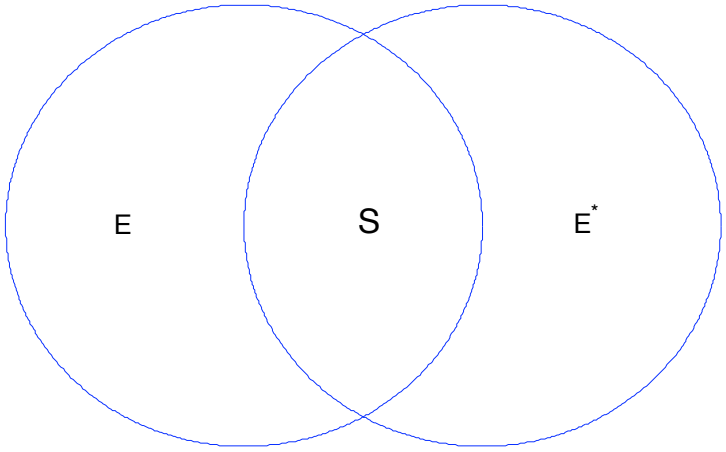


Figure 17.6: Critical region C as determined by the Neyman-Pearson lemma is indicated by the circle on the left. The circle on the right C^* is the critical region is for a second α level test. Thus, $C = S \cup E$ and $C^* = S \cup E^*$.

- The contribution to type I errors from data in S and for type II errors from data outside $E \cup E^*$ are the same for both tests. Consequently, we can focus on differences in types of error by examining the case in which the data land in either E and E^*.
- Because both test have level α, the probability that the data land in E or in E^* are the same under the null hypothesis.
- Under the likelihood ratio critical region, the null hypothesis is not rejected in E^*.
- Under the second test, the null hypothesis is not rejected in E.
- E^* is outside likelihood ratio critical region. So, under the alternative hypothesis, the probability that the data land in E is at most k_α times as large as it is under the null hypothesis. This contributes to the type II error for the likelihood ratio based test.
- E is in the likelihood ratio critical region. So, under the alternative hypothesis, the probability that the data land in E is at least k_α times as large as it is under the null hypothesis. This contributes a larger amount to the type II error for the second test than is added from E^* to the likelihood ratio based test.
- Thus, the type II error for the likelihood ratio based test is smaller than the type II error for the second test.

To carry out the proof, first consider the parameter value θ_0 and subtract from both sides in (17.10) the probability $P_{\theta_0}\{X \in S\}$ that the data land in the shared critical regions and thus would be rejected by both tests to obtain

$$P_{\theta_0}\{X \in E^*\} \geq P_{\theta_0}\{X \in E\}$$

or

$$P_{\theta_0}\{X \in E^*\} - P_{\theta_0}\{X \in E\} \geq 0. \tag{17.11}$$

Moving to the parameter value θ_1, the difference in the corresponding type II error probabilities is

$$\begin{aligned}\beta^* - \beta &= P_{\theta_1}\{X \notin C^*\} - P_{\theta_1}\{X \notin C\} \\ &= (1 - P_{\theta_1}\{X \in C^*\}) - (1 - P_{\theta_1}\{X \in C\}) = P_{\theta_1}\{X \in C\} - P_{\theta_1}\{X \in C^*\}.\end{aligned}$$

Now subtract from both of the integrals the quantity $P_{\theta_1}\{X \in S\}$, the probability that the hypothesis would be falsely rejected by both tests to obtain

$$\beta^* - \beta = P_{\theta_1}\{X \in E\} - P_{\theta_1}\{X \in E^*\} \tag{17.12}$$

We can use the likelihood ratio criterion on each of the two integrals above.

- For $\mathbf{x} \in E$, then $\mathbf{x}$ is in the critical region and consequently $L(\theta_1|\mathbf{x}) \geq k_\alpha L(\theta_0|\mathbf{x})$ and

$$P_{\theta_1}\{X \in E\} = \int_E L(\theta_1|\mathbf{x})\, d\mathbf{x} \geq k_\alpha \int_E L(\theta_0|\mathbf{x})\, d\mathbf{x} = k_\alpha P_{\theta_0}\{X \in E\}.$$

- For $\mathbf{x} \in E^*$, then $\mathbf{x}$ is not in the critical region and consequently $L(\theta_1|\mathbf{x}) \leq k_\alpha L(\theta_0|\mathbf{x})$ and

$$P_{\theta_1}\{X \in E^*\} = \int_{E*} L(\theta_1|\mathbf{x})\, d\mathbf{x} \leq k_\alpha \int_{E^*} L(\theta_0|\mathbf{x})\, d\mathbf{x} = k_\alpha P_{\theta_0}\{X \in E^*\}.$$

Apply these two inequalities to (17.12)

$$\beta^* - \beta \geq k_\alpha(P_{\theta_0}\{X \in E^*\} - P_{\theta_0}\{X \in E\}).$$

This difference is at least 0 by (17.11) and consequently $\beta^* \geq \beta$, i. e., the critical region C^* has at least as large type II error probability as that given by the likelihood ratio test.

17.6 An Brief Introduction to the Bayesian Approach

As with other aspects of the Bayesian approach to statistics, hypothesis testing is closely aligned with Bayes theorem. For a simple hypothesis, we begin with a **prior probability** for each of the competing hypotheses.

$$\pi\{\theta_0\} = P\{H_0 \text{ is true}\} \quad \text{and} \quad \pi\{\theta_1\} = P\{H_1 \text{ is true}\}.$$

Naturally, $\pi\{\theta_0\} + \pi\{\theta_1\} = 1$. Although this is easy to state, the choice of a prior ought to be grounded in solid scientific reasoning.

As before, we collect data and with it compute the **posterior probabilities** of the two parameter values θ_0 and θ_1. This gives us the posterior probabilities that H_0 is true and H_1 is true.

We can see, in its formulation, the wide difference in perspective between the Bayesian and classical approaches.

- In the Bayesian approach, we begin with a prior probability that H_0 is true. In the classical approach, the assumption is that H_0 is true.
- In the Bayesian approach, we use the data and Bayes formula to compute the posterior probability that H_1 is true. In the classical approach, we use the data and a significance level to make a decision to reject H_0. The question: *What is the probability that H_1 is true?* has *no* meaning in the classical setting.
- The decision to reject H_0 in the Bayesian setting is based on minimizing risk using presumed losses for type I and type II errors. In classical statistics, the choice of type I error probability is used to construct a critical region. This choice is made with a view to making the type II error probability as small as possible. We reject H_0 whenever the data fall in the critical region.

Both approaches use as a basic concept, the likelihood function $L(\theta|\mathbf{x})$ for the data $\mathbf{x}$. Let $\tilde{\Theta}$ be a random variable taking on one of the two values θ_0, θ_1 and having a distribution equal to the prior probability π. Thus,

$$\pi\{\theta_i\} = P\{\tilde{\Theta} = \theta_i\}, i = 0, 1.$$

Recall Bayes formula for events A and C,

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|C^c)P(C^c)},$$

we set C to be the event that the alternative hypothesis is true and A to be the event that the data take on the value $\mathbf{x}$. In symbols,

$$C = \{\tilde{\Theta} = \theta_1\} = \{H_1 \text{ is true}\} \quad \text{and} \quad A = \{X = \mathbf{x}\}.$$

Focus for the moment on the case in which the data are discrete, we have the conditional probabilities for the alternative hypothesis.

$$P(A|C) = P_{\theta_1}\{X = \mathbf{x}\} = f_X(\mathbf{x}|\theta_1) = L(\theta_1|\mathbf{x}).$$

Similarly, for the null hypothesis,

$$P(A|C^c) = P_{\theta_0}\{X = \mathbf{x}\} = f_X(\mathbf{x}|\theta_0) = L(\theta_0|\mathbf{x}).$$

The posterior probability that H_1 is true can be written symbolically in several ways.

$$f_{\tilde{\Theta}|X}(\theta_1|\mathbf{x}) = P\{H_1 \text{ is true}|X = \mathbf{x}\} = P\{\tilde{\Theta} = \theta_1|X = \mathbf{x}\}$$

Returning to Bayes formula, we make the substitutions,

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|C^c)P(C^c)},$$
$$f_{\tilde{\Theta}|X}(\theta_1|\mathbf{x}) = \frac{L(\theta_1|\mathbf{x})\pi\{\theta_1\}}{L(\theta_0|\mathbf{x})\pi\{\theta_0\} + L(\theta_1|\mathbf{x})\pi\{\theta_1\}}.$$

By making a similar argument involving limits, we can reach the same identity for the density of continuous random variables. The formula for the posterior probability can be more easily understood if we rewrite the expression above in terms of odds, i. e., as the ratio of probabilities.

$$\frac{f_{\tilde{\Theta}|X}(\theta_1|\mathbf{x})}{f_{\tilde{\Theta}|X}(\theta_0|\mathbf{x})} = \frac{P\{H_1 \text{ is true}|X=\mathbf{x}\}}{P\{H_0 \text{ is true}|X=\mathbf{x}\}} = \frac{P\{\tilde{\Theta}=\theta_1|X=\mathbf{x}\}}{P\{\tilde{\Theta}=\theta_0|X=\mathbf{x}\}} = \frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} \cdot \frac{\pi\{\theta_1\}}{\pi\{\theta_0\}}. \tag{17.13}$$

With this expression we see that the posterior odds are equal to the likelihood ratio times the prior odds.

The decision whether or not to reject H_0 depends on the values assigned for the loss obtained in making an incorrect conclusion. We begin by setting values for the loss. This can be a serious exercise in which a group of experts weighs the evidence for either adverse outcome. We will take a loss of 0 for making a correct decision, a loss of ℓ_{I} for a type I error and ℓ_{II} for a type II error. We summarize this in a table.

loss function table		
decision	H_0 is true	H_1 is true
H_0	0	ℓ_{II}
H_1	ℓ_{I}	0

The Bayes procedure is to make the decision that has the smaller posterior expected loss, also known as the **risk**. If the decision is H_0, the loss $\mathcal{L}_0(\mathbf{x})$ takes on two values

$$\mathcal{L}_0(\mathbf{x}) = \begin{cases} 0 & \text{with probability } P\{H_0 \text{ is true}|X=\mathbf{x}\}, \\ \ell_{\mathrm{II}} & \text{with probability } P\{H_1 \text{ is true}|X=\mathbf{x}\}. \end{cases}$$

The expected loss

$$E\mathcal{L}_0(\mathbf{x}) = \ell_{\mathrm{II}} P\{H_1 \text{ is true}|X=\mathbf{x}\} = \ell_{\mathrm{II}}(1 - P\{H_0 \text{ is true}|X=\mathbf{x}\}) \tag{17.14}$$

is simply the product of the loss and the probability of incorrectly choosing H_1.

If the decision is H_1, the loss $\mathcal{L}_1(\mathbf{x})$ also takes on two values

$$\mathcal{L}_1(\mathbf{x}) = \begin{cases} \ell_{\mathrm{I}} & \text{with probability } P\{H_0 \text{ is true}|X=\mathbf{x}\}, \\ 0 & \text{with probability } P\{H_1 \text{ is true}|X=\mathbf{x}\}. \end{cases}$$

In this case, the expected loss

$$E\mathcal{L}_1(\mathbf{x}) = \ell_{\mathrm{I}} P\{H_0 \text{ is true}|X=\mathbf{x}\} \tag{17.15}$$

is a product of the loss and the probability of incorrectly choosing H_0.

We can now express the Bayesian procedure in symbols using the criterion of smaller posterior expected loss:

$$\text{decide on } H_1 \quad \text{if and only if} \quad E\mathcal{L}_1(\mathbf{x}) \le E\mathcal{L}_0(\mathbf{x}).$$

Now substituting for $E\mathcal{L}_0(\mathbf{x})$ and $E\mathcal{L}_1(\mathbf{x})$ in (17.14) and (17.15), we find that we make the decision on H_1 and reject H_0 if and only if

$$\begin{aligned} \ell_{\mathrm{I}} P\{H_0 \text{ is true}|X=\mathbf{x}\} &\le \ell_{\mathrm{II}}(1 - P\{H_0 \text{ is true}|X=\mathbf{x}\}) \\ (\ell_{\mathrm{I}} + \ell_{\mathrm{II}}) P\{H_0 \text{ is true}|X=\mathbf{x}\} &\le \ell_{\mathrm{II}} \\ P\{H_0 \text{ is true}|X=\mathbf{x}\} &\le \frac{\ell_{\mathrm{II}}}{\ell_{\mathrm{I}} + \ell_{\mathrm{II}}} \end{aligned}$$

or stated in terms of odds

$$\frac{P\{H_1 \text{ is true}|X=\mathbf{x}\}}{P\{H_0 \text{ is true}|X=\mathbf{x}\}} \ge \frac{\ell_{\mathrm{I}}}{\ell_{\mathrm{II}}}, \tag{17.16}$$

we reject H_0 whenever the posterior odds exceed the ratio of the losses for each type of error.

Exercise 17.16. *Show that (17.16) follows from the inequality above it.*

As we saw in (17.13), this ratio of posterior odds is dependent on the ratio of prior odds. Taking this into account, we see that the criterion for rejecting H_0 is a level test for the likelihood ratio:

Reject H_0 if and only if

$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} \geq \frac{\ell_{\mathrm{I}}/\pi\{\theta_1\}}{\ell_{\mathrm{II}}/\pi\{\theta_0\}}. \tag{17.17}$$

Thus, the criterion for rejecting H_0 is a level test on the likelihood ratio. This is exactly the same type of criterion as that used in classical statistics. However, the rationale, thus the value for the ratio necessary to reject, is quite different. For example, the lower the value of the prior odds, the higher the likelihood ratio needed to reject H_0 under the Bayesian framework.

Example 17.17. *For normal observations with means μ_0 for the null hypothesis and μ_1 for the alternative hypothesis. If the variance has a known value, σ_0, we have from Example 17.4, the likelihood ratio*

$$\frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} = \exp -\frac{\mu_0-\mu_1}{2\sigma_0^2}\sum_{i=1}^{n}(2x_i-\mu_1-\mu_0) = \exp\left(-\frac{\mu_0-\mu_1}{2\sigma_0^2}n(2\bar{x}-\mu_1-\mu_0)\right). \tag{17.18}$$

For Example 17.3 on the model and mime butterfly species, $\mu_0 = 10$, $\mu_1 = 7$, $\sigma_0 = 3$, and sample mean $\bar{x} = 8.931$ based on $n = 16$ observations, we find the likelihood ratio 9.96. Thus,

$$\frac{P\{H_1 \text{ is true}|X=\mathbf{x}\}}{P\{H_0 \text{ is true}|X=\mathbf{x}\}} = \frac{P\{\tilde{M}=\mu_1|X=\mathbf{x}\}}{P\{\tilde{M}=\mu_0|X=\mathbf{x}\}} = 9.96\frac{\pi\{\mu_1\}}{\pi\{\mu_0\}}.$$

where $\tilde{M}$ is a random variable having a distribution equal to the prior probability π for the model and mimic butterfly wingspan. Consequently, the posterior odds for the mimic vs. mime species is approximately ten times the prior odds.

Finally, the decision will depend on the ratio of $\ell_{\mathrm{II}}/\ell_{\mathrm{I}}$, i. e., the ratio of the loss due to letting the invasion by the mimic take its course versus eradication of the resident model species.

Exercise 17.18. *Substitute the likelihood ratio in 17.18 into 17.17 and solve in terms of $\bar{x}$. Use this to determine threshold values for $barx$ to reject H_0 for prior probabilities $\pi\{\theta_0\} = 0.0, 0.10, 0.20$ and lost ratios $\ell_{\mathrm{II}}/\ell_{\mathrm{I}} = 1/2, 1, 2$. What situations give the lowest and highest threshold values for $\bar{x}$? Explain your answer.*

Exercise 17.19. *Returning to a previous example, give the likelihood ratios for $n = 20$ Bernoulli trials with $p_0 = 0.6$ and $p_1 = 0.7$ for values $x = 0, \ldots, 20$ for the number of successes. Give the values for the number of successes in which the number of successes change the prior odds by a factor of 5 or more as given by the posterior odds.*

17.7 Answers to Selected Exercises

17.2 Flip a biased coin in which the probability of heads is α under both the null and alternative hypotheses and reject whenever heads turns up. Then

$$\alpha = P_{\theta_0}\{\text{heads}\} = P_{\theta_1}\{\text{heads}\} = 1-\beta.$$

Thus, the receiver operating characteristic curve is the line through the origin having slope 1.

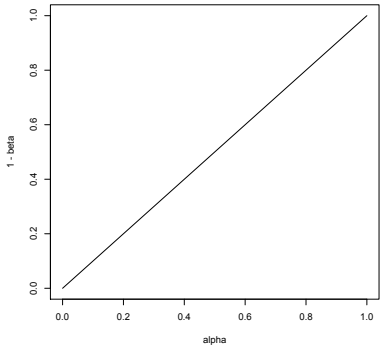


Figure 17.7: Receiver operating Characteristic based on a biased coin toss. Thus, any viable ROC should be above the line the the graph.

17.4. The likelihood ratio

$$\frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} = \exp -\frac{\mu_0-\mu_1}{2\sigma_0^2}\sum_{i=1}^{n}(2x_i-\mu_1-\mu_0) \geq k_\alpha.$$

Thus,

$$\begin{aligned}\tfrac{\mu_1-\mu_0}{2\sigma_0^2}\textstyle\sum_{i=1}^n(2x_i-\mu_1-\mu_0) &\geq \ln k_\alpha\\ \textstyle\sum_{i=1}^n(2x_i-\mu_1-\mu_0) &\geq \tfrac{2\sigma_0^2}{\mu_1-\mu_0}\ln k_\alpha\\ 2\bar{x}-\mu_1-\mu_0 &\leq \tfrac{2\sigma_0^2}{n(\mu_1-\mu_0)}\ln k_\alpha\\ \bar{x} &\leq \tfrac{1}{2}\left(\tfrac{2\sigma_0^2}{n(\mu_1-\mu_0)}\ln k_\alpha+\mu_1+\mu_0\right)=\tilde{k}_\alpha\end{aligned}$$

Notice that since $\mu_1 < \mu_0$, division by $\mu_1 - \mu_0$ changes the direction of the inequality.

17.6. If c_α is the critical value in expression in (17.3) then

$$\frac{\mu_1-\mu_0}{2\sigma_0^2}\sum_{i=1}^n(2x_i-\mu_1-\mu_0)\geq c_\alpha$$

SInce $\mu_1 > \mu_0$, division by $\mu_1 - \mu_0$ does not change the direction of the inequality. The rest of the argument proceeds as before. we obtain that $\bar{x} \geq \tilde{k}_\alpha$.

17.7. If power means easier to distinguish using the data, then this is true when the means are farther apart, the measurements are less variable or the number of measurements increases. This can be seen explicitly is the power equation (17.5).

17.8. We shall do the case $\mu_1 > \mu_0$. the other case is similar. From equation (17.5),

$$\beta=\Phi\left(z_\alpha-\frac{|\mu_1-\mu_0|}{\sigma_0/\sqrt{n}}\right)$$

The goal is to choose n so that the argument argument $z_\alpha + \frac{|\mu_1-\mu_0|}{\sigma_0/\sqrt{n}}$ has probability β. However, we have that $-z_\beta$ has *lower* tail probability β. In other words, $\beta = \Phi(-z_\beta)$. Because Φ, the cumulative distribution function for the standard normal, is one-to-one,

$$\begin{aligned}-z_\beta &= z_\alpha-\frac{|\mu_1-\mu_0|}{\sigma_0/\sqrt{n}}\\ \sqrt{n}\frac{|\mu_1-\mu_0|}{\sigma_0} &= z_\alpha+z_\beta\\ \sqrt{n} &= \frac{\sigma_0}{|\mu_1-\mu_0|}(z_\alpha+z_\beta)\\ n &= \frac{\sigma_0^2}{(\mu_1-\mu_0)^2}(z_\alpha+z_\beta)^2\end{aligned}$$

Thus, n^*, any integer al least as large as n will have the desired type I and type II errors.

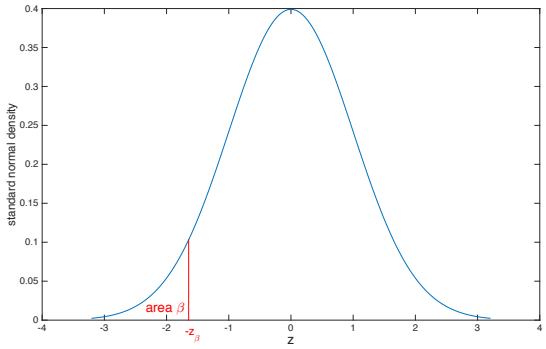


Figure 17.8: Plot of standard normal density function The value $-z_\beta$ has lower tail probabiility β. ($\beta = 0.05$ is shown.)

17.9. For $\mu_0 > \mu_1$,

$$\begin{aligned}\beta &= P_{\mu_1}\{X\notin C\}=P_{\mu_1}\{\bar{X}>\mu_0-\frac{\sigma_0}{\sqrt{n}}z_\alpha\}\\ &= P_{\mu_1}\left\{\frac{\bar{X}-\mu_1}{\sigma_0/\sqrt{n}}>-z_\alpha-\frac{\mu_1-\mu_0}{\sigma_0/\sqrt{n}}\right\}=1-\Phi\left(-z_\alpha-\frac{\mu_1-\mu_0}{\sigma_0/\sqrt{n}}\right)\end{aligned}$$

and the power

$$1-\beta=\Phi\left(-z_\alpha-\frac{\mu_1-\mu_0}{\sigma_0/\sqrt{n}}\right).$$

17.10. Interpreting equation (17.5) in R, we find that

```
> mu0<-10
> sigma0<-3
> n<-16
> zalpha<-qnorm(0.995)
> mu1<-c(9,8,7)
> power<-1-pnorm(zalpha-abs(mu1-mu0)/(sigma0/sqrt(n)))
> data.frame(mu1,power)
  mu1     power
1   9 0.1070268
2   8 0.5361891
3   7 0.9228015
```

Notice that the power has decreased from the case $\alpha = 0.05$. This could be anticipated. In reducing the significance level from $\alpha = 0.05$ to $\alpha = 0.01$, we make the criterion for rejecting more stringent by reducing he critical region C. The effect can be seen in FIgure 17.2. On the left, the vertical dashed line is moved left to reduce the area under the black curve to the left of the dashed line. This, in turn, reduces the area under the other curves to the left of the dashed line. On the right figure, the vertical dashed line is the line $\alpha = 0.01$ and, because the ROC curve is increasing, the values for the power is decreased

17.12. For the likelihood ratio (17.7), take the logarithm to obtain

$$\ln\left(\frac{L(p_1|\mathbf{x})}{L(p_0|\mathbf{x})}\right) = n\ln\left(\frac{1-p_1}{1-p_0}\right) + (x_1 + \cdots + x_n)\ln\left(\left(\frac{p_1}{1-p_1}\right) \Big/ \left(\frac{p_0}{1-p_0}\right)\right) \geq \ln k_\alpha.$$

If $p_0 < p_1$ then the ratio in the expression for the logarithm in the second term is greater than 1 and consequently, the logarithm is positive. Thus, we isolate the sum $\sum_{i=1}^{n} x_i$ to give the test (17.8). For $p_0 > p_1$, the logarithm is negative and the direction of the inequality in (17.8) is reversed.

17.13. If we take $\alpha = 0.10$, then

$$P\{N \geq 15\} = 1 - P\{N \leq 14\} = 1 - 0.8744 = 0.1256 > 0.10$$
$$P\{N \geq 16\} = 1 - P\{N \leq 15\} = 1 - 0.9491 = 0.0509 < 0.10$$

Consequently, we need to have at least 16 successes in order to reject H_0. If we take $\alpha = 0.01$, then

$$P\{N \geq 17\} = 1 - P\{N \leq 16\} = 1 - 0.9840 = 0.0160 > 0.01$$
$$P\{N \geq 18\} = 1 - P\{N \leq 17\} = 1 - 0.9964 = 0.0036 < 0.01$$

Consequently, we need to have at least 18 successes in order to reject H_0. For $C = \{16, 17, 18, 19, 20\}$,

$$P\{N \in C\} = 1 - P\{N \leq 15\} = 1 - 0.9491 = 0.0509.$$

Thus, α must be less that 0.0509 for C to be a critical region. In addition, $P\{N \geq 17\} = 0.0160$. Consequently, if we take any value for $\alpha < 0.0160$, then the critical region will be smaller than C.

17.16. Begin with

$$P\{H_0 \text{ is true}|X = \mathbf{x}\} \leq \frac{\ell_{\mathrm{I}}}{\ell_{\mathrm{I}} + \ell_{\mathrm{II}}}.$$

or, taking reciprocals,

$$\frac{1}{P\{H_0 \text{ is true}|X = \mathbf{x}\}} \geq \frac{\ell_{\mathrm{I}} + \ell_{\mathrm{II}}}{\ell_{\mathrm{I}}}.$$

Because $P\{H_0 \text{ is true}|X = \mathbf{x}\} + P\{H_1 \text{ is true}|X = \mathbf{x}\} = 1$, we have that

$$P\{H_1 \text{ is true}|X = \mathbf{x}\} \geq \frac{\ell_{\text{II}}}{\ell_{\text{I}} + \ell_{\text{II}}}.$$

Now multiply the two inequalities above. Because all the quantities are positive, the product of the left sides is greater than the product of the right sides and so the inequality (17.16) holds.

17.18. Making the substitution of 17.18 into 17.17, we have

$$\begin{aligned}
\exp\left(-\frac{\mu_0 - \mu_1}{2\sigma_0^2} n(2\bar{x} - \mu_1 - \mu_0)\right) &\geq \frac{\ell_{\text{I}}/\pi\{\theta_1\}}{\ell_{\text{II}}/\pi\{\theta_0\}} \\
\frac{\mu_1 - \mu_0}{2\sigma_0^2} n(2\bar{x} - \mu_1 - \mu_0) &\geq \ln\left(\frac{\ell_{\text{I}}/\pi\{\theta_1\}}{\ell_{\text{II}}/\pi\{\theta_0\}}\right) \\
2\bar{x} - \mu_1 - \mu_0 &\leq \frac{2\sigma_0^2}{n(\mu_1 - \mu_0)} \ln\left(\frac{\ell_{\text{I}}/\pi\{\theta_1\}}{\ell_{\text{II}}/\pi\{\theta_0\}}\right) \\
\bar{x} &\leq \frac{1}{2}\left(\frac{2\sigma_0^2}{n(\mu_1 - \mu_0)} \ln\left(\frac{\ell_{\text{I}}/\pi\{\theta_1\}}{\ell_{\text{II}}/\pi\{\theta_0\}}\right) + \mu_1 + \mu_0\right)
\end{aligned}$$

```
> mu0<-10;mu1<-7;sigma<-3;n<-16
> pi0<-c(0.05,0.10,0.20)
> lr<-1/2
> threshold<-(2*sigma^2/(n*(mu1-mu0))*log(lr*pi0/(1-pi0))+mu1+mu0)/2
> data.frame(pi0,threshold)
   pi0 threshold
1 0.05  9.182047
2 0.10  9.041945
3 0.20  8.889895
> lr<-1
> threshold<-(2*sigma^2/(n*(mu1-mu0))*log(lr*pi0/(1-pi0))+mu1+mu0)/2
> data.frame(pi0,threshold)
   pi0 threshold
1 0.05  9.052082
2 0.10  8.911980
3 0.20  8.759930
> lr<-2
> threshold<-(2*sigma^2/(n*(mu1-mu0))*log(lr*pi0/(1-pi0))+mu1+mu0)/2
> data.frame(pi0,threshold)
   pi0 threshold
1 0.05  8.922117
2 0.10  8.782015
3 0.20  8.629965
```

The lowest threshold value $\bar{x} = 8.62$ is for the case $\pi\{\theta_0\} = 0.20$ and $\ell_{\text{I}}/\ell_{\text{II}} = 2$. This is the highest prior probability and the highest relative loss for a type I error. These both require stronger evidence to reject H_0 and thus need a more extreme and thus lower value for $\bar{x}$ to reject H_0.

The highest threshold value $\bar{x} = 9.18$ is for the case $\pi\{\theta_0\} = 0.05$ and $\ell_{\text{I}}/\ell_{\text{II}} = 1/2$. This is the lowest prior probability and the lowest relative loss for a type I error. These both require less evidence to reject H_0 and thus need a less extreme and thus higher value for $\bar{x}$ to reject H_0.

17.19. Using the likelihood ratio formula in Example 17.9.

```
> x<-c(0:20)
> n<-20
> p0<-0.6
> p1<-0.7
> ratio<-((1-p1)/(1-p0))^n*((p1/(1-p1))/(p0/(1-p0)))^x
> data.frame(x[1:7],ratio[1:7],x[8:14],ratio[8:14],x[15:21],ratio[15:21])

  x.1.7.  ratio.1.7. x.8.14. ratio.8.14. x.15.21. ratio.15.21.
1      0 0.003171212       7  0.06989143       14     1.540361
2      1 0.004932996       8  0.10872001       15     2.396118
3      2 0.007673550       9  0.16912001       16     3.727294
4      3 0.011936633      10  0.26307558       17     5.798013
5      4 0.018568096      11  0.40922867       18     9.019132
6      5 0.028883705      12  0.63657794       19    14.029761
7      6 0.044930208      13  0.99023235       20    21.824072
```

Thus, values $x \geq 17$ increase the posterior odds in favor of H_1 by a factor greater than 5. values $x \leq 9$ increase the posterior odds in favor of H_0 by a factor greater than 5.

Topic 18

Composite Hypotheses

Simple hypotheses limit us to a decision between one of two possible states of nature. This limitation does not allow us, under the procedures of hypothesis testing to address the basic question:

> *Does the length, the reaction rate, the fraction displaying a particular behavior or having a particular opinion, the temperature, the kinetic energy, the Michaelis constant, the speed of light, mutation rate, the melting point, the probability that the dominant allele is expressed, the elasticity, the force, the mass, the parameter value θ_0 increase, decrease or change at all under under a different experimental condition?*

18.1 Partitioning the Parameter Space

This leads us to consider **composite hypotheses**. In this case, the parameter space Θ is divided into two disjoint regions, Θ_0 and Θ_1. The hypothesis test is now written

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1.$$

Again, H_0 is called the **null hypothesis** and H_1 the **alternative hypothesis**.

For the three alternatives to the question posed above, let θ be one of the components in the parameter space, then

- increase would lead to the choices $\Theta_0 = \{\theta; \theta \leq \theta_0\}$ and $\Theta_1 = \{\theta; \theta > \theta_0\}$,
- decrease would lead to the choices $\Theta_0 = \{\theta; \theta \geq \theta_0\}$ and $\Theta_1 = \{\theta; \theta < \theta_0\}$, and
- change would lead to the choices $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta; \theta \neq \theta_0\}$

for some choice of parameter value θ_0. The effect that we are meant to show, here the nature of the change, is contained in Θ_1. The first two options given above are called **one-sided tests**. The third is called a **two-sided test**,

Rejection and failure to reject the null hypothesis, critical regions, C, and type I and type II errors have the same meaning for a composite hypotheses as it does with a simple hypothesis. Significance level and power will necessitate an extension of the ideas for simple hypotheses.

18.2 The Power Function

Power is now a function of the parameter value θ. If our test is to reject H_0 whenever the data fall in a **critical region** C, then the **power function** is defined as

$$\pi(\theta) = P_\theta\{X \in C\}.$$

that gives the probability of rejecting the null hypothesis for a given value of the parameter.

- For $\theta \in \Theta_0$, $\pi(\theta)$ is the probability of making a type I error, i.e., rejecting the null hypothesis when it is indeed true.
- For $\theta \in \Theta_1$, $1 - \pi(\theta)$ is the probability of making a type II error, i.e., failing to reject the null hypothesis when it is false.

The ideal power function has

$$\pi(\theta) \approx 0 \text{ for all } \theta \in \Theta_0 \text{ and } \pi(\theta) \approx 1 \text{ for all } \theta \in \Theta_1$$

With this property for the power function, we would rarely reject the null hypothesis when it is true and rarely fail to reject the null hypothesis when it is false.

In reality, incorrect decisions are made. Thus, for $\theta \in \Theta_0$,

$$\pi(\theta) \text{ is the probability of making a type I error,}$$

i.e., rejecting the null hypothesis when it is indeed true. For $\theta \in \Theta_1$,

$$1 - \pi(\theta) \text{ is the probability of making a type II error,}$$

i.e., failing to reject the null hypothesis when it is false.

The goal is to make the chance for error small. The traditional method is analogous to that employed in the Neyman-Pearson lemma. Fix a **(significance) level** α, now defined to be the largest value of $\pi(\theta)$ in the region Θ_0 defined by the null hypothesis. In other words, by focusing on the value of the parameter in Θ_0 that is most likely to result in an error, we insure that the probability of a type I error is no more that α irrespective of the value for $\theta \in \Theta_0$. Then, we look for a critical region that makes the power function as large as possible for values of the parameter $\theta \in \Theta_1$

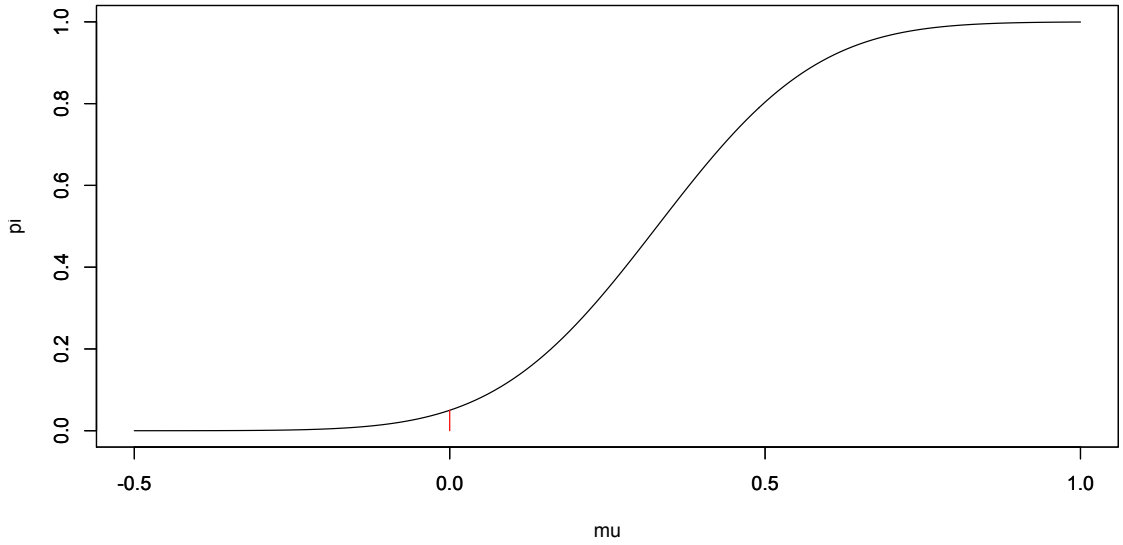


Figure 18.1: Power function for the one-sided test with alternative "greater". The size of the test α is given by the height of the red segment. Notice that $\pi(\mu) < \alpha$ for all $\mu < \mu_0$ and $\pi(\mu) > \alpha$ for all $\mu > \mu_0$

Example 18.1. *Let $X_1, X_2, \ldots, X_n$ be independent $N(\mu, \sigma_0)$ random variables with σ_0 known and μ unknown. For the composite hypothesis for the* **one-sided test**

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0,$$

we use the test statistic from the likelihood ratio test and reject H_0 if the statistic $\bar{x}$ is too large. Thus, the critical region

$$C = \{\mathbf{x}; \bar{x} \geq k(\mu_0)\}.$$

If μ is the **true mean***, then the power function*

$$\pi(\mu) = P_\mu\{X \in C\} = P_\mu\{\bar{X} \geq k(\mu_0)\}.$$

As we shall see soon, the value of $k(\mu_0)$ depends on the level of the test.

As the actual mean μ increases, then the probability that the sample mean $\bar{X}$ exceeds a particular value $k(\mu_0)$ also increases. In other words, π is an increasing function. Thus, the maximum value of π on the set $\Theta_0 = \{\mu; \mu \leq \mu_0\}$ takes place for the value μ_0. Consequently, to obtain level α for the hypothesis test, set

$$\alpha = \pi(\mu_0) = P_{\mu_0}\{\bar{X} \geq k(\mu_0)\}.$$

We now use this to find the value $k(\mu_0)$. When μ_0 is the value of the mean, we standardize to give a standard normal random variable

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}.$$

Choose z_α so that $P\{Z \geq z_\alpha\} = \alpha$. Thus

$$P_{\mu_0}\{Z \geq z_\alpha\} = P_{\mu_0}\{\bar{X} \geq \mu_0 + \frac{\sigma_0}{\sqrt{n}} z_\alpha\}$$

and $k(\mu_0) = \mu_0 + (\sigma_0/\sqrt{n})z_\alpha$.

If μ is the true state of nature, then

$$Z = \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}$$

is a standard normal random variable. We use this fact to determine the power function for this test.

$$\pi(\mu) = P_\mu\{\bar{X} \geq \frac{\sigma_0}{\sqrt{n}} z_\alpha + \mu_0\} = P_\mu\{\bar{X} - \mu \geq \frac{\sigma_0}{\sqrt{n}} z_\alpha - (\mu - \mu_0)\} \tag{18.1}$$

$$= P_\mu\left\{\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \geq z_\alpha - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right\} = 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right) \tag{18.2}$$

where Φ is the distribution function for a standard normal random variable.

We have seen the expression above in several contexts.

- *If we fix n, the number of observations and the alternative value $\mu = \mu_1 > \mu_0$ and determine the power $1 - \beta$ as a function of the significance level α, then we have the receiver operating characteristic as in Figure 17.2.*
- *If we fix μ_1 the alternative value and the significance level α, then we can determine the power as a function of the number of observations as in Figure 17.3.*
- *If we fix n and the significance level α, then we can determine the power function $\pi(\mu)$, the power as a function of the alternative value μ. An example of this function is shown in Figure 18.1.*

Exercise 18.2. *If the alternative is less than, show that*

$$\pi(\mu) = \Phi\left(-z_\alpha - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right).$$

Returning to the example with a model species and its mimic. For the plot of the power function for $\mu_0 = 10$, $\sigma_0 = 3$, and $n = 16$ observations,

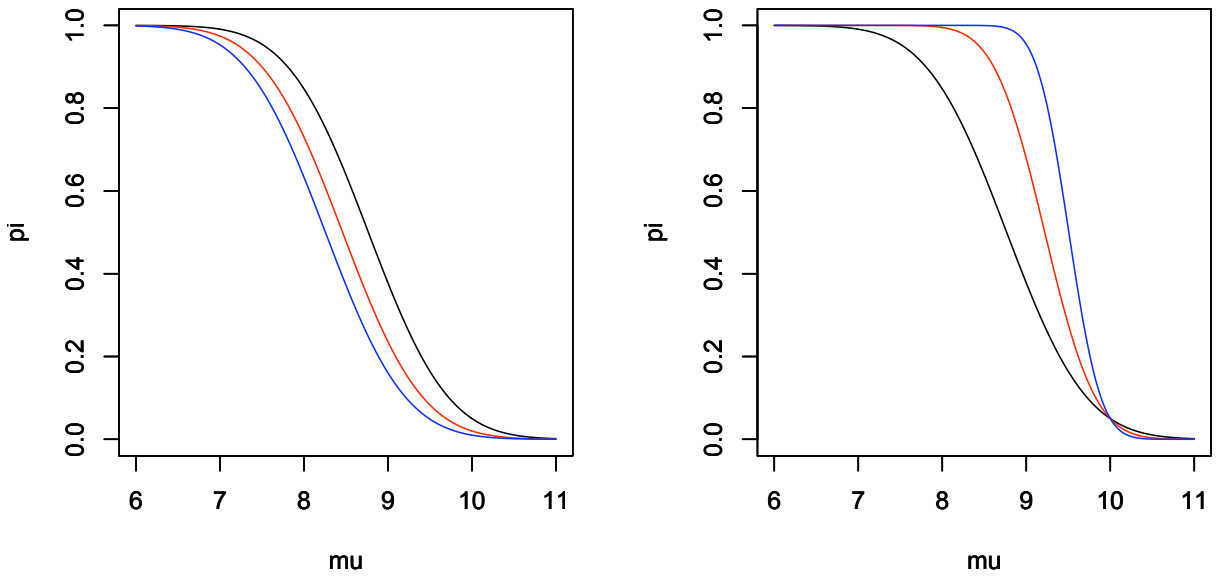


Figure 18.2: Power function for the one-sided test with alternative "less than". $\mu_0 = 10$, $\sigma_0 = 3$. Note, as argued in the text that π is a decreasing function. **(left)** $n = 16$, $\alpha = 0.05$ (black), 0.02 (red), and 0.01 (blue). Notice that lowering significance level α reduces power $\pi(\mu)$ for each value of μ. **(right)** $\alpha = 0.05$, $n = 15$ (black), 40 (red), and 100 (blue). Notice that increasing sample size n increases power $\pi(\mu)$ for each value of $\mu \leq \mu_0$ and decreases type I error probability for each value of $\mu > \mu_0$. For all 6 power curves, we have that $\pi(\mu_0) = \alpha$.

```
> zalpha<-qnorm(0.95)
> mu0<-10
> sigma0<-3
> mu<-(600:1100)/100
> n<-16
> z<--zalpha - (mu-mu0)/(sigma0/sqrt(n))
> pi<-pnorm(z)
> plot(mu,pi,type="l")
```

In Figure 18.2, we vary the values of the significance level α and the values of n, the number of observations in the graph of the power function π

Example 18.3 (mark and recapture)**.** *We may want to use mark and recapture as an experimental procedure to test whether or not a population has reached a dangerously low level. The variables in mark and recapture are*

- *t be the number captured and tagged,*
- *k be the number in the second capture,*
- *r the the number in the second capture that are tagged, and let*
- *N be the total population.*

If N_0 is the level that a wildlife biologist say is dangerously low, then the natural hypothesis is one-sided.

$$H_0 : N \geq N_0 \quad \text{versus} \quad H_1 : N < N_0.$$

The data are used to compute r, the number in the second capture that are tagged. The likelihood function for N is the hypergeometric distribution,

$$L(N|r) = \frac{\binom{t}{r}\binom{N-t}{k-r}}{\binom{N}{k}}.$$

The maximum likelihood estimate is $\hat{N} = [tk/r]$. Thus, higher values for r lead us to lower estimates for N. Let R be the (random) number in the second capture that are tagged, then, for an α level test, we look for the minimum value r_α so that

$$\pi(N) = P_N\{R \geq r_\alpha\} \leq \alpha \text{ for all } N \geq N_0. \tag{18.3}$$

As N increases, then recaptures become less likely and the probability in (18.3) decreases. Thus, we should set the value of r_α according to the parameter value N_0, the minimum value under the null hypothesis. Let's determine r_α for several values of α using the example from the topic, Maximum Likelihood Estimation, and consider the case in which the critical population is $N_0 = 2000$.

```
> N0<-2000; t<-200; k<-400
> alpha<-c(0.05,0.02,0.01)
> ralpha<-qhyper(1-alpha,t,N0-t,k)
> data.frame(alpha,ralpha)
  alpha ralpha
1  0.05     49
2  0.02     51
3  0.01     53
```

For example, we must capture al least 49 that were tagged in order to reject H_0 at the $\alpha = 0.05$ level. In this case the estimate for N is $\hat{N} = [kt/r_\alpha] = 1632$. As anticipated, r_α increases and the critical regions shrinks as the value of α decreases.

Using the level r_α determined using the value N_0 for N, we see that the power function

$$\pi(N) = P_N\{R \geq r_\alpha\}.$$

R is a hypergeometric random variable with mass function

$$f_R(r) = P_N\{R = r\} = \frac{\binom{t}{r}\binom{N-t}{k-r}}{\binom{N}{k}}.$$

The plot for the case $\alpha = 0.05$ is given using the R *commands*

```
> N<-c(1300:2100)
> pi<-1-phyper(49,t,N-t,k)
> plot(N,pi,type="l",ylim=c(0,1))
```

We can increase power by increasing the size of k, the number the value in the second capture. This increases the value of r_α. For $\alpha = 0.05$, we have the table.

```
> k<-c(400,600,800)
> N0<-2000
> ralpha<-qhyper(0.95,t,N0-t,k)
> data.frame(k,ralpha)
    k ralpha
1 400     49
2 600     70
3 800     91
```

We show the impact on power $\pi(N)$ of both significance level α and the number in the recapture k in Figure 18.3.

Exercise 18.4. *Determine the type II error rate for $N = 1600$ with*

- *$k = 400$ and $\alpha = 0.05, 0.02$, and 0.01, and*
- *$\alpha = 0.05$ and $k = 400, 600$, and 800.*

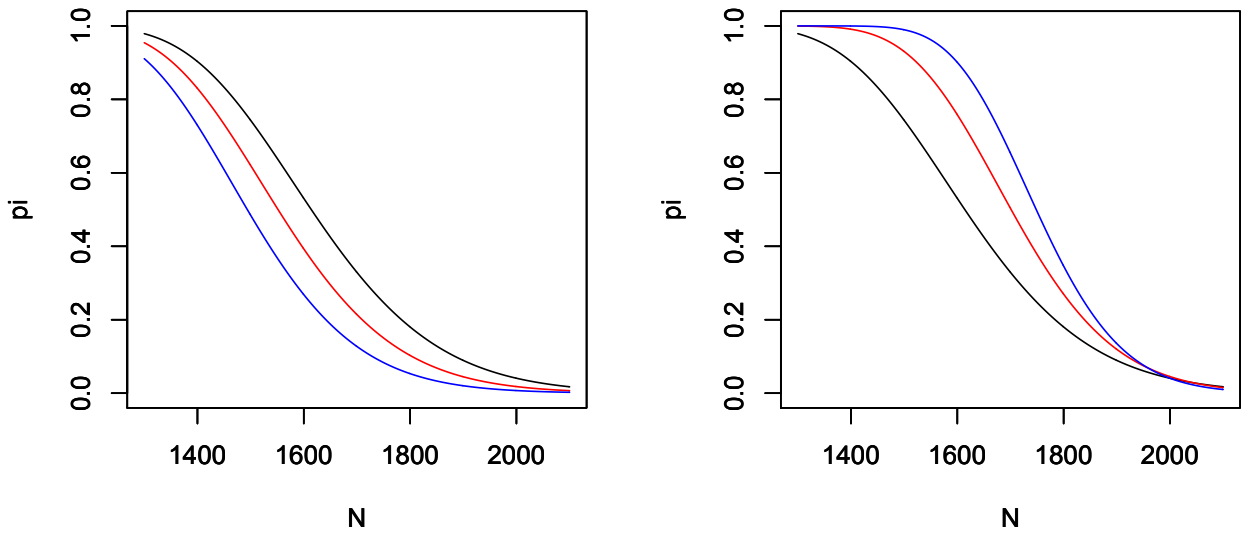


Figure 18.3: Power function for Lincoln-Peterson mark and recapture test for population $N_0 = 2000$ and $t = 200$ captured and tagged. **(left)** $k = 400$ recaptured $\alpha = 0.05$ (black), 0.02 (red), and 0.01 (blue). Notice that lower significance level α reduces power. **(right)** $\alpha = 0.05$, $k = 400$ (black), 600 (red), and 800 (blue). As expected, increased recapture size increases power.

Example 18.5. *For a* **two-sided test**

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

In this case, the parameter values for the null hypothesis Θ_0 consist of a single value, μ_0. We reject H_0 if $|\bar{X} - \mu_0|$ is too large. Under the null hypothesis,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

is a standard normal random variable. For a significance level α, choose $z_{\alpha/2}$ so that

$$P\{Z \geq z_{\alpha/2}\} = P\{Z \leq -z_{\alpha/2}\} = \frac{\alpha}{2}.$$

Thus, $P\{|Z| \geq z_{\alpha/2}\} = \alpha$. For data $\mathbf{x} = (x_1, \ldots, x_n)$, this leads to a critical region

$$C = \left\{ \mathbf{x}; \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \geq z_{\alpha/2} \right\}.$$

If μ is the actual mean, then

$$\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}$$

is a standard normal random variable. We use this fact to determine the power function for this test

$$\begin{aligned}
\pi(\mu) &= P_\mu\{X \in C\} = 1 - P_\mu\{X \notin C\} = 1 - P_\mu\left\{ \left| \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \right| < z_{\alpha/2} \right\} \\
&= 1 - P_\mu\left\{ -z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} < z_{\alpha/2} \right\} = 1 - P_\mu\left\{ -z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} < z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} \right\} \\
&= 1 - \Phi\left(z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} \right) + \Phi\left(-z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} \right)
\end{aligned}$$

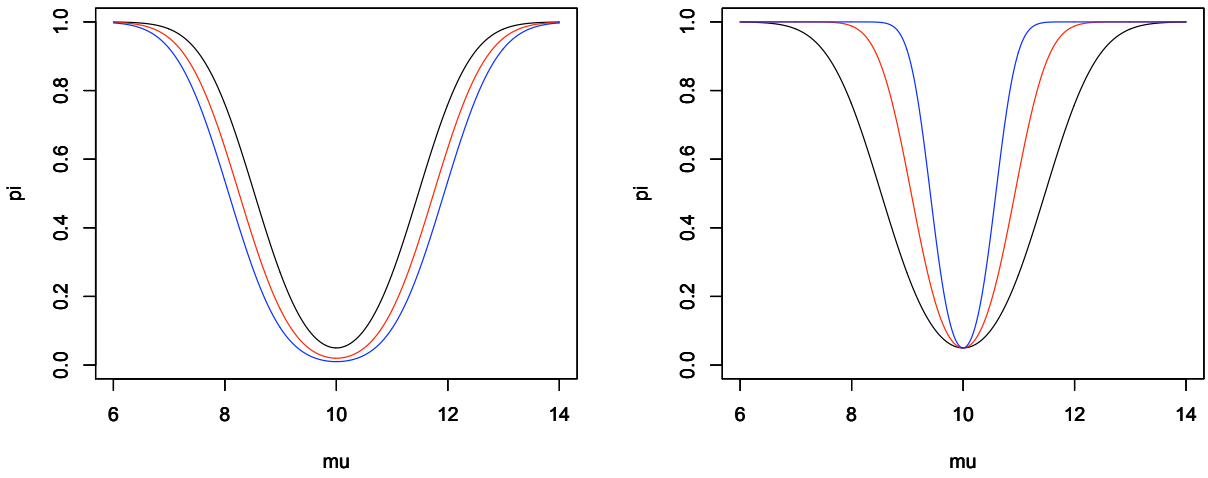


Figure 18.4: Power function for the two-sided test. $\mu_0 = 10$, $\sigma_0 = 3$. **(left)** $n = 16$, $\alpha = 0.05$ (black), 0.02 (red), and 0.01 (blue). Notice that lower significance level α reduces power. **(right)** $\alpha = 0.05$, $n = 15$ (black), 40 (red), and 100 (blue). As before, decreased significance level reduces power and increased sample size n increases power.

If we do not know if the mimic is larger or smaller that the model, then we use a two-sided test. Below is the R *commands for the power function with* $\alpha = 0.05$ *and* $n = 16$ *observations.*

```
> zalpha = qnorm(.975)
> mu0<-10
> sigma0<-3
> mu<-(600:1400)/100
> n<-16
> pi<-1-pnorm(zalpha-(mu-mu0)/(sigma0/sqrt(n)))
  +pnorm(-zalpha-(mu-mu0)/(sigma0/sqrt(n)))
> plot(mu,pi,type="l")
```

We shall see in the the next topic how these tests follow from extensions of the likelihood ratio test for simple hypotheses.

The next example is unlikely to occur in any genuine scientific situation. It is included because it allows us to compute the power function explicitly from the distribution of the test statistic. We begin with an exercise.

Exercise 18.6. *For* $X_1, X_2, \ldots, X_n$ *independent* $U(0, \theta)$ *random variables,* $\theta \in \Theta = (0, \infty)$. *The density*

$$f_X(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_{(n)}$ *denote the maximum of* $X_1, X_2, \ldots, X_n$, *then* $X_{(n)}$ *has distribution function*

$$F_{X_{(n)}}(x) = P_\theta\{X_{(n)} \le x\} = \left(\frac{x}{\theta}\right)^n.$$

Example 18.7. *For* $X_1, X_2, \ldots, X_n$ *independent* $U(0, \theta)$ *random variables, take the null hypothesis that* θ *lands in some normal range of values* $[\theta_L, \theta_R]$. *The alternative is that* θ *lies outside the normal range.*

$$H_0 : \theta_L \le \theta \le \theta_R \quad \text{versus} \quad H_1 : \theta < \theta_L \text{ or } \theta > \theta_R.$$

Because θ is the highest possible value for an observation, if any of our observations X_i are greater than θ_R, then we are certain $\theta > \theta_R$ and we should reject H_0. On the other hand, all of the observations could be below θ_L and the maximum possible value θ might still land in the normal range.

Consequently, we will try to base a test based on the statistic $X_{(n)} = \max_{1 \le i \le n} X_i$ and reject H_0 if $X_{(n)} > \theta_R$ and too much smaller than θ_L, say $\tilde{\theta}$. We shall soon see that the choice of $\tilde{\theta}$ will depend on n the number of observations and on α, the size of the test.

The power function

$$\pi(\theta) = P_\theta\{X_{(n)} \le \tilde{\theta}\} + P_\theta\{X_{(n)} \ge \theta_R\}$$

We compute the power function in three cases - low, middle and high values for the parameter θ. The second case has the values of θ under the null hypothesis. The first and the third cases have the values for θ under the alternative hypothesis. An example of the power function is shown in Figure 18.5.

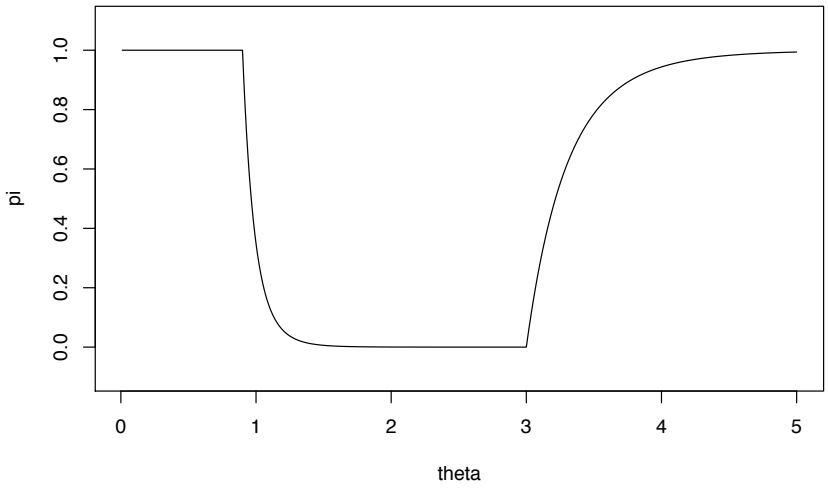


Figure 18.5: Power function for the test above with $\theta_L = 1, \theta_R = 3, \tilde{\theta} = 0.9$, and $n = 10$. The size of the test is $\pi(1) = 0.3487$.

Case 1. $\theta \le \tilde{\theta}$.

In this case all of the observations X_i must be less than θ which is in turn less than $\tilde{\theta}$. Thus, $X_{(n)}$ is certainly less than $\tilde{\theta}$ and

$$P_\theta\{X_{(n)} \le \tilde{\theta}\} = 1 \text{ and } P_\theta\{X_{(n)} \ge \theta_R\} = 0$$

and therefore $\pi(\theta) = 1$.

Case 2. $\tilde{\theta} < \theta \le \theta_R$.

Here $X_{(n)}$ can be less that $\tilde{\theta}$ but never greater than θ_R.

$$P_\theta\{X_{(n)} \le \tilde{\theta}\} = \left(\frac{\tilde{\theta}}{\theta}\right)^n \text{ and } P_\theta\{X_{(n)} \ge \theta_R\} = 0$$

and therefore $\pi(\theta) = (\tilde{\theta}/\theta)^n$.

Case 3. $\theta > \theta_R$.

Repeat the argument in Case 2 to conclude that

$$P_\theta\{X_{(n)} \le \tilde{\theta}\} = \left(\frac{\tilde{\theta}}{\theta}\right)^n$$

and that

$$P_\theta\{X_{(n)} \ge \theta_R\} = 1 - P_\theta\{X_{(n)} < \theta_R\} = 1 - \left(\frac{\theta_R}{\theta}\right)^n$$

and therefore $\pi(\theta) = (\tilde{\theta}/\theta)^n + 1 - (\theta_R/\theta)^n$.

The size of the test is the maximum value of the power function under the null hypothesis. This is case 2. Here, the power function

$$\pi(\theta) = \left(\frac{\tilde{\theta}}{\theta}\right)^n$$

decreases as a function of θ*. Thus, its maximum value takes place at* θ_L *and*

$$\alpha = \pi(\theta_L) = \left(\frac{\tilde{\theta}}{\theta_L}\right)^n$$

To achieve this level, we solve for $\tilde{\theta}$*, obtaining* $\tilde{\theta} = \theta_L \sqrt[n]{\alpha}$*. Note that* $\tilde{\theta}$ *increases with* α*. Consequently, we must expand the critical region in order to reduce the significance level. Also,* $\tilde{\theta}$ *increases with* n *and we can reduce the critical region while maintaining significance if we increase the sample size.*

The assessment of statistical power is an important aspect of experimental design. In practical terms, we can *increase* power by either *increasing effort* or *asking a less stringent question*. For example, we can increase effort

- **(mathematics)** by applying a more powerful test or a more rigorous design,
- **(engineering)** by designing a better measuring devise, reducing variance, or
- **(exersion)** by increasing sample size

We can ask a less stringent question

- by increasing the significance level and thus the ability to reject the null hypothesis or
- by increasing the difference between null value and the alternative value for detection of difference

These practical considerations will be useful in understanding the change in power resulting from a change in the experimental design and hypothesis testing.

18.3 The p-value

The report of *reject* the null hypothesis does not describe the strength of the evidence because it fails to give us the sense of whether or not a small change in the values in the data could have resulted in a different decision. Consequently, one common method is not to choose, in advance, a significance level α of the test and then report "reject" or "fail to reject", but rather to report the value of the test statistic and to give all the values for α that would lead to the rejection of H_0. The p-value is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. In this way, we provide an assessment of the strength of evidence against H_0. Consequently, a very low p-value indicates strong evidence against the null hypothesis.

Example 18.8. *For the one-sided hypothesis test to see if the mimic had invaded,*

$$H_0 : \mu \geq \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0.$$

with $\mu_0 = 10$ cm, $\sigma_0 = 3$ cm and $n = 16$ observations. The test statistics is the sample mean $\bar{x}$ and the critical region is $C = \{\mathbf{x}; \bar{x} \leq k\}$

Our data had sample mean $\bar{x} = 8.93125$ cm. The maximum value of the power function $\pi(\mu)$ for μ in the subset of the parameter space determined by the null hypothesis occurs for $\mu = \mu_0$. Consequently, the p-value is

$$P_{\mu_0}\{\bar{X} \leq 8.93125\}.$$

With the parameter value $\mu_0 = 10$ cm, $\bar{X}$ has mean 10 cm and standard deviation $3/\sqrt{16} = 3/4$. We can compute the p-value using R.

```
> pnorm(8.93125,10,3/4)
[1] 0.0770786
```

If the p-value is below a given significance level α, then we say that the result is **statistically significant** at the level α. For the previous example, we could not have rejected H_0 at the $\alpha = 0.05$ significance level. Indeed, we could not have rejected H_0 at any level below the p-value, 0.0770786. On the other hand, we would reject H_0 for any significance level above this value.

Many statistical software packages (including R, see the example below) do not need to have the significance level in order to perform a test procedure. This is especially important to note when setting up a hypothesis test for the purpose of deciding whether or not to reject H_0. In these circumstances, the significance level of a test is a value that should be decided *before* the data are viewed. After the test is performed, a report of the p-value adds information beyond simply saying that the results were or were not significant.

It is tempting to associate the p-value to a statement about the probability of the null or alternative hypothesis being true. Such a statement would have to be based on knowing which value of the parameter is the true state of nature. Assessing whether of not this parameter value is in Θ_0 is the reason for the testing procedure and the p-value was computed in knowledge of the data and our choice of Θ_0.

In the example above, the test is based on having a test statistic $S(\mathbf{x})$ (namely $-\bar{x}$) exceed a level k_α, i.e., we have decision

$$\text{reject } H_0 \text{ if and only if } S(\mathbf{x}) \geq k_\alpha.$$

This choice of k_α is based on the choice of significance level α and the choice of $\theta_0 \in \Theta_0$ so that $\pi(\theta_0) = P_{\theta_0}\{S(X) \geq k_\alpha\} = \alpha$, the lowest value for the power function under the null hypothesis. If the *observed* data $\mathbf{x}$ takes the value $S(\mathbf{x}) = k$, then the p-value equals

$$P_{\theta_0}\{S(X) \geq k\}. \tag{18.4}$$

This is the lowest value for the significance level that would result in rejection of the null hypothesis *if we had chosen it in advance of seeing the data*.

Example 18.9. *Returning to the example on the proportion of hives that survive the winter, the appropriate composite hypothesis test to see if more that the usual normal of hives survive is*

$$H_0 : p \leq 0.7 \quad \textit{versus} \quad H_1 : p > 0.7.$$

The R *output shows a p-value of 3%.*

```
>  prop.test(88,112,0.7,alternative="greater")

1-sample proportions test with continuity correction

data:  88 out of 112, null probability 0.7
X-squared = 3.5208, df = 1, p-value = 0.0303
```

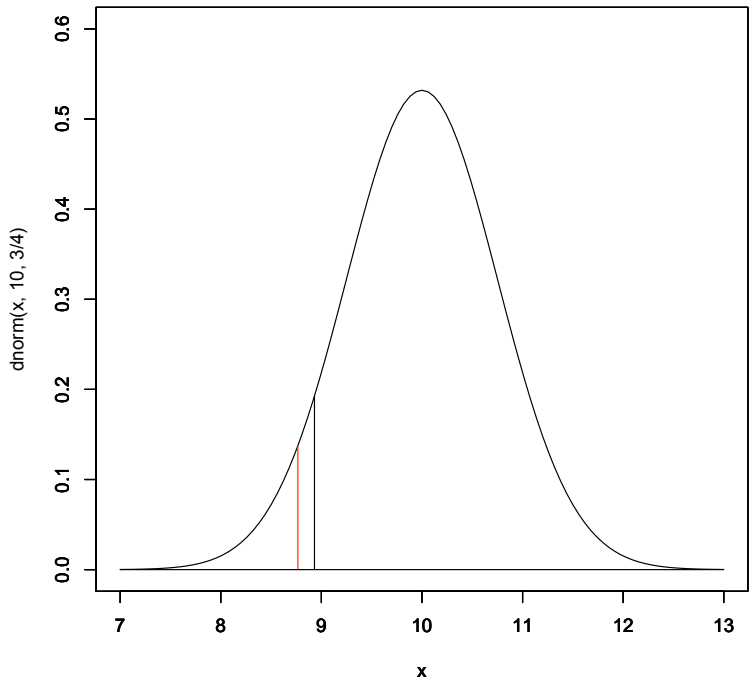


Figure 18.6: Under the null hypothesis, $\bar{X}$ has a normal distribution mean $\mu_0 = 10$ cm, standard deviation $3/\sqrt{16} = 3/4$ cm. The p-value, 0.077, is the area under the density curve to the left of the observed value of 8.931 for $\bar{x}$, The critical value, 8.767, for an $\alpha = 0.05$ level test is indicated by the red line. Because the p-vlaue is greater than the significance level, we cannot reject H_0.

```
alternative hypothesis: true p is greater than 0.7
95 percent confidence interval:
 0.7107807 1.0000000
sample estimates:
        p
0.7857143
```

Exercise 18.10. *Is the hypothesis test above significant at the 5% level? the 1% level?*

18.4 Multiple Hypothesis Testing

We now consider testing *multiple* hypotheses. This is common in the world of "big data" with thousands of hypothesis on many issues in subjects including genomics, internet searches, or financial transactions. For m hypotheses, let $p_1, \ldots, p_m$ be the p-values for m hypothesis tests.

18.4.1 Familywise Error Rate

The **familywise error rate** (FWER) is the probability of making even one type I error. If we set α_B for the significance level for a single test, then the simplest strategy is to employ the **Bonferroni correction**. This uses the Bonferroni inequality,

$$P(A_1 \cup \cdots \cup A_m) \leq P(A_1) + \cdots + P(A_m)$$

for events $A_1, \ldots, A_m$.

If A_i is the event of rejecting the null hypothesis when it is true, then $A_1 \cup \cdots \cup A_m$ is the event that at least one of the hypotheses is rejected when it is true. For each i, $P(A_i) = \alpha_B$ and so $\alpha = P(A_1 \cup \cdots \cup A_m) \leq m\alpha_B$. Thus,

the Bonferroni correction is to reject if

$$p_i \le \frac{\alpha}{m} \quad \text{for all } i.$$

Exercise 18.11. *For m independent, α_I level hypothesis tests, show that the familywise error $\alpha = 1 - (1 - \alpha_I)^m$. Thus, $(1 - \alpha)^{1/m} = 1 - \alpha_I$ and $\alpha_I = 1 - (1 - \alpha)^{1/m}$ is the level necessary to obtain an α familywise error rate.*

This gives a cautionary take, if we take $\alpha = 0.05$ and $m = 20$, then the probability of one or more false positive tests, $1 - (1 - 0.05)^{20} \approx 0.64$, is well above 1/2. The Bonferroni correction, $\alpha_B = 0.05/20 = 0.0025$ and the independence correction, $\alpha_I = 1 - (1 - 0.05)^{1/20} = 0.0256$ will guarantee a familywise error rate $\alpha = 0.05$

Note that the second method allows for slightly higher values of α than the Bonferroni correction. However, it is far less general. For independent test statistics, **Fisher's method** for testing multiple works directly with the p-values. We begin with the following exercise.

Exercise 18.12. *Let θ_0 be the true state of nature. Assume that the distribution function for the text statistic $S(X)$ is continuous and strictly increasing for all possible values. Show that the p-value is uniformly distributed on the interval $[0, 1]$.*

In this circumstance, if the null hypothesis is true for all m hypotheses, then

$$p_1, \dots, p_m \text{ are independent } U(0, 1) \text{ random variables.}$$

Recall from the use of the probability transform that

$$-2 \ln p_1, \dots, -2 \ln p_m \text{ are independent } Exp(1/2) \text{ random variables.}$$

So their sum

$$-2 \ln p_1 - \cdots - 2 \ln p_m \text{ is a } \Gamma(1/2, m) \text{ random variable.}$$

Thus this Γ random variable can serve as a test statistic for the multiple hypothesis that all the null hypotheses are true, rejecting if the sum above is sufficiency large. Traditionally, we use the fact that $\Gamma(1/2, m)$ is also a member of the chi-square family, namely, χ^2_{2m} and then use this as the distribution of $-2 \ln p_1 - \cdots - 2 \ln p_m$ under the multiple hypothesis that all m null hypotheses hold.

Example 18.13. *For 10 independent test consider the p-values*

```
> p
 [1] 0.0086 0.0164 0.6891 0.7671 0.2967 0.5465 0.0247 0.8235 0.9603 0.0041
```

The test statistic for Fisher's method

```
 > -2*sum(log(p))
[1] 41.5113
```

gives a p-value of 0.3% for the multiple test for all 10 hypotheses.

```
> 1-pchisq(-2*sum(log(p)),2*10)
[1] 0.003200711
```

18.4.2 False Discovery Rate

When the number of tests becomes very large, then having all hypotheses true is an extremely strict criterion. A more relaxed and often more valuable criterion is the **false discovery rate**.. Let's consider this in the case of identical simple hypotheses.

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

Fix a parameter value θ_1 under the alternative hypothesis. For each value of α, the power

$$1 - \beta(\alpha) = P_{\theta_1}\{S(X) \geq k_\alpha\}.$$

In other words, $1 - \beta(\alpha)$ is the probability, under θ_1 that the p-value is less than α. Making this consideration for each value of α leads to the following exercise.

Exercise 18.14. *Show that the receiver operating characteristic gives the distribution function for the p-values for a fixed alternative $\theta_1 \in \Theta_1$.*

Thus, we can model question *Is the null hypothesis hypothesis true?* as a sequence of Bernoulli trials. Call the success parameter π. Thus, with probability π, the null hypothesis is true and the p-values follow F_U, the uniform distribution on the interval $[0, 1]$. With probability $1 - \pi$, the null hypothesis is false and the p-values follow F_R, the distribution of the receiver operating characteristic. Taken together, we say that the p-values are distributed according to the mixture

$$F(x) = \pi F_U(x) + (1 - \pi)F_R(x) = \pi x + (1 - \pi)F_R(x). \tag{18.5}$$

Thus, if use reject if the p-value is below a chosen value α, then the type I error probability is α. From this we determine the false discovery rate, here defined as

$$q = P\{H_0 \text{ is true}|\text{reject } H_0\}.$$

Using Bayes formula

$$q = \frac{P\{\text{reject } H_0|H_0 \text{ is true}\}P\{H_0 \text{ is true}\}}{P\{\text{reject } H_0\}} = \frac{\alpha\pi_0}{F(\alpha)}.$$

An estimate of the false discovery rate can be determined from an estimate of π. This is determined by looking at the p-values and estimating the mixture in (18.5).

Example 18.15. *Consider a simple hypothesis*

$$H_0 : \mu = 0 \quad \textit{versus} \quad H_1 : \mu = 1.$$

for the mean μ based on 16 observations of normal random variable, variance 1. Thus, either the effect is not present ($\mu = 0$), or it is ($\mu = 1$). If we take the significance level $\alpha = 0.01$, then based on $n = 16$ observations, the test statistic $\bar{X}$ has standard deviation $1/\sqrt{16} = 1/4$,

```
> alpha<-0.01
> (kalpha<-qnorm(1-alpha,0,1/4))
[1] 0.581587
```

and, thus, we reject H_0 if the sample mean $\bar{x} > k_\alpha = 0.581587$. The power, i.e.., the probability that we reject H_0 when H_1 is true,

```
> (p_1<-1-pnorm(xbar,1,1/4))
[1] 0.9529005
```

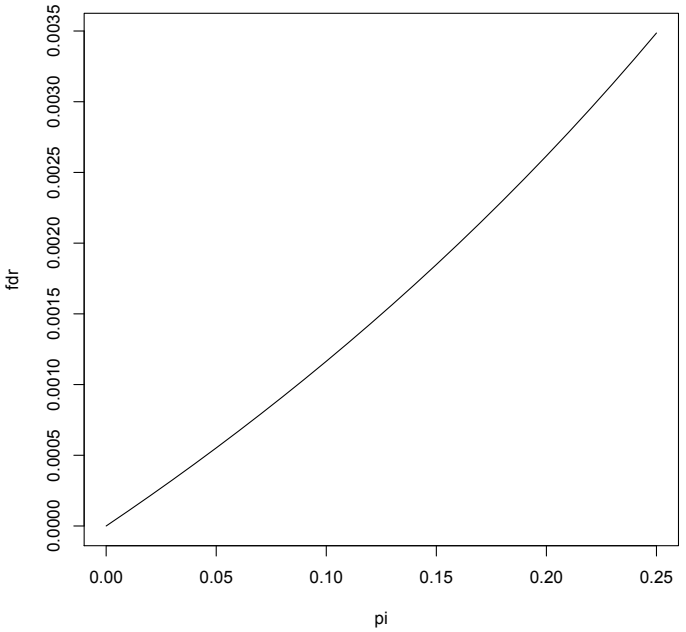


Figure 18.7: False discovery rate versus π. Here the significance level $\alpha = 0.01$, the power, $\beta = 0.953$.

If we plot the false discovery rate versus π, the probability H_0 is true, then

```
> pi<-seq(0,0.25,0.01)
> fdr<-alpha*pi/(alpha*pi+p_1*(1-pi))
> plot(pi,fdr,type="l")
```

In this case, for $\pi = 0.10$, we have a false discovery rate $q = 0.00116$, For 10,000 hypothesis, we have a mean of 11.6 false discoveries.

18.5 Answers to Selected Exercises

18.2. In this case the critical regions is $C = \{\mathbf{x}; \bar{x} \leq k(\mu_0)\}$ for some value $k(\mu_0)$. To find this value, note that

$$P_{\mu_0}\{Z \leq -z_\alpha\} = P_{\mu_0}\{\bar{X} \leq -\frac{\sigma_0}{\sqrt{n}} z_\alpha + \mu_0\}$$

and $k(\mu_0) = -(\sigma_0/\sqrt{n})z_\alpha + \mu_0$. The power function

$$\begin{aligned}\pi(\mu) &= P_\mu\{\bar{X} \leq -\frac{\sigma_0}{\sqrt{n}} z_\alpha + \mu_0\} = P_\mu\{\bar{X} - \mu \leq -\frac{\sigma_0}{\sqrt{n}} z_\alpha - (\mu - \mu_0)\} \\ &= P_\mu\left\{\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \leq -z_\alpha - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right\} = \Phi\left(-z_\alpha - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right).\end{aligned}$$

18.4. The type II error rate β is $1 - \pi(1600) = P_{1600}\{R < r_\alpha\}$. This is the distribution function of a hypergeometric random variable and thus these probabilities can be computed using the `phyper` command

- For varying significance, we have the R commands:

```
> t<-200;N<-1600
> k<-400
> alpha<-c(0.05,0.02,0.01)
> ralpha<-c(49,51,53)
> beta<-1-phyper(ralpha-1,t,N-t,k)
> data.frame(alpha,beta)
  alpha      beta
1  0.05 0.5993010
2  0.02 0.4609237
3  0.01 0.3281095
```

  Notice that the type II error probability is high for $\alpha = 0.05$ and increases as α decreases.

- For varying recapture size, we continue with the R commands:

```
> k<-c(400,600,800)
> ralpha<-c(49,70,91)
> beta<-1-phyper(ralpha-1,t,N-t,k)
> data.frame(k,beta)
    k      beta
1 400 0.5993010
2 600 0.8043988
3 800 0.9246057
```

Notice that increasing recapture size has a significant impact on type II error probabilities.

18.6. The i-th observation satisfies

$$P\{X_i \leq x\} = \int_0^x \frac{1}{\theta}\, d\tilde{x} = \frac{x}{\theta}$$

Now, $X_{(n)} \leq x$ occurs precisely when all of the n-independent observations X_i satisfy $X_i \leq x$. Because these random variables are independent,

$$\begin{aligned}F_{X_{(n)}}(x) &= P_\theta\{X_{(n)} \leq x\} = P_\theta\{X_1 \leq x, X_1 \leq x, \ldots, X_n \leq x\} \\ &= P_\theta\{X_1 \leq x\}P\{X_1 \leq x\}, \cdots P\{X_n \leq x\} = \left(\frac{x}{\theta}\right)\left(\frac{x}{\theta}\right)\cdots\left(\frac{x}{\theta}\right) = \left(\frac{x}{\theta}\right)^n\end{aligned}$$

18.10. Yes, the p-value is below 0.05. No, the p-value is above 0.01.

18.11. By the complement rule, de Morgan's law, and independence of the A_i, we have

$$\begin{aligned}\alpha &= P(A_1 \cup \cdots \cup A_m) = 1 - P((A_1 \cup \cdots \cup A_m)^c) = 1 - P(A_1^c \cap \cdots \cap A_m^c) \\ &= 1 - P(A_1^c) \cdots P(A_m^c) = 1 - (1 - P(A_1)) \cdots (1 - P(A_m)) = 1 - (1 - \alpha_I)^m.\end{aligned}$$

18.12. Let $F_{S(x)}(s|\theta_0)$ be the distribution function for $S(X)$ under θ_0 and note that the conditions on function $F_{S(x)}(s|\theta_0)$ insure that it is one to one and thus has an inverse. By 18.4, the p-value is $1 - F_{S(x)}(S(X)|\theta_0)$. Choose u in the interval $[0, 1]$. Then,

$$\begin{aligned}P_{\theta_0}\{1 - F_{S(x)}(S(X)|\theta_0) \leq u\} &= P_{\theta_0}\{F_{S(x)}(S(X)|\theta_0) \geq 1 - u\} = P_{\theta_0}\{S(X) \geq F^{-1}_{S(x)}(1 - u|\theta_0)\} \\ &= 1 - F_{S(x)}(F^{-1}_{S(x)}(1 - u|\theta_0)|\theta_0) = 1 - (1 - u) = u,\end{aligned}$$

showing that $1 - F_{S(x)}(S(X)|\theta_0)$ is uniformly distributed on the interval $[0, 1]$.

18.14. By the definition of k_α, $F_{S(x)}(k_\alpha|\theta_0)|\theta_0) = 1 - \alpha$. The distribution of p-values under θ_1,

$$\begin{aligned}P_{\theta_1}\{1 - F_{S(x)}(S(X)|\theta_0) \leq \alpha\} &= P_{\theta_1}\{F_{S(x)}(S(X)|\theta_0) \geq 1 - \alpha\} = P_{\theta_1}\{F_{S(x)}(S(X)|\theta_0) \geq F_{S(x)}(k_\alpha|\theta_0)\} \\ &= P_{\theta_1}\{S(X) \geq k_\alpha\} = 1 - \beta(\alpha),\end{aligned}$$

the power as a function of the significance level α. This is the receiver operating characteristic.

Topic 19

Extensions on the Likelihood Ratio

We begin with a composite hypothesis test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

with Θ_0 and Θ_1 a partition of the parameter space Θ ($\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$). Let C be the critical region for an α level test, i .e, we reject the null hypothesis whenever the data $\mathbf{x}$ fall in the critical region. Thus, the power function

$$\pi(\theta) = P_\theta\{X \in C\}$$

has the property that

$$\pi(\theta) \leq \alpha \quad \text{for all } \theta \in \Theta_0$$

and that α is the maximum value of the power function on Θ_0, the parameter values associated to the null hypothesis.

We have seen several critical regions that were defined by taking a statistic $T(\mathbf{x})$ and defining the critical region by have this statistic either be more or less that a critical value. For a one-sided test, we have seen critical regions

$$\{T(\mathbf{x}) \geq \tilde{k}_\alpha\} \quad \text{or} \quad \{T(\mathbf{x}) \leq \tilde{k}_\alpha\}.$$

For a two-sided test, we saw

$$\{|T(\mathbf{x})| \geq \tilde{k}_\alpha\}.$$

where $\tilde{k}_\alpha$ is determined by the level α. We thus use the commands `qnorm`, `qbinom`, or `qhyper` when the test statistic has, respectively, a normal, binomial, or hypergeometric distribution under a appropriated choice of $\theta \in \Theta_0$. Here we will examine extensions of the likelihood ratio test for simple hypotheses that have desirable properties for a critical region.

19.1 One-Sided Tests

Let's collect a simple random sample of independent normal observations with unknown mean and known variance σ_0^2. We noticed, in the case of a simple hypothesis test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu = \mu_1$$

that the critical region as determined by the Neyman-Pearson lemma depended only on whether or not μ_1 was greater than μ_0. For example, if $\mu_1 > \mu_0$, then the critical region $C = \{\mathbf{x}; \bar{x} \geq \tilde{k}_\alpha\}$ shows that we reject H_0 whenever the sample mean is higher than some threshold value $\tilde{k}_\alpha$ *irrespective of the difference between* μ_0 *and* μ_1. An analogous situation occurs in the case that $\mu_1 < \mu_0$

We will examine the idea that if a test is most powerful against each possible alternative in a simple hypothesis test, when we can say that this test is in some sense best overall for a composite hypothesis. Stated in terms of the

power function, we are asking if a test has the property that its power function π is greater for every value of $\theta \in \Theta_1$ than the power function of *any* other test. Such a test is called **uniformly most powerful**. In general, a hypothesis will not have a uniformly most powerful test. However, we can hope for such a test for procedures involving simple hypotheses in which the test statistic that emerged from the likelihood test did not depend on the specific value of the alternative. This was seen in the example above using independent normal data. In this case, the power function $\pi(\mu) = P_\mu\{\bar{X} \geq k_\alpha\}$ increases as μ increases and so the test has the intuitive property of becoming more powerful with increasing μ.

In general, we look for a test statistic $T(\mathbf{x})$ (like $\bar{x}$ in the example above). Next, we check that the likelihood ratio,

$$\frac{L(\theta_2|\mathbf{x})}{L(\theta_1|\mathbf{x})}, \quad \theta_1 < \theta_2. \tag{19.1}$$

depends on the data $\mathbf{x}$ only through the value of statistic $T(\mathbf{x})$ and, in addition, this ratio is a monotone increasing function of $T(\mathbf{x})$. The Karlin-Rubin theorem states:

If these conditions hold, then for an appropriate value of $\tilde{k}_\alpha$, $C = \{\mathbf{x}; T(\mathbf{x}) \geq \tilde{k}_\alpha\}$ is the critical region for a uniformly most powerful α level test for the one-sided alternative hypothesis

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

A corresponding criterion holds for the one sided test with the inequalities reversed:

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0.$$

Exercise 19.1 (mark and recapture)**.** *Use the hypothesis*

$$H_0 : N \geq N_0 \quad \text{versus} \quad H_1 : N < N_0$$

on the population size N and let the data be r, the number in the second capture that are tagged. Give the corresponding criterion to (19.1) and verify that it holds for the likelihood function $L(N|r)$ and the test statistic $T(r) = r$.

These conditions are satisfied for the case above as well as the tests for p, the probability of success in Bernoulli trials.

Exercise 19.2 (One sample one proportion z-test)**.** *For $X = (X_1, \ldots, X_n)$ is a sequence of Bernoulli trials with unknown success probability p, we can have and the* **one-sided tests** *with the alternative is greater*

$$H_0 : p \leq p_0 \quad \textit{versus} \quad H_1 : p > p_0$$

or less

$$H_0 : p \geq p_0 \quad \textit{versus} \quad H_1 : p < p_0.$$

Show that (19.1) holds with $T(x) = \bar{x}$ when the alternative is greater than.

Example 19.3. *We return to the example of the survivability of bee hives over a given winter. The probability of survival is $p_0 = 0.7$. The one-sided alternative for a mild winter is that this survival probability has increased. This leads us to consider the hypotheses*

$$H_0 : p \leq 0.7 \quad \textit{versus} \quad H_1 : p > 0.7.$$

for a test of the probability that a feral bee hive survives a winter. If the expected number of successes, np, and failures, $n(1-p)$, are both above 10, we can employ a test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

derived from the central limit theorem. For an α level test, the critical value is z_α where α is the probability that a standard normal is at least z_α

For this study, 112 colonies have been selected with 88 surviving. Thus $\hat{p} = 0.7875$ and $z = 1.979$. If the significance level is $\alpha = 0.05$, then we will reject H_0 because $z = 1.979 > 1.645 = z_\alpha$. Previously, we performed this test in R *and found a p-value of 0.0303.*

A direct appeal to the central limit theorem gives a slightly different p-value, namely

```
> 1-pnorm(1.979)
[1] 0.02390800
```

This is expressed in the R *output as a* **continuity correction**. *In the topic Central Limit Theorem, we learned that we are approximating probability for the outcome $P\{X \geq x\}$ for X the number of successes in Bernoulli trials by $P\{Y \geq x+1/2\}$ where Y is a normal random variable. This correction can be seen by looking at the area associated to a histogram for the mass function for X and the density function for Y. (See Figure 11.5.)*

Because $P\{X \geq x\} = P\{X > x-1\} = 1 - P\{X \leq x-1\}$, the R *command for $P\{X \geq x\}$ is* `1-pbinom(x-1,n,p)`. *The table below compares computing the P-value using the binomial directly* `binompvalue`, *the normal approximation* `normpvalue`, *and the normal approximation with the continuity correction* `normpvaluecc`. *The number for the test above are shown on line 9.*

```
> n<-112
> p<-0.7
> x<-c(80:92)
> binompvalue<-1-pbinom(x-1,n,p)
> normpvalue<-1-pnorm(x,n*p,sqrt(n*p*(1-p)))
> normpvaluecc<-1-pnorm(x+0.5,n*p,sqrt(n*p*(1-p)))
> data.frame(x,binompvalue,normpvalue,normpvaluecc)
    x  binompvalue   normpvalue normpvaluecc
1  80 0.4154655712 0.3707323063 0.4102833420
2  81 0.3367097745 0.2959409315 0.3325027711
3  82 0.2641118385 0.2289505277 0.2613434153
4  83 0.2000721958 0.1714364652 0.1989421672
5  84 0.1460628585 0.1241065395 0.1464910316
6  85 0.1025553369 0.0867731109 0.1042322648
7  86 0.0691142614 0.0585464089 0.0715976128
8  87 0.0446167293 0.0380903233 0.0474407060
9  88 0.0275341591 0.0238806213 0.0303009849
10 89 0.0162104857 0.0144193968 0.0186445384
11 90 0.0090854776 0.0083811681 0.0110460276
12 91 0.0048368618 0.0046873842 0.0062982479
13 92 0.0024402066 0.0025215113 0.0034547541
```

Exercise 19.4. *Use the central limit theorem to show that the power function, π, for a one-sided level α test with a greater than alternative. Using the critical region,*

$$C = \left\{ \mathbf{x}; \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z_\alpha \right\},$$

show that the power is

$$\pi(p) = 1 - \Phi\left(z_\alpha \sqrt{\frac{p_0(1-p_0)}{p(1-p)}} + \frac{p_0 - p}{\sqrt{p(1-p)/n}} \right) \tag{19.2}$$

where Φ is the distribution function for a standard normal and z_α is the critical value for an upper critical probability α for the standard normal.

To explore the properties of the power function in the case of overwintering of bee hives, we first keep the number of hives at $n = 112$ and considering increasing values $p = 0.75, 0.80, 0.85, 0.90$ for the alternative to see the power increase from about 30% to nearly 100%.

```
> n<-112
> p0<-0.7
> zalpha<-qnorm(0.95)
> p<-c(0.75,0.80,0.85,0.90)
> power<-1-pnorm(zalpha*sqrt(p0*(1-p0)/(p*(1-p))) + (p0-p)/sqrt(p*(1-p)/n))
> data.frame(p,power)
     p     power
1 0.75 0.3019748
2 0.80 0.7767714
3 0.85 0.9902226
4 0.90 0.9999972
```

Power increases with increasing sample size. Here we fix the alternative at $p = 0.8$ and choose n from 40 to 240. The power for these values increases from 38% to more than 97%.

```
> n<-c(1:6)*40
> p0<-0.7
> zalpha<-qnorm(0.95)
> p<-0.8
> power<-1-pnorm(zalpha*sqrt(p0*(1-p0)/(p*(1-p))) + (p0-p)/sqrt(p*(1-p)/n))
> data.frame(n,power)
    n     power
1  40 0.3808391
2  80 0.6374501
3 120 0.8035019
4 160 0.8993508
5 200 0.9506427
6 240 0.9766255
```

Example 19.5. *For a test of hive survivability over a harsh winter, we have*

$$H_0 : p \geq 0.7 \quad \textit{versus} \quad H_1 : p < 0.7.$$

If we have 26 observations, then we are reluctant to use the central limit theorem and appeal directly to the binomial distribution. If 16 hives survive, then we use the binomial test as follows.

```
> binom.test(16,26,0.7,alternative=c("less"))

Exact binomial test

data:  16 and 26
number of successes = 16, number of trials = 26, p-value = 0.2295
alternative hypothesis: true probability of success is less than 0.7
95 percent confidence interval:
 0.0000000 0.7743001
sample estimates:
probability of success
             0.6153846
```

and we do not reject for any significance level α below 0.2295.

Exercise 19.6. *Use the* R *command* `pbinom` *to compute the p-value in the binomial test above.*

Exercise 19.7. *Use the* R *command* `qbinom` *to compute the critical value for a* $\alpha = 0.05$ *test.*

19.2 Likelihood Ratio Tests

The likelihood ratio test is a popular choice for composite hypothesis when Θ_0 is a subspace of the whole parameter space. The rationale for this approach is that the null hypothesis is unlikely to be true if the maximum likelihood on Θ_0 is sufficiently smaller that the likelihood maximized over Θ, the entire parameter space. In symbols, let $\hat{\theta}_0$ be the parameter value that maximizes the likelihood for $\theta \in \Theta_0$ and $\hat{\theta}$ be the parameter value that maximizes the likelihood for $\theta \in \Theta$. Then the **likelihood ratio**

$$\Lambda(\mathbf{x}) = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}. \tag{19.3}$$

Note that this ratio is the reciprocal from the version given by the Neyman-Pearson lemma. Thus, the critical region consists of those values that are below a critical value.

The critical region for an α-level **likelihood ratio test** is

$$\{\Lambda(\mathbf{x}) \le \lambda_\alpha\} \tag{19.4}$$

As with any α level test, λ_α is chosen so that

$$P_\theta\{\Lambda(X) \le \lambda_\alpha\} \le \alpha \text{ for all } \theta \in \Theta_0.$$

This in the end may result in a procedure that many take several steps to develop. First, we must determine the likelihood $L(\theta|\mathbf{x})$ for the parameter θ based on the data $\mathbf{x}$. We have two optimization problems - maximize $L(\theta|\mathbf{x})$ on the parameter space Θ and on the null hypothesis space Θ_0. We evaluate the likelihood at these values and form the ratio. This generally give us a complex test statistic which we then simplify. We show this in some detail for a two-sided test for the mean based on normal data.

Example 19.8. *Let* $\Theta = \mathbb{R}$ *and consider the* **two-sided hypothesis**

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

Here the data are n *independent* $N(\mu, \sigma_0)$ *random variables* $X_1 \dots, X_n$ *with known variance* σ_0^2*. The parameter space* Θ *is one dimensional giving the value* μ *for the mean. As we have seen before* $\hat{\mu} = \bar{x}$*.* Θ_0 *is the single point* $\{\mu_0\}$ *and so* $\hat{\mu}_0 = \mu_0$*. Using the information, we find that*

$$L(\hat{\mu}_0|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp -\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i - \mu_0)^2, \quad L(\hat{\mu}|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp -\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i - \bar{x})^2$$

and

$$\Lambda(\mathbf{x}) = \exp -\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n ((x_i - \mu_0)^2 - (x_i - \bar{x})^2)\right) = \exp -\frac{n}{2\sigma_0^2}(\bar{x} - \mu_0)^2.$$

Now notice that

$$-2\ln\Lambda(\mathbf{x}) = \frac{n}{\sigma_0^2}(\bar{x} - \mu_0)^2 = \left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right)^2.$$

Then, critical region (19.4),

$$\{\Lambda(\mathbf{x}) \le \lambda_\alpha\} = \left\{\left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right)^2 \ge -2\ln\lambda_\alpha\right\}$$

Because $(\bar{X}-\mu_0)/(\sigma_0/\sqrt{n})$ is a standard normal random variable, $-2\ln\Lambda(X)$ is the square of a single standard normal. This is the defining property of a χ-square random variable with 1 degree of freedom.

Naturally we can use both

$$\left(\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}\right)^2 \quad \textit{and} \quad \left|\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}\right|.$$

as a test statistic. For the first, the critical value is just the square of the critical value for the second choice. We have seen the second choice in the section on Composite Hypotheses using the example of a possible invasion of a model butterfly by a mimic.

Exercise 19.9 (Bernoulli trials)**.** *Consider the* **two-sided hypothesis**

$$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0.$$

Use the linear approximation of the logarithm to show that

$$-\ln\Lambda(\mathbf{x}) \approx \frac{(\hat{p}-p_0)^2}{p_0(1-p_0)/n}.$$

Again, this is approximately the square of a standard normal. Thus, we can either use

$$\frac{(\hat{p}-p_0)^2}{p_0(1-p_0)/n} \quad \text{or} \quad \left|\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}\right|$$

for the test statistic. Under the null hypothesis, this statistic has approximately, respectively, the square and the absolute value of a standard normal distribution.

Returning to the example on the proportion of hives that survive the winter, for a two-sided test and a 98% confidence interval, notice that the output give the χ^2 statistic.

```
> prop.test(88,112,0.7,alternative=c("two.sided"),conf.level = 0.98)

1-sample proportions test with continuity correction

data:  88 out of 112, null probability 0.7
X-squared = 3.5208, df = 1, p-value = 0.0606
alternative hypothesis: true p is not equal to 0.7
98 percent confidence interval:
 0.6785906 0.8652397
sample estimates:
        p
0.7857143
```

to obtain the interval $(0.676, 0.8652)$.

Exercise 19.10. *Why is 0.0606 the p-value for the two-sided test equal to twice the value of the p-value for the corresponding one-sided test? Is the test significant at the 10& level? 5% level? Explain.*

Exercise 19.11. *For the two-sided two-sample α-level proportion test*

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 \neq p_2,$$

based on n_1 Bernoulli trials, $x_{1,1}, x_{i,2}, \ldots, x_{1,n_1}$ from the first population and, independently, n_2 Bernoulli trials, $x_{2,1}, x_{i,2}, \ldots, x_{2,n_2}$ from the second, the likelihood ratio test is equivalent to the critical region

$$|z| \geq z_{\alpha/2}$$

where

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{19.5}$$

with $\hat{p}_i$, the sample proportion of successes from the observations from population i and $\hat{p}_0$, the **pooled proportion**

$$\hat{p}_0 = \frac{1}{n_1+n_2}\left((x_{1,1}+\cdots+x_{1,n_1})+(x_{2,1}+\cdots+x_{2,n_2})\right) = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}. \tag{19.6}$$

By a variant of the central limit theorem, z has, under the null hypothesis, approximately a standard normal random variable.

A one-sided two-sample proportion test

$$H_0 : p_1 \le p_2 \quad \text{versus} \quad H_1 : p_1 \ge p_2, \tag{19.7}$$

also uses the z-statistic (19.5) provided that the central limit theorem is applicable. One standard rule of thumb is that the both the expected number of success and the expected number of failures in both sets of trails each exceeds 10 for both sets of observations.

Exercise 19.12. *The next winter was consider harsher than the one with 88 out of 112 hive surviving. In this more severe winter, we find that only 64 out of 99 random selecting hives surviving. State an appropriate hypothesis test and carry out the test, stating the evidence against the null hypothesis. Note that the rule of thumb on sample sizes necessary for the z test is satisfied.*

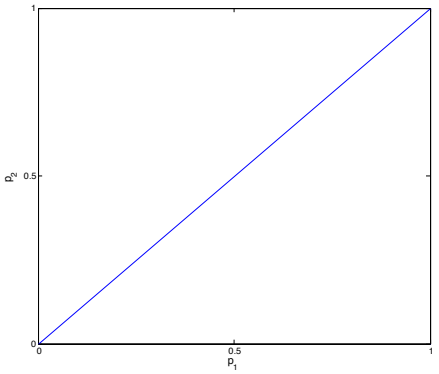


Figure 19.1: For the two-sided two-sample α-level likelihood ratio test for proportions p_1 and p_2, we maximize the likelihood over $\Theta_0 = \{(p_1, p_2); p_1 = p_2\}$ (shown as the blue line) and over $\Theta = [0,1]\times[0,1]$, the entire parameter space, shown as the square, and then take the ratio (19.3).

R handles this easily. Because R employs a continuity correction, the p-value is slightly different from the one in the exercise.

```
> prop.test(c(88,64),c(112,99),alternative="greater")

2-sample test for equality of proportions with continuity correction

data:  c(88, 64) out of c(112, 99)
X-squared = 4.3909, df = 1, p-value = 0.01807
alternative hypothesis: greater
95 percent confidence interval:
 0.02818133 1.00000000
sample estimates:
   prop 1    prop 2
0.7857143 0.6464646
```

Power analyses for two sample tests for proportions can be executed in R using the `power.prop.test` command. For example, if we want to be able to detect a difference between two proportions $p_1 = 0.7$ and $p_2 = 0.6$ in a one-sided test with a significance level of $\alpha = 0.05$ and power $1-\beta = 0.8$, then we will need a sample of $n = 281$ from each group.

```
> power.prop.test(p1=0.70,p2=0.6,sig.level=0.05,power=0.8,
  alternative = c("one.sided"))

     Two-sample comparison of proportions power calculation

              n = 280.2581
```

```
            p1 = 0.7
            p2 = 0.6
     sig.level = 0.05
         power = 0.8
   alternative = one.sided

NOTE: n is number in *each* group
```

Exercise 19.13. *What is the power for 100 observations in a test with significance level $\alpha = 0.10$.*

Exercise 19.14. *For the Salk vaccine trial, in the treatment group 56 out of 20000 contracted polio. From the control group, 142 out of 20000 contracted polio. Give an appropriate hypothesis test, find a p-value for the test, and assess the evidence against your null hypothesis.*

19.3 Chi-square Tests

This exact computation for normal data for a two-sided test of the mean shows that the test statistic has a χ^2 distribution with 1 degree of freedom. For the two-sided sample proportion test, we used the central limit theorem to assert that our test statistic is approximately the square of a standard normal random variable, and hence is a χ^2 random variable with one degree of freedom.

These ideas can be extended to the case in which Θ is a d-dimensional parameter space and k of these parameters are, under the null hypothesis, assume to have fixed values. Thus, Θ_0 is $d-k$-dimensional.

Theorem 19.15. *Whenever the maximum likelihood estimator has an asymptotically normal distribution, let $\Lambda(\mathbf{x})$ be the likelihood ratio (19.3) for an d-dimensional parameter space:*

$$H_0 : \theta_i = c_i \text{ for all } i = 1, \ldots, k \quad \text{versus} \quad H_1 : \theta_1 \neq c_1 \text{ for some } i = 1, \ldots, k$$

Then under H_0, the distributions of

$$-2 \ln \Lambda_n(X)$$

converge to a χ^2_k distribution as the sample size $n \to \infty$.

More generally, if the test is based on a d-dimensional parameter space and Θ_0 is defined has by k linear constraints, then the we can obtain the test above by a linear change of variables. Thus,

$$\text{degrees of freedom} = \dim(\Theta) - \dim(\Theta_0).$$

Typically we can ascertain the degrees of freedom by counting free parameters in both Θ and Θ_0 and subtracting.

Exercise 19.16. *Use a second order Taylor series for $\ln L(\theta|\mathbf{x})$ and the asymptotic normality of maximum likelihood estimators to outline the argument for the case $d = k = 1$,*

The basic approach taken for this case extends to the general case. If we expand $\ln L(\mathbf{c}|\mathbf{x})$ in a Taylor series about the parameters $\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2, \cdots, \theta_k = \hat{\theta}_k$ the maximum likelihood estimators, then the first order terms in the expansion of $\ln L(\mathbf{c}|\mathbf{x})$ vanish. The second order derivatives are the entries of the Fisher information matrix evaluated at the maximum likelihood estimator. These terms converge by the law of large numbers. A multidimensional central limit theorem applies to the vector of terms $\sqrt{n}(\hat{\theta}_1 - c_1, \ldots, \hat{\theta}_k - c_k)$. The result is the Fisher information matrix and its inverse multiplying to give the identity matrix and resulting the sum of the squares of k approximately normal random variables. This is the definition of a χ^2_k distribution.

Example 19.17. *During meiosis, paired chromosomes experience* **crossing over** *events in the formation of gametes. During prophase I, the four available chromatids (two from each parent) are in tightly aligned allowing breaks and reattachments of homologous sites (called chiasmata) on two chromatids. (See Figure 19.1.)*

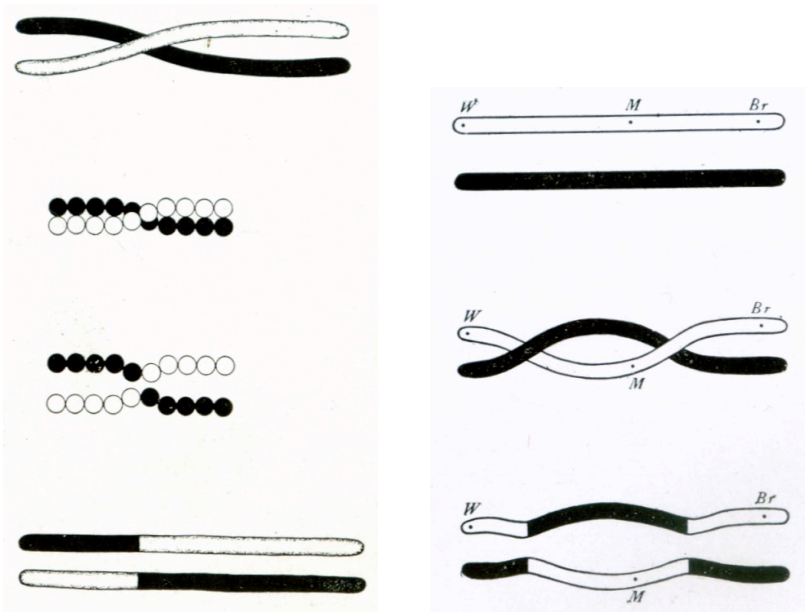


Figure 19.2: Thomas Morgan's 1916 drawings of a (**left**) single and (**right**) double crossing over event. The places where homologous non-sister chromatids (newly replicated chromosomes) exchange genetic material during chromosomal crossover during meiosis are called chaismata (singular: chiasma)

Recombination can occur with a small probability at any location along chromosome. As described in Topic 9 in the discussion that moved us from Bernoulli trials to a Poisson random variable, the number of crossing over events can be modeled as Poisson random variable. The mean number of cross overs for a given chromosomal segment is called its **genetic length** *with Morgans as the unit of measurement. This name is in honor of Thomas Morgan who won the Nobel Prize in Physiology or Medicine in 1933 for discoveries relating the role the chromosome plays in heredity.*

We are now collecting whole genome sequences for trios - an individual along with both parents. Consequently, we can determine on both the father's and the mother's chromosome the number of crossing over events and address the question: Are these processes different in the production of sperm and eggs? One simple question is: Are the number of crossing over events different in sperm and in eggs? Using the subscript m for male and f for female, this leads to the hypothesis

$$H_0 : \lambda_m = \lambda_f \quad \text{versus} \quad H_1 : \lambda_m \neq \lambda_f$$

where λ_m and λ_f is the parameter in the Poisson random variable that gives the number of crossing over events in the human chromosome across all 22 autosomes. (We will not look at the sex chromosomes X and Y in this circumstance.)

The data are n_m, n_f the number of crossing over events for each parent's chromosome. Thus, assuming that the recombination sites are independent on the two parents, the likelihood function is

$$L(\lambda_m, \lambda_f | n_m, n_f) = \frac{\lambda_m^{n_m}}{n_m!} e^{-\lambda_m} \cdot \frac{\lambda_f^{n_f}}{n_f!} e^{-\lambda_f}.$$

Exercise 19.18. *Show that the maximum likelihood estimates for the likelihood function above is*

$$\hat{\lambda}_m = n_m \quad \textit{and} \quad \hat{\lambda}_f = n_f.$$

Thus,

$$L(\hat{\lambda}_m, \hat{\lambda}_f | n_m, n_f) = \frac{n_m^{n_m}}{n_m!} \cdot \frac{n_f^{n_f}}{n_f!} e^{-(n_m + n_f)}.$$

Under the null hypothesis, λ_m and λ_f have a common value. Let's denote this by λ_0. Then the likelihood function is

$$L(\lambda_0|n_m, n_f) = \frac{\lambda_0^{n_m}}{n_m!}e^{-\lambda_0} \cdot \frac{\lambda_0^{n_f}}{n_f!}e^{-\lambda_0} = \frac{\lambda_0^{n_m+n_f}}{n_m!n_f!}e^{-2\lambda_0}.$$

Exercise 19.19. *Show that the maximum likelihood estimate for the likelihood function above is*

$$\hat{\lambda}_0 = \frac{n_m + n_f}{2}.$$

Thus,

$$L(\hat{\lambda}_0|n_m, n_f) = \frac{((n_m + n_f)/2)^{n_m+n_f}}{n_m!n_f!}e^{-(n_m+n_f)}.$$

The likelihood ratio, after canceling the factorial and exponential factors, is

$$\Lambda(n_m, n_f) = \frac{L(\hat{\lambda}_0|n_m, n_f)}{L(\hat{\lambda}_m, \hat{\lambda}_f|n_m + n_f)} = \frac{(n_m + n_f)^{n_m+n_f}}{2^{n_m+n_f}n_m^{n_m}n_f^{n_f}}.$$

For the parameter space, $d = 2$ and $k = 1$. Our data for two individuals sharing the same parents are $n_m = 56$ and $n_f = 107$. Thus,

$$-2\ln\Lambda(n_m, n_f) = -2((n_m + n_f)(\ln(n_m + n_f) - \ln 2) - n_f \ln n_f - n_m \ln n_m) = 16.228.$$

To compute the p-value

```
> nm<-56; nf<-107; n<-nm+nf
> 1-pchisq(-2*(n*(log(n)-log(2))-nf*log(nf)-nm*log(nm)),1)
[1] 5.615274e-05
```

This very low p-value, 0.0056%, allow us to reject the null hypothesis.

Exercise 19.20. *A similar set up for gibbon, a primate species whose habitat is much of southeast Asia, has $n_m = 51$ and $n_f = 120$. Give the p-value for a likelihood ratio test. In addition, test if the proportion of recombination events is the same in humans and gibbons.*

Exercise 19.21. *Consider the two sided hypotheses*

$$H_0 : \lambda = \lambda_0 \quad \text{versus} \quad H_1 : \lambda \neq \lambda_0$$

based on n independent observations from an exponential random variable parameter λ. Show that

$$-2\ln\Lambda(\mathbf{x}) = -2n(\ln(\lambda_0\bar{x}) - (\lambda_0\bar{x} - 1)). \tag{19.8}$$

Simulate 20 random variables with $\lambda = 15$ and perform the χ^2 test with $\lambda_0 = 10$ and report the p-value.

For $\lambda = 11, 12, 13, 14$, and 15, use simulations to estimate the power based on $n = 20, 40$ and 80 observations and a significance level $\alpha = 0.05$. Use that fact that that S_n, $\Gamma(n, \lambda)$ is the sum of n independent $Exp(\lambda)$ random variables to compute power. Comment on what you see.

For the sample mean of n independent $Exp(\lambda_0)$ random variables,

$$E_{\lambda_0}\bar{X} = \frac{1}{\lambda_0} \quad \text{and} \quad \text{Var}_{\lambda_0}(\bar{X}) = \frac{1}{\lambda_0^2 n}.$$

Thus, by the law of large numbers,

$$\bar{x} \approx \frac{1}{\lambda_0} \quad \text{or} \quad \lambda_0\bar{x} \approx 1.$$

A second order Taylor series expansion of the logarithm about the point $y = 1$ yields

$$\ln y \approx (y-1) - \frac{1}{2}(y-1)^2 \quad \text{and} \quad \ln y - (y-1) \approx -\frac{1}{2}(y-1)^2.$$

Substitiuting $y = \lambda_0 \bar{x}$ into this approximation and using (19.8), we find that

$$-2\ln \Lambda(\mathbf{x}) \approx n(\lambda_0 \bar{x} - 1)^2 = \left(\frac{\bar{x} - 1/\lambda_0}{1/(\lambda_0\sqrt{n})} \right)^2. \tag{19.9}$$

Consequently, $-2\ln\Lambda(\mathbf{x})$ is approximately equal to the square of the standardized score. By the central limit theorem, the standardized score is approximately normally distributed. In this way, we also see that $-2\ln\Lambda(\mathbf{x})$ is approximately the square of a standard normal random variable, i.e., is approximately χ^2_1 as promised by Theorem 19.15.

Exercise 19.22. *Use the normal approximation in (19.9) to compute the power for the values of n and λ and λ_0 in the previous exercise.*

19.4 Answers to Selected Exercises

19.1. The critical region takes the form $C\{r; r \le r_\alpha\}$ for an appropriate value r_α for an α-level test.

The likelihood function for the population, N. is the hypergeometric distribution.

$$L(N|r) = \frac{\binom{t}{r}\binom{N-t}{k-r}}{\binom{N}{k}}$$

Recall that

- t be the number captured and tagged,
- k be the number in the second capture,

In this case, we show that, for $N_2 < N_1$, $L(N_2|r)/L(N_1|r)$ increases with r.

$$\begin{aligned}\frac{L(N_2|r)}{L(N_1|r)} &= \frac{\binom{t}{r}\binom{N_2-t}{k-r}/\binom{N_2}{k}}{\binom{t}{r}\binom{N_1-t}{k-r}/\binom{N_1}{k}} = \frac{\binom{N_2-t}{k-r}\binom{N_1}{k}}{\binom{N_1-t}{k-r}\binom{N_2}{k}} \\ &= \frac{(N_2-t)_{k-r}}{(N_1-t)_{k-r}} \cdot \frac{(N_1)_k}{(N_2)_k}\end{aligned}$$

Increase r by 1 to obtain

$$\frac{L(N_2|r+1)}{L(N_1|r+1)} = \frac{(N_2-t)_{k-r-1}}{(N_1-t)_{k-r-1}} \cdot \frac{(N_1)_k}{(N_2)_k}$$

To see if this increases with r, we look at the ratio of ratios,

$$\begin{aligned}\frac{L(N_2|r+1)}{L(N_1|r+1)} \Big/ \frac{L(N_2|r)}{L(N_1|r)} &= \frac{(N_2-t)_{k-r-1}}{(N_2-t)_{k-r}} \Big/ \frac{(N_1-t)_{k-r-1}}{(N_1-t)_{k-r}} \\ &= \frac{N_1-t-k+r-1}{N_2-t-k+r-1} > 1\end{aligned}$$

because the denominator is smaller than the numerator and thus, $L(N_2|r)/L(N_1|r)$ increases with r showing that, by the Karlin-Rubin theorem, the level test is uniformly most powerful.

19.2. The likelihood

$$\begin{aligned}L(p|\mathbf{x}) &= p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \\ &= p^n \left(\frac{p}{1-p}\right)^{x_1+\cdots+x_n} = p^n \left(\frac{p}{1-p}\right)^{n\bar{x}}.\end{aligned}$$

Thus,

$$\frac{L(p_2|\mathbf{x})}{L(p_1|\mathbf{x})} = \left(\frac{p_2}{p_1}\right)^n \left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right)^{n\bar{x}}.$$

If $p_2 > p_1$, then

$$\frac{p_2}{p_1} > 1 \quad \text{and} \quad \frac{1-p_1}{1-p_2} > 1.$$

and so the product is greater than 1. Consequently, $L(p_2|\mathbf{x})/L(p_1|\mathbf{x})$ is a monotone increasing function of $\bar{x}$.

19.5. To find $\pi(p)$, we need to rewrite this expression so that we can create an expression

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

that is approximately a standard normal under the parameter value p. Beginning with the expression defining the critical region, we have that

$$\begin{aligned}
\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} &\geq z_\alpha \\
\hat{p} - p_0 &\geq z_\alpha \sqrt{p_0(1-p_0)/n} \\
\hat{p} - p &\geq z_\alpha \sqrt{p_0(1-p_0)/n} + p_0 - p \\
\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} &\geq z_\alpha \frac{\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}} + \frac{p_0 - p}{\sqrt{p(1-p)/n}} \\
z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} &\geq z_\alpha \sqrt{\frac{p_0(1-p_0)}{p(1-p)}} + \frac{p_0 - p}{\sqrt{p(1-p)/n}}
\end{aligned}$$

Take the probability to see that $\pi(p)$ has the expression in (19.2).

19.6. For a $\alpha = 0.05$ level test

```
> 1+qbinom(0.95,26,0.7)
[1] 23
```

Notice using the command `pbinom`

```
> 1-pbinom(c(21,22),26,0.7)
[1] 0.07330250 0.02595652
```

to see that the complement of the set $\{N \leq 22\}$ where N is the number of successes in 26 Bernoulli trials.

19.7. The p-value is $P_{0.7}\{N \leq 16\}$ where N is $Bin(26, 0.7)$. We find this in R by typing

```
> pbinom(16,26,0.7)
[1] 0.2295268
```

This matches the value in `binom.test`.

19.9. The likelihood is

$$\mathbf{L}(p|\mathbf{x}) = (1-p)^{n-(x_1+\cdots+x_n)} p^{x_1+\cdots+x_n}.$$

Using the definition of the likelihood ratio, we find that, under the null hypothesis, $\hat{p}_0 = p_0$ and $\hat{p}$, the sample proportion, is the maximum likelihood estimator. Thus,

$$\Lambda(\mathbf{x}) = \frac{(1-p_0)^{n(1-\hat{p})} p_0^{n\hat{p}}}{(1-\hat{p})^{n(1-\hat{p})} \hat{p}^{n\hat{p}}}.$$

Let's repeat the strategy that we used for normal data in the previous example:

$$-\ln \Lambda(\mathbf{x}) = n((1-\hat{p})(\ln(1-\hat{p}) - \ln(1-p_0)) + \hat{p}(\ln \hat{p} - \ln p_0)).$$

Next, let's replace the logarithms with their linear approximation:

$$\ln(1-\hat{p}) - \ln(1-p_0)) \approx -\frac{\hat{p}-p_0}{1-p_0} \quad \ln \hat{p} - \ln p_0 \approx \frac{\hat{p}-p_0}{p_0}.$$

Then,

$$\begin{aligned} -\ln \Lambda(\mathbf{x}) &= n(\left((1-\hat{p})(-\frac{\hat{p}-p_0}{1-p_0}) + \hat{p}(\frac{\hat{p}-p_0}{p_0})\right) = n(\hat{p}-p_0)\left(-\frac{1-\hat{p}}{1-p_0} + \frac{\hat{p}}{p_0}\right) \\ &= n(\hat{p}-p_0)\left(\frac{\hat{p}-p_0}{p_0(1-p_0)}\right) = \frac{(\hat{p}-p_0)^2}{p_0(1-p_0)/n} \end{aligned}$$

19.10. For a one-side test with alternative $H_1 : p > p_0$, the p-value is the area under the standard normal density above the z-score of the data. For the two-sided test, the p-value is the area under the standard normal density farther from zero than the z-score of the data. This is the sum of the area under the curve above this z-score and the area under the curve below the negative of the z-score. Because the standard normal density is symmetric about 0, these two areas are equal. FIgure 19.2 demonstrates the p-value as the area of the two regions under the density curves outside the black vertical lines.

The test is significant at the 10% level because the p-value is below 0.10. Correspondingly, the test is not significant at the 5% level because the p-value is above 0.05. These values are indicated by the area under the density outside the inner (for10%) or outer (for 5%) red vertical lines.

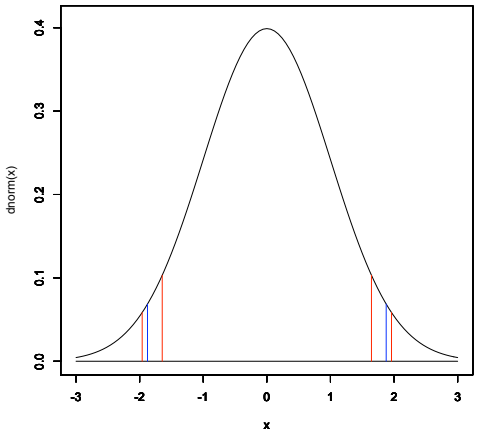


Figure 19.3: The critical values and z-test statistic value used in Exercise 19.9.

19.11. For n_i Bernoulli trials, $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n_i})$, $i = 1, 2$, we have the likelihood

$$\begin{aligned} L(p_1, p_2|\mathbf{x}_1, \mathbf{x}_2) &= p_1^{x_{1,1}}(1-p_1)^{1-x_{1,1}} \cdots p_1^{x_{1,n_1}}(1-p_1)^{1-x_{1,n_1}} \cdot p_2^{x_{2,1}}(1-p_2)^{1-x_{2,1}} \cdots p_2^{x_{2,n_2}}(1-p_2)^{1-x_{2,n_2}} \\ &= p_1^{(x_{1,1}+\cdots+x_{1,n_1})}(1-p_1)^{n_1-(x_{1,1}+\cdots x_{1,n_1})} p_2^{(x_{2,1}+\cdots+x_{2,n_2})}(1-p_2)^{n_2-(x_{2,1}+\cdots x_{2,n_2})} \end{aligned}$$

To find the maximum likelihood estimator, take logarithms and derivatives with respect to p_1 and p_2 to obtain

$$\hat{p}_1 = \frac{1}{n_1}(x_{1,1} + \cdots + x_{1,n_1}) \quad \text{and} \quad \hat{p}_2 = \frac{1}{n_2}(x_{2,1} + \cdots + x_{2,n_2}).$$

Then,

$$L(\hat{p}_1, \hat{p}_2|\mathbf{x}_1, \mathbf{x}_2) = \hat{p}_1^{n_1\hat{p}_1}(1-\hat{p}_1)^{n_1(1-\hat{p}_1)} \hat{p}_2^{n_2\hat{p}_2}(1-\hat{p}_2)^{n_2(1-\hat{p}_2)}$$

Under the null hypothesis, $p_1 = p_2$. We set this equal to p_0 to write the likelihood

$$\begin{aligned} L(p_0|\mathbf{x}_1, \mathbf{x}_2) &= p_0^{(x_{1,1}+\cdots+x_{1,n_1})}(1-p_0)^{n_1-(x_{1,1}+\cdots x_{1,n_1})} \\ &\quad \cdot p_0^{(x_{2,1}+\cdots+x_{2,n_2})}(1-p_0)^{n_2-(x_{2,1}+\cdots x_{2,n_2})} \\ &= p_0^{(x_{1,1}+\cdots+x_{1,n_1})+(x_{2,1}+\cdots+x_{2,n_2})} \\ &\quad \cdot (1-p_0)^{n_1-(x_{1,1}+\cdots x_{1,n_1})+n_2-(x_{2,1}+\cdots x_{2,n_2})} \\ &= p_0^{n_1\hat{p}_1+n_2\hat{p}_2}(1-p_0)^{n_(1-\hat{p}_1)+n_2-(1-\hat{p}_2)} \end{aligned}$$

Again, take logarithms and derivatives with respect to p_0 to obtain $\hat{p}_0$ in equation (19.6), the proportion obtained by pooling the data. Here,

$$L(\hat{p}_0|\mathbf{x}_1, \mathbf{x}_2) = \hat{p}_0^{n_1\hat{p}_1}(1-\hat{p}_0)^{n_1(1-\hat{p}_1)}\hat{p}_0^{n_2\hat{p}_2}(1-\hat{p}_0)^{n_2(1-\hat{p}_2)}$$

Thus, the likelihood ratio,

$$\Lambda(\mathbf{x}_1, \mathbf{x}_2) = \frac{\hat{p}_0^{n_1\hat{p}_1}(1-\hat{p}_0)^{n_1(1-\hat{p}_1)}\hat{p}_0^{n_2\hat{p}_2}(1-\hat{p}_0)^{n_2(1-\hat{p}_2)}}{\hat{p}_1^{n_1\hat{p}_1}(1-\hat{p}_1)^{n_1(1-\hat{p}_1)}\hat{p}_2^{n_2\hat{p}_2}(1-\hat{p}_2)^{n_2(1-\hat{p}_2)}}.$$

Again, replace the logarithms with their linear approximation:

$$\ln(1-\hat{p}_i) - \ln(1-\hat{p}_0)) \approx -\frac{\hat{p}_i - \hat{p}_0}{1-\hat{p}_0} \quad \ln\hat{p}_i - \ln\hat{p}_0 \approx \frac{\hat{p}_i - \hat{p}_0}{\hat{p}_0}.$$

$$\begin{aligned}
-\ln\Lambda(\mathbf{x}_1, \mathbf{x}_2) &= n_1\hat{p}_1(\ln\hat{p}_1 - \ln\hat{p}_0) + n_1(1-\hat{p}_1)(\ln(1-\hat{p}_1) - \ln(1-\hat{p}_0)) \\
&\quad + n_2\hat{p}_2(\ln\hat{p}_2 - \ln\hat{p}_0) + n_2(1-\hat{p}_2)(\ln(1-\hat{p}_2) - \ln(1-\hat{p}_0)) \\
&\approx n_1\hat{p}_1\left(\frac{\hat{p}_1-\hat{p}_0}{\hat{p}_0}\right) - n_1(1-\hat{p}_1)\left(\frac{\hat{p}_1-\hat{p}_0}{1-\hat{p}_0}\right) + n_2\hat{p}_2\left(\frac{\hat{p}_2-\hat{p}_0}{\hat{p}_0}\right) - n_2(1-\hat{p}_2)\left(\frac{\hat{p}_2-\hat{p}_0}{1-\hat{p}_0}\right) \\
&= n_1(\hat{p}_1-\hat{p}_0)\left(\frac{\hat{p}_1}{\hat{p}_0} - \frac{1-\hat{p}_1}{1-\hat{p}_0}\right) + n_1(\hat{p}_2-\hat{p}_0)\left(\frac{\hat{p}_2}{\hat{p}_0} - \frac{1-\hat{p}_2}{1-\hat{p}_0}\right) \\
&= n_1(\hat{p}_1-\hat{p}_0)\left(\frac{\hat{p}_1-\hat{p}_0}{\hat{p}_0(1-\hat{p}_0)}\right) + n_2(\hat{p}_2-\hat{p}_0)\left(\frac{\hat{p}_2-\hat{p}_0}{\hat{p}_0(1-\hat{p}_0)}\right) \\
&= \frac{n_1(\hat{p}_1-\hat{p}_0)^2 + n_2(\hat{p}_2-\hat{p}_0)^2}{\hat{p}_0(1-\hat{p}_0)}
\end{aligned}$$

Now note that

$$n_1(\hat{p}_1-\hat{p}_0)^2 = n_1\left(\frac{(n_1+n_2)\hat{p}_1 - (n_1\hat{p}_1 + n_2\hat{p}_2)}{n_1+n_2}\right)^2 = n_1n_2^2\left(\frac{\hat{p}_1-\hat{p}_2}{n_1+n_2}\right)^2.$$

Perform a similar computation for the second term

$$\begin{aligned}
-\ln\Lambda(\mathbf{x}_1, \mathbf{x}_2) &\approx \frac{(n_1n_2^2 + n_2n_1^2)((\hat{p}_1-\hat{p}_2)/(n_1+n_2))^2}{\hat{p}_0(1-\hat{p}_0)} = \frac{n_1n_2(\hat{p}_1-\hat{p}_2)^2/(n_1+n_2)}{\hat{p}_0(1-\hat{p}_0)} \\
&= \frac{(\hat{p}_1-\hat{p}_2)^2}{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \left|\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right|^2
\end{aligned}$$

19.12. Beginning with the population parameters,

- p_1 is proportion of all hives that would have survived the first, less harsh, winter, and
- p_2 is proportion of all hives that would have survived the second, harsher, winter.

Then we can write the hypothesis as (19.7). The sample proportions for each group and the pooled proportion are, respectively,

$$\hat{p}_1 = \frac{88}{112} = 0.7867, \quad \hat{p}_2 = \frac{64}{99} = 0.6598, \quad \hat{p}_0 = \frac{88+64}{112+99} = 0.7204.$$

Thus,

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.7867 - 0.6598}{\sqrt{0.7204(1-0.7204)\left(\frac{1}{112}+\frac{1}{99}\right)}} = 2.249$$

The p-value

```
> 1-pnorm(2.249)
[1] 0.01225625
```

ia sufficiently law to say that we have moderate evidence against the null hypothesis and say that the harsher winter reduced the proportion of surviving bee hives.

19.13. We see that the power $1 - \beta = 0.580$.

```
> power.prop.test(n=100,p1=0.70,p2=0.6,sig.level=0.10,alternative = c("one.sided"))

     Two-sample comparison of proportions power calculation

              n = 100
             p1 = 0.7
             p2 = 0.6
      sig.level = 0.1
          power = 0.5800652
    alternative = one.sided

 NOTE: n is number in *each* group
```

19.14. If p_c is the proportion in the population that have the control and contract polio and p_t is the proportion in the population that have the treatment and contract polio. Then, we have the hypothesis test, We want to show that the vaccine reduces the rate of polio infection. thus,

$$H_0 : p_t \geq p_c \quad \text{versus} \quad H_1 : p_t < p_c.$$

```
> prop.test(c(56,142),c(200000,200000),alternative=c("less"))

2-sample test for equality of proportions with continuity
correction

data:  c(56, 142) out of c(2e+05, 2e+05)
X-squared = 36.508, df = 1, p-value = 7.602e-10
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000000 -0.0003093083
sample estimates:
 prop 1  prop 2
0.00028 0.00071
```

With the very low p-value of 7.6×10^{-10}, we can reject H_0 and say that the vaccine reduces the rate of polio,

19.16. For the case in which $d = k = 1$. Then, for n observations, and maximum likelihood estimator $\hat{\theta}$,

$$\sqrt{n}(c - \hat{\theta})$$

converges in distribution to a normal random variable with variance $1/I(c)$, the reciprocal of the Fisher information. Thus,

$$\frac{c-\hat{\theta}}{1/\sqrt{nI(c))}} = \frac{\sqrt{n}(c-\hat{\theta})}{1/\sqrt{I(c))}}$$

has approximately a standard normal distribution and its square

$$\frac{n(c-\hat{\theta})^2}{1/I(c)} \tag{19.10}$$

has approximately a χ^2_1 distribution.

We next apply Taylor's theorem to obtain the quadratic approximation

$$\begin{aligned} -2\ln\Lambda_1(X) &= -2\ln L(c|X) + 2\ln L(\hat{\theta}|X) \approx -2(c-\hat{\theta})\frac{d}{d\theta}\ln L(\hat{\theta}_1|X) - (c-\hat{\theta})^2\frac{d^2}{d\theta^2}\ln L(\hat{\theta}|X) \\ &= -n(c-\hat{\theta})^2\frac{d^2}{d\theta^2}\ln L(X|c) \end{aligned}$$

The linear term vanishes because $d\ln L(\hat{\theta}|X)/d\theta = 0$ at $\hat{\theta}$, the maximum likelihood estimator. Recall that the likelihood

$$L(X|\theta) = f_X(x_1|\theta)\cdots f_X(x_n|\theta)$$

and using the properties of the logarithm and the derivative, we see that

$$\frac{d^2}{d\theta^2}\ln L(X|\theta) = \frac{d^2}{d\theta^2}\ln f_X(x_1|\theta) + \cdots + \frac{d^2}{d\theta^2}\ln f_X(x_n|\theta)$$

We now apply the law of large numbers to see that for large n

$$\frac{1}{n}\frac{d^2}{d\theta^2}\ln L(c|X) = \frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}f_X(X_i|c) \approx E_\theta\left[\frac{d^2}{d\theta^2}f_X(X_1|c)\right] = -I(c).$$

Thus,

$$-2\ln\Lambda_1(X) \approx -n(c-\hat{\theta})^2 \times (-I(c)) = \frac{n(c-\hat{\theta})^2}{1/I(c)}$$

which in (19.10) we have noted has approximately a χ^2_1 distribution.

19.18. Taking logarithms, we find

$$\ln L(\lambda_m, \lambda_f|n_m, n_f) = n_m\ln\lambda - \ln(n_m!) - \lambda_m + n_f\ln\lambda - \ln(n_f!) - \lambda_f.$$

The derivative with respect to λ_m is

$$\frac{\partial}{\partial\lambda_m}\ln L(\lambda|n_m, n_f) = \frac{n_m}{\lambda_m} - 1.$$

Now set this equal to 0 and solve for λ_m. Because the second derivative with respect to λ_m is negative, this is a maximum. A nearly identical computation can be used to fiind $\hat{\lambda}_f$.

19.19. Taking logarithms, we find the score function,

$$\ln L(\lambda|n_m, n_f) = (n_m + n_f)\ln\lambda - \ln(n_m!n_f!) - 2\lambda.$$

The derivative with respect to λ is

$$\frac{\partial}{\partial\lambda}\ln L(\lambda|n_m, n_f) = \frac{n_m + n_f}{\lambda} - 2.$$

Now set this equal to 0 and solve for λ. Because the second derivative with respect to λ is negative, this is a maximum.

19.20. For the hypothesis on the number of crossing over events we have the test statistic and p-value

```
> nm<-51; nf<-120; n<-nm+nf
> 1-pchisq(-2*(n*(log(n)-log(2))-nf*log(nf)-nm*log(nm)),1)
[1] 8.664093e-08
```

This is also a very low p-value and we can reject the hypothesis of equal rtes of crossing over events.

The second question is a two-sample two-sided proportion test. Here is the R output.

```
> prop.test(c(56,51),c(163,171))

2-sample test for equality of proportions with continuity correction

data:  c(56, 51) out of c(163, 171)
X-squared = 0.5926, df = 1, p-value = 0.4414
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06076261  0.15138795
sample estimates:
   prop 1    prop 2
0.3435583 0.2982456
```

This gives a p-value of 44%, much too high to reject a hypothesis of equal proportion of crossing over events derived from the females in the two species, human and gibbon.

19.21. The likelihood function for observations $\mathbf{x} = (x_1 \ldots, x_n)$ is

$$L(\lambda|\mathbf{x}) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i) = \lambda^n \exp(-\lambda n \bar{x}).$$

We have seem that the maximum likelihood estimate $\hat{\lambda} = 1/\bar{x}$. Thus the likelihood ratio,

$$\begin{aligned}\Lambda(\mathbf{x}) = \frac{L(\lambda_0|\mathbf{x})}{L(\hat{\lambda}|\mathbf{x})} &= \frac{\lambda_0^n \exp(-\lambda_0 n\bar{x})}{(1/\bar{x}^n)\exp(-(1/\bar{x})n\bar{x})} \\ &= (\lambda_0\bar{x})^n \frac{\exp(-\lambda_0 n\bar{x})}{\exp(-n)} = (\lambda_0\bar{x})^n \exp(-n(\lambda_0\bar{x} - 1)).\end{aligned}$$

and

$$-2\ln\Lambda(\mathbf{x}) = -2n(\ln\lambda_0\bar{x} - (\lambda_0\bar{x} - 1)).$$

```
> x<-rexp(20,15)
> (xbar<-mean(x))
[1] 0.06126583
> -2*20*(log(10*xbar)-(10*xbar-1))
[1] 4.509281
> 1-pchisq(4.509281,1)
[1] 0.03371141

> x<-rexp(20,15)
> (xbar<-mean(x))
[1] 0.07144714
> -2*20*(log(20*xbar)-(20*xbar-1))
[1] 2.880317
> 1-pchisq(2.880317,1)
[1] 0.08966837
```

For the two simulations, the p-values are 0.0337 and 0.0897.

For the simulation, we also take advantage of the fact that $S_n, \Gamma(n, \lambda)$ is the sum of n independent $Exp(\lambda)$ random variables. So, we simulation S_n/n 10,000 times for each value of λ and n and find the proportion of times that we reject $\lambda_0 = 10$ using the χ^2 statistic.

```
> lambda0<-10;N<-10000; lambda<-11:15
> n<-20;power20<-rep(0,5)
> for (i in 1:5){xbar<-rgamma(N,n,lambda[i])/n;
  chisqstat<--2*n*(log(lambda0*xbar)-(lambda0*xbar-1));
  pvalue<-1-pchisq(chisqstat,1);power20[i]<-length(pvalue[pvalue<0.05])/N}
> n<-40;power40<-rep(0,5)
> for (i in 1:5){xbar<-rgamma(N,n,lambda[i])/n;
  chisqstat<--2*n*(log(lambda0*xbar)-(lambda0*xbar-1));
  pvalue<-1-pchisq(chisqstat,1);power40[i]<-length(pvalue[pvalue<0.05])/N}
> n<-80;power80<-rep(0,5)
> for (i in 1:5){xbar<-rgamma(N,n,lambda[i])/n;
  chisqstat<--2*n*(log(lambda0*xbar)-(lambda0*xbar-1));
  pvalue<-1-pchisq(chisqstat,1);power80[i]<-length(pvalue[pvalue<0.05])/N}
> data.frame(lambda,power20,power40,power80)
  lambda power20 power40 power80
1     11  0.0692  0.0932  0.1358
2     12  0.1187  0.2017  0.3646
3     13  0.2021  0.3601  0.6385
4     14  0.3100  0.5378  0.8508
5     15  0.4093  0.7187  0.9507
```

We can use the $\Gamma(n, \lambda)$ distribution as a test statistic to compute power directly.

```
> n<-20;upper<-qgamma(0.975,n,lambda0); lower<-qgamma(0.025,n,lambda0)
> power20<-1-pgamma(upper,n,lambda)+pgamma(lower,n,lambda)
> n<-40;upper<-qgamma(0.975,n,lambda0); lower<-qgamma(0.025,n,lambda0)
> power40<-1-pgamma(upper,n,lambda)+pgamma(lower,n,lambda)
> n<-80;upper<-qgamma(0.975,n,lambda0); lower<-qgamma(0.025,n,lambda0)
> power80<-1-pgamma(upper,n,lambda)+pgamma(lower,n,lambda)
> data.frame(lambda,power20,power40,power80)
  lambda    power20   power40   power80
1     11 0.06283839 0.0832438 0.1254101
2     12 0.10841061 0.1856479 0.3434286
3     13 0.17993890 0.3413916 0.6217434
4     14 0.27219452 0.5214876 0.8383838
5     15 0.37812771 0.6897081 0.9489048
```

Notice that the power increases with n and with distance $!\lambda - \lambda_0|$ from the value under the null hypothesis.

19.22, We first find the upper and lower rejection regions for $\bar{X}$, here approximated as a $N(1/\lambda_0, 1/\sqrt{\lambda_0 n})$ and then determine the probability of rejection when $\bar{X}$ is $N(1/\lambda, 1/\sqrt{\lambda n})$

```
> n<-20
> upper<-qnorm(0.975,1/lambda0,1/(lambda0*sqrt(n)))
> lower<-qnorm(0.025,1/lambda0,1/(lambda0*sqrt(n)))
> power20<-1-pnorm(upper,1/lambda,1/(lambda*sqrt(n)))
  +pnorm(lower,1/lambda,1/(lambda*sqrt(n)))
```

```
> n<-40
> upper<-qnorm(0.975,1/lambda0,1/(lambda0*sqrt(n)))
> lower<-qnorm(0.025,1/lambda0,1/(lambda0*sqrt(n)))
> power40<-1-pnorm(upper,1/lambda,1/(lambda*sqrt(n)))
  +pnorm(lower,1/lambda,1/(lambda*sqrt(n)))
> n<-80
> upper<-qnorm(0.975,1/lambda0,1/(lambda0*sqrt(n)))
> lower<-qnorm(0.025,1/lambda0,1/(lambda0*sqrt(n)))
> power80<-1-pnorm(upper,1/lambda,1/(lambda*sqrt(n)))
  +pnorm(lower,1/lambda,1/(lambda*sqrt(n)))
> data.frame(lambda,power20,power40,power80)
  lambda    power20    power40   power80
1     11 0.04836719 0.06646455 0.1047011
2     12 0.07306953 0.13865740 0.2866999
3     13 0.11389874 0.25766108 0.5538240
4     14 0.16976769 0.41522390 0.7977917
5     15 0.24075449 0.58797216 0.9372622
```

These values are lower than that found in the simulation. The normal approximation removes the skewness found in the exponential distribution. This results in rejection occurring less frequently using the normal approximation and thus lower power. Notice that as n increases, $\bar{X}$ becomes less skewed and so power computations more closely agree.

Topic 20

t Procedures

> *A curve has been found representing the frequency distribution of values of the means of such samples, when these values are measured from the mean of the population in terms of the standard deviation of the sample....* - William Sealy Gosset. 1908, The Probable Error of a Mean, *Biometrika*

The z-score is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

taken under the assumption that the population standard deviation is known.

If we are forced to replace the unknown σ^2 with its unbiased estimator s^2, then the statistic is known as t:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

The term $s/\sqrt{n}$ which estimates the standard deviation of the sample mean is called the **standard error**.

We have previously noted that for independent normal random variables the distribution of the t statistic can be determined **exactly**. Because we approximate σ with s, the t-statistic has a higher level of uncertainty than the corresponding z-statistic. This uncertainty decreases with n, the number of observations. Thus, when using the t distribution to construct a confidence interval for the population mean μ, we saw that the margin of error decreased as the number of observations increased. Typically, we do not use the number of observations n to describe this but rather **degrees of freedom** $n - 1$ to match the division by $n - 1$ in the computation of the sample variance, s^2.

We now turn to using the t-statistic as a test statistic for hypothesis tests of the population mean. As with several other procedures we have seen, the two-sided t test is a likelihood ratio test. We will save showing this result into the last section and instead focus on the applications of this widely used set of procedures.

20.1 Guidelines for Using the t Procedures

- Except in the case of small samples, the assumption that the data are a simple random sample from the population of interest is more important that the population distribution is normal.
- For sample sizes less than 15, use t procedures if the data are close to normal.
- For sample sizes at least 15 use t procedures except in the presence of outliers or strong skewness.
- The t procedures can be used even for clearly skewed distributions when the sample size is large, typically over 40 observations.

These criteria are designed to ensure that $\bar{x}$ is a sample from a nearly normal distribution. When these guidelines fail to be satisfied, then we can turn to alternatives that are not based on the central limit theorem, but rather use the rankings of the data. These alternatives, the Mann-Whitney or Wilcoxon rank sum test and the Wilcoxon signed-ranked test, are discussed at the end of this topic.

20.2 One Sample t Tests

We will later explain that the likelihood ratio test for the two sided hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$

based on independent **normal** observations $X_1 \ldots, X_n$ with unknown mean μ and **unknown** variance σ^2 is a t-test.

So, compute the t statistic $T(\mathbf{x})$ from the data $\mathbf{x}$. Then, the critical region

$$C = \{|T(\mathbf{x})| > t_{n-1,\alpha/2}\}.$$

where $t_{n-1,\alpha/2}$ is the upper $\alpha/2$ tail probability of the t distribution with $n-1$ degrees of freedom.

Example 20.1. *Radon is a radioactive, colorless, odorless, tasteless noble gas, occurring naturally as the decay product of uranium. It is one of the densest substances that remains a gas under normal conditions.*

Radon is responsible for the majority of the public exposure to ionizing radiation and is the most variable from location to location. Radon gas from natural sources can accumulate in buildings, especially in confined areas such as attics, and basements. Epidemiological evidence shows a clear link between breathing high concentrations of radon and incidence of lung cancer. According to the United States Environmental Protection Agency, radon is the second most frequent cause of lung cancer, after cigarette smoking, causing 21,000 lung cancer deaths per year in the United States.

*To check the reliability of radon detector, a university placed 12 detectors in a chamber having 105 picocuries of radon. (*1 *picocurie is* 3.7×10^{-2} *decays per second. This is roughly the activity of 1 picogram of the radium 226.)*

The two-sided hypothesis

$$H_0 : \mu = 105 \quad \textit{versus} \quad H_1 : \mu \neq 105,$$

where μ *is the actual amount of radon radiation. In other words, we are checking to see if the detector is biased either upward or downward.*

The detector readings were:

91.9 97.8 111.4 122.3 105.4 95.0 103.8 99.6 96.6 119.3 104.8 101.7

Using R*, we find for an* $\alpha = 0.05$ *level significance test:*

```
> radon<-c(91.9,97.8,111.4,122.3,105.4,95.0,103.8,99.6,96.6,119.3,104.8,101.7)
> hist(radon)
> mean(radon)
[1] 104.1333
> sd(radon)
[1] 9.39742
> length(radon)
[1] 12
> (tstar<-qt(0.975,11))
[1] 2.200985
```

Thus, the t*-statistic is*

$$t = \frac{105 - 104.1333}{9.39742/\sqrt{12}} = -0.3195.$$

Thus, for a 5% significance test, $|t| < 2.200985$*, the critical value and we fail to reject* H_0*.* R *handles this procedure easily.*

```
> t.test(radon,alternative=c("two.sided"),mu=105)

One Sample t-test
```

```
data:  radon
t = -0.3195, df = 11, p-value = 0.7554
alternative hypothesis: true mean is not equal to 105
95 percent confidence interval:
  98.1625 110.1042
sample estimates:
mean of x
 104.1333
```

The output also gives the 95% confidence interval

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.025,11}.$$

The **power** is the probability of rejecting when the parameter value is μ

$$\pi(\mu) = P_\mu \left\{ \left| \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \right| \geq t_{n-1,\alpha/2} \right\}$$

To determine the power curve, we begin with the following exercise.

Exercise 20.2. *If the observations $X_1, \ldots, X_n$ are independent normal random variables with mean μ then*

$$\tilde{T} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

has a t distribution with $n-1$ degrees of freedom and **non-centrality parameter**

$$a = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}.$$

Thus, the power function

$$\begin{aligned} \pi(\mu) &= P_\mu \left\{ |\tilde{T}| \geq t_{n-1,\alpha/2} \right\} \\ &= 1 - P_\mu \left\{ |\tilde{T}| < t_{n-1,\alpha/2} \right\} \\ &= 1 - P_\mu \left\{ -t_{n-1,\alpha/2} < \tilde{T} < t_{n-1,\alpha/2} \right\} \end{aligned}$$

Thus, we use the `qt` command to set the critical values $t_{n-1,\alpha/2}$ and the `pt` command to find the power using the appropriate non-centrality parameter. This same function can be used to construct the receiver operating characteristic, determine sample size to achieve disired type I and type II errors, and to look at power as a function of the number of samples. In this regards, note that the non-centrality parameter increases with the number of observations and consequently power increases.

Returning to the example of the radon detector, by design, $\pi(\mu_0) = \alpha$, the significance level. We estimate a by replacing σ with the sample standard deviation s and, as an example, estimate the power against an alternative of a change in $\Delta = \mu - \mu_0 = 5$ picocuries is

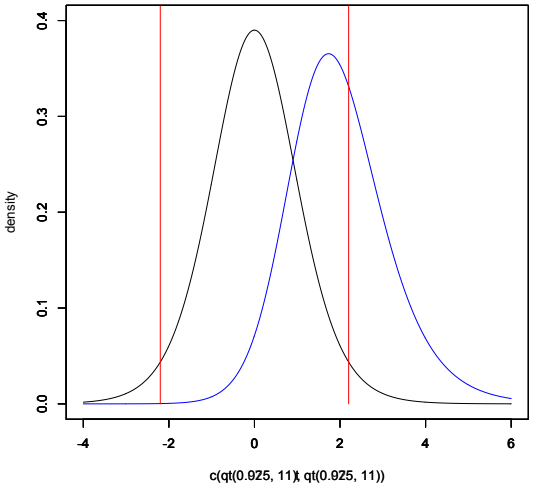


Figure 20.1: I black, t density with 11 degrees of freedom. Red vertical lines at $t_{n-1,\alpha/2} = 2.2010$ are the critical values for a two-sided test with $\alpha = 0.05$, In blue, t density with 11 degrees of freedom and non-centrality parameter $a = 1.8431$. The power is the area outside the red lines under the blue density function. The `power.t.test` command considers only the larger area on the right side.

```
> delta<-5
> (a<-delta/(sd(radon)/sqrt(length(radon))))
[1] 1.843113
> 1-(pt(tstar,11,a)-pt(-tstar,11,a))
[1] 0.390913
```

Exercise 20.3. *Draw the power function for this circumstance with significance level* $\alpha = 0.05$*,* $\mu_0 = 105$ *and* $n = 12$*. Use the standard deviation obtained from the data.*

R makes this computation using the command `power.t.test`.

```
> power.t.test(n=12,delta=5,sd=sd(radon),type=c("one.sample"))

     One-sample t test power calculation

              n = 12
          delta = 5
             sd = 9.39742
      sig.level = 0.05
          power = 0.3907862
    alternative = two.sided
```

Notice that this command give a different value for the power. This is due to the fact that this R command accounts only for the larger area in FIgure 20.1.

```
> 1-(pt(qt(0.975,11),11,a))
[1] 0.3907862
```

The `power.t.test` command consider both one and two sample t procedures. It will also handle both one-sided and two-sided tests. The command considers five issues - sample size `n`, the difference between the null and a fixed value of the alternative `delta`, the standard deviation s, the significance level α, and the power. We can use `power.t.test` to drop out any one of these five and use the remaining four to determine the remaining value. For example, if we want to assure an 80% power against an alternative of 110, then we need to make 30 measurements.

```
> power.t.test(power=0.80,delta=5,sd=sd(radon),type=c("one.sample"))

     One-sample t test power calculation

              n = 29.70383
          delta = 5
             sd = 9.39742
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

In these types of application, we often use the terms **specificity** and **sensitivity**. Recall that setting the significance level α is the same as setting the false positive rate or type I error probability. The specificity of the test is equal to $1 - \alpha$, the probability that the test is not rejected when the null hypothesis is true. The sensitivity is the same as the power, one minus the type II error rate, $1 - \beta$.

Exercise 20.4. *Plot the receiver operating characteristic for* $n = 6, 12$ *and* 24 *observations, using the standard deviation obtained from the data and an alternative* $\mu = 110$*.*

20.3 Correspondence between Two-Sided Tests and Confidence Intervals

For a two-sided t-test, we have the following list of equivalent conditions:

$$\text{fail to reject with significance level } \alpha.$$

$$|t| < t_{n-1,\alpha/2}$$

$$\left|\frac{\mu_0 - \bar{x}}{s/\sqrt{n}}\right| < t_{n-1,\alpha/2}$$

$$-t_{n-1,\alpha/2} < \frac{\mu_0 - \bar{x}}{s/\sqrt{n}} < t_{n-1,\alpha/2}$$

$$-t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu_0 - \bar{x} < t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu_0 < \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

$$\mu_0 \text{ is in the } \gamma = 1 - \alpha \text{ confidence interval}$$

This is displayed in Figure 20.1 with the green $\bar{x}$ and the horizontal green line indicating the γ-level confidence interval containing μ_0. In addition, *reject the hypothesis with significance level* α is equivalent to μ_0 *is not in the confidence interval.* This is displayed in Figure 1 with the red $\bar{x}$ and the horizontal line indicating the γ-level confidence interval that fails to contain μ_0.

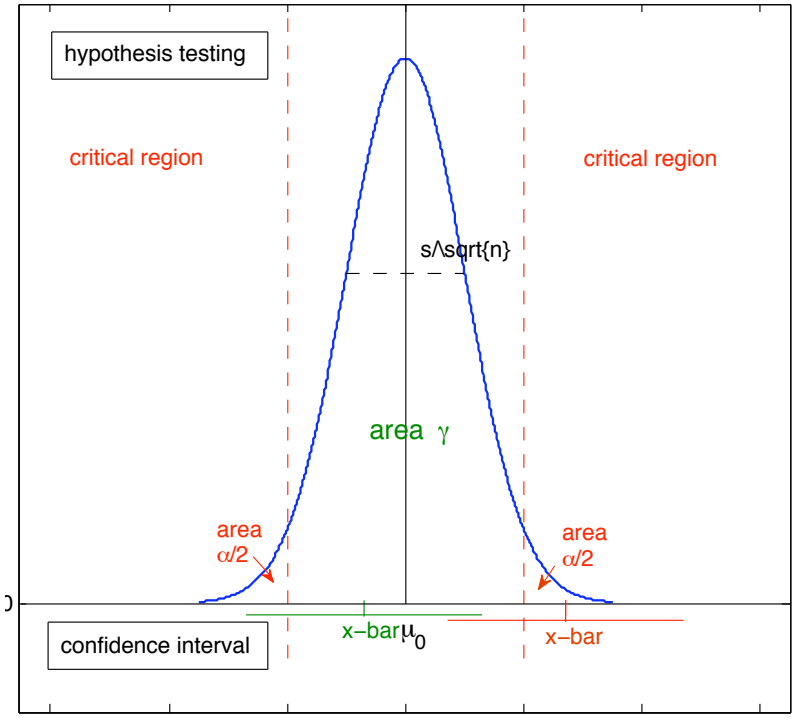


Figure 20.2: γ-level confidence intervals and α level hypothesis tests. $\gamma = 1 - \alpha$. The blue density curve is density of the sampling distribution under the null hypothesis. The red vertical dashes show the critical values for a two-sided test. γ is the area under the density curve between the vertical critical value lines. α is the area under the density curve outside the vertical critical value lines. The green $\bar{x}$ shows the case of *fails to reject* is equivalent to *the confidence interval contains* μ_0. The red $\bar{x}$ show the case *reject* is equivalent to *the confidence interval fails to contain* μ_0.

20.4 Matched Pairs Procedures

A **matched pair procedure** is called for when a pair of quantitative measurements from a simple random sample

$$X_1, X_2, \ldots, X_n, \quad \text{and} \quad Y_1, Y_2, \ldots, Y_n$$

are made on the same subjects. The alternative can be either one-sided or two sided. Underlying this assumption is that the populations are the same under the null hypothesis.

Thus, when H_0 holds and if in addition, if the data are normal, then $\bar{X} - \bar{Y}$ is also normal and so

$$T = \frac{\bar{X} - \bar{Y}}{S_{X-Y}/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom.

The γ-level confidence interval for the difference in the population means is

$$\bar{x} - \bar{y} \pm \frac{s_{X-Y}}{n} t_{n-1,(1-\gamma)/2,\cdot}$$

Example 20.5. *Researchers are concerned about the impact of vitamin C content reduction due to storage and shipment. To test this, researchers randomly chose a collection of bags of wheat soy blend bound for Haiti, marked them, and measured vitamin C from a sample of the contents. Five months later, the bags were opened and a second sample was measured for vitamin C content. The units are milligrams of vitamin C per 100g of wheat soy blend.*

Factory	Haiti	Factory	Haiti	Factory	Haiti	Factory	Haiti
44	40	45	38	39	43	50	37
50	37	32	40	52	38	40	34
48	39	47	35	45	38	39	38
44	35	40	38	37	38	39	34
42	35	38	34	38	41	37	40
47	41	41	35	44	40	44	36
49	37	40	34	43	35		

Here is the R *output with the 95% confidence interval for $\mu_F - \mu_H$ where*

- *μ_F is the mean vitamin C content of the wheat soy blend at the factory and*
- *μ_H is the mean vitamin C content of the wheat soy blend in Haiti.*

```
> factory<-c(44,50,48,44,42,47,49,45,32,47,40,38,41,40,39,52,45,37,38,44,43,
+50,40,39,39,37,44)
> haiti<-c(40,37,39,35,35,41,37,38,40,35,38,34,35,34,43,38,38,38,41,40,35,37,
+34,38,34,40,36)
> boxplot(factory,haiti)
> t.test(factory, haiti, alternative = c("two.sided"),mu = 0, paired = TRUE)

Paired t-test

data:  factory and haiti
t = 4.9589, df = 26, p-value = 3.745e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.122616 7.544050
sample estimates:
mean of the differences
              5.333333
```

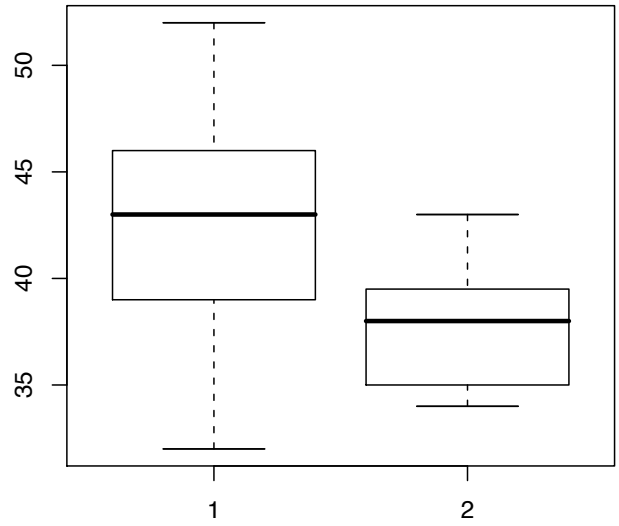


Figure 20.3: Vitamin C content in milligrams per 100 grams, measured at the factory and measured 5 month later in Haiti.

The input

```
> t.test(factory - haiti, alternative = c("two.sided"),mu = 0)
```

gives essentially the same output.

In addition, the output

```
> t.test(haiti, alternative = c("less"),mu = 40)

One Sample t-test

data:  haiti
t = -5.3232, df = 26, p-value = 7.175e-06
alternative hypothesis: true mean is less than 40
95 percent confidence interval:
     -Inf 38.23811
sample estimates:
mean of x
 37.40741
```

shows that we would reject the one sided test

$$H_0 : \mu \geq 40 \quad \textit{versus} \quad H_1 : \mu < 40,$$

based on a goal of having 40mg/100g vitamin C in the wheat soy blend consumed by the Haitians.

We have used R primarily to compute a confidence interval. If the goal of the program is to have reduction in vitamin C be less that a given amount c, then we have the hypothesis

$$H_0 : \mu_F - \mu_H \geq c \quad \text{versus} \quad H_1 : \mu_F - \mu_H < c.$$

We can test this using R by replacing `mu=0` with `mu=c`.

20.5 Two Sample Procedures

Now we consider the situation in which the two samples

$$X_1, X_2, \ldots, X_{n_X}, \quad \text{and} \quad Y_1, Y_2, \ldots, Y_{n_Y}$$

are independent but are not paired. In particular, the number of observations n_X and n_Y in the two samples could be different. If the first sample has common mean μ_X and variance σ_X^2 and the second sample has common mean μ_Y and variance σ_Y^2, then

$$E[\bar{X} - \bar{Y}] = \mu_X - \mu_Y \quad \text{and} \quad \text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}.$$

For the two sided hypothesis test

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y,$$

The corresponding t-statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \tag{20.1}$$

with s_X^2 and s_Y^2 the unbiased sample variances. Unlike the match pairs procedures, the test statistic (20.1) does not have a t distribution under the null hypothesis. Indeed, the density and the distribution of this statistic are difficult to compute.

In this circumstance, we now make what is commonly known in statistics as a *conservative* approximation. We replace the actual distribution of the t statistic in (20.1) with one which has slightly bigger tails. Thus, the computed p-value which are just integrals of the density function will be slightly larger. In this way, a conservative procedures is one that does not decrease the type I error probability.

This goal can be accomplished by approximating an ordinary Student's t distribution with the effective degrees of freedom ν calculated using the **Welch-Satterthwaite** equation:

$$\nu = \frac{(s_X^2/n_X + s_Y^2/n_Y)^2}{(s_X^2/n_X)^2/(n_X - 1) + (s_Y^2/n_Y)^2/(n_Y - 1)}. \tag{20.2}$$

As we saw in our discussion on Interval Estimation, this also gives a γ-level confidence interval for the difference in the means μ_x and μ_Y.

$$\bar{x} - \bar{y} \pm t_{(1-\gamma)/2,\nu} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}.$$

We also learned that the effective degrees of freedom are largest when the two sample variances are nearly equal. In this case the number of degrees of freedom is 2 fewer than the sum of the two sets of observations.

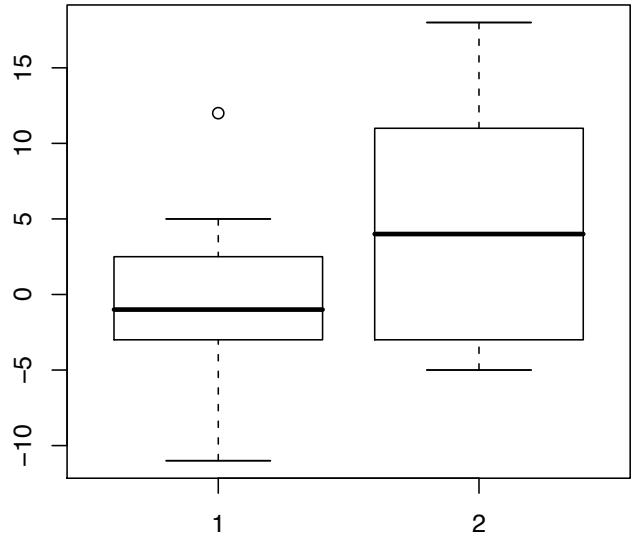
Example 20.6. *To investigate the effect on blood pressure of added calcium in the diet, a researchers conducts a double blind randomized experiment. In the treatment group, each individual receives a calcium supplement. In the control group, the individual takes a placebo. The response variable is the decrease in systolic blood pressure, measured in millimeters of mercury, after 12 weeks. The test subjects are all male.*

```
> calcium<-c(7,-4,18,17,-3,-5,1,10,11,-2)
> mean(calcium)
[1] 5
> sd(calcium)
[1] 8.743251
> placebo<-c(-1,12,-1,-3,3,-5,5,2,-11,-1,-3)
> mean(placebo)
```

```
[1] -0.2727273
> sd(placebo)
[1] 5.900693
> boxplot(placebo,calcium)
```

The null hypothesis is that the treatment did not reduce μ_t the mean blood pressure of the treatment any more than it did the mean μ_c for the control group. The alternative is that it did reduce blood pressure more. Formally the hypothesis test is

$$H_0 : \mu_c \leq \mu_t \quad \text{versus} \quad H_1 : \mu_c > \mu_t.$$



The t-statistic is

$$t = \frac{5.000 + 0.273}{\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}} = 1.604.$$

```
> t.test(calcium,placebo,alternative = c("greater"))

Welch Two Sample t-test

data:  calcium and placebo
t = 1.6037, df = 15.591, p-value = 0.06442
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.476678       Inf
sample estimates:
 mean of x  mean of y
 5.0000000 -0.2727273
```

Thus. the evidence against the null hypothesis is modest with a p-value of about 6%. Notice that the effective degrees of freedom is $\nu = 15.591$. The maximum possible is value for degrees of freedom is 19.

To see a 90% confidence interval remove the "greater than" alternative" and set the confidence level.

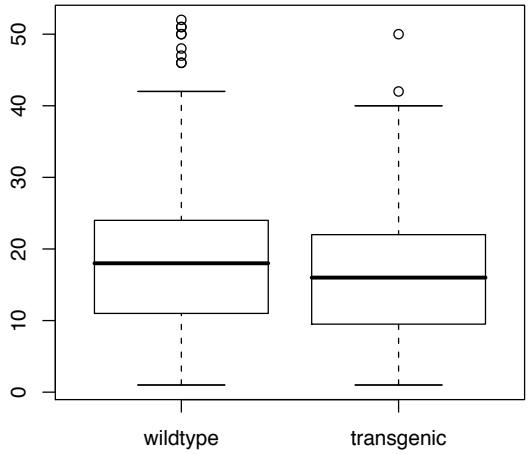
```
> t.test(calcium,placebo,conf.level = 0.9)

Welch Two Sample t-test

data:  calcium and placebo
t = 1.6037, df = 15.591, p-value = 0.1288
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -0.476678 11.022133
sample estimates:
 mean of x  mean of y
 5.0000000 -0.2727273
```

Example 20.7. *The life span in days of 88 wildtype and 99 transgenic mosquitoes is given in the following data set.*

```
> mosquitoes<-read.delim("http://math.arizona.edu/~jwatkins/mosquitoes.txt")
> boxplot(mosquitoes)
```



The goal is to see if overstimulation of the insulin signaling cascade in the mosquito midgut reduces the μ_t, the mean life span of these transgenic mosquitoes from that of the wild type μ_{wt}.

$$H_0 : \mu_{wt} \le \mu_t \quad \text{versus} \quad H_1 : \mu_{wt} > \mu_t.$$

```
> wildtype<-mosquitoes[1:88,1]
> transgenic<-mosquitoes[,2]
> t.test(transgenic,wildtype,alternative = c("less"))

Welch Two Sample t-test
```

```
data:  transgenic and wildtype
t = -2.4106, df = 169.665, p-value = 0.008497
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.330591
sample estimates:
mean of x mean of y
 16.54545  20.78409
```

To determine a 98% confidence interval, we again remove the alternative command.

```
> t.test(transgenic,wildtype,conf.level=0.98)

Welch Two Sample t-test

data:  transgenic and wildtype
t = -2.4106, df = 169.665, p-value = 0.01699
alternative hypothesis: true difference in means is not equal to 0
98 percent confidence interval:
 -8.3680812 -0.1091915
sample estimates:
mean of x mean of y
 16.54545  20.78409
```

Exercise 20.8. *Notice that the 98% confidence interval,* $(-8.3680812, -0.1091915)$ *does not contain 0. What can be said about a two-sided test at the 2% significance level? What can be said about the p-value for a one-sided test?*

The two-sample procedure assumes that the two sets of observations are independent and so it is *not* an appropriate procedure when the the observations are paired. Moreover, the two sample procedure is yields less powerful test when the two sets of observations are positively correlated.

Example 20.9. *Previously we investigated the relationship of age of parents to the* de novo *mutations in the offspring for the 78 Icelandic trios. Now, we address the simpler question: are fathers older than mothers at the time of the child's birth? State as a hypothesis, we have*

$$H_0 : \mu_f \leq \mu_m \quad \text{versus} \quad H_1 : \mu_f > \mu_m.$$

where μ_m *is the mean age of Icelandic fathers at the time of birth and* μ_m *is the mean age of Icelandic mothers. We anticipate that the two parents' ages are positively correlcate - older fathers with older mothers and younger fathers with younger mothers.*

The appropriate matched-pair test has R *commands.*

```
> t.test(father,mother,alternative=c("greater"),paired=TRUE)

Paired t-test

data:  father and mother
t = 6.6209, df = 77, p-value = 2.151e-09
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.838564      Inf
sample estimates:
mean of the differences
               2.456197
```

Notice that $t = 6.6209$. If we had performed (inappropriately) the two-sample t test, we find

```
> t.test(father,mother,alternative=c("greater"))

Welch Two Sample t-test

data:  father and mother
t = 2.7834, df = 153.252, p-value = 0.003028
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.9958999       Inf
sample estimates:
mean of x mean of y
 29.65919  27.20299
```

and t has the much lower value 2.7834.

To make the comparison, we rewrite the two t statistics as

$$t = \frac{\bar{x}_m - \bar{x}_f}{\sqrt{s^2_{x_m - x_f}/n}} \quad \textit{and} \quad t = \frac{\bar{x}_m - \bar{x}_f}{\sqrt{(s^2_{x_m} + s^2_{x_f})/n}}. \tag{20.3}$$

Notice that the numerators are the same. Thus, the change in the values for the t statistics must arise from a difference in the values in the denominator. Recall that the law of cosines for variance requires the variances in the expressions above as well as their correlation r to obtain

$$s^2_{x_m - x_f} = s^2_{x_m} + s^2_{x_f} - 2r s_{x_m} s_{x_f}.$$

Thus, if $r > 0$,

$$s^2_{x_m - x_f} < s^2_{x_m} + s^2_{x_f}$$

and the denominator for the expression (20.3) on the left is smaller than the expression on the right. Consequently, the t statistic is larger and the test is more powerful

Exercise 20.10. *Use the* R *output above and the values*

```
> cor(father,mother)
[1] 0.8252784
> sd(father);sd(mother)
[1] 5.700042
[1] 5.314777
```

to compute the two t statistics in (20.3).

Example 20.11 (pooled two-sample t-test)**.** *Sometimes, the two-sample procedure is based on the assumption of a common value σ^2 for the variance of the two-samples. In this case, we* **pool** *the data to compute an unbiased estimate for the variance:*

$$s_p^2 = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2\right).$$

Thus, we weight the variance from each of the two samples by the number of degrees of freedom. If we modify the t-statistics above to

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}}} = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

This indeed has the t-distribution with $n_X + n_Y - 2$ degrees of freedom. This is accomplished in R *by adding* `var.equal=TRUE` *command.*

```
> t.test(calcium,placebo,alternative = c("greater"),var.equal=TRUE)

Two Sample t-test

data:  calcium and placebo
t = 1.6341, df = 19, p-value = 0.05935
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.3066129        Inf
sample estimates:
 mean of x  mean of y
 5.0000000 -0.2727273
```

Note that this increases the number of degrees of freedom and lowers the P-value from 6.4% to 5.9%. A natural question to ask at this point is *How do we compare the mean of more that two groups?* As we shall soon learn, this leads us to a test procedure, analysis of variance, that is a generalization of the pooled two-sample procedure?

20.6 Summary of Tests of Significance

The procedures we have used to perform hypothesis tests are based on some quantity θ generally expressed as a value in a parameter space Θ. In setting the hypothesis, we partition the parameter space into two parts Θ_0 for the null hypothesis, H_0, and Θ_1 for the alternative, H_1. Our strategy is to look for generalizations of the Neyman-Pearson paradigm. For example, the Karlin-Rubin criterion provides a condition for one-sided tests that allows us to say the we have a uniformly most powerful test.

For two-sided tests, we look to the likelihood ratio approach. For this approach, we first maximize the likelihood $L(\theta|\mathbf{x})$ both over Θ_0 and over Θ and then compute the ratio $\Lambda(\mathbf{x})$. If the data, $\mathbf{x}$, lead to a ratio that is sufficiently small, then likelihood for all values of $\theta \in \Theta_0$ are less likely then some values in Θ_1. This leads us to reject the null hypothesis in favor of the alternative. If the number of observations is large, then we can approximate, under the null hypothesis, the distribution of the test-statistic $-2\ln\Lambda(\mathbf{x})$ with a χ^2 distribution. This leads to a critical region $C = \{-2\ln\Lambda(\mathbf{x}) \geq \tilde{k}_\alpha\}$ for an α-level test.

In practice, much of our inference is for population proportions and the population means. In these cases, we often reserve the test for those cases in which the central limit theorem applies and thus the estimates, the sample proportions and the sample means, have approximately a normal distribution. We summarize these procedures below.

20.6.1 General Guidelines

- Hypotheses are stated in terms of a *population parameter*.
- The null hypothesis H_0 is a statement that no effect is present.
- The alternative hypothesis H_1 is a statement that a parameter differs from its null value in a specific direction (one-sided alternative) or in either direction (two-sided alternative).
- A test statistic is designed to assess the strength of evidence against H_0.
- If a decision must be made, specify the significance level α.
- Assuming H_0 is true, the p-value is the probability that the test statistic would take a value as extreme or more extreme than the value observed.
- If the p-value is smaller than the significance level α, then H_0 is rejected and the data are said to be *statistically significant at level* α.

20.6.2 Test for Population Proportions

The design is based on Bernoulli trials . For this we have

- A fixed number of trials n.
- The outcome of each trial is independent of the other trials.
- Each trial has one of two outcomes **success** and **failure**.
- The probability of success p is the same for each trial.

This test statistic is the z-score. and thus is based the applicability of the central limit theorem on using the standard normal distribution. This procedure is considered valid is the sample is small ($< 10\%$) compared to the total population and both np_0 and $n(1 - p_0)$ is at least 10. Otherwise, use the binomial distribution directly for the test statistic.

The statistics $\hat{p}$ for a one-proportion procedure and $\hat{p}_1$, $\hat{p}_2$ for a two-sample procedure, is the appropriate proportions of success.

	null hypothesis		
sample proportions	one-sided	two-sided	test statistic
single proportion	$H_0 : p \geq p_0$ $H_0 : p \leq p_0$	$H_0 : p = p_0$	$z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$
two proportions	$H_0 : p_1 \geq p_2$ $H_0 : p_1 \leq p_2$	$H_0 : p_1 = p_2$	$z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$

The pooled sample proportion $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$ where x_i is the number of successes in the n_i Bernoulli trials from group i.

20.6.3 Test for Population Means

- Use the z-statistic when the standard deviations are known.
- Use the t-statistic when the standard deviations are computed from the data.

	null hypothesis	
t or z-procedure	one-sided	two-sided
single sample	$H_0 : \mu \leq \mu_0$ $H_0 : \mu \geq \mu_0$	$H_0 : \mu = \mu_0$
two samples	$H_0 : \mu_1 \leq \mu_2$ $H_0 : \mu_1 \geq \mu_2$	$H_0 : \mu_1 = \mu_2$

The test statistic

$$t = \frac{\text{estimate} - \text{parameter}}{\text{standard error}}.$$

The p-value is determined by the distribution of a random variable having a t distribution with the appropriate number of degrees of freedom. For one-sample and two-sample z procedures, replace the values s with σ and s_1 and s_2 with σ_1 and σ_2, respectively. Use the normal distribution for these tests.

t-procedure	parameter	estimate	standard error	degrees of freedom
one sample	μ	$\bar{x}$	$\frac{s}{\sqrt{n}}$	$n-1$
two sample	$\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	ν in equation (20.2)
pooled two sample	$\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2$	$s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	n_1+n_2-2

20.7 A Note on the Delta Method

For a one sample test hypothesizing a value for $g(\mu)$, we use the t statistic

$$t=\frac{g(\bar{x})-g(\mu_0)}{|g'(\bar{x})|s/\sqrt{n}}$$

and base the test on the t distribution with $n-1$ degrees of freedom.

For a test that compare a function of the mean of a two samples $g(\mu_X)$ and $g(\mu_Y)$ we can use the test statistic

$$t=\frac{g'(\bar{x})-g(\bar{y})}{\sqrt{\frac{(g'(\bar{x})s_X)^2}{n_X}+\frac{(g'(\bar{y})s_Y)^2}{n_Y}}}$$

The degrees of freedom ν can be computed from the Welch-Satterthwaite equation specialized to this circumstance.

$$\nu=\frac{(g(\bar{x})s_X)^2/n_X+(g'(\bar{y})s_Y)^2/n_Y)^2}{((g'(\bar{x})s_X)^2/n_X)^2/(n_X-1)+((g'(\bar{y})s_Y)^2/n_Y)^2/(n_Y-1)}.$$

20.8 The t Test as a Likelihood Ratio Test

Again, we begin with independent normal observations $X_1\dots,X_n$ with unknown mean μ and unknown variance σ^2. We show that the critical region $C=\{\mathbf{x};|T(\mathbf{x})|>t_{n-1,\alpha/2}\}$ is a consequence of the criterion given by a likelihood ratio test with significance level α.

The likelihood function

$$L(\mu,\sigma^2|\mathbf{x})=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\ln L(\mu,\sigma^2|\mathbf{x})=-\frac{n}{2}(\ln 2\pi+\ln\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2|\mathbf{x})=-\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

Thus, $\hat{\mu}=\bar{x}$.

$$\frac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2|\mathbf{x})=-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Thus,

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

For the hypothesis

$$H_0:\mu=\mu_0\quad\text{versus}\quad H_1:\mu\neq\mu_0,$$

the **likelihood ratio test**

$$\Lambda(x) = \frac{L(\mu_0, \hat{\sigma}_0^2|\mathbf{x})}{L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x})}$$

where the value

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

gives the maximum likelihood on the set $\mu = \mu_0$.

$$L(\mu_0, \hat{\sigma}_0^2|\mathbf{x}) = \frac{1}{(2\pi\hat{\sigma}_0^2)^{n/2}}\exp -\frac{1}{2\hat{\sigma}_0^2}\sum_{i=1}^{n}(x_i - \mu_0)^2 = \frac{1}{(2\pi\hat{\sigma}_0^2)^{n/2}}\exp -\frac{2}{n},$$

$$L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x})) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}}\exp -\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}}, \exp -\frac{2}{n},$$

and the likelihood ratio is

$$\Lambda(\mathbf{x}) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left(\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{-n/2}$$

The critical region $\lambda(\mathbf{x}) \leq \lambda_0$ is equivalent to the fraction in parenthesis above being sufficiently large. In other words for some value c,

$$\begin{aligned} c &\leq \frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x}) + (\bar{x} - \mu_0)\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \mu_0) + \sum_{i=1}^{n}(\bar{x} - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

In a now familiar strategy, we have added and subtracted $\bar{x}$ to decompose the variation. Continuing we find that

$$(c-1)(n-1) \leq \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)} = \frac{(\bar{x} - \mu_0)^2}{s^2/n}$$

or

$$(c-1)(n-1) \leq T(\mathbf{x})^2 \tag{20.4}$$

where

$$T(\mathbf{x}) = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and s is the square root of the *unbiased* estimator of the variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Taking square roots in (20.4), we have the critical region

$$C = \left\{\mathbf{x}; \sqrt{(c-1)(n-1)} \leq |T(\mathbf{x})|\right\}$$

Now take $\sqrt{(c-1)(n-1)} = t_{n-1,\alpha/2}$.

20.9 Non-parametric alternatives

The strategy for hypothesis testing is to choose a test statistic with high power and to determine its distribution under the null hypothesis. We then compute the value of the test statistic for our data and make an assessment. For the t test, thus assessment is based on the t distribution. This, in turn, relies on the ability to say that the distributions of sample means are nearly normally distributed. This is typically assured upon appeal to the central limit theorem. In some circumstances, we then choose a significance level α and then the decision to reject the null hypotheses is based on relating the test statistics to this critical value, rejecting if the test statistic is more extreme than this standard. In other circumstances, we compute the p-value and use that to indicate the strength of the evidence against the null hypothesis. We then follow up by deciding to reject the null hypothesis if the p-value is below the significance level α.

20.9.1 Permutation Test

In the example of a test for the value of added calcium in the diet to lower blood pressure, the boxplot showed an outlier in the blood pressure of one individual in the group that received a calcium supplement. As a consequence, the assumption that the central limit theorem holds for the distribution of the sample means is in doubt. These data have a difference in sample means:

```
> (diffdata<-mean(calcium)-mean(placebo))
[1] 5.272727
```

When the central limit theorem applies, we can assess this difference by dividing by the standard error of the sample means and use the value of a t statistic with the appropriate number of degrees of freedom.

Permutation tests present an alternative to finding the sampling distribution for any test statistic, here the difference in sample means. The procedures begin by asking, *If the null hypothesis is true, then what shuffles of the data are consistent with this statement?*

In this case, we can think of the null hypothesis as stating that all of the observations for blood pressure are actually derived from a single group and thus the assignment of an individual to a group, either placebo or calcium, should have no effect on the outcome. For the two groups, sizes n_1 and n_2, we have

$$\binom{n_1+n_2}{n_1}$$

ways to divide the entire sample into two groups, retaining the sizes found in the observations that constitute our data. Each of these choices provides an additional value for the difference, under the null hypothesis, of sample means. These can be used to create an empirical sampling distribution. This methodology is feasible if n_1 and n_2 are not too large. For larger sample sizes, we can also use a **resampling technique**, similar to the bootstrap, to *randomly* create the two groups and compute the difference in the means of these randomly chosen resampled groups.

```
> bp<-c(placebo,calcium)
> diff<-rep(0,10000)
> groups<-c(rep(1,length(calcium)),
  rep(2,length(placebo)))
> for(i in 1:10000){groupperm<-sample(groups);
   diff[i]<-mean(bp[groupperm==1]
   -mean(bp[groupperm==2]))}
```

Because the alternative is *greater than*, the p-value (which will vary a bit from one simulation to the next) is the fraction of the simulations greater than what is found in the data. (See Figure 20.3.)

```
> (pvalue<-length(diff[diff>diffdata])/length(diff))
[1] 0.0556
```

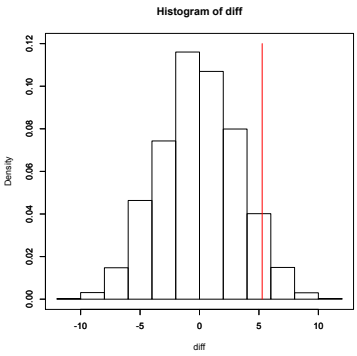


Figure 20.4: Histogram of mean differences for 10,000 simulated samples. The p-value, 0.0556, is the area to the right of the vertical red line.

This is slightly lower than the p-value 0.06442 found using the two-sample t-test.

Exercise 20.12. *Create a permutation test for a matched pair procedure and apply it to the data set vitamin C in the wheat soy blend, (*Hint; *Under the null hypothesis the difference in the measurements, in either order have the same distribution that is symmetric about 0. The procedure uses independent Bernoulli random variables to choose the order of subtraction.)*

When the assumption of the normal distribution cannot be said to hold in a two-sample t-procedure, another option is to use a test that does not depend on the numerical values of the observations but rather on the *ranks* of the data. Because this test is based on the ranks of the data, we cannot base the hypothesis test on the means of the data, but rather on a parameter that uses only the ranks of the data. For these **non-parametric tests** the hypotheses are often stated in terms of medians.

20.9.2 Mann-Whitney or Wilcoxon Rank Sum Test

We will explain this procedure more carefully in the case of the data on the lifetime of wildtype and transgenic mosquitoes. In this case our data are $x_1, x_2, \ldots, x_{n_1}$ are the lifetimes in days for the wildtype mosquitoes and $y_1, y_2, \ldots, y_{n_2}$ are the lifetimes in days for the transgenic mosquitoes.

For the **Wilcoxon signed-rank test**, the hypothesis is based on the median values m_1 and m_2. Here

- m_1 is the median lifetime in days for wildtype mosquitoes, and
- m_2 is the median lifetime in days for transgenic mosquitoes.

$$H_0 : m_2 \geq m_1 \quad \text{versus} \quad H_1 : m_2 < m_1.$$

The following identity will be useful in our discussion.

Exercise 20.13. *The sums of the first m integers*

$$\sum_{j=1}^{m} j = 1 + 2 + \cdots + m = \frac{m(m+1)}{2}$$

Let's look at a small pilot data set to get a sense of the procedure. The values are the lifespan in days.

```
> wildtype
[1] 31 36 18 11 33  9 34
> transgenic
[1]  3  8 21 24 25
```

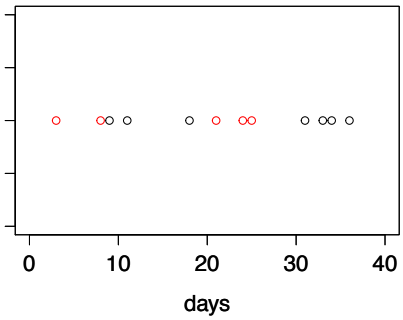


Figure 20.5: Life span in days for 5 transgenic (in red) and 7 wildtype (in black) mosquitoes.

From these date, we see that the ranks

- transgenic - 1 2 6 7 8
- wildtype - 3 4 5 9 10 11 12

The strategy for the test to to see if there is a significant difference in the ranks of the data. The basic statistic is the sum of the ranks of one of the samples. For the transgenic mosquitoes, this sum is

$$R_1 = 1 + 2 + 6 + 7 + 8 = 24$$

for $n_1 = 5$ observations

We can now compare this sum of ranking to all $\binom{12}{5} = 792$ possible rankings of the data. This is accomplished using the `wilcox.test` command in R.

```
> wilcox.test(transgenic,wildtype,alternative=c("less"))

Wilcoxon rank sum test

data:  transgenic and wildtype
W = 9, p-value = 0.101
alternative hypothesis: true location shift is less than 0
```

Thus, the 10.1% of the ranks below the given value of 24 give us the p-value. This small amount of data gives a hint that the transgenic mosquito may have a shorter lifespan. The U statistic is related to the sum of the ranks by subtracting the minimum possible value as shown in Exercise 20.6.

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}.$$

R uses another variant W of the $R_!$ statistic.

If the entire data set is used, then we cannot carry out all of the comparisons without a long computational time. However, we have a version of the central limit theorem that gives the mean and standard deviation of the R_1, U_1 or W-statistic. Thus, we can use the normal distribution to determine a p-value. As with the binomial distribution, R uses a continuity correction to deal with the fact that the test statistic W is a discrete random variable.

```
> wilcox.test(transgenic,wildtype,alternative=c("less"))

Wilcoxon rank sum test with continuity correction

data:  transgenic and wildtype
W = 3549.5, p-value = 0.0143
alternative hypothesis: true location shift is less than 0
```

Notice the value $W = 3549.5$ is not an integer. This is a result of the fact that ties are resolved by giving fractional values to ties. For example, if the third and fourth values are equal, they are both given the rank 3 1/2. If the seventh, eighth, and ninth values are equal. they are all given the rank (7+8+9)/3 = 8.

Exercise 20.14. *Define R_2 to be the sum of the ranks for the n_2 observations in the second sample and set*

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}.$$

Then, $U_1 + U_2 = n_1 n_2$.

20.9.3 Wilcoxon Signed-Rank Test

For the **Wilcoxon signed-rank test**, the hypothesis is based on the median values m_x and m_y for an experimental procedure in which pairs are matched. This gives an alternative to the measurement of the amount of vitamin C in wheat soy blend at the factory and 5 months later in Haiti. Our data are

$$x_1, x_2, \ldots, x_n$$

are the measurements of vitamin C content of the wheat soy blend at the factory and

$$y_1, y_2, \ldots, y_n$$

are the measurements of vitamin C content of the wheat soy blend in Haiti.

For the hypothesis test to see if vitamin C content decreases due to shipping and shelf time, set

- m_F is the median vitamin C content of the wheat soy blend at the factory and
- m_H is the median vitamin C content of the wheat soy blend in Haiti.

To perform the test

$$H_0 : m_F \le m_H \quad \text{versus} \quad H_1 : m_F > m_H, ,$$

we use a test statistic based on both the sign of the difference $y_i - x_i$ in the paired observations and in the ranks of $|y_i - x_i|$ Here is the R command and output. Note the choice `paired=TRUE` for the signed-rank test.

```
> wilcox.test(factory, haiti, alternative = c("greater"),paired=TRUE)

Wilcoxon signed rank test with continuity correction

data:  factory and haiti
V = 341, p-value = 0.0001341
alternative hypothesis: true location shift is greater than 0

Warning message:
In wilcox.test.default(factory, haiti, alternative = c("greater"),  :
  cannot compute exact p-value with ties
```

20.10 Answers to Selected Exercises

20.2. Recall that for $Z_1, Z_2, \dots Z_n$, independent standard normal random variables with standard deviation s_Z,

$$\tilde{T} = \frac{\sqrt{n}\bar{Z} - a}{s_Z}$$

has a t distribution with $n-1$ degrees of freedom and non-centrality parameter a.

If we reproduce the calculation we made in determining power for a one-sample two-sided z-test, then, in the denominator, we add and subtract the mean μ to obtain

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{(\bar{X} - \mu) + (\mu - \mu_0)}{s/\sqrt{n}}$$

If the X_i have common standard deviation σ, we standardize the variables, writing

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

Thus, $s_Z = s/\sigma$ is the standard deviation of the Z_i and upon dividing each term by σ,

$$\begin{aligned}\frac{(\bar{X} - \mu) + (\mu - \mu_0)}{s/\sqrt{n}} &= \frac{\sqrt{n}((\bar{X} - \mu)/\sigma + (\mu - \mu_0)/\sigma)}{s/\sigma} \\ &= \frac{\sqrt{n}\bar{Z} - \sqrt{n}(\mu_0 - \mu)/\sigma}{s_Z}\end{aligned}$$

which has t distribution with non-centrality parameter $a = \sqrt{n}(\mu - \mu_0)/\sigma$.

20.3. To plot the power function, π, we first enter the data.

```
> radon<-c(91.9,97.8,111.4,122.3,105.4,95.0,
  103.8,99.6,96.6,119.3,104.8,101.7)
> mu0<-105
> mu<-seq(95,115,0.01)
> a<-(mu0-mu)/(sd(radon)/sqrt(length(radon)))
> tstar<-qt(0.975,11)
> pi<-1-(pt(tstar,11,a)-pt(-tstar,11,a))
> plot(mu,pi,type="l")
```

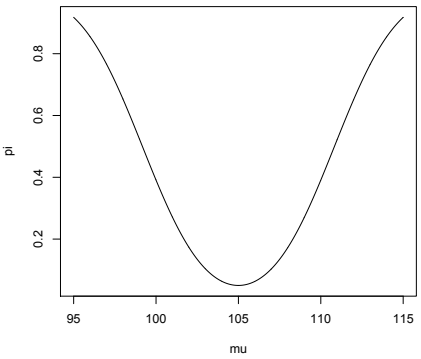


Figure 20.6: Power curve for radon detector.

20.4. Recall that the receiver operating characteristic is a plot of α the significance level versus the power.

```
> alpha<-seq(0,1,0.01) delta<-5
> n<-6; talpha<-qt(1-alpha/2,n-1); a<-delta/(sd(radon)/sqrt(n-1))
> power6<-1-(pt(talpha,n-1,a)-pt(-talpha,n-1,a))
> plot(alpha,power6,type="l",xlim=c(0,1),ylim=c(0,1),
  xlab=c("significance"),ylab=c("power"))
> par(new=TRUE)
> n<-12; talpha<-qt(1-alpha/2,n-1); a<-delta/(sd(radon)/sqrt(n-1))
> power12<-1-(pt(talpha,n-1,a)-pt(-talpha,n-1,a))
> par(new=TRUE)
> plot(alpha,power12,type="l",xlim=c(0,1),ylim=c(0,1),xlab=c(""),ylab=c(""),col="red")
> n<-24; talpha<-qt(1-alpha/2,n-1); a<-delta/(sd(radon)/sqrt(n-1))
> power24<-1-(pt(talpha,n-1,a)-pt(-talpha,n-1,a))
> par(new=TRUE)
> plot(alpha,power24,type="l",xlim=c(0,1),ylim=c(0,1),xlab=c(""),ylab=c(""),col="blue")
```

We compare the ROCs for low significance levels by using the `head` command. Notice how the power increases with sample size.

```
> head(data.frame(alpha,power6,power12,power24))
  alpha     power6   power12   power24
1  0.00 0.00000000 0.0000000 0.0000000
2  0.01 0.04304033 0.1435099 0.4179953
3  0.02 0.07812389 0.2199330 0.5299071
4  0.03 0.10924263 0.2773645 0.5988122
5  0.04 0.13759358 0.3241511 0.6480282
6  0.05 0.16380961 0.3638693 0.6858441
```

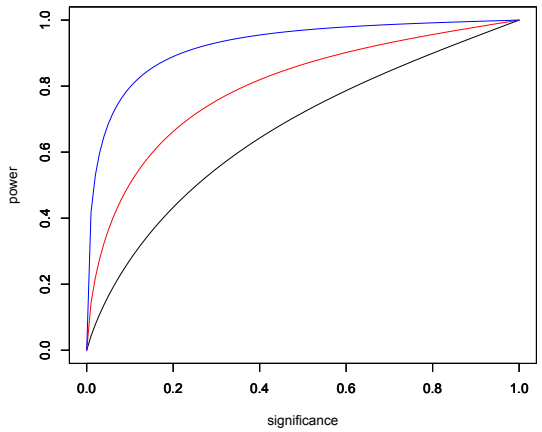


Figure 20.7: Receiver operating characteristic for the radon detector with 6 (black), 12 (red) and 24 (blue) observations and a null $\mu_0 - 105$ and an alternative $\mu = 110$.

20.8. This means that we can reject the null hypothesis that transgenic and wildtype mosquitoes have the seem mean lifetime. For a one-sided test, the p-value is 0.01699/2 = 0.008549.

20.9. By the correspondence between two-sided hypothesis tests and confidence intervals, the fact that 0 is not in the confidence interval, indicates that the test *is* significant at the 2% level and thus the p-value < 0.02. Notice that the output shows a p-value of 0.01699. For a one-sided test, we know that the p-value is half that of a two-sided test and thus below 0.01. Notice that the R give a p-value of 0.008497 for a one-sided test.

20.11. The numerator in the test statistics

$$\bar{x}_m - \bar{x}_f = 29.65919 - 27.20299 = 2.4562.$$

For the unpaired two-sample test.

$$s^2_{x_m} + s^2_{x_f} = 5.700042^2 + 5.314777^2 = 60.73733$$

and

$$t = \frac{\bar{x}_m - \bar{x}_f}{\sqrt{s^2_{x_m - x_f}/n}} = \frac{2.4562}{\sqrt{60.73733/78}} = 2.783448,$$

matching the R output. For the paired two-sample test

$$s^2_{x_m - x_f} = s^2_{x_m} + s^2_{x_f} - 2rs_{x_m}s_{x_f} = 5.700042^2 + 5.314777^2 - 2(0.3571538)(5.700042)(5.314777) = 10.73462.$$

and

$$t = \frac{\bar{x}_m - \bar{x}_f}{\sqrt{s^2_{x_m - x_f}/n}} = \frac{2.4562}{\sqrt{10.7346278/78}} = 6.62091,$$

again matching the output.

20.12. Under the null hypothesis, the two sets of observations, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ have the same distribution. Thus, $\Delta_k = Y_k - X_k$, $k = 1, \ldots, n$ are independent and their distribution is symmetric about zero. Thus, we can randomly take the sign $\pm\Delta_k$ as the basis for the permutation test. Here is the R code. We use one million simulations to be able to determine a small p-value

```
> delta<-factory-haiti
> deltamean<-mean(delta)
> deltasim<-numeric(1000000)
> for (i in 1:1000000){deltasim[i]<-mean(rbinom(length(delta),1,0.5)*delta)}
> length(deltasim[deltasim>deltamean])/1000000
[1] 4e-05
```

This yields a permutation test p-value of 4×10^{-5}. The t-test procedure gave a very similar p-value, 3.745×10^{-5}

20.13. We prove this using mathematical induction. For the case $m = 1$, we have $1 = \frac{1(1+1)}{2}$ and the identity holds true. Now assume that the identity holds for $m = k$. We then check that it also holds for $m = k + 1$

$$1 + 2 + \cdots + k + (k+1) = \frac{k(k+1)}{2} + (k+1) = \left(\frac{k}{2} + 1\right)(k+1) = \frac{k+2}{2}(k+1) = \frac{(k+1)(k+2)}{2}.$$

So, by the principle of mathematical induction, we have the identity for all non-negative integers.

20.14. By Exercise 20.10, the sum of the ranks

$$R_1 + R_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

Thus,

$$\begin{aligned} U_1 + U_2 &= \left(R_1 - \frac{n_1(n_1+1)}{2}\right) + \left(R_2 - \frac{n_2(n_2+1)}{2}\right) \\ &= \frac{(n_1+n_2)(n_1+n_2+1)}{2} - \frac{n_1(n_1+1)}{2} - \frac{n_2(n_2+1)}{2} \\ &= \frac{1}{2}(n_1(n_1+1) + n_1n_2 + n_2(n_2+1) + n_1n_2 - n_1(n_1+1) - n_2(n_2+1)) = n_1n_2. \end{aligned}$$

Topic 21

Goodness of Fit

The object of this paper is to investigate a criterion of the probability on any theory of an observed system of errors, and to apply it to the determination of goodness of fit. - Karl Pearson. 1900, On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling, *Philosophical Magazine*

21.1 Fit of a Distribution

Goodness of fit tests examine the case of a sequence of independent observations each of which can have 1 of k possible categories. For example, each of us has one of 4 possible of blood types, O, A, B, and AB. The local blood bank has good information from a national database of the fraction of individuals having each blood type,

$$\pi_O, \pi_A, \pi_B, \text{ and } \pi_{AB}.$$

The actual fraction p_O, p_A, p_B, and p_{AB} of these blood types in the community for a given blood bank may be different than what is seen in the national database. As a consequence, the local blood bank may choose to alter its distribution of blood supply to more accurately reflect local conditions.

To place this assessment strategy in terms of formal hypothesis testing, let $\pi = (\pi_1, \ldots, \pi_k)$ be postulated values of the probability

$$P_\pi\{\text{individual is a member of } i\text{-th category}\} = \pi_i$$

and let $\mathbf{p} = (p_1, \ldots, p_k)$ denote the possible states of nature. Then, the parameter space is

$$\Theta = \{\mathbf{p} = (p_1, \ldots, p_k); p_i \geq 0 \text{ for all } i = 1, \ldots, k, \sum_{i=1}^{k} p_i = 1\}.$$

This parameter space has $k - 1$ free parameters. Once these are chosen, the remaining parameter value is determined by the requirement that the sum of the p_i equals 1. Thus, $\dim(\Theta) = k - 1$.

The hypothesis is

$$H_0 : p_i = \pi_i, \text{ for all } i = 1, \ldots, k \quad \text{versus} \quad H_1 : p_i \neq \pi_i, \text{ for some } i = 1, \ldots, k. \tag{21.1}$$

The parameter space for the null hypothesis is a single point $\pi = (\pi_1, \ldots, \pi_k)$. Thus, $\dim(\Theta_0) = 0$. Consequently, the likelihood ratio test will have a chi-square test statistic with $\dim(\Theta) - \dim(\Theta_0) = k - 1$ degrees of freedom. The data $\mathbf{x} = (x_1, \ldots, x_n)$ are the categories for each of the n observations.

Let's use the likelihood ratio criterion to create a test for the distribution of human blood types in a given population. For the data

$$\mathbf{x} = \{O, B, O, A, A, A, A, A, O, AB\}$$

for the blood types of tested individuals, then, in the case of independent observations, the likelihood is

$$L(\mathbf{p}|\mathbf{x}) = p_O \cdot p_B \cdot p_O \cdot p_A \cdot p_A \cdot p_A \cdot p_A \cdot p_A \cdot p_O \cdot p_{AB} = p_O^3 p_A^5 p_B p_{AB}.$$

Notice that the likelihood has a factor of p_i whenever an observation take on the value i. In other words, if we summarize the data using

$$n_i = \#\{\text{observations from category } i\}$$

to create $\mathbf{n} = (n_1, n_2, \cdots, n_k)$, a vector that records the number of observations in each category, then, the **likelihood function**

$$L(\mathbf{p}|\mathbf{n}) = p_1^{n_1} \cdots p_k^{n_k}. \tag{21.2}$$

The **likelihood ratio** is the ratio of the maximum value of the likelihood under the null hypothesis and the maximum likelihood for any parameter value. In this case, the numerator is the likelihood evaluated at π.

$$\Lambda(\mathbf{n}) = \frac{L(\pi|\mathbf{n})}{L(\hat{\mathbf{p}}|\mathbf{n})} = \frac{\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_k^{n_k}}{\hat{p}_1^{n_1}\hat{p}_2^{n_2}\cdots\hat{p}_k^{n_k}} = \left(\frac{\pi_1}{\hat{p}_1}\right)^{n_1} \cdots \left(\frac{\pi_k}{\hat{p}_k}\right)^{n_k}. \tag{21.3}$$

To find the maximum likelihood estimator $\hat{p}$, we, as usual, begin by taking the logarithm in (21.2),

$$\ln L(\mathbf{p}|\mathbf{n}) = \sum_{i=1}^{k} n_i \ln p_i.$$

Because not every set of values for p_i is admissible, we cannot just take derivatives, set them equal to 0 and solve. Indeed, we must find a maximum under the constraint

$$s(\mathbf{p}) = \sum_{i=1}^{k} p_i = 1.$$

The maximization problem is now stated in terms of the method of **Lagrange multipliers**. This method tells us that at the maximum likelihood estimator $(\hat{p}_1, \ldots, \hat{p}_k)$, the gradient of $\ln L(\mathbf{p}|\mathbf{n})$ is proportional to the gradient of the constraint $s(\mathbf{p})$. To explain this briefly, recall that the gradient of a function is a vector that is perpendicular to a level set of that function. In this case,

$$\nabla_{\hat{\mathbf{p}}} s(\mathbf{p}) \quad \text{is perpendicular to the level set} \quad \{\mathbf{p}; s(\mathbf{p}) = 1\}.$$

Now imagine walking along the set of parameter values of $\mathbf{p}$ given by the constraint $s(\mathbf{p}) = 1$, keeping track of the values of the function $\ln L(\mathbf{p}|\mathbf{n})$. If the walk takes us from a value of this function below ℓ_0 to values above ℓ_0 then (See Figure 21.1.), the level surfaces

$$\{\mathbf{p}; s(\mathbf{p}) = 1\}$$

and

$$\{\ln L(\mathbf{p}|\mathbf{n} = \ell_0\}$$

intersect. Consequently, the gradients

$$\nabla_{\hat{\mathbf{p}}} s(\mathbf{p}) \quad \text{and} \quad \nabla_{\hat{\mathbf{p}}} \ln L(\mathbf{p}|\mathbf{n})$$

point in different directions on the intersection of these two surfaces. At a local maximum or minimum of the log-likelihood function, the level surfaces are tangent and the two gradients are parallel. In other words, the these two gradients vectors are related by a constant of proportionality, λ, known as the **Lagrange multiplier**. Consequently, at extreme values,

$$\nabla_{\mathbf{p}} \ln L(\hat{\mathbf{p}}|\mathbf{n}) = \lambda \nabla_{\hat{\mathbf{p}}} s(\mathbf{p}).$$

$$\left(\frac{\partial}{\partial p_1} \ln L(\hat{\mathbf{p}}|\mathbf{n}), \ldots, \frac{\partial}{\partial p_k} \ln L(\hat{\mathbf{p}}|\mathbf{n})\right) = \lambda \left(\frac{\partial}{\partial p_1} s(\mathbf{p}), \ldots, \frac{\partial}{\partial p_k} s(\mathbf{p})\right)$$

$$\left(\frac{n_1}{\hat{p}_1}, \ldots, \frac{n_k}{\hat{p}_k}\right) = \lambda(1, \ldots, 1)$$

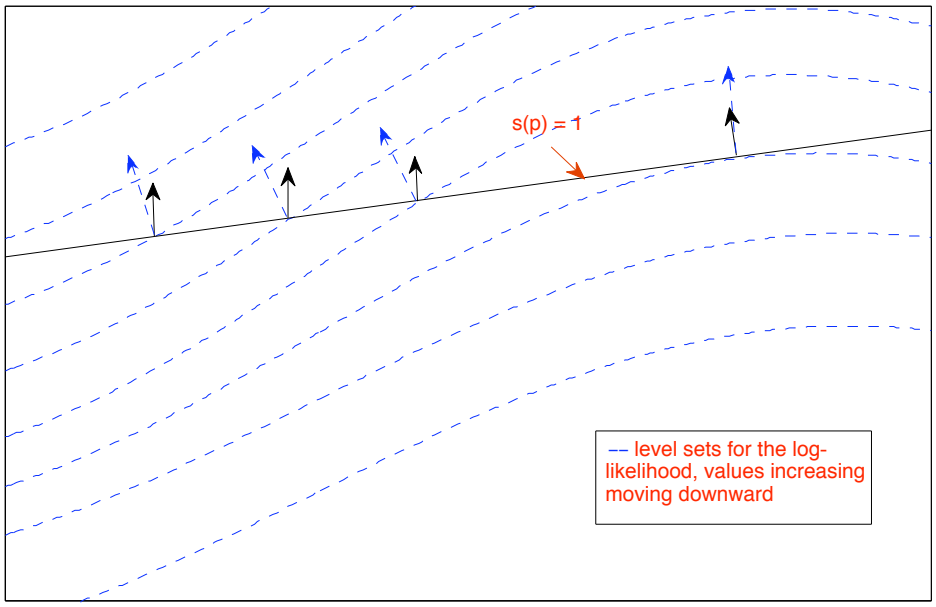


Figure 21.1: Lagrange multipliers Level sets of the log-likelihood function.shown in dashed blue. The level set $\{s(\mathbf{p}) = 1\}$ shown in black. The gradients for the log-likelihood function and the constraint are indicated by dashed blue and black arrows, respectively. At the maximum, these two arrows are parallel. Their ratio λ is called the Lagrange multiplier. If we view the blue dashed lines as elevation contour lines and the black line as a trail, crossing contour line indicates walking either up or down hill. When the trail reaches its highest elevation, the trail is tangent to a contour line and the gradient for the hill is perpendicular to the trail.

Each of the components of the two vectors must be equal. In other words,

$$\frac{n_i}{\hat{p}_i} = \lambda, \quad n_i = \lambda \hat{p}_i \quad \text{for all } i = 1, \dots, k. \tag{21.4}$$

Now sum this equality for all values of i and use the constraint $s(\mathbf{p}) = 1$ to obtain

$$n = \sum_{i=1}^{k} n_i = \lambda \sum_{i=1}^{k} \hat{p}_i = \lambda s(\hat{\mathbf{p}}) = \lambda.$$

Returning to (21.4), we have that

$$\frac{n_1}{\hat{p}_i} = n \quad \text{and} \quad \hat{p}_i = \frac{n_i}{n}. \tag{21.5}$$

This is the answer we would guess - the estimate for p_i is the fraction of observations in category i. Thus, for the introductory example,

$$\hat{p}_O = \frac{3}{10}, \quad \hat{p}_A = \frac{5}{10}, \quad \hat{p}_B = \frac{1}{10}, \quad \text{and} \quad \hat{p}_O = \frac{1}{10}.$$

Next, we substitute the maximum likelihood estimates $\hat{p}_i = n_i/n$ into the likelihood ratio (21.3) to obtain

$$\Lambda(\mathbf{n}) = \frac{L(\pi|\mathbf{n})}{L(\hat{\mathbf{p}}|\mathbf{n})} = \left(\frac{\pi_1}{n_1/n}\right)^{n_1} \cdots \left(\frac{\pi_k}{n_k/n}\right)^{n_k} = \left(\frac{n\pi_1}{n_1}\right)^{n_1} \cdots \left(\frac{n\pi_k}{n_k}\right)^{n_k}. \tag{21.6}$$

Recall that we reject the null hypothesis if this ratio is too low, i.e, the maximum likelihood under the null hypothesis is sufficiently smaller than the maximum likelihood under the alternative hypothesis.

Let's review the process. the random variables $X_1, X_2, \dots, X_n$ are independent, taking values in one of k categories each having distribution π. In the example, we have 4 categories, namely the common blood types O, A, B, and AB. Next, we organize the data into

$$N_i = \#\{j; X_j = i\},$$

the number of observations in category i. Next, create the vector $\mathbf{N} = (N_1, \ldots, N_k)$ to be the vector of observed number of occurrences for each category i. In the example we have the vector (3,5,1,1) for the number of occurrences of the 4 blood types.

When the null hypothesis holds true, $-2\ln\Lambda(\mathbf{N})$ has approximately a χ^2_{k-1} distribution. Using (21.6) we obtain the the likelihood ratio test statistic

$$-2\ln\Lambda(\mathbf{N}) = -2\sum_{i=1}^{k} N_i \ln\frac{n\pi_i}{N_i} = 2\sum_{i=1}^{k} N_i \ln\frac{N_i}{n\pi_i}$$

The last equality uses the identity $\ln(1/x) = -\ln x$ for the logarithm of reciprocals.

The test statistic $-2\ln\Lambda_n(\mathbf{n})$ is generally rewritten using the notation $O_i = n_i$ for the number of **observed** occurrences of i and $E_i = n\pi_i$ for the number of **expected** occurrences of i as given by H_0. Then, we can write the test statistic as

$$-2\ln\Lambda_n(\mathbf{O}) = 2\sum_{i=1}^{k} O_i \ln\frac{O_i}{E_i} \tag{21.7}$$

This is called the G^2 **test statistic**. Thus, we can perform our inference on the hypothesis (21.1) by evaluating G^2. The p-value will be the probability that the a χ^2_{k-1} random variable takes a value greater than $-2\ln\Lambda_n(\mathbf{O})$

The traditional method for a test of goodness of fit, we use, instead of the G^2 statistic, the chi-square statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}. \tag{21.8}$$

Thus, large values for χ^2 result from large differences between the observed values O_i and the expected values E_i. Consequently. large values of χ^2 is evidence against the null hypothesis.

The χ^2 test statistic was introduced between 1895 and 1900 by Karl Pearson and consequently has been in use for longer that the concept of likelihood ratio tests. Indeed, R output call the test `Pearson's Chi-squared test` in the case of contingency tables, our next topic.

We establish the relation between (21.7) and (21.8), through the following two exercises.

Exercise 21.1. *Define*

$$\delta_i = \frac{O_i - E_i}{E_i} = \frac{O_i}{E_i} - 1.$$

Show that

$$\sum_{i=1}^{k} E_i\delta_i = 0 \quad \textit{and} \quad E_i(1+\delta_i) = O_i.$$

Exercise 21.2. *Show the relationship between the G^2 and χ^2 statistics in (21.7) and (21.8) by applying the quadratic Taylor polynomial approximation for the natural logarithm,*

$$\ln(1+\delta_i) \approx \delta_i - \frac{1}{2}\delta_i^2$$

and keeping terms up to the square of δ_i

To compute either the G^2 or χ^2 statistic, we begin by creating a table.

i	1	2	$\cdots$	k
observed	O_1	O_2	$\cdots$	O_k
expected	E_1	E_2	$\cdots$	E_k

We show this procedure using a larger data set on blood types.

Example 21.3. *The Red Cross recommends that a blood bank maintains 44% blood type O, 42% blood type A, 10% blood type B, 4% blood type AB. You suspect that the distribution of blood types in Tucson is not the same as the recommendation. In this case, the hypothesis is*

$$H_0 : p_O = 0.44, p_A = 0.42, p_B = 0.10, p_{AB} = 0.04 \quad \textit{versus} \quad H_1 : \textit{at least one } p_i \textit{ is unequal to the given values}$$

Based on 400 observations, we observe 228 for type O, 124 for type A, 40 for type B and 8 for type AB by computing $400 \times p_i$ using the values in H_0. This gives the table

type	O	A	B	AB
observed	228	124	40	8
expected	176	168	40	16

Using this table, we can compute the value of either (21.7) and (21.8). The `chisq.test` *command in* R *uses (21.8). The program computes the expected number of observations.*

```
> chisq.test(c(228,124,40,8),p=c(0.44,0.42,0.10,0.04))

Chi-squared test for given probabilities

data:  c(228, 124, 40, 8)
X-squared = 30.8874, df = 3, p-value = 8.977e-07
```

The number of degrees of freedom is $4 - 1 = 3$. Note that the p-value is very low and so the distribution of blood types in Tucson is very unlikely to be the same as the national distribution. We can also perform the test using the G^2-statistic in (21.7):

```
> O<-c(228,124,40,8)
> E<-sum(O)*c(0.44,0.42,0.10,0.04)
> G2stat<-2*sum(O*log(O/E))
> G2stat
[1] 31.63731
> 1-pchisq(G2stat,3)
[1] 6.240417e-07
```

One way to visualize the discrepancies from from the null hypothesis is to display them with a **hanging chi-gram**. *This plots category i with a bar of height of the* **standardized residuals** *(also known as* **Pearson residuals***).*

$$\frac{O_i - E_i}{\sqrt{E_i}}. \tag{21.9}$$

Note that these values can be either positive or negative.

```
> resid<-(O-E)/sqrt(E)
> barplot(resid, names.arg=c("O","A","B","AB"),
  xlab="chigram for blood donation data")
```

Example 21.4. *Is sudden infant death syndrome seasonal (SIDS)? Here we are hypothesizing that 1/4 of each of the occurrences of sudden infant death syndrome take place in the spring, summer, fall, and winter. Let p_1, p_2, p_3, and p_4 be the respective probabilities for these events. Then the hypothesis takes the form*

$$H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4}, \quad \textit{versus} \quad H_1 : \textit{at least one } p_i \textit{ is unequal to } \frac{1}{4}.$$

To test this hypothesis, public health officials from King County, Washington, collect data on $n = 322$ cases, finding

$$n_1 = 78, \quad n_2 = 71, \quad n_3 = 87, \quad n_4 = 86$$

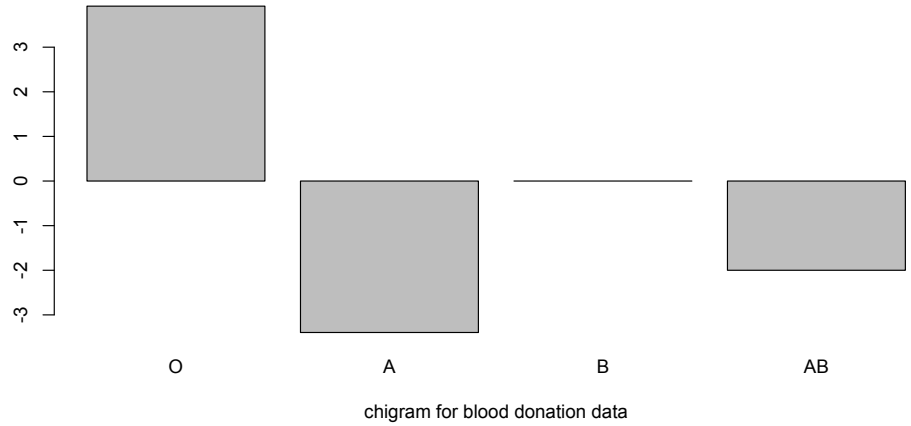


Figure 21.2: The heights of the bars for each category are the standardized residuals (21.9). Thus, blood type O is overrepresented and types A and AB are underrepresented compare to the expectations under the null hypothesis.s

for deaths in the spring, summer, fall, and winter, respectively. Thus, we find more occurrences of SIDS in the fall and winter. Is this difference statistical significant or are these difference better explained by chance fluctuations?

We carry out the chi square test. In this case, each of the 4 categories is equally probable. Because this is the default value in R*, we need not include this in the command.*

```
> chisq.test(c(78,71,87,86))

Chi-squared test for given probabilities

data:  c(78, 71, 87, 86)
X-squared = 2.0994, df = 3, p-value = 0.552
```

This p-value is much to high to reject the null hypothesis.

Example 21.5 (Hardy-Weinberg equilibrium)**.** *As we saw with Gregor Mendel's pea experiments, the two-allele Hardy-Weinberg principle states that after two generations of random mating the genotypic frequencies can be represented by a binomial distribution. So, if a population is segregating for two alleles A_1 and A_2 at an autosomal locus with frequencies p_1 and p_2, then random mating would give a proportion*

$$p_{11} = p_1^2 \text{ for the } A_1A_1 \text{ genotype}, \; p_{12} = 2p_1p_2 \text{ for the } A_1A_2 \text{ genotype, and } p_{22} = p_2^2 \text{ for the } A_2A_2 \text{ genotype}. \tag{21.10}$$

Then, with both genes in the homozygous genotype and half the genes in the heterozygous genotype, we find that

$$p_1 = p_{11} + \frac{1}{2}p_{12} \quad p_2 = p_{22} + \frac{1}{2}p_{12}. \tag{21.11}$$

Our parameter space $\Theta = \{(p_{11}, p_{12}, p_{22}); p_{11} + p_{12} + p_{22} = 1\}$ is 2 dimensional. Θ_0, the parameter space for the null hypothesis, are those values p_1, p_2 that satisfy (21.11). With the choice of p_1, the value p_2 is determined because $p_1 + p_2 = 1$. Thus, $\dim(\Theta_0) = 1$. Consequently, the chi-square test statistic will have 2-1=1 degree of freedom. Another way to see this is the following.

McDonald et al. (1996) examined variation at the CVJ5 locus in the American oyster, Crassostrea virginica*. There were two alleles, L and S, and the genotype frequencies in Panacea, Florida were 14 LL, 21 LS, and 25 SS. So,*

$$\hat{p}_{11} = \frac{14}{60}, \quad \hat{p}_{12} = \frac{21}{60}, \quad \hat{p}_{22} = \frac{25}{60}.$$

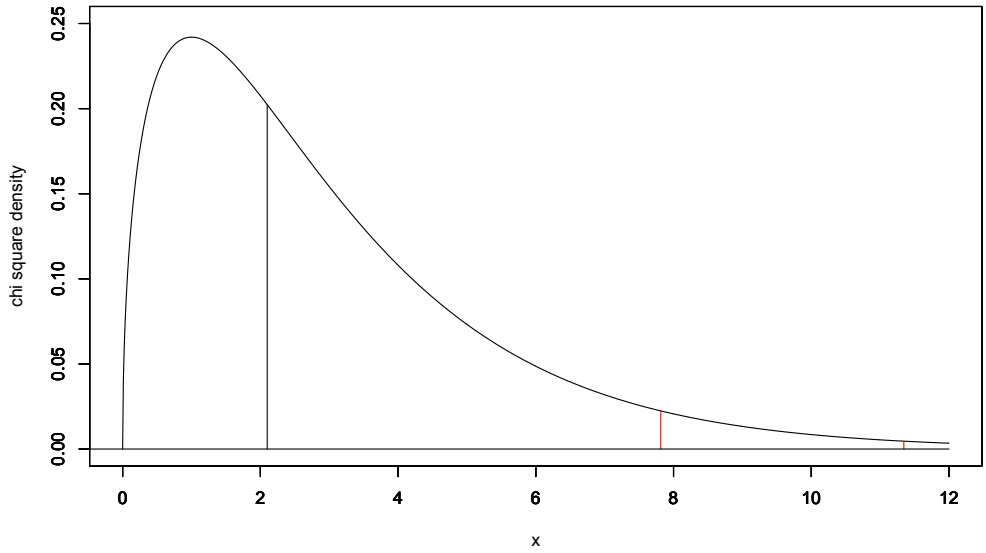


Figure 21.3: Plot of the chi-square density function with 3 degrees of freedom. The black vertical bar indicates the value of the test statistic in Example 21.3. The area 0.552 under the curve to the right of the vertical line is the p-value for this test. This is much to high to reject the null hypothesis. The red vertical lines show the critical values for a test with significance $\alpha = 0.05$ (to the left) and $\alpha = 0.01$ (to the right). Thus, the area under the curve to the right of these vertical lines is 0.05 and 0.01, respectively. These values can be found using `qchisq(1-α,3)`. We can also see that the test statistic value of 30.8874 in Example 21.3 has a very low p-value.

So, the estimate of p_1 and p_2 are

$$\hat{p}_1 = \hat{p}_{11} + \frac{1}{2}\hat{p}_{12} = \frac{49}{120}, \quad \hat{p}_2 = \hat{p}_{22} + \frac{1}{2}\hat{p}_{12} = \frac{71}{120}.$$

So, the expected number of observations is

$$E_{11} = 60\hat{p}_1^2 = 10.00417, \quad E_{12} = 60 \times 2\hat{p}_1\hat{p}_2 = 28.99167, \quad E_{22} = 60\hat{p}_2^2 = 21.00417.$$

The chi-square statistic

$$\chi^2 = \frac{(14-10)^2}{10} + \frac{(21-29)^2}{29} + \frac{(25-21)^2}{21} = 1.600 + 2.207 + 0.762 = 4.569$$

The p-value

```
> 1-pchisq(4.569,1)
[1] 0.03255556
```

Thus, we have moderate evidence against the null hypothesis of a Hardy-Weinberg equilibrium. Many forces may be the cause of this - non-random mating, selection, or migration to name a few possibilities.

Exercise 21.6. *Perform the chi-squared test using the G^2 statistic for the example above.*

21.2 Contingency tables

Contingency tables, also known as **two-way tables** or **cross tabulations** are a convenient way to display the frequency distribution from the observations of two categorical variables. For an $r \times c$ contingency table, we consider two factors A and B for an experiment. This gives r categories

$$A_1, \ldots A_r$$

for factor A and c categories

$$B_1, \ldots B_c$$

for factor B.

Here, we write O_{ij} to denote the number of occurrences for which an individual falls into both category A_i and category B_j. The results is then organized into a two-way table.

	B_1	B_2	$\cdots$	B_c	total
A_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1\cdot}$
A_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot c}$	n

Example 21.7. *Returning to the study of the smoking habits of 5375 high school children in Tucson in 1967, here is a two-way table summarizing some of the results.*

	student smokes	student does not smoke	total
2 parents smoke	400	1380	1780
1 parent smokes	416	1823	2239
0 parents smoke	188	1168	1356
total	1004	4371	5375

For a contingency table, the null hypothesis we shall consider is that the factors A and B are independent. For the experimental design, we assume that the number of observations n is fixed but the marginal distributions (row and column totals) are not.

To set the parameters for this model, we define

$$p_{ij} = P\{\text{an individual is simultaneously a member of category } A_i \text{ and category } B_j\}.$$

Then, we have the parameter space

$$\Theta = \{\mathbf{p} = (p_{ij}, 1 \le i \le r, 1 \le j \le c); p_{ij} \ge 0 \text{ for all } i, j = 1, \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1\}.$$

Write the **marginal distribution**

$$p_{i\cdot} = \sum_{j=1}^{c} p_{ij} = P\{\text{an individual is a member of category } A_i\}$$

and

$$p_{\cdot j} = \sum_{i=1}^{r} p_{ij} = P\{\text{an individual is a member of category } B_j\}.$$

The null hypothesis of independence of the categories A and B can be written

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}, \text{ for all } i, j \quad \text{versus} \quad H_1 : p_{ij} \ne p_{i\cdot}p_{\cdot j}, \text{ for some } i, j.$$

Write

$$\mathbf{n} = \{n_{ij}, 1 \le i \le r, 1 \le j \le c\}$$

where

$$n_{ij} = \#\{\text{observations simultaneously in category } A_i \text{ and category } B_j\}.$$

As with 21.2, the **likelihood function**

$$L(\mathbf{p}|\mathbf{n}) = \prod_{i=1}^{r}\prod_{j=1}^{c} p_{ij}^{n_{ij}}. \tag{21.12}$$

Follow the procedure as before for the goodness of fit test to end with a G^2 and its corresponding χ^2 test statistic. The G^2 statistic follows from the likelihood ratio test criterion. The χ^2 statistics is a second order Taylor series approximation to G^2.

$$-2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}\ln\frac{E_{ij}}{O_{ij}} \approx \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}. \tag{21.13}$$

The null hypothesis $p_{ij} = p_{i\cdot}p_{\cdot j}$ can be written in terms of observed and expected observations as

$$\frac{E_{ij}}{n} = \frac{O_{i\cdot}}{n}\frac{O_{\cdot j}}{n}.$$

or

$$E_{ij} = \frac{O_{i\cdot}O_{\cdot j}}{n}.$$

The test statistic, under the null hypothesis, has a χ^2 distribution. To determine the number of degrees of freedom, consider the following. Start with a contingency table with no entries but with the prescribed marginal values.

	B_1	B_2	$\cdots$	B_c	total
A_1					$O_{1\cdot}$
A_2					$O_{2\cdot}$
$\vdots$					$\vdots$
A_r					$O_{r\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot c}$	n

The number of degrees of freedom is the number of values that we can place in the contingency table before all the remaining values are determined. To begin, fill in the first row with values $E_{11}, E_{12}, \ldots, E_{1,c-1}$. The final value $E_{1,c}$ in this determined by the other values in the row and the constraint that the row sum must be $O_{1\cdot}$. Continue filling the rows, noting that the value in column c is determined by the constraint on the row sum. Finally, when the time comes to fill in the bottom row r, notice that all the values are determined by the constraint on the row sums $O_{\cdot j}$. Thus, we can fill $c-1$ values in each of the $r-1$ rows before the remaining values are determined. Thus, the number of degrees of freedom is $(r-1)\times(c-1)$.

Exercise 21.8. *Verify that G^2 statistic in (21.13) is the likelihood ratio test statistic.*

Exercise 21.9. *Give* $\dim(\Theta)$ *and* $\dim(\Theta_0)$*, the dimensions, respectively, of the parameter space and the null hypothesis space. Show that the difference is* $(r-1)\times(c-1)$*,*

Example 21.10. *Returning to the data set on smoking habits in Tucson, we find that the expected table is*

	student smokes	student does not smoke	total
2 parents smoke	332.49	1447.51	1780
1 parent smokes	418.22	1820.78	2239
0 parents smoke	253.29	1102.71	1356
total	1004	4371	5375

For example,

$$E_{11} = \frac{O_{1\cdot}O_{\cdot 1}}{n} = \frac{1780\cdot 1004}{5375} = 332.49.$$

To compute the chi-square statistic

$$\begin{aligned}
& \frac{(400-332.49)^2}{332.49} + \frac{(1380-1447.51)^2}{1447.51} \\
+ \quad & \frac{(416-418.22)^2}{418.22} + \frac{(1823-1820.78)^2}{1820.78} \\
+ \quad & \frac{(188-253.29)^2}{253.29} + \frac{(1168-1102.71)^2}{1102.71} \\
= \quad & 13.71 + 3.15 \\
+ \quad & 0.012 + 0.003 \\
+ \quad & 16.83 + 3.866 \\
= \quad & 37.57
\end{aligned}$$

The number of degrees of freedom is $(r-1)\times(c-1) = (3-1)\times(2-1) = 2$*. This can be seen by noting that one the first two entries in the "student smokes" column is filled, the rest are determined. Thus, the p-value*

```
> 1-pchisq(37.57,2)
[1] 6.946694e-09
```

is very small and leads us to reject the null hypothesis. Thus, we conclude that children smoking habits are not independent of their parents smoking habits. An examination of the individual cells shows that the children of parents who do not smoke are less likely to smoke and children who have two parents that smoke are more likely to smoke. Under the null hypothesis, each cell has a mean approximately 1 and so values much greater than 1 show contribution that leads to the rejection of H_0*.*

R *does the computation for us using the* `chisq.test` *command*

```
> smoking<-matrix(c(400,416,188,1380,1823,1168),nrow=3)
> smoking
     [,1] [,2]
[1,]  400 1380
[2,]  416 1823
[3,]  188 1168
> chisq.test(smoking)

Pearson's Chi-squared test

data:  smoking
X-squared = 37.5663, df = 2, p-value = 6.959e-09
```

We can look at the residuals $(O_{ij}-E_{ij})/\sqrt{E_{ij}}$ *for the entries in the* χ^2 *test as follows.*

```
> smokingtest<-chisq.test(smoking)
> residuals(smokingtest)
           [,1]        [,2]
[1,]  3.7025160 -1.77448934
[2,] -0.1087684  0.05212898
[3,] -4.1022973  1.96609088
```

Notice that if we square these values, we obtain the entries found in computing the test statistic.

```
> residuals(smokingtest)^2
            [,1]       [,2]
[1,] 13.70862455 3.14881241
[2,]  0.01183057 0.00271743
[3,] 16.82884348 3.86551335
```

Exercise 21.11. *Make three horizontally placed chigrams that summarize the residuals for this χ^2 test in the example above.*

Exercise 21.12 (two-by-two tables)**.** *Here is the contingency table can be thought of as two sets of Bernoulli trials as shown.*

	group 1	group 2	total
successes	x_1	x_2	$x_1 + x_2$
failures	$n_1 - x_1$	$n_2 - x_2$	$(n_1 + n_2) - (x_1 + x_2)$
total	n_1	n_2	$n_1 + n_2$

Show that the chi-square test is equivalent to the two-sided two sample proportion test.

21.3 Applicability and Alternatives to Chi-squared Tests

The chi-square test uses the central limit theorem and so is based on the ability to use a normal approximation. One criterion, the **Cochran conditions** requires no cell has count zero, and more than 80% of the cells have counts at least 5. If this does not hold, then **Fisher's exact test** uses the hypergeometric distribution (or its generalization) directly rather than normal approximation.

For example, for the 2×2 table,

	B_1	B_2	total
A_1	O_{11}	O_{12}	$O_{1\cdot}$
A_2	O_{21}	O_{22}	$O_{2\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	n

The idea behind Fisher's exact test is to begin with an empty table:

	B_1	B_2	total
A_1			$O_{1\cdot}$
A_2			$O_{2\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	n

and a null hypothesis that uses equally likely outcomes to fill in the table. We will use as an analogy the model of mark and recapture. Normally the goal is to find n, the total population. In this case, we assume that this population size is known and will consider the case that the individuals in the two captures are independent. This is assumed in the mark and recapture protocol. Here we test this independence.

In this regard,

- A_1 - an individual in the first capture and thus tagged.
- A_2 - an individual not in the first capture and thus not tagged.
- B_1 - an individual in the second capture.
- B_2 - an individual not in the second capture

Then, from the point of view of the A classification:

- We have $O_{1\cdot}$ from a population n with the A_1 classification (tagged individuals). This can be accomplished in

$$\binom{n}{O_{1\cdot}} = \frac{n!}{O_{1\cdot}!O_{2\cdot}!}$$

ways. The remaining $O_{2\cdot} = n - O_{1\cdot}$ have the A_2 classification (untagged individuals). Next, we fill in the values for the B classification

- From the $O_{\cdot 1}$ belonging to category B_1 (individuals in the second capture), O_{11} also belong to A_1 (have a tag). This outcome can be accomplished in

$$\binom{O_{\cdot 1}}{O_{11}} = \frac{O_{\cdot 1}!}{O_{11}!O_{21}!}$$

ways.

- From the $O_{\cdot 2}$ belonging to category B_2 (individuals not in the second capture), O_{12} also belong to A_1 (have a tag). This outcome can be accomplished in

$$\binom{O_{\cdot 2}}{O_{21}} = \frac{O_{\cdot 2}!}{O_{12}!O_{22}!}$$

ways.

Under the null hypothesis that every individual can be placed in any group, provided we have the given marginal information. In this case, the probability of the table above has the formula from the hypergeometric distribution

$$\frac{\binom{O_{1\cdot}}{O_{11}}\binom{O_{2\cdot}}{O_{21}}}{\binom{n}{O_{\cdot 1}}} = \frac{O_{\cdot 1}!/(O_{11}!O_{21}!)\cdot O_{\cdot 2}!/(O_{12}!O_{22}!)}{n!/(O_{1\cdot}!O_{2\cdot}!)} = \frac{O_{\cdot 1}!O_{\cdot 2}!O_{1\cdot}!O_{2\cdot}!}{O_{11}!O_{12}!O_{21}!O_{22}!n!}. \tag{21.14}$$

Notice that the formula is symmetric in the column and row variables. Thus, if we had derived the hypergeometric formula from the point of view of the B classification we would have obtained exactly the same formula (21.14).

To complete the exact test, we rely on statistical software to do the following:

- compute the hypergeometric probabilities over all possible choices for entries in the cells that result in the given marginal values, and
- rank these probabilities from most likely to least likely.
- Find the ranking of the actual data.
- For a one-sided test of too rare, the p-value is the sum of probabilities of the ranking lower than that of the data.

A similar procedure applies to provide the Fisher exact test for $r \times c$ tables.

Example 21.13. *As a test of the assumptions for mark and recapture. We examine a small population of 120 fish. The assumption are that each group of fish are equally likely to be capture in the first and second capture and that the two captures are independent. This could be violated, for example, if the tagged fish are not uniformly dispersed in the pond.*

Twenty-five are tagged and returned to the pond. For the second capture of 30, seven are tagged. With this information, given in red in the table below, we can complete the remaining entries.

	in 2nd capture	not in 2nd capture	total
in 1st capture	7	18	25
not in 1st capture	23	72	95
total	30	90	120

Fisher's exact test show a much too high p-value to reject the null hypothesis.

```
> fish<-matrix(c(7,23,18,72),ncol=2)
> fisher.test(fish)

Fisher's Exact Test for Count Data

data:  fish
p-value = 0.7958
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3798574 3.5489546
sample estimates:
odds ratio
  1.215303
```

Exercise 21.14. *Perform the χ^2 test on the data set above and report the findings.*

Example 21.15. *We now return to a table on hemoglobin genotypes on two Indonesian islands. Recall that heterozygotes are protected against malaria.*

genotype	AA	AE	EE
Flores	128	6	0
Sumba	119	78	4

We noted that heterozygotes are rare on Flores and that it appears that malaria is less prevalent there since the heterozygote does not provide an adaptive advantage. Here are both the chi-square test and the Fisher exact test.

```
> genotype<-matrix(c(128,119,6,78,0,4),nrow=2)
> genotype
     [,1] [,2] [,3]
[1,]  128    6    0
[2,]  119   78    4
> chisq.test(genotype)

Pearson's Chi-squared test

data:  genotype
X-squared = 54.8356, df = 2, p-value = 1.238e-12

Warning message:
In chisq.test(genotype) : Chi-squared approximation may be incorrect
```

and

```
> fisher.test(genotype)

Fisher's Exact Test for Count Data

data:  genotype
p-value = 3.907e-15
alternative hypothesis: two.sided
```

Note that R cautions against the use of the chi-square test with these data.

21.4 Answer to Selected Exercise

21.1. For the first identity, using $\delta_i = (O_i - E_i)/E_i$.

$$\sum_{i=1}^{k} E_i \delta_i = \sum_{i=1}^{k} E_i \frac{O_i - E_i}{E_i} = \sum_{i=1}^{k} (O_i - E_i) = n - n = 0$$

and for the second

$$E_i(1 + \delta_i) = E_i \left(\frac{E_i}{E_i} + \frac{O_i - E_i}{E_i} \right) = E_i \frac{O_i}{E_i} = O_i.$$

21.2. We apply the quadratic Taylor polynomial approximation for the natural logarithm,

$$\ln(1 + \delta_i) \approx \delta_i - \frac{1}{2}\delta_i^2,$$

and use the identities in the previous exercise. Keeping terms up to the square of δ_i, we find that

$$\begin{aligned} -2 \ln \Lambda_n(\mathbf{O}) &= 2 \sum_{i=1}^{k} O_i \ln \frac{O_i}{E_i} = 2 \sum_{i=1}^{k} E_i(1 + \delta_i) \ln(1 + \delta_i) \\ &\approx 2 \sum_{i=1}^{k} E_i(1 + \delta_i)(\delta_i - \frac{1}{2}\delta_i^2) \approx 2 \sum_{i=1}^{k} E_i(\delta_i + \frac{1}{2}\delta_i^2) \\ &= 2 \sum_{i=1}^{k} E_i \delta_i + \sum_{i=1}^{k} E_i \delta_i^2 \\ &= 0 + \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}. \end{aligned}$$

21.6. Here is the R output.

```
> O<-c(14,21,25)
> phat<-c(O[1]+O[2]/2,O[3]+O[2]/2)/sum(O)
> phat
[1] 0.4083333 0.5916667
> E<-sum(O)*c(phat[1]^2,2*phat[1]*phat[2],phat[2]^2)
> E
[1] 10.00417 28.99167 21.00417
> sum(E)
[1] 60
> G2stat<-2*sum(O*log(O/E))
> G2stat
[1] 4.572896
> 1-pchisq(G2stat,1)
[1] 0.03248160
```

21.8. First we maximize the likelihood $L(\mathbf{p}|\mathbf{n})$ in (21.12). As with (21.5), we find that the maximum likelihood estimate

$$\hat{p}_{ij} = \frac{n_{ij}}{n}$$

is simply the fraction of observations that are simultaneously in categories A_i and B_j. Here n is the total number of observations. Thus,

$$\log L(\hat{\mathbf{p}}|\mathbf{n}) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log \frac{n_{ij}}{n}.$$

To maximize under the null hypothesis note that $p_{0,ij} = p_{i\cdot} \cdot p_{\cdot j}$ and therefore

$$L(\mathbf{p}_0|\mathbf{n}) = \prod_{i=1}^{r}\prod_{j=1}^{c}(p_{i\cdot}p_{\cdot j})^{n_{ij}} = \prod_{i=1}^{r}\prod_{j=1}^{c} p_{i\cdot}^{n_{ij}} p_{\cdot j}^{n_{ij}} = \prod_{i=1}^{r} p_{i\cdot}^{n_{i\cdot}} \cdot \prod_{j=1}^{c} p_{\cdot j}^{n_{\cdot j}}.$$

We now have two maximization problems, for $p_{i\cdot}$ and $p_{\cdot j}$. Again, we return to the strategy to determine the maximum likelihood estimate (21.5) to see that

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \text{and} \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}.$$

Thus, the maximum likelihood estimate under the null hypothesis

$$\hat{p}_{0,ij} = \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}$$

and

$$\log L(\hat{\mathbf{p}}_0|\mathbf{n}) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log\left(\frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}\right).$$

Next we subtract to find the logarithm of the likelihood ratio.

$$\log \Lambda(\mathbf{n}) = \log L(\hat{\mathbf{p}}_0|\mathbf{n}) - \log L(\hat{\mathbf{p}}|\mathbf{n}) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\left(\log\left(\frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}\right) - \log \frac{n_{ij}}{n}\right).$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\left(\log \frac{n_{i\cdot}n_{\cdot j}}{n} - \log n_{ij}\right) = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \log \frac{E_{ij}}{O_{ij}}$$

Multiply by -2 to obtain the desired expression for G^2 as the likelihood ratio test statistic.

21.9. For the parameter space Θ, we have $r \times c$ probabilities p_{ij} with the single constraint that their sum is 1. Thus, $\dim(\Theta) = rc - 1$. For the null hypothesis space Θ_0, we have r row probabilities $p_{i\cdot}$ with the constraint that the sum is 1 and c column probabilities $p_{\cdot j}$ with the constraint that the sum is 1. Thus, $\dim(\Theta_0) = (r-1) + (c-1)$. Finally,

$$\dim(\Theta) - \dim(\Theta_0) = rc - 1 - (r-1) - (c-1) = rc - r - c + 1 = (r-1)(c-1).$$

21.11. Here is the R output

```
> resid<-residuals(smokingtest)
> colnames(resid)<-c("smokes","does not smoke")
> par(mfrow=c(1,3))
> barplot(resid[1,],main="2 parents",ylim=c(-4.5,4.5))
> barplot(resid[2,],main="1 parent",ylim=c(-4.5,4.5))
> barplot(resid[3,],main="0 parents",ylim=c(-4.5,4.5))
```

21.12. The table of expected observations

	group 1	group 2	total
successes	$\frac{n_1(x_1+x_2)}{n_1+n_2}$	$\frac{n_2(x_1+x_2)}{n_1+n_2}$	$x_1 + x_2$
failures	$\frac{n_1((n_1+n_2)-(x_1+x_2))}{n_1+n_2}$	$\frac{n_2((n_1+n_2)-(x_1+x_2))}{n_1+n_2}$	$(n_1+n_2)-(x_1+x_2)$
total	n_1	n_2	n_1+n_2



Figure 21.4: Chigram for the data on teen smoking in Tucson, 1967. R commands found in Exercise 21.9.

Now, write $\hat{p}_i = x_i/n_i$ for the sample proportions from each group, and

$$\hat{p}_0 = \frac{x_1 + x_2}{n_1 + n_2}$$

for the pooled sample proportion. Then we have the table of observed and expected observations

observed	group 1	group 2	total
successes	$n_1\hat{p}_1$	$n_2\hat{p}_2$	$(n_1+n_2)\hat{p}_0$
failures	$n_1(1-\hat{p}_1)$	$n_2(1-\hat{p}_2)$	$(n_1+n_2)(1-\hat{p}_0)$
total	n_1	n_2	n_1+n_2

expected	group 1	group 2	total
successes	$n_1\hat{p}_0$	$n_2\hat{p}_0$	$(n_1+n_2)\hat{p}_0$
failures	$n_1(1-\hat{p}_0)$	$n_2(1-\hat{p}_0)$	$(n_1+n_2)(1-\hat{p}_0)$
total	n_1	n_2	n_1+n_2

The chi-squared test statistic

$$\begin{array}{rcl}
 & \frac{(n_1(\hat{p}_1-\hat{p}_0))^2}{n_1\hat{p}_0} + \frac{(n_2(\hat{p}_2-\hat{p}_0))^2}{n_2\hat{p}_0} & \\
+ & \frac{(n_1((1-\hat{p}_1)-(1-\hat{p}_0)))^2}{n_1(1-\hat{p}_0)} + \frac{(n_2((1-\hat{p}_2)+(1-\hat{p}_0)))^2}{n_2(1-\hat{p}_0)} & \\
= & n_1\frac{(\hat{p}_1-\hat{p}_0)^2}{\hat{p}_0} + n_2\frac{(\hat{p}_2-\hat{p}_0)^2}{\hat{p}_0} & \\
+ & n_1\frac{(\hat{p}_1-\hat{p}_0)^2}{(1-\hat{p}_0)} + n_2\frac{(\hat{p}_2-\hat{p}_0)^2}{(1-\hat{p}_0)} & \\
= & n_1(\hat{p}_1-\hat{p}_0)^2\frac{1}{\hat{p}_0(1-p_0)} + n_2(\hat{p}_2-\hat{p}_0)^2\frac{1}{\hat{p}_0(1-p_0)} & \\
= & \frac{n_1(\hat{p}_1-\hat{p}_0)^2+n_2(\hat{p}_2-\hat{p}_0)^2}{\hat{p}_0(1-p_0)} = -\ln\Lambda(\mathbf{x}_1,\mathbf{x}_2) &
\end{array}$$

from the likelihood ratio computation for the two-sided two sample proportion test.

21.14. The R commands follow:

```
> chisq.test(fish)

Pearson's Chi-squared test with Yates' continuity correction

data:  fish
X-squared = 0.0168, df = 1, p-value = 0.8967
```

The p-value is notably higher for the χ^2 test.

Topic 22

Analysis of Variance

The above property of the variance, by which each independent cause makes its own contribution to the total, enables us to analyze *the total, and to assign, with more or less of accuracy, the several portions to their appropriate causes, or groups of causes. In Table II is shown the analysis of the total variance for each plot, divided as it may be ascribed . . . .* - Ronald Fisher. 1921, Studies in Crop Variation. I. An examination of the yield of dressed grain from Broadbalk, *Journal of Agricultural Science*

22.1 Overview

Two-sample t procedures are designed to compare the means of two populations. Our next step is to compare the means of several populations. We shall explain the methodology through an example. Consider the data set gathered from the forests in Borneo.

Example 22.1 (Rain forest logging)**.** *The data on 30 forest plots in Borneo are the number of trees per plot.*

	never logged	logged 1 year ago	logged 8 years ago
n_i	12	12	9
$\bar{y}_i$	23.750	14.083	15.778
s_i	5.065	4.981	5.761

We compute these statistics from the data $y_{11}, \ldots y_{1n_1}$, $y_{21}, \ldots y_{2n_2}$ and $y_{31}, \ldots y_{2n_2}$

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \quad \text{and} \quad s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_i)^2$$

One way analysis of variance (ANOVA) is a statistical procedure that allows us to test for the differences in means for two or more independent groups. In the situation above, we have set our design so that the data in each of the three groups is a random sample from within the groups. The basic question is: *Are these means the same (the null hypothesis) or not (the alternative hypothesis)?*

As the case with the t procedures, the appropriateness of one way analysis of variance is based on the applicability of the central limit theorem. As with t procedures, ANOVA has an alternative, the Kruskal-Wallis test, based on the ranks of the data for circumstances in which the central limit theorem does not apply.

The basic idea of the test is to examine the ratio of s^2_{between}, the variance between the groups 1, 2, and 3. and s^2_{residual}, a statistic that measures the variances within the groups. If the resulting ratio test statistic is sufficiently large, then we say, based on the data, that the means of these groups are distinct and we are able to reject the null hypothesis. Even though the boxplots use different measures of center (median vs. mean) and spread (quartiles vs. standard deviation), this idea can be expressed by examining in Figure 22.1 the fluctuation in the centers of boxes compared to the width of the boxes.

As we have seen before, this decision to reject H_0 will be the consequence a sufficiently high value of a test statistic - in this case the F statistic. The distribution of this test statistic will depend on the number of groups (3 in the example above) and the number of total observations (33 in the example above). Consequently, variances between groups that are not statistically significant for small sample sizes can become significant as the sample sizes and, with it, the power increase.

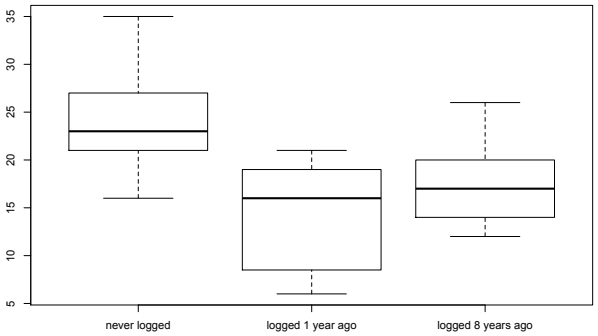


Figure 22.1: Side-by-side boxplots of the number of trees per plot. The groups will be considered different if the differences between the groups (indicated by the variation in the center lines of the boxes) is large compared to the width of the boxes in the boxplot.

22.2 One Way Analysis of Variance

For one way analysis of variance, we expand to more than the two groups seen for t procedures and ask whether or not the means of all the groups are the same. The hypothesis in this case is

$$H_0 : \mu_j = \mu_k \text{ for all } j, k \quad \text{and} \quad H_1 : \mu_j \neq \mu_k \text{ for some } j, k.$$

The data $\{y_{ij}, 1 \leq i \leq n_j, 1 \leq j \leq q\}$ represents that we have n_i observation for the i-th group and that we have q groups. The total number of observations is denoted by $n = n_1 + \cdots + n_q$. The model is

$$y_{ij} = \mu_j + \epsilon_{ij}.$$

where ϵ_{ij} are independent $N(0, \sigma)$ random variables with σ^2 unknown. This allows us to define the likelihood and to use that to determine the analysis of variance F test as a likelihood ratio test. Notice that the model for analysis requires a common value σ for all of the observations.

In order to develop the F statistic at the test statistic, we will need to introduce two types of sample means:

- The **within group means** is simply the sample mean of the observations inside each of the groups,

$$\overline{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \quad j = 1. \ldots, q.$$

  These are given in the table in Example 22.1 for the Borneo rains forest. The within group mean $\overline{y}_j$ is the maximum likelihood estimate of μ_j under H_0.

- The mean of the data taken as a whole, known as the **grand mean**,

$$\overline{\overline{y}} = \frac{1}{n} \sum_{j=1}^{q} \sum_{i=1}^{n_j} y_{ij} = \frac{1}{n} \sum_{j=1}^{q} n_j \bar{y}_j.$$

  This is the weighted average of the $\bar{y}_i$ with weights n_i, the sample size in each group. The Borneo rain forest example has an overall mean

$$\overline{\overline{y}} = \frac{1}{n} \sum_{j=1}^{3} n_j \bar{y}_j = \frac{1}{12 + 12 + 9}(12 \cdot 23.750 + 12 \cdot 14.083 + 9 \cdot 15.778) = 18.06055.$$

  The grand mean $\overline{\overline{y}}$ is the maximum likelihood estimate of the common values for the μ_j under H_1.

source of variation	degrees of freedom	sums of squares	mean square
between groups	$q-1$	SS_{between}	$s^2_{\text{between}} = SS_{\text{between}}/(q-1)$
residuals	$n-q$	SS_{residual}	$s^2_{\text{residual}} = SS_{\text{residual}}/(n-q)$
total	$n-1$	SS_{total}	

Table I: Table for one way analysis of variance

Analysis of variance uses the **total sum of squares**

$$SS_{\text{total}} = \sum_{j=1}^{q}\sum_{i=1}^{n_j}(y_{ij} - \overline{\overline{y}})^2, \tag{22.1}$$

the total square variation of individual observations from their grand mean. SS_{total} appears because SS_{total}/n is the the maximum likelihood estimate for σ^2 under H_1.

However, the test statistic is determined by decomposing SS_{total}. We start with a bit of algebra to rewrite the interior sum in (22.1) as

$$\sum_{i=1}^{n_j}(y_{ij} - \overline{\overline{y}})^2 = \sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2 = (n_j - 1)s_j^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2. \tag{22.2}$$

Here, s_j^2 is the unbiased estimator of the variance based on the observations in the j-th group.

Exercise 22.2. *Show the first equality in (22.2). (Hint: Begin with the difference in the two sums.)*

Together (22.1) and (22.2) yields the decomposition of the variation

$$SS_{\text{total}} = SS_{\text{residual}} + SS_{\text{between}}$$

with

$$SS_{\text{residual}} = \sum_{j=1}^{q}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2 = \sum_{j=1}^{q}(n_j - 1)s_j^2 \quad \text{and} \quad SS_{\text{between}} = \sum_{j=1}^{q} n_j(\overline{y}_j - \overline{\overline{y}})^2.$$

SS_{residual}/n is the the maximum likelihood estimate for σ^2 under H_0.

For the rain forest example, we find that

$$SS_{\text{between}} = \sum_{j=1}^{3} n_j(\overline{y}_j - \overline{\overline{y}})^2 = 12\cdot(23.750 - \overline{\overline{y}})^2 + 12\cdot(14.083 - \overline{\overline{y}})^2 + 9\cdot(15.778 - \overline{\overline{y}})^2) = 625.1793$$

and

$$SS_{\text{residual}} = \sum_{j=1}^{3}(n_j - 1)s_j^2 = (12-1)\cdot 5.065^2 + (12-1)\cdot 4.981^2 + (9-1)\cdot 5.761^2 = 820.6234$$

From this, we obtain the general form for one-way analysis of variance as shown in Table I.

- The $q-1$ degrees of freedom between groups is derived from the q groups minus one degree of freedom used to compute $\overline{\overline{y}}$.

- The $n - q$ degrees of freedom within the groups is derived from the $n_j - 1$ degree of freedom used to compute the variances s_j^2. Add these q values for the degrees of freedom to obtain $n - q$.

The test statistic

$$F = \frac{s^2_{\text{between}}}{s^2_{\text{residual}}} = \frac{SS_{\text{between}}/(q-1)}{SS_{\text{residual}}/(n-q)}.$$

is, under the null hypothesis, a constant multiple of the ratio of two independent χ^2 random variables with parameter $q-1$ for the numerator and $n-q$ for the denominator. This ratio is called an F **random variable** with $q-1$ numerator degrees of freedom and $n - q$ denominator degrees of freedom.

Using Table II, we find the value of the test statistic for the rain forest data is

$$F = \frac{s^2_{\text{between}}}{s^2_{\text{residual}}} = \frac{312.6}{27.4} = 11.43.$$

and the p-value (calculated below) is 0.0002. The critical value for an $\alpha = 0.01$ level test is 5.390. So, we do reject the null hypothesis that mean number of trees does not depend on the history of logging.

```
> 1-pf(11.43,2,30)
[1] 0.0002041322
> qf(0.99,2,30)
[1] 5.390346
```

Confidence intervals are determined using the data from all of the groups as an unbiased estimate for the variance, σ^2. Using all of the data allows us to increase the number of degrees of freedom in the t distribution and thus reduce the upper critical value for the t statistics and with it the margin of error.

The variance $s^2_{\text{residuals}}$ is given by the expression $SS_{\text{residuals}}/(n - q)$, shown in the table in the "mean square" column and the "residuals" row. The standard deviation s_{residual} is the square root of this number. For example, the γ-level confidence interval for μ_j is

$$\bar{y}_j \pm t_{(1-\gamma)/2,n-q}\frac{s_{\text{residual}}}{\sqrt{n_j}}.$$

The confidence for the difference in $\mu_j - \mu_k$ is similar to that for a pooled two-sample t confidence interval and is given by

$$\bar{y}_j - \bar{y}_k \pm t_{(1-\gamma)/2,n-q}s_{\text{residual}}\sqrt{\frac{1}{n_j} + \frac{1}{n_k}}.$$

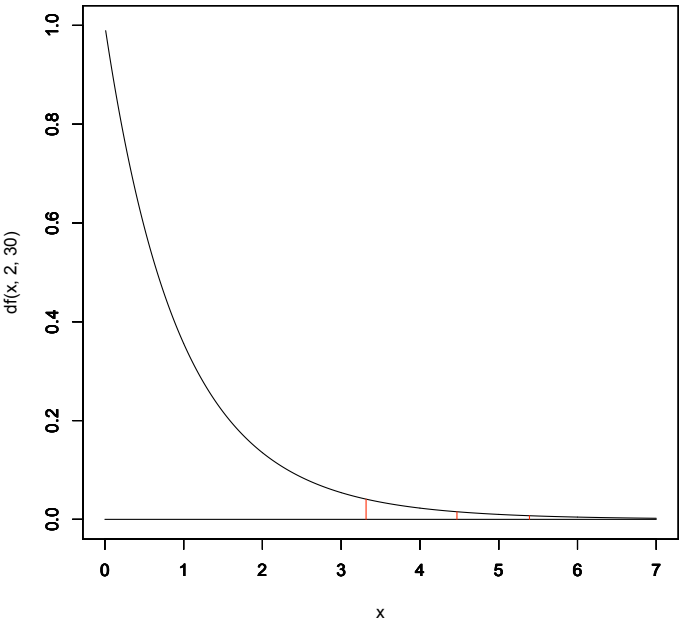


Figure 22.2: Upper tail critical values. The density for an F random variable with numerator degrees of freedom, 2, and denominator degrees of freedom, 30. The indicated values 3.316, 4.470, and 5.390 are critical values for significance levels $\alpha = 0.05$, 0.02, and 0.01, respectively.

source of variation	degrees of freedom	sums of squares	mean square
between groups	2	625.2	312.6
residuals	30	820.6	27.4
total	32	1445.8	

Table II: Analysis of variance information for the Borneo rain forest data

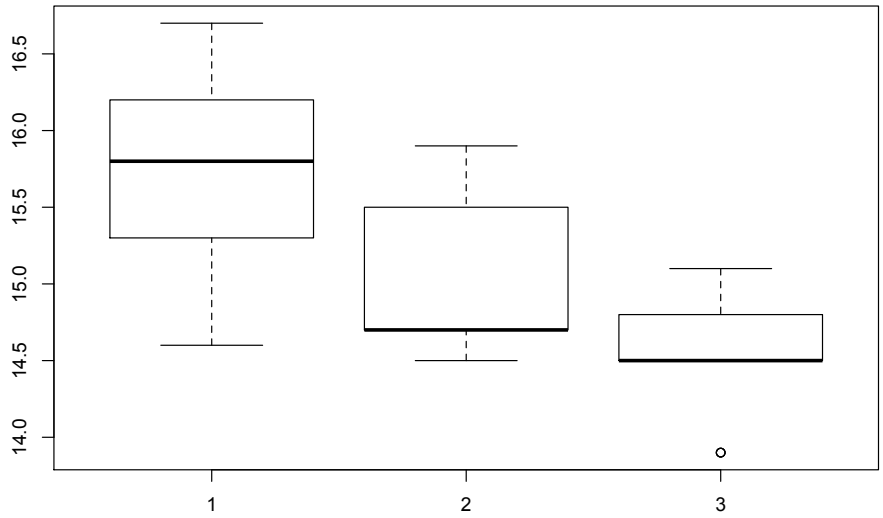


Figure 22.3: Side-by-side boxplot of queen development times. The time is measured in days. the plots show cool (1) medium (2) and warm (3) hive temperatures.

In this case, the 95% confidence interval for the mean number of trees on a lot "logged 1 year ago" has $n - q = 33 - 3$, $t_{0.025,30} = 2.042$, $s_{\text{residual}} = \sqrt{27.4} = 5.234$ and the confidence interval is

$$14.083 \pm 2.042\frac{\sqrt{27.4}}{\sqrt{12}} = 14.083 \pm 4.714 = (9.369, 18.979).$$

Exercise 22.3. *Give the 95% confidence intervals for the difference in trees between plots never logged and plots logged 8 years ago.*

Example 22.4. *The development time for a European queen in a honey bee hive is suspected to depend on the temperature of the hive. To examine this, queens are reared in a low temperature hive (*31.1° *C), a medium temperature hive (*32.8° *C) and a high temperature hive (*34.4° *C). The hypothesis is that higher temperatures increase metabolism rate and thus reduce the time needed from the time the egg is laid until an adult queen honey bee emerges from the cell. The hypothesis is*

$$H_0 : \mu_{\text{low}} = \mu_{\text{med}} = \mu_{\text{high}} \quad \text{versus} \quad H_1 : \mu_{\text{low}}, \mu_{\text{med}}, \mu_{\text{high}} \textit{ differ}$$

where $\mu_{\text{low}}, \mu_{\text{med}}$, *and* μ_{high} *are, respectively, the mean development time in days for queen eggs reared in a low, a medium, and a high temperature hive. Here are the data and a boxplot:*

```
> ehblow<-c(16.2,14.6,15.8,15.8,15.8,15.8,16.2,16.7,15.8,16.7,15.3,14.6,
  15.3,15.8)
> ehbmed<-c(14.5,14.7,15.9,15.5,14.7,14.7,14.7,15.5,14.7,15.2,15.2,15.9,
  14.7,14.7)
> ehbhigh<-c(13.9,15.1,14.8,15.1,14.5,14.5,14.5,14.5,13.9,14.5,14.8,14.8,
  13.9,14.8,14.5,14.5,14.8,14.5,14.8)
> boxplot(ehblow,ehbmed,ehbhigh)
```

The commands in R *to perform analysis and the output are shown below. The first command puts all of the data in a single vector,* `ehb`*. Next, we label the groups with the variable or factor name* `temp`*. Expressed in this way, this variable is considered by* R *as a numerical vector. We then need to inform* R *to convert these numbers into factors and list the factors in the vector* `ftemp`*. Without this, the command* `anova(lm(ehb~temp))` *would attempt to do linear regression with* `temp` *as the explanatory variable.*

```
> ehb<-c(ehblow,ehbmed,ehbhigh)
> temp<-c(rep(1,length(ehblow)),rep(2,length(ehbmed)),rep(3,length(ehbhigh)))
> ftemp<-factor(temp,c(1:3))
> anova(lm(ehb~ftemp))
Analysis of Variance Table

Response: ehb
          Df Sum Sq Mean Sq F value    Pr(>F)
ftemp      2 11.222  5.6111  23.307 1.252e-07 ***
Residuals 44 10.593  0.2407
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The anova output shows strong evidence against the null hypothesis. The p-value is 1.252×10^{-7}. *The values in the table can be computed directly from the formulas above.*

For the sums of square between groups, SS_{between},

```
> length(ehblow)*(mean(ehblow)-mean(ehb))^2
  + length(ehbmed)*(mean(ehbmed)-mean(ehb))^2
  + length(ehbhigh)*(mean(ehbhigh)-mean(ehb))^2
[1] 11.22211
```

and within groups, SS_{residual},

```
> sum((ehblow-mean(ehblow))^2)+sum((ehbmed-mean(ehbmed))^2)
  + sum((ehbhigh-mean(ehbhigh))^2)
[1] 10.59278
```

For confidence intervals we use $s^2_{resid} = 0.2407$, $s_{resid} = 0.4906$ *and the* t*-distribution with 44 degrees of freedom. For the medium temperature hive, the 95% confidence interval for* μ_{med} *can be computed*

```
> mean(ehblow)
[1] 15.74286
> qt(0.975,44)
[1] 2.015368
> length(ehblow)
[1] 14
```

Thus, the intverval is

$$\bar{y}_{med} \pm t_{0.025,44}\frac{s_{resid}}{\sqrt{n_{med}}} = 15.742 \pm 2.0154\frac{0.4906}{\sqrt{14}} = (15.478, 16.006).$$

22.3 Contrasts

After completing a one way analysis of variance, resulting in rejecting the null hypotheses, a typical follow-up procedure is the use of **contrasts**. Contrasts use as a null hypothesis that some linear combination of the means equals to zero.

Example 22.5. *If we want to see if the rain forest has seen recovery in logged areas over the past 8 years. This can be written as*

$$H_0 : \mu_2 = \mu_3 \quad \text{versus} \quad H_1 : \mu_2 \neq \mu_3.$$

or

$$H_0 : \mu_2 - \mu_3 = 0 \quad \text{versus} \quad H_1 : \mu_2 - \mu_3 \neq 0$$

Under the null hypothesis, the test statistic

$$t = \frac{\bar{y}_2 - \bar{y}_3}{s_{residual}\sqrt{\frac{1}{n_2} + \frac{1}{n_3}}},$$

has a t-distribution with $n - q$ degrees of freedom. Here

$$t = \frac{14.083 - 15.778}{5.234\sqrt{\frac{1}{12} + \frac{1}{9}}} = -0.7344,$$

with $n - q = 33 - 3$ degrees of freedom, the p-value for this 2-sided test is

```
> 2*pt(-0.7344094,30)
[1] 0.4684011
```

is considerably too high to reject the null hypothesis.

Example 22.6. *To see if the mean queen development medium hive temperature is midway between the time for the high and low temperature hives, we have the contrast,*

$$H_0 : \frac{1}{2}(\mu_{low} + \mu_{high}) = \mu_{med} \quad \text{versus} \quad H_1 : \frac{1}{2}(\mu_{low} + \mu_{high}) \neq \mu_{med}$$

or

$$H_0 : \frac{1}{2}\mu_{low} - \mu_{med} + \frac{1}{2}\mu_{high} = 0 \quad \text{versus} \quad H_1 : \frac{1}{2}\mu_{low} - \mu_{med} + \frac{1}{2}\mu_{high} \neq 0$$

Notice that, under the null hypothesis

$$E\left[\frac{1}{2}\bar{Y}_{low} - \bar{Y}_{med} + \frac{1}{2}\bar{Y}_{high}\right] = \frac{1}{2}\mu_{low} - \mu_{med} + \frac{1}{2}\mu_{high} = 0$$

and

$$\text{Var}\left(\frac{1}{2}\bar{Y}_{low} - \bar{Y}_{med} + \frac{1}{2}\bar{Y}_{high}\right) = \frac{1}{4}\frac{\sigma^2}{n_{low}} + \frac{\sigma^2}{n_{med}} + \frac{1}{4}\frac{\sigma^2}{n_{high}} = \sigma^2\left(\frac{1}{4n_{low}} + \frac{1}{n_{med}} + \frac{1}{4n_{high}}\right).$$

This leads to the test statistic

$$t = \frac{\frac{1}{2}\bar{y}_{low} - \bar{y}_{med} + \frac{1}{2}\bar{y}_{high}}{s_{residual}\sqrt{\frac{1}{4n_{low}} + \frac{1}{n_{med}} + \frac{1}{4n_{high}}}} = \frac{\frac{1}{2}15.743 - 15.043 + \frac{1}{2}14.563}{0.4906\sqrt{\frac{1}{4\cdot 14} + \frac{1}{14} + \frac{1}{4\cdot 19}}} = 0.7005.$$

The p-value,

```
> 2*(1-pt(0.7005,44))
[1] 0.487303
```

again, is considerably too high to reject the null hypothesis.

Exercise 22.7. *Under the null hypothesis appropriate for one way analysis of variance, with n_i observations in group $i = 1, \ldots, q$ and $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$,*

$$E[c_1\bar{Y}_1 + \cdots + Y_q\mu_q] = c_1\mu_1 + \cdots + c_q\mu_q, \quad \text{Var}(c_1\bar{Y}_1 + \cdots + c_qY_q) = \frac{c_1^2\sigma^2}{n_1} + \cdots + \frac{c_q^2\sigma^2}{n_q}.$$

In general, a contrast begins with a linear combination of the means

$$\psi = c_1\mu_1 + \cdots + c_q\mu_q.$$

The hypothesis is

$$H_0 : \psi = 0 \quad \text{versus} \quad H_1 : \psi \neq 0$$

For sample means, $\bar{y}_1, \ldots, \bar{y}_q$, the test statistic is

$$t = \frac{c_1\bar{y}_1 + \cdots + c_q\bar{y}_q}{s_{residual}\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_q^2}{n_q}}}.$$

Under the null hypothesis the t statistic has a t distribution with $n - q$ degrees of freedom.

22.4 Two Sample Procedures

We now show that the t-sample procedure results from a likelihood ratio test. We keep to two groups in the development of the F test. The essential features can be found in this example without the extra notation necessary for an arbitrary number of groups.

Our hypothesis test is based on two independent samples of normal random variables. The data are

$$y_{ij} = \mu_j + \epsilon_{ij}.$$

where ϵ_{ij} are independent $N(0, \sigma)$ random variables with σ unknown. Thus, we have n_j independent $N(\mu_j, \sigma)$ random variables $Y_{1j} \ldots, Y_{n_j j}$ with unknown *common* variance σ^2, $j = 1$ and 2. The assumption of a common variance is critical to the ability to compute the test statistics.

Consider the **two-sided hypothesis**

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2.$$

Thus, the parameter space is

$$\Theta = \{(\mu_1, \mu_2, \sigma^2); \mu_1, \mu_2 \in \mathbb{R}, \sigma^2 > 0\}.$$

For the null hypothesis, the possible parameter values are

$$\Theta_0 = \{(\mu_1, \mu_2, \sigma^2); \mu_1 = \mu_2, \sigma^2 > 0\}$$

Step 1. Determine the log-likelihood. To find the test statistic derived from a likelihood ratio test, we first write the likelihood and its logarithm based on observations $\mathbf{y} = (y_{11}, \ldots, y_{n_1 1}, y_{12}, \ldots, y_{n_2 2})$.

$$L(\mu_1.\mu_2, \sigma^2|\mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_1+n_2} \exp -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{i1} - \mu_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \mu_2)^2\right) \tag{22.3}$$

$$\ln L(\mu_1.\mu_2, \sigma^2|\mathbf{y}) = -\frac{(n_1 + n_2)}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{i1} - \mu_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \mu_2)^2\right) \tag{22.4}$$

Step 2. Find the maximum likelihood estimates and the maximum value of the likelihood. By taking partial derivatives with respect to μ_1 and μ_2 we see that with two independent samples, the maximum likelihood estimate for the mean μ_j for each of the samples is the sample mean $\bar{y}_j$.

$$\hat{\mu}_1 = \bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{i1}, \quad \hat{\mu}_2 = \bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{i2}.$$

Now differentiate (22.4) with respect to σ^2

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu_1, \mu_2, \sigma^2 | \mathbf{x}) = -\frac{n_1 + n_2}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \left(\sum_{i=1}^{n_1} (y_{i1} - \mu_1)^2 + \sum_{i=1}^{n_2} (y_{i2} - \mu_2)^2 \right).$$

Thus, the maximum likelihood estimate of the variance is the *weighted* average, weighted according to the sample size, of the maximum likelihood estimator of the variance for each of the respective samples.

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{n_1} (y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{i2} - \bar{y}_2)^2 \right).$$

Now, substitute these values into the likelihood (22.3) to see that the maximum value for the likelihood is

$$\begin{aligned} L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2 | \mathbf{x}) &= \frac{1}{(2\pi\hat{\sigma}^2)^{(n_1+n_2)/2}} \exp - \frac{1}{2\hat{\sigma}^2} \left(\sum_{i=1}^{n_1} (y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{i2} - \bar{y}_2)^2 \right) \\ &= \frac{1}{(2\pi\hat{\sigma}^2)^{(n_1+n_2)/2}} \exp - \frac{n_1 + n_2}{2} \end{aligned}$$

Step 3. Find the parameters that maximize the likelihood under the null hypothesis and then find the maximum value of the likelihood on Θ_0. Next, for the likelihood ratio test, we find the maximum likelihood under the null hypothesis. In this case the two means have a common value which we shall denote by μ.

$$L(\mu, \sigma^2 | \mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^{n_1+n_2} \exp - \frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_1} (y_{i1} - \mu)^2 + \sum_{i=1}^{n_2} (y_{i2} - \mu)^2 \right) \tag{22.5}$$

$$\ln L(\mu, \sigma^2 | \mathbf{x}) = -\frac{(n_1 + n_2)}{2} (\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_1} (y_{i1} - \mu)^2 + \sum_{i=1}^{n_2} (y_{i2} - \mu)^2 \right) \tag{22.6}$$

The μ derivative of (22.6) is

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2 | \mathbf{x}) = \frac{1}{\sigma^2} \left(\sum_{i=1}^{n_1} (y_{i1} - \mu) + \sum_{i=1}^{n_2} (y_{i2} - \mu) \right).$$

Set this to 0 and solve to realize that the maximum likelihood estimator under the null hypothesis is the **grand sample mean** $\overline{\overline{y}}$ obtained by considering all of the data being derived from one large sample

$$\hat{\mu}_0 = \overline{\overline{y}} = \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{n_1} y_{i1} + \sum_{i=1}^{n_2} y_{i2} \right) = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2}.$$

Intuitively, if the null hypothesis is true, then all of our observations are independent and have the same distribution and thus, we should use all of the data to estimate the common mean of this distribution.

The value for σ^2 that maximizes (22.5) on Θ_0, is also the maximum likelihood estimator for the variance obtained by considering all of the data being derived from one large sample:

$$\hat{\sigma}_0^2 = \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{n_1} (y_{i1} - \overline{\overline{y}})^2 + \sum_{i=1}^{n_2} (y_{i2} - \overline{\overline{y}})^2 \right).$$

We can find that the maximum value on Θ_0 for the likelihood is

$$\begin{aligned} L(\hat{\mu}_0, \hat{\sigma}_0^2 | \mathbf{x}) &= \frac{1}{(2\pi\hat{\sigma}_0^2)^{(n_1+n_2)/2}} \exp - \frac{1}{2\hat{\sigma}_0^2} \left(\sum_{i=1}^{n_1} (y_{i1} - \overline{\overline{y}})^2 + \sum_{i=1}^{n_2} (y_{i2} - \overline{\overline{y}})^2 \right) \\ &= \frac{1}{(2\pi\hat{\sigma}_0^2)^{(n_1+n_2)/2}} \exp - \frac{n_1 + n_2}{2} \end{aligned}$$

Step 4. Find the likelihood statistic $\Lambda(\mathbf{y})$. From steps 2 and 3, we find a likelihood ratio of

$$\Lambda(\mathbf{y}) = \frac{L(\hat{\mu}_0, \hat{\sigma}_0^2|\mathbf{x})}{L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x})} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{-(n_1+n_2)/2} = \left(\frac{\sum_{i=1}^{n_1}(y_{i1}-\overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{i2}-\overline{\overline{y}})^2}{\sum_{i=1}^{n_1}(y_{i1}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2}-\bar{y}_2)^2}\right)^{-(n_1+n_2)/2}. \tag{22.7}$$

This is the ratio, SS_{total}, of the variation of individuals observations from the grand mean and $SS_{residuals}$. the variation of these observations from the mean of its own groups.

Step 5. Simplify the likelihood statistic to determine the test statistic F. Traditionally, the likelihood ratio is simplified by looking at the differences of these two types of variation, the numerator in (22.7)

$$SS_{\text{total}} = \sum_{i=1}^{n_1}(y_{i1}-\overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{i2}-\overline{\overline{y}})^2$$

and the denominator in (22.7)

$$SS_{\text{residuals}} = \sum_{i=1}^{n_1}(y_{i1}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2}-\bar{y}_2)^2$$

Exercise 22.8. *Show that* $SS_{\text{total}} - SS_{\text{residuals}} = n_1(\bar{y}_1-\overline{\overline{y}})^2 + n_2(\bar{y}_2-\overline{\overline{y}})^2$.

In words, SS_{total} the sums of squares of the differences of an individual observation from the overall mean $\overline{\overline{y}}$, is the sum of two sources. The first is the sums of squares of the difference of the average of each group mean and the overall mean,

$$SS_{\text{between}} = n_1(\overline{\overline{y}}-\bar{y}_1)^2 + n_2(\overline{\overline{y}}-\bar{y}_2)^2.$$

The second is the sums of squares of the difference of the individual observations with its own group mean, $SS_{\text{residuals}}$. Thus, we can write

$$SS_{\text{total}} = SS_{\text{residual}} + SS_{\text{between}}$$

Now, the likelihood ratio (22.7) reads

$$\Lambda(\mathbf{y}) = \left(\frac{SS_{\text{residual}} + SS_{\text{between}}}{SS_{\text{residuals}}}\right) = \left(1 + \frac{SS_{\text{between}}}{SS_{\text{residuals}}}\right)^{-(n_1+n_2)/2}$$

Due to the negative power in the exponent, the critical region $\Lambda(\mathbf{y}) \le \lambda_0$ is equivalent to

$$\frac{SS_{\text{between}}}{SS_{\text{residuals}}} = \frac{n_1(\overline{\overline{y}}-\bar{y}_1)^2 + n_2(\overline{\overline{y}}-\bar{y}_2)^2}{(n_1-1)s_1^2 + (n_2-1)s_2^2} \ge c \tag{22.8}$$

for an appropriate value c. The ratio in (22.8) is, under the null hypothesis, a multiple of an F-distribution. The last step to divide both the numerator and denominator by the degrees of freedom. Thus, we see, as promised, we reject if the F-statistics is too large, i.e., the variation between the groups is sufficiently large compared to the variation within the groups.

Exercise 22.9 (pooled two-sample t-test)**.** *For an α level test, show that the test above is equivalent to*

$$|T(\mathbf{y})| > t_{\alpha/2, n_1+n+2}.$$

where

$$T(\mathbf{y}) = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}.$$

and s_p is the standard deviation of the data pooled into one sample.

$$s_p^2 = \frac{1}{n_1+n_2-2}\left((n_1-1)s_1^2 + (n_2-1)s_2^2\right)$$

Exercise 22.10. *Generalize the formulae for $\overline{\overline{y}}$, SS_{between} and $SS_{\text{residuals}}$ from the case $q = 2$ to an arbitrary number of groups.*

Thus, we can use the two-sample procedure to compare any two of the three groups. For example, to compared the never logged forest plots to those logged 8 years ago., we find the pooled variance

$$s_p^2 = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2) = \frac{1}{19}(11 \cdot 5.065^2 + 8 \cdot 5.761^2) = 28.827$$

and $s_p = 5.37$. Thus, the t-statistic

$$t = \frac{23.750 - 15.778}{5.37\sqrt{\frac{1}{12} + \frac{1}{9}}} = 7.644.$$

```
> 1-pt(7.644,19)
[1] 1.636569e-07
```

Thus, the p-value at 1.64×10^7 is strong evidence against the null hypothesis.

22.5 Kruskal-Wallis Rank-Sum Test

The Kruskal-Wallis test is an alternative to one-way analysis of variance in much the same way that the Wilcoxon rank-sum test is a alternative to two-sample t procedures. Like the Wilcoxon test, we replace the actual data with their ranks. This non-parametric alternative obviates the need to use the normal distribution arising from an application of the central limit theorem. The H test statistic has several analogies with the F statistic. To compute this statistic:

- Replace the data $\{y_{ij}, 1 \le i \le n_j, 1 \le j \le q\}$ for n_i observations for the i-th group from each of the q groups with $\{r_{ij}, 1 \le i \le n_j, 1 \le j \le q\}$, the ranks of the data taking all of the groups together. For ties, average the ranks.
- The total number of observations $n = n_1 + \cdots + n_q$.
- The average rank within the groups

$$\bar{r}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{ij}, \quad i = 1, \ldots, q.$$

- The grand average of the ranks

$$\overline{\overline{r}} = \frac{1}{n}(1 + \cdots + n) = \frac{1}{n}n(n+1) = \frac{n+1}{2}.$$

  (See Exercise 20.6.)

- The Kruskal-Wallis test statistic looks at the sums of squares of ranks between groups and the total sum of squares of ranks

$$H = \frac{SSR_{\text{between}}}{SSR_{\text{total}}/(n-1)} = \frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \overline{\overline{r}})^2}{\sum_{i=1}^{q}\sum_{j=1}^{n_i}(r_{ij} - \overline{\overline{r}})^2/(n-1)},$$

- For larger data sets (each $n_i \ge 5$), the p-value is approximately the probability that a χ^2_{q-1} random variable exceeds the value of the H statistic.
- For smaller data sets, more sophisticated procedures are necessary.
- The test can be followed by using a procedure analogous to contrasts based on the Wilcoxon rank-sum test.

Exercise 22.11. *For the case of no ties, show that*

$$SSR_{\text{total}} = \frac{(n-1)n(n+1)}{12}$$

In this case,

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{q} n_i \left(\bar{r}_i - \frac{n+1}{2} \right)^2 = \frac{12}{n(n+1)} \sum_{i=1}^{q} n_i \bar{r}_i^2 - 3(n+1).$$

The Kruskal-Wallis test also gives a very small p-value to the queen development times for Africanized honey bees. Begin with the R commands in Example 22.4 to enter the data and create the temperature factors `ftemp`.

```
> kruskal.test(ehb~ftemp)

Kruskal-Wallis rank sum test

data:  ehb by ftemp
Kruskal-Wallis chi-squared = 20.4946, df = 2, p-value = 3.545e-05
```

22.6 Answer to Selected Exercises

22.2. Let's look at this difference for each of the groups.

$$\sum_{j=1}^{n_i} (y_{ij} - \overline{\overline{y}})^2 - \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{j=1}^{n_i} \left((y_{ij} - \overline{\overline{y}})^2 - (y_{ij} - \bar{y}_i)^2 \right)$$
$$= \sum_{j=1}^{n_i} (2y_{ij} - \overline{\overline{y}} - \bar{y}_i)(-\overline{\overline{y}} + \bar{y}_i) = n_i(2\bar{y}_i - \overline{\overline{y}} - \bar{y}_i)(-\overline{\overline{y}} + \bar{y}_i) = n_i(\bar{y}_i - \overline{\overline{y}})^2$$

Now the numerator in (22.7) can be written to show the decomposition of the variation into two sources - the within group variation and the between group variation.

$$\sum_{i=1}^{n_1} (y_{i1} - \overline{\overline{y}})^2 + \sum_{i=1}^{n_2} (y_{i2} - \overline{\overline{y}})^2 = \sum_{i=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{2j} - \bar{y}_2)^2 + n_1(\overline{\overline{y}} - \bar{y}_1)^2 + n_2(\overline{\overline{y}} - \bar{y}_2)^2.$$
$$= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + n_1(\overline{\overline{y}} - \bar{y}_1)^2 + n_2(\overline{\overline{y}} - \bar{y}_2)^2.$$

22.3. Here, we are looking for a confidence interval for $\mu_1 - \mu_3$. From the summaries, we need

$$n_1 = 12, \quad \bar{y}_1 = 23.750, \quad n_3 = 9, \quad \bar{y}_3 = 17.778.$$

From the computation for the test, we have $s_{\text{residual}} = \sqrt{27.4} = 5.234$ and using the `qt(0.975,30)` command we find $t_{0.025,30} = 2.042$. Thus,

$$(\bar{y}_1 - \bar{y}_3) \qquad \pm t_{(0.975,30)} s_{\text{residual}} \sqrt{\frac{1}{n_1} + \frac{1}{n_3}}$$
$$= (23.750 - 17.778) \pm 2.042 \cdot 5.234 \sqrt{\frac{1}{12} + \frac{1}{9}}$$
$$= \qquad 5.972 \qquad \pm 2.079 = (3.893, 8.051)$$

22.7. This follows from the fact that expectation is a linear functional and the generalized Pythagorean identity for the variance of a linear combination of independent random variables.

22.8. Look at the solution to Exercise 22.2.

22.9. We will multiply the numerator in (22.8) by $(n_1 + n_2)^2$ and note that $(n_1 + n_2)\overline{\overline{y}} = n_1\bar{y}_1 + n_2\bar{y}_2$. Then,

$$\begin{aligned}(n_1+n_2)^2(n_1(\overline{\overline{y}}-\bar{y}_1)^2+n_2(\overline{\overline{y}}-\bar{y}_2)^2 &= n_1((n_1+n_2)\overline{\overline{y}}-(n_1+n_2)\bar{y}_1)^2+n_2((n_1+n_2)\overline{\overline{y}}-(n_1+n_2)\bar{y}_2)^2\\ &= n_1(n_1\bar{y}_1+n_2\bar{y}_2-(n_1+n_2)\bar{y}_1)^2+n_2(n_1\bar{y}_1+n_2\bar{y}_2-(n_1+n_2)\bar{y}_2)^2\\ &= n_1(n_2(\bar{y}_2-\bar{y}_1))^2+n_2(n_1(\bar{y}_1-\bar{y}_2))^2\\ &= (n_1n_2^2+n_2n_1^2)(\bar{y}_1-\bar{y}_2)^2 = n_1n_2(n_1+n_2)(\bar{y}_1-\bar{y}_2)^2\end{aligned}$$

Consequently

$$(n_1(\overline{\overline{y}}-\bar{y}_1)^2+n_2(\overline{\overline{y}}-\bar{y}_2)^2 = \frac{n_1n_2}{n_1+n_2}(\bar{y}_1-\bar{y}_2)^2 = (\bar{y}_1-\bar{y}_2)^2/\left(\frac{1}{n_1}+\frac{1}{n_2}\right).$$

The denominator

$$\sum_{j=1}^{n_1}(y_{i1}-\bar{y}_1)^2+\sum_{j=1}^{n_2}(y_{i2}-\bar{y}_2)^2 = (n_1+n_2-2)s_p^2.$$

The ratio

$$\frac{SS_{\text{between}}}{SS_{\text{residuals}}} = \frac{(\bar{y}_1-\bar{y}_2)^2}{(n_1+n_2-2)s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \frac{T(\mathbf{y})^2}{n_1+n_2-2}.$$

Thus, the test is a constant multiple of the square of the t-statistic. Take the square root of both sides to create a test using a threshold value for $|T(\mathbf{y})|$ for the critical region.

22.10. For observations, $y_{i1},\dots y_{in_i}$ in group $i = 1,\dots,q$, let $n = n_1 + \cdots + n_q$ be the total number of observations, then the grand mean

$$\overline{\overline{y}} = \frac{1}{n}\left(n_1\bar{y}_1+\cdots+n_q\bar{y}_q\right)$$

where $\bar{y}_i$ is the sample mean of the observations in group i. The sums of squares are

$$SS_{\text{between}} = \sum_{i=1}^{q} n_i(\bar{y}_i-\overline{\overline{y}})^2 \quad \text{and} \quad SS_{\text{residuals}} = \sum_{i=1}^{q}(n_i-1)s_i^2$$

where s_i^2 is the sample variance of the observations in group i.

22.11. In anticipation of its need, let's begin by showing that

$$\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6}.$$

Notice that the formula holds for the case $n = 1$ with

$$1^2 = \frac{1(1+1)(2\cdot 1+1)}{6} = \frac{6}{6} = 1.$$

Now assume that the identity holds for $n = k$. We then check that it also holds for $n = k + 1$

$$\begin{aligned}1^2 + 2^2 + \cdots + k^2 + (k+1)^2 &= \frac{k(k+1)(2k+1)}{6} + (k+1)^2 \\ &= \frac{k+1}{6}(k(2k+1) + 6(k+1)) = \frac{k+1}{6}(2k^2 + 7k + 6) \\ &= \frac{(k+1)(k+2)(2k+3)}{6}\end{aligned}$$

This is the formula for $n = k + 1$ and so by the mathematical induction, we have the identity for all non-negative integers.

With no ties, each rank appears once and

$$\begin{aligned}SSR_{\text{total}} &= \sum_{j=1}^{n}\left(j - \frac{n+1}{2}\right)^2 = \sum_{j=1}^{n} j^2 - 2\sum_{j=1}^{n} j\frac{n+1}{2} + \sum_{j=1}^{n}\left(\frac{n+1}{2}\right)^2 \\ &= \frac{n(n+1)(2n+1)}{6} - 2\frac{n(n+1)}{2}\frac{n+1}{2} + n\left(\frac{n+1}{2}\right)^2 \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} \\ &= \frac{n(n+1)}{12}(2(2n+1) - 3(n+1)) = \frac{(n-1)n(n+1)}{12}.\end{aligned}$$

Appendix A: A Sample R Session

The purpose of this appendix is to become accustomed to the R and the way if responds to line commands. R can be downloaded from

`http://cran.r-project.org/`

Be sure to download the version of R corresponding to your operating system - linux, MacOS, or windows.

As you progress, you will learn the statistical ideas behind the commands that ask for graphs or computation. Note that R only prints output when it is requested. On a couple of occasions, you will see a plus (+) sign at the beginning of the line. This is supplied by R and you will not type this on your R console.

- Learn how to access help. Type

  ```
  > help.start()
  ```

  to access on-line manuals, references, and other material.

- Find fundamental constants

  ```
  > pi
  > exp(1)
  ```

- The `<-` is used to indicate an assignment. Type

  ```
  > x<-rnorm(50)
  > length(x)
  > hist(x)
  > mean(x)
  > sd(x)
  > summary(x)
  ```

- To see what the `sort` command does type

  ```
  > ?sort
  ```

  Next, sort the values in `x`, first in increasing and then decreasing order.

  ```
  > sort(x)
  > sort(x,decreasing=TRUE)
  ```

  The first command gives 50 independent standard normal random variables and stores them as a vector `x`. It then gives the number of entries in `x`, creates a histogram, computes the mean and standard deviation, and gives a summary of the data. The last two commands give the values of `x` sorted from bottom to top and then from top to bottom

- To prepare for a scatterplot, enter

```
> (y<-rnorm(x))
```

  This gives 50 additional independent standard normal random variables and stores them as a vector `y`. When the command is placed in parentheses, R prints out the value of the variable.

- To make a scatterplot of these data, type

```
> plot(x,y)
```

  A graphics window will appear automatically.

- To find the correlation between `x` and `y`.

```
> cor(x,y)
```

- To preform a t-test, type

```
> t.test(x,y)
> t.test(x,y,alternative="greater")
```

  Notice the difference in `p-value`.

- To check to see what is in your workspace, type

```
> ls()
```

- To remove a variable `x`

```
> rm(x)
```

  Now type `ls()` again to see that `x` has been removed.

- To make a variety of graphs of $\sin(\theta)$

```
> theta<-seq(0,2*pi,length=100)
> plot(theta,sin(theta))
> par(new=TRUE)
> plot(theta,sin(theta),type="h")
> plot(theta,sin(theta),type="l")
> plot(theta,sin(theta),type="s")
> theta<-seq(0,2*pi,length=10)
> plot(theta,sin(theta),type="l")
> plot(theta,sin(theta),type="b")
```

  To see what these commands mean, type

```
> help(plot)
```

- To make some simple arithmetic and repeating sequences, type

```
> c(1:25)
> seq(1,25)
> seq(25,1,-1)
> seq(1,25,2)
> seq(1,25,length=6)
> seq(0,2,0.1)
> rep(0,25)
> rep(1,25)
```

- Make a vector of integers from 1 to 25

```
> n<-c(1:25)
```

- Choose 10 without replacement.

```
> sample(n,10)
```

- Choose 30 with replacement.

```
> samp<-sample(n,30,replace=TRUE)
> samp
```

- Turn this into a 10×3 matrix and view it.

```
>(A<-matrix(samp,ncol=10))
>(B<-matrix(samp,nrow=3))
```

  Notice that these give the same matrix. The entries are filled by moving down the columns from left to right.

- Check the dimension.

```
> dim(A)
```

- View it as a spreadsheet.

```
> fix(A)
```

  You will need to close the window before entering the next command into R.

- Find the transpose.

```
> t(A)
```

- View the first row.

```
> A[1,]
```

  the second, third and fourth column,

```
> A[,2:4]
```

  all but the second, third and fourth column,

```
> A[,-(2:4)]
```

and the 1,4 entry

```
> A[1,4]
```

- Turn this into a 3×10 matrix.

```
> matrix(samp,ncol=3)
```

- Make a segmented bar plot of these numbers.

```
> barplotdata<-matrix(samp,nrow=3)
> barplot(barplotdata)
```

- Perform a chi-squared test..

```
> chisq.test(barplotdata)
```

- Make a column of weight vectors equal to the square root of `n`.

```
> w<-sqrt(n)
```

- Simulate some response variables, and display them in a table.

```
> r<- n + rnorm(n)*w
> data.frame(n,r)
```

- Create a regression line, display the results, create a scatterplot, and draw the regression line on the plot in red.

```
> regress.rn<-lm(r~n)
> summary(regress.rn)
> plot(n,r)
> abline(regress.rn,col="red")
```

Note that the order of `r` and `n` for the regression line is reversed from the order in the plot.

- Plot the residuals and put labels on the axes.

```
> plot(fitted(regress.rn), resid(regress.rn),xlab="Fitted values",
+ ylab="Residuals",main="Residuals vs Fitted")
```

- Simulate 100 tosses of a fair coin and view the results

```
> x<-rbinom(100,1,0.5)
> x
```

Next, keep a running total of the number of heads, plot the result with steps (`type = "s"`)

```
> c<-cumsum(x)
> plot(c,type="s")
```

- Roll a fair dice 1000 times, look at a summary, and make a table.

```
> fair<-sample(c(1:6),1000,replace=TRUE)
> summary(fair)
> table(fair)
```

- Roll a biased dice 1000 times, look at a summary, and make a table.

```
> biased<-sample(c(1:6),1000,replace=TRUE,prob=c(1/12,1/12,1/12,1/4,1/4,1/4))
> summary(biased)
> table(biased)
```

- The next data set arise from the famous Michaelson-Morley experiment. To see the data set, type

```
> morley
```

  There are five experiments (column `Expt`) and each has 20 runs (column `Run`) and `Speed` is the recorded speed of light minus 290,000 km/sec.

- The data in the first two columns are labels, type

```
> morley$Expt <- factor(morley$Expt)
```

  so that the experiment number will be a factor

- Now make a labeled boxplot of the speed in column 3

```
> boxplot(morley[,3]~morley$Expt,main="Speed of Light Data", xlab="Experiment",
+ ylab="Speed")
```

- Perform an analysis of variance to see if the speed are measured speeds are significantly different between experiments. .

```
> anova.mm<-aov(Speed~Expt,data=morley)
> summary(anova.mm)
```

- Draw a cubic.

```
> x<-seq(-2,2,0.01)
> plot(x,x^3-3*x,type="l")
```

- Draw a bell curve.

```
> curve(dnorm(x),-3,3)
```

- Look at the probability mass function for a binomial distribution.

```
> x<-c(0:100)
> prob<-dbinom(x,100,0.5)
> plot(x,prob,type="h")
```

- To plot a parameterized curve, start with a sequence and give the x and y values.

```
> angle<-seq(-pi,pi,0.01)
> x<-sin(3*angle)
> y<-cos(4*angle)
> plot(x,y,type="l")
```

The `type ="l"` (the letter ell, not the number one) command connects the values in the sequence with lines.

- Now we will plot contour lines and a surface. First, we give a sequence of values. This time we specify the number of terms.

```
> x<-seq(-pi, pi, len=150)
> y<-x
```

Then, we define a function for these x and y values and draw a contour map. Then, choose the number of levels.

```
> f<-outer(x,y,function(x,y) cos(y)/(1+x^2))
> contour(x,y,f)
> contour(x,y,f,nlevels=20)
```

- For a color coded "heat map",

```
> image(x,y,f)
```

- To draw a surface plot,

```
> persp(x,y,f,col="orange")
```

and change the viewing angle

```
> persp(x,y,f,col="orange",theta=-30, phi=45)
```

Index

R command
- D, 117
- abline, 40, 59, 398
- abs, 171, 199
- anova, 385
- aov, 399
- array, 7, 68
- atan, 278
- barplot, 7, 369, 378
- beta, 157, 162
- binom.test, 326
- binom, 149, 160
- boxplot, 24, 29, 65, 350, 385
- chisq.test, 369, 370
- chisq, 159, 162
- chol2inv, 61
- choose, 87, 88, 92, 124, 145
- colnames, 9, 65
- colors, 6
- cor, 36
- cumsum, 124, 169, 171, 178, 290
- curve, 113, 156, 399
- data.frame, 40, 85, 92, 115, 128, 188
- dbinom, 161
- det, 61
- dgamma, 257
- dgeom, 115
- dim, 397
- dpois, 199
- eval, 113
- expression, 113
- exp, 155, 162
- factor, 385
- fisher.test, 376
- fix, 397
- for, 85, 128, 174, 176, 223, 226
- function, 173, 178, 400
- f, 160, 162
- gamma, 156, 162
- geom, 150, 160
- gnls, 51
- head, 28
- help.start, 395
- help, 396
- hist, 10, 140, 221, 344
- hyper, 153, 160, 250, 311
- integrate, 173, 178
- kruskal.test, 392
- legend, 6
- length, 22, 129, 140
- lines, 123
- lm, 40, 41, 46, 49, 55, 58, 252, 385, 398
- lnorm, 162
- log, 45, 55
- matrix, 377, 397
- mean, 22, 140, 173, 174, 176
- mfrow, 58, 178
- nbinom, 150
- nmle, 51
- norm, 158, 162, 169, 266
- numeric, 89, 174
- order, 290
- outer, 179, 400
- par, 122, 129, 140, 178
- pbinim, 325
- pbinom, 188, 334
- pchisq, 332, 369, 371
- persp, 400
- pexp, 113
- pf, 384
- pgeom, 115
- phyper, 320
- pie, 6
- pi, 395
- plot, 12, 14, 55, 122, 124, 129, 169, 171, 178, 250
- pnorm, 187, 188, 199, 294, 309, 325
- points, 28
- pois, 151
- power.prop.test, 329, 337
- power.t.test, 346
- ppois, 199
- predict, 40, 59
- prop.test, 316, 328, 329
- qbeta, 280

qbinom, 334
qf, 384
qhyper, 311
qnorm, 199, 292, 309, 326
qqnorm, 140
qqplot, 29
qt, 267, 280, 344, 386
quantile, 27, 28, 199, 279
rbeta, 177
rbinom, 229
rcauchy, 171, 178
replicate, 89
rep, 85, 128, 174, 176, 199, 221, 223
reside, 59, 398
residuals, 374
resid, 40, 46, 56
rm, 396
rnorm, 36, 158
round, 6
rownames, 9, 65
runif, 122, 155, 173, 176, 178, 179, 199, 221
sample, 67, 120, 128, 152, 161, 223, 288, 397
scale, 28
sd, 25, 140, 174, 176
seq, 188
sort, 12, 21, 29, 129
sqrt, 122, 173
summary, 24, 46, 58, 176, 252
sum, 22, 92, 120, 128, 145, 223
t.test, 344, 348, 351
table, 121, 398
t, 61, 160, 162
unif, 155, 162
var, 25, 174
weighted.mean, 23
wilcox.test, 360

addition principle, 81
allele, 72
  dominant, 72
  recessive, 72
alternative hypothesis, 287
analysis of variance, 381
Archaeopteryx, 252
Archeopteryx, 13, 36
area under the curve, 291
asymptotic normality, 260
  maximum likelihood estimator, 256
AUC, 291
axioms of probability, 80

balancing selection, 10
bar graph, 7
  segmented, 9
Bayes factor, 100
Bayes formula, 98, 208, 319
Bayesian statistics, 208
  hypothesis testing, 300
  interval estimation, 279
Bernoulli random variable, 149, 160, 239
Bernoulli trials, 134, 188, 196, 206, 210, 249, 295, 356
best region, 288
beta function, 157
beta random variable, 157, 162, 208, 210
bias, 222, 229
bias-variance decomposition, 230
binomial random variable, 135, 149, 160, 185
binomial test, 326
binomial theorem, 87, 135
bird fecundity, 193
birthday problem, 85
blood type, 365, 369
Bonferroni correction, 317
Bonferroni inequality, 83, 317
bootstrap, 278
Borneo forests, 381
boxplot, 24, 382
Bradley effect, 66

categorical data, 4
categorical variable, 3
Cauchy random variable, 171
Cauchy-Schwarz inequality, 34
causality, 63
census, 63
central limit theorem, 184, 195, 233
central moments, 141
chain rule, 96
characteristic function, 142
chi-squared random variable, 159, 162, 236
chi-squared test, 330
  contingency table, 371
  likelihood ratio, 330
  two-way table, 371
chiasma, 331
choose function, 86
chromotid, 331
Cochran conditions, 375
coefficient of determination, 43
coefficient of linear expansion, 190, 192
coefficient of volume expansion, 193
combination, 86
complement, 82, 94

complement rule, 82
composite hypothesis, 307
    one-sided test, 307, 308, 310
    two-sided test, 307
conditional distribution, 64
conditional probability, 95
confidence interval, 208, 265, 347
    analysis of variance, 384
    delta method, 276
    interpretation, 274
    linear regression slope, 272
    matched pair t, 270
    maximum likelihood estimation, 277
    sample proportion, 272
    summary, 273
    t, 267
    two sample proportions, 273
    two sample t, 270
    two sample z, 270
    two sided test, 347
    z, 266
confidence level, 265
confounding variable, 63
conjugate pair, 214
conservative procedure, 350
consistency, 237, 260
    maximum likelihood estimator, 256
contingency table, 7, 371
    chi-squared test, 371
    Cochran conditions, 375
continuity correction, 187
    proportion test, 325
continuity property, 83
continuous random variable, 112, 117, 136, 140
contrasts, 386
controlled experiment, 65
convex function, 233
correlation, 34
    distributional, 143
    simulated, 36
covariance, 33
    distributional, 143
covariance matrix, 144
covariate, 37
Cramér-Rao bound, 238, 239
credible interval, 279
critical region, 288
    best, 288
    most powerful, 288
cross tabulation, 63, 371
cross tabulation table, 7
cross tabulation table , 7
cumulative distribution function, 110
Current Population Survey, 69

de Morgan's law, 84
de Morgan, August, 131
degrees of freedom, 267, 343
    contingency table, 374
    denominator for F statistic, 383
    goodness of fit, 369
    numerator for F statistic, 383
delta method, 192, 196, 222, 224, 357
    confidence interval, 276
    multidimensional, 196
density
    posterior, 208
    prior, 208
density function, 117, 207
    joint, 119, 120
    marginal, 120
design matrix, 52
difference rule, 82
differentiation, 117
digamma function, 257
discrete random variable, 111, 131, 140
discrete uniform random variable, 152, 160
discriptor variable, *see* explanatory variable
distribution
    conditional, 64
    joint, 119
    Maxwell-Boltzmann, 159
distribution function, 110
distribution of fitness effects, 224, 257
dominant allele, 72
Doob, Joseph, 109
double blind procedure, 66
double reciprocal plot, 49
Doyle, Arthur Conan, 169
Dravet sydome, 28

effective degrees of freedom, 271, 350
efficiency, 238, 260
    maximum likelihood estimator, 256
element, 80, 94
empirical cumulative distribution function, 11
    survival function, 12
equally likely outcomes, 80, 95, 153
estimate, 206
estimator, 206
    interval, 208
    maximum likelihood, 232, 247

method of moments, 219
point, 208
summary of maximum likelihood, 255
unbiased, 190, 230
ethical guidelines, 64
event, 79, 94
expected value, 131, 133
continuous random variable, 136
discrete random variable, 134
experiment, 65
controlled, 65
Heshey-Chase, 73
Michelson-Morley, 71
natural, 68
explanatory variable, 13, 37, 63
exploratory data analysis, 63
exponential family, 241
exponential random variable, 113, 123, 137, 142, 155, 162, 186, 189, 240, 332
extrapolation, 41

F random variable, 160, 162, 384
F test, 383
factor, 3, 66, 371
factorial moment, 141
falling factorial, 85
falling function, 85
false discovery rate, 318
false negative, 100, 288
false positive, 101, 287
false positive fraction, 290
familywise error rate, 317
first quartile, 24, 31
Fisher information, 239
matrix, 258
Fisher's exact test, 375
Fisher's method, 318
Fisher, Ronald, 381
fit, 38
fitness, 224
five number summary, 24
floor function, 10, 110
Fourier transform, 142
fundamental principle of counting, 79, 84

G square test statistic, 368
Galton, Sir Francis, 33
gamma function, 156
gamma random variable, 156, 162, 208, 224, 257
Garfield, James A., 63
GenBank, 70
gene ontology, 5
generating function
probability, 142
genetic recombination, 330
geometric random variable, 115, 150, 160
geometric sequence, 115
Global Health Observatory, 69
goodness of fit, 365
google, 109
Gossett, William Sealy, 267
gradient descent, 51
grand mean, 382, 389
Gusset, William Sealy, 343

hanging chi-gram, 369, 374, 378
Hardy-Weinberg principle, 370
hat matrix, 53
Hershy-Chase experiment, 73
heteroscedasticity, 40
heterozygosity, 233
histogram, 10
homoscedasticity, 40
honey bees, 66, 67, 316, 326, 385
hybridization, 291
hypergeometric random variable, 97, 153, 160, 223, 235, 333
hypothesis, 287
alternative, 287
Bayesian approach, 300
composite, 307
null, 287
simple, 287, 318

ideal gas, 159
importance sampling, 176
function, 176
weights, 176
INCAP, 75
inclusion-exclusion rule, 83
independence, 102
random variable, 143
independent random variables, 119
individual, 3
information, 214
observed, 256
information inequality, 239
input variable, *see also* explanatory variable
Intergovernmental Panel on Climate Change, 17
interquartile range, 24, 31
intersection, 94
interval estimation, 265

Bayesian, 279
interval estimator, 208

Jensen's inequality, 233
joint density, 207
joint density function, 119, 120
joint distribution, 119
joint mass function, 143
joint probability density function, 119

Kölreute, Josef Gottlieb, 72
Karlin-Rubin theorem, 324, 355
Kolmogorov, Andrey, 79
Kruskal-Wallis test, 391
kurtosis, 141

Lagrange multipliers, 366
Laplace transform, 142
law of blending, 72
law of cosines, 34
law of independent assortment, 103
law of large numbers, 169, 219, 237
law of mass action, 47
law of segregation, 103
law of total probability, 97, 118
least squares regression, *see* linear regression
leptokurtic, 141
level, 3, 66
likelihood function, 207
likelihood ratio, 366
likelihood ratio test, 289, 327, 382
chi-squared test, 330
t test, 357
two sided test, 327
Lincoln-Peterson method, 222, 235, 249
linear regression, 37, 206, 250, 251
fit, 38
heteroscedasticity, 40
homoscedasticity, 40
multiple, 51
nonlinear regression, 51
residual, 38
slope confidence interval, 272
weighted least squares, 252
Lineweaver-Burke plot, 49
link function, 45
linkage disequilibrium, 104
log likelihood surface, 256
log-normal random variable, 158
log-odds, 245
logarithm, 16, 45
lurking variable, 63

malaria, 377
margin of error, 188, 265, 269
marginal density, 207
marginal density function, 120
marginal distribution, 9, 372
two-way table, 9
marginal mass function, 119
mark and recapture, 222, 235, 249, 310, 376
Lincoln Peterson method, 311
power function, 311
two-sided test, 312
mass function, 114
joint, 143
marginal, 119
matched pair procedures
t interval, 270
t test, 348
matrix, 51
determinant, 52
inverse, 52
transpose, 52
maximum likelihood estimation
confidence interval, 277
maximum likelihood estimator, 232, 247
asymptotic properties, 255
efficiency, 256
summary, 255
Maxwell-Boltzmann distribution, 159
mean, 131
distributional, 141
grand, 382
posterior, 210
sample, 21, 31, 185, 196
trimmed, 23
truncated, 23
weighted, 22
within group, 382
mean square error, 230
median, 21, 31
memorylessness, 115, 156
Mendel, Gregor, 72, 103, 273, 370
method of moments, 171, 219, 259
method of steepest descent, 51
Michaelis constant, 48
Michaelis-Menten kinetics, 47
Michaelis-Metens kinetic equation, 48
Michelson-Morley experiment, 71
Millennium Development Goals, 69
mimic, 291
mixture, 118, 319
continuous, 208

model, 291
    parametric, 64
    probability, 79
moment generating function, 142
moments, 141
    method of, 219
monotonicity rule, 83
Monte Carlo integration, 172
Morgan, Thomas Hunt, 73
    genetic recombination, 330
mosquitoes, 352
multidimensional
    Fisher information, 256
multidimensional delta method, 196
multiple regression, 51
multiplication principle, 96
muon, 222
mutually exclusive, 81

National Foundation for Infantile Paralysis, 74
natural experiment, 68
natural parameter, 241
negative binomial random variable, 150, 160
neutrino, 222
Neyman, Jerzy, 265, 287
Neyman-Pearson lemma, 288, 298
non-parametric tests, 360
nonlinear regression, 51
normal equations, 53
normal probability plot, 140
normal random variable, 138, 158, 162, 186, 207, 214, 240, 249
normal score, *see* standardized version
null hypothesis, 287

Obama, Barack, 63
observational study, 64
observed information, 256
odds, 84
    posterior, 301
    prior, 301
oil production, 15, 45
one-sided test, 307, 308
    power function, 310
    uniformly most powerful, 324
order statistic, 21, 157, 253
outcome, 79, 94
oyster, 370

p-value, 315, 325
parallel axis theorem, 26
parameter space, 206
parametric model, 64
Pareto random variable, 159, 220, 233, 235, 240
partition, 97
Pascal's triangle, 88
Pearson residual, 369
Pearson, Egon Sharpe, 287
Pearson, Karl, 365, 368
permutation, 85
permutation test, 359
pie chart, 4
pion, 222
placebo effect, 66
platykurtic, 141
plot
    residual, 40
Pochhammer symbol, 85
point estimation, 265
point estimator, 208
Poisson random variable, 151, 160, 189, 199, 242
Polya, George, 183
polynomial regression, 54
pooled two-sample t procedure, 390
pooled two-sample t procedures, 354
population, 3, 63
population means
    tests, 356
population proportions
    tests, 356
posterior density, 208
posterior mean, 210
posterior probability, 301
power, 67, 288
    t test, 346
power function, 307
    lilelihood ratio, 332
    Lincoln Peterson method, 311
    mark and recapture, 311
    one-sided test, 310
    proportion test, 326
    simulated, 332
    t test, 345
    two-sided test, 312
power law distribution, 159
predictor variable, *see also* explanatory variable
prior density, 208
prior probability, 300
probability, 79
    axioms, 80
    conditional, 95
probability density function, 117
    joint , 119

probability generating function, 142
probability mass function, 114
probability model, 79
probability transform, 122, 221
propagation of error, 189
    multidimensional, 193
propagation of uncertainty, 189
proportion, 186
    sample, 186
proportion test, 325, 356
    binomial test, 326
    continuity correction, 325
    power function, 326
proposal density, 176
Public Health Service, 74
Punnett square, 103
Punnett, Reginald, 72
Pythagorean identity, 34, 169
    distributional, 143
Pythagoreanf identity, 183

quadratic identity for variance, 142, 169
quantile, 27
quantile function, 122
quantile plot, 140
quantile-quantile plot, 28
quantitative variable, 3
quartile, 24
    first, 24, 31
    third, 24, 31

radon, 344
random sample, 67
    simple, 67
    stratified, 68
random variable, 109, 160
    Bernoulli, 149, 239
    beta, 157, 208, 210
    binomial, 135, 149, 185
    Cauchy, 171
    chi-squared, 159, 236
    continuous, 112, 117, 136, 140
    dartboard, 113, 122, 137
    discrete, 111, 131, 140
    discrete uniform, 152
    exponential, 113, 123, 137, 142, 155, 186, 189, 240, 332
    F, 160, 384
    gamma, 156, 208, 224, 257
    geometric, 115, 150
    hypergeometric, 97, 153, 223, 235, 333
    log-normal, 158
    negative binomial, 150
    normal, 138, 158, 186, 207, 214, 240, 249
    Pareto, 159, 220, 233, 235, 240
    Poisson, 151, 189, 199, 242
    t, 160, 267
    uniform, 122, 155, 206, 253
random variables
    independent, 119
receiver operating characteristic, 291, 295, 319
recessive allele, 72
regression
    polynomial, 54
reliability engineering, 212
residual, 38
    Pearson, 369
    standardized, 369
residual plot, 40, 46
residual sum of squares, 383
resistant measure, 23
response variable, 13, 37, 63
rising factorial, 85
risk, 301

Salk vaccine field trials, 74
sample, 63
sample mean, 21, 31, 185, 196
sample proportion, 186, 196, 206
sample proportion interval, 272
sample size determination, 294
sample space, 79, 94
sample standard deviation, 25, 31
sample sum, 184, 195
sample variance, 25, 31
    unbiased estimator, 230
sampling with replacement, 120, 154
sampling without replacement, 97, 120, 153, 223
scatterplot, 13
score function, 239, 247, 260
segmented bar graph, 9
sensitivity, 100, 288, 346
set theory, 94
significance level, 287, 308
Silver, Nate, 95
simple hypothesis, 287, 318
simple random sample, 67, 207
simulating discrete random variables, 120
skewness, 141, 186
smoking, 372
social desirability bias, 66
specificity, 101, 288, 346

speed of light, 71
spin, 222
standard deviation, 25, 141
    sample, 25, 31
standard error, 267, 343
standard score, *see* standardized version
standardized residual, 369
standardized score, 195, 270
standardized version, 28, 141
state of nature, 79, 207
state space, 110
statistic, 205
steepest descent, 51
Stirling approximation, 237
stratification, 63
stratified random sample, 68
studentized score, 270
Sturges' formula, 10
subject, 65
subset, 80, 94
sudden infant death syndrome, 369
summary
    confidence interval, 273
    estimators, 255
    normal approximations, 195
    properties of random variables, 161
    random variables and expectation, 140
    significance tests, 355
    simple hypothesis, 297
supercedure, 67
survival function, 12, 23, 111, 137

t interval, 267
t procedures
    guidelines, 343
    pooled two-sample, 354
    two sample, 350
t random variable, 160, 162, 267
t test, 343
    matched pair procedures, 348
    one sample, 344
    power, 346
    power function, 345
target variable, 13
telescoping sum, 90
thermal expansion, 192
third quartile, 24, 31
time plot, 15
total sum of squares, 383
transform
    probability, 122
tree diagram, 99
trigamma function, 258
true negatiive probability, 287
true positive, 100
true positive fraction, 290
truncated mean, 23
Tufte, Edward, 3
two sample proportion interval, 273
two sample t interval, 270
two sample t procedures, 350, 388
two sample z interval, 270
two sided test
    confidence interval, 347
    likelihood ratio test, 327
two-sided test, 307
two-way table, 7, 371
    chi-squared test, 371
type I error, 287, 308
type II error, 288, 308

unbiased estimator, 190, 230
    linear regression, 252
uniform random variable, 122, 155, 162, 206, 253
uniformly most powerful test, 324
union, 94
United States census, 69
universal set, 80, 94
upper tail probability, 266

variable, 3
    categorical, 3
    confounding, 63
    descriptor, 13
    explanatory, 37, 63
    input, 13
    lurking, 63
    predictor, 13
    quantitative, 3
    response, 13, 37, 63
    target, 13
variance, 25, 141
    distributional, 141
    quadratic identity, 27
    sample, 25, 31
vitamin C wheat soy blend, 348

weighted least squares, 252
weighted mean, 22
Welch-Satterthwaite equation, 271, 350
Wells, H.G., vii
Wilcoxon signed rank test, 360, 361
Wilkes, Samuel, vii

within group mean, 382
Women's Health Initiative, 70

z interval, 266
z-score, *see* standardized version
z-value, *see* standardized version