Probabilistic reasoning and statistical inference: An introduction (for linguists and philosophers)

NASSLLI 2012 Bootcamp
June 16-17

Lecturer: Daniel Lassiter
Computation & Cognition Lab
Stanford Psychology

(Combined handouts from days 1-2)

> The theory of probabilities is nothing but good sense reduced to calculation; it allows one to appreciate with exactness what accurate minds feel by a sort of instinct, without often being able to explain it. (Pierre Laplace, 1814)

> Probable evidence, in its very nature, affords but an imperfect kind of information; and is to be considered as relative only to beings of limited capacities. For nothing which is the possible object of knowledge, whether past, present, or future, can be probable to an infinite Intelligence .... But to us, probability is the very guide of life. (Bishop Joseph Butler, 1736)

0 Overview

This course is about foundational issues in probability and statistics:

- The practical and scientific importance of reasoning about uncertainty (§1)
- Philosophical interpretations of probability (§2)
- Formal semantics of probability, and ways to derive it from more basic concepts (§3)
- More on probability and random variables: Definitions, math, sampling, simulation (§4)
- Statistical inference: Frequentist and Bayesian approaches (§5)

The goal is to gain intuitions about how probability works, what it might be useful for, and how to identify when it would be a good idea to consider building a probabilistic model to help understand some phenomenon you're interested in. (Hint: almost anytime you're dealing with uncertain information, or modeling agents who are.)

In sections 4 and 5 we'll be doing some simple simulations using the free statistical software R (available at http://www.r-project.org/). I'll run them in class and project the results, and you can follow along on a laptop by typing in the code in boxes marked "R code" or by downloading the code from http://www.stanford.edu/~danlass/NASSLLI-R-code.R. The purpose of these simulations

is to connect the abstract mathematical definitions with properties of data sets that we can control (because we built the models and generated the data ourselves) and that we can inspect to check that the math makes intuitive sense. It shouldn't matter if you're not familiar with R or any other programming language, since we'll only be using very simple features and everything will be explained along the way.

There are lots of important and interesting topics in probability and statistics that we won't talk about much or at all:

- Statistical techniques used in practical data analysis (e.g. t-tests, ANOVA, regression, correlation; if we have extra time at the end we may cover the important topics of correlation and regression briefly, though.)
- The use of probabilistic models in psychology and linguistics (see Goodman's and Lappin's courses)
- Other logical representations of uncertainty and a comparison of advantages and disadvantages (see e.g. Baltag & Smets' course for some candidates)
- Machine learning and computational linguistics/NLP (see Lappin, Lopez courses)
- Measure theory (in fact, almost anything involving infinite sets or continuous sample spaces)

This course should, however, give you a foundation for exploring these more advanced topics with an appreciation for the meaning(s) of probability, an appreciation of what assumptions are being made in building models and drawing inferences, and how you could go about discerning whether these assumptions are appropriate.

1 Uncertainty and uncertain reasoning

1.1 Intuition warmup: some examples

You already know a lot about how to make smart inferences from uncertain information. If you didn't, you wouldn't be here ...

Ex. 1

Crossing the street in traffic. We've all done this: you're in a hurry, so instead of waiting for the "walk" sign you look both ways and see that the nearest cars are far enough away that you can cross safely before they arrive where you are. You start walking and (I'm guessing) make it across just fine.

Q1: Did you **know** (with absolute certainty) that the cars you saw in the distance weren't moving fast enough to hit you? If so, how did you come to know this? If not, how could you possibly justify making a decision like this, given the extremely high stakes? After all, you were literally betting your life ...

Q2: Can logic help us understand how a rational person could make a risky decision like this, despite not having perfect knowledge of all relevant factors?

The street-crossing example is chosen for the vivid consequences of making a wrong decision, but less dramatic examples (tying shoelaces, chopping vegetables) would make the point. We almost never know with absolute certainty what the consequences of our actions will be, but we usually manage to make reasonably confident decisions nonetheless — and most of the time we choose right. This needs explaining.

Ex. 2

The cop and the man in the window. You're a police officer out on patrol late at night. You hear an alarm go off and follow the sound to a jewelry store. When you arrive, you see a broken window and a man crawling out of it wearing black clothes and a mask, carrying a sack which turns out to be full of jewelry. (Jaynes 2003: ch.1)

Q1: What will you conclude?

Q2: Can you find a way to justify this conclusion using logic (without pretending to have certain knowledge that you don't actually have)?

Q3: The man says he's the owner, has just returned from a costume party where he was dressed as a burglar, couldn't find his keys when he got home, broke the window to get in, and then realized he'd better clear out the stock so that someone else doesn't crawl in through the broken window and take it. Is this plausible? Why or why not? What would you (the cop) do at this point?

Suppose we wanted a logic that would explain how to justify as rational the decision to cross the street or the cop's judgment about the honesty of the man in the window. What would that logic need to look like? In other words, what formal tools do we need to understand rational inference and rational decision-making in the presence of uncertainty?

Ex. 3

Medical diagnosis #1. Suppose we observe a person coughing, and we consider three hypotheses as explanations: the person has a cold (h_1), lung disease (h_2), or heartburn (h_3). (Tenenbaum, Kemp, Griffiths & Goodman 2011)

Q1: Which of these hypotheses is most reasonable?

Q2: Can you explain the intuitive basis of this judgment?

Q3: Consider the following simple theory: information is represented as a set of possibilities. Inferences from information gain proceed by eliminating possibilities incompatible with the evidence you have, and drawing conclusions that follow logically from the updated information state (i.e., conclusions that are true in every remaining possibility). What would such a theory of inference predict about the status of h_1-h_3? What kind of assumptions would you need to add to the theory in order to get the intuitively correct result?

Ex. 4

Medical diagnosis #2. A particular disease affects 300,000 people in the U.S., or about 1 in 1,000. There is a very reliable test for the disease: on average, if we test 100 people that have the disease, 99 will get a positive result; and if we test 100 people that do not have the disease, 99 will get a negative result. (Gigerenzer 1991)

Q1: Suppose we test 100,000 people chosen at random from the U.S. population. How many of them, on average, will have the disease? How many will not? How many of those who have the disease will test positive? How many who do not have the disease will test positive?

Q2: Suppose I test positive. How worried should I be?

Ex. 5

Hair color in Ausländia.

(a) You've just arrived in the capital city of a faraway country, Ausländia, that you don't know much about. The first person that you see has red hair. How likely is it that the second person you see will have red hair? (Please assume that there is no uncertainty about what counts as red hair.)

(b) The second, third, and fourth people you see have red hair too. How likely is is that the fifth person will?

(c) Being the fastidious person you are, you keep records. Of the 84 people you see on your first day in Ausländia, 70 have red hair. If you had to guess a number, what proportion of Ausländers would you say have red hair? Can you think of a range of proportions for the whole population that might be reasonable, given what you've observed?

(d) You stay in the capital city throughout your trip. Of the 1,012 people you see during there, 923 have red hair. What proportion of Ausländers would now you guess have red hair? What is a believable range of proportions that might be reasonable, give what you've observed?

(e) Suppose, on your return, you read that hair color is not consistent in different parts of Ausländia; in some parts most people have black hair, in some parts most have red, and in some parts most have brown. Will you revise your guess about the proportion of Ausländers who have red hair? If so, what is your new guess? If not, does anything else change about your guess?

Ex. 6

The number game.You've been teamed up with a partner who has been given a set of numbers between 1 and 100. These are the "special" numbers. The game goes like this: your partner can pick a maximum of 3 examples of special numbers, and your job is to guess what the set of special numbers is. Here are the examples he picks: 9, 49, and 81. (Tenenbaum 1999)

Q1: Is 4 special? 8? 36? 73?

Q2: What do you think the special numbers are?

Q3: Think of some alternative hypotheses that are also consistent with the examples. (There are many!) Why didn't you guess these as the special numbers? In other words, can you explain why

the answer you chose initially is a better guess, even though these data are logically consistent with various other hypotheses? (Hint: there are at least two, somewhat different reasons.)

1.2 Motivations

Doing science requires the ability to cope with uncertainty.

- Science generally: we need good procedures for uncertain inference because we want to formulate and justify scientific theories even though our data are almost always incomplete and noisy. (Data = e.g. instrumental measurements, information in dictionaries and grammars, testimony from others, or just whatever we happen to have encountered in the world).
  - Familiar **deductive logic** is great for reasoning about things that are known to be true or false, but not directly applicable to information that is merely *likely* or *plausible*.
  - To do science, we need a procedure for determining which conclusions to draw (albeit tentatively) from incomplete data and how and when to withdraw old conclusions when we get new evidence. According to some, this should take the form of an **inductive logic**. (See http://plato.stanford.edu/entries/induction-problem/.)

Ex. 7

Swans. Philosophers have worried for a long time about whether epistemically limited agents can ever know with certainty that a logically contingent universal statement is true. In early modern philosophy in Europe, an example used to make the case that we could have such knowledge was "All swans are white", a universal generalization whose truth had supposedly been established by observation of many, many white swans and no non-white swans. This was before Europeans went to Australia. When they got there, they discovered that Australian swans are black. D'oh!

- Cognitive sciences (e.g. linguistics, psychology, AI, philosophy of mind & epistemology): we need a theory of uncertain reasoning because we're trying to understand human intelligence, and much of human intelligence is about using uncertain information to make (hopefully) reasonable inferences that aid us in decision-making.

We can even up the ante by combining the two motivations: we need a theory of uncertain reasoning that will help cognitive scientists figure out which theory of reasoning best describes how humans make intelligent inferences using noisy and uncertain information.

2 What does probability mean?

On face, the apparatus of probability allows us to give content to statements like "The probability that a fair coin will come up heads is $\frac{1}{2}$" or "The probability that it will rain tomorrow is .8" or "The probability that an individual whose test for disease x is positive actually has the disease is p". But

really, we don't know what these statements mean unless we know what probabilities themselves are, and this is a matter of some controversy.

Before we get into the more technical material, it will help to have a glimpse of the major interpretations of probability, each of which gives us a different answer to the question of what probability statements are about. There are several major possibilities, not all mutually exclusive:

- Objective interpretations
  - Frequency interpretation
  - Propensity interpretation
- Bayesianism: Probability as a measure of belief/weight of evidence

There is a further "logical" interpretation of probability associated in particular with Carnap (1950). We won't discuss it, in part because it's not widely considered viable today, and in part because I don't feel like I understand it well.

Theorists' choices about how to interpret probability have numerous consequences for the material we'll see later: for example, advocates of the frequentist and Bayesian interpretations of probability tend to prefer different ways of motivating the use of probability (§3). Likewise, much of modern statistics was developed with a frequentist interpretation of probability in mind, and the recent flourishing of Bayesian methods has led to many new methods of statistical analysis and a rejection of many traditional ideas (§5).

As a running example, we'll use the well-worn but useful example of flipping a fair coin. Different philosophies of probability will give different contents to the following statements:

(1) a. The probability that a flip of this fair coin is heads is .5.
    b. The probability that the next flip of this fair coin is heads is .5.

This is a rich topic, and we'll cover it pretty briskly. See Hacking 2001; Mellor 2005 and Hájek's SOEP article "Interpretations of probability" for more detail on the debates covered in this section and further references.

2.1 Objective approaches

2.1.1 Relative Frequency

According to *frequentists*, the probability of an event is defined as the relative frequency of the event in some reference class. The meaning of "The probability that a flip of this fair coin is heads is .5" is that, if I flip the coin enough times, half of the flips will come up heads. More generally, frequentists think of probabilities as properties that can attach only to "random experiments" — experiments whose outcome can't be predicted in advance, but which can be repeated many times under the same conditions.

The frequentist interpretation of probability has the advantage of concreteness, and has sometimes been argued to be supported by evidence from cognitive psychology or quantum physics.

However, there are several problems. One is that the probability of an event becomes dependent on the choice of a reference class. Hájek puts it nicely:

> Consider a probability concerning myself that I care about — say, my probability of living to age 80. I belong to the class of males, the class of non-smokers, the class of philosophy professors who have two vowels in their surname, ... Presumably the relative frequency of those who live to age 80 varies across (most of) these reference classes. What, then, is my probability of living to age 80? It seems that there is no single frequentist answer.

Another problem is that the interpretation of probability as relative frequency can't make intuitive sense of the fact that probabilities can attach to non-repeatable events, e.g. the probability that the next flip of this fair coin will be heads or the probability that the Heat will win the 2012 NBA finals. According to the frequentist definition, the probability of an event that can only happen once is either 1 (if it happens) or 0 (if it doesn't). Some frequentists (e.g. von Mises (1957)) simply deny that probability statements about single events are meaningful. But (1b) certainly doesn't *feel* nonsensical or trivially false.

A further problem with the relative frequency interpretation is that it seems to tie probabilities too closely to contingent facts about the world. Suppose I toss a coin 50 times and get 35 heads. This could easily happen, even if the coin is fair. According to the relative frequency interpretation, the probability of heads is now .7. But we want to be able to say that the fact that more heads than tails occurred was just chance, and that it doesn't *really* make the probability of heads .7.

A variant of frequentism associated in particular with von Mises (1957) claims that the probability of heads should be identified with the relative frequency of heads in a hypothetical sequence generating by flipping the coin an infinite number of times. This helps with the puzzle just mentioned, but creates problems of its own. For instance, by rearranging the order of flips we can give the same coin any probability between 0 and 1. This approach also abandons much of the empiricist appeal of frequentism, since it ties the meaning of a probability statement to the properties of a counterfactual (what would happen if ...). This apparently makes probability statements non-verifiable in principle.

Note that much of the apparatus of mainstream statistics was developed in the heyday of frequentist interpretations of probability, and this philosophy is still adopted *de facto* in many fields that make use of statistical models. (§5)

The next objectivist theory that we'll consider was designed to deal with those problems (and some problems in quantum mechanics that we can safely ignore here).

2.1.2 Propensity

Like so many important ideas, the propensity interpretation of probability originated in the work of C.S. Pierce, but went unnoticed and was independently rediscovered later. The philosopher of science Karl Popper (e.g., 1959) is its most prominent proponent. He explains (p.30):

> Propensities may be explained as possibilities (or as measures of 'weights' of possibilities) which are endowed with tendencies or dispositions to realise themselves,

> and which are taken to be responsible for the statistical frequencies with which they will in fact realize themselves in long sequences of repetitions of an experiment.

There is an important distinction between the relative frequency and propensity interpretations, then: a fair coin has a certain propensity to land heads or tails, but this is a non-observable feature of the coin, rather than a fact about a sequence of flips of coins. The coin has this property whether or not it is ever actually flipped. However, if we flip such a coin repeatedly, on average it will come up heads 50% of the time.

Suppose I hatch a devious plan to mint a fair coin, flip it once, and then destroy it. On the frequentist interpretation, it either doesn't make sense to talk about the probability that the single flip will land heads, or the probability is trivial (1 or 0, depending on what actually happens). On the propensity interpretation, the probability is non-trivial: it is a fact about the coin and its interactions with its environment that its propensity to come up heads when flipped is the same as its propensity to come up tails. Similarly, we might think of "The probability that the Heat will win the NBA finals is .4" as describing an objective but unobservable feature of the basketball team and their environment — a propensity, attaching to the team right now, to a certain critical number of basketball games in a particular series against a particular opponent. This propensity exists regardless of who actually ends up winning.

Perhaps not accidentally, the relative frequency interpretation was dominant during the heyday of logical positivism, the doctrine that the only meaningful statements are those that are verifiable or can be reduced to statements that are verifiable. The propensity interpretation started to become popular around the same time that logical positivism started to be unpopular.

One objection that has been made to the propensity interpretation is that it is trivializing. Quoting Hájek again:[1]

> There is some property of this coin tossing arrangement such that this coin would land heads with a certain long-run frequency, say. But as Hitchcock (2002) points out, "calling this property a 'propensity' of a certain strength does little to indicate just what this property is." Said another way, propensity accounts are accused of giving empty accounts of probability, à la Molière's 'dormitive virtue' ...

2.2 Bayesianism

The "Bayesian" interpretation of probability is probably most justly attributed not to the Reverend Thomas Bayes but to Ramsey (1926) and de Finetti (1937). The basic idea is that probability is a measure of a rational agent's degree of belief in a proposition. For instance, my degrees of belief that the coin will come up heads on the next toss and that it won't should add up to 1, on pain of irrationality. Ramsey's famous argument for the irrationality of failing to align your beliefs this way is called a "Dutch Book argument", and we'll discuss it briefly in §3. Note that Bayesianism does

1 The reference to Molière is to *Le Malade Imaginaire*, in which a physician explains helpfully in Latin: "Quare Opium facit dormire: ... Quia est in eo Virtus dormitiva" (The reason why opium induces sleep: because it has in it a dormitive virtue).

not necessarily exclude the possibility that real agents may sometimes assign degrees of belief that don't conform to the rules of probability; it's just that such an agent will be judged to be irrational.

All Bayesians, it seems, agree about two things. One is the centrality of conditionalization in belief update: your degree of belief in hypothesis h once you've received evidence E should be equal to the conditional degree of belief in h given E that you had before observing E. (Discussion question: Why does this make sense?) The second is the practical importance of Bayes' rule as a way of updating prior beliefs in light of new information. The basic formula for updating the probability of h upon receipt of evidence E is:

$$(\text{posterior probability of } h \text{ given } E) \propto (\text{probability of } E \text{ given } h) \times (\text{prior probability of } h)$$

For Bayesians, this update rule is a crucial part of the normatively correct method of updating prior to posterior degrees of belief.

There are two rough categories of Bayesians. Thoroughgoing subjective Bayesians like Ramsey (1926); de Finetti (1937); Savage (1954); Jeffrey (2004) argue that there are no constraints on a rational agent's degrees of belief except that they obey the rules of the probability calculus. Less-subjective Bayesians such as Jaynes (2003) and Jon Williamson (2009) think that thoroughgoing subjective Bayesianism is too permissive: not any assignment of probabilities is rational. They argue that Bayesianism can be combined with rational rules of probability assignment in the face of evidence.

One of the main areas of concern for less-subjective Bayesians is whether there are general principles of how probabilities should be assigned in cases when an agent has very little information with which to calibrate an estimate of probabilities. There are several approaches, but they are mostly technical variations on what Keynes (1921) dubbed the "Principle of indifference": if you don't have information favoring one outcome over another, assign them the same probability.

There are, in turn, many arguments which suggest that this principle isn't sufficient in itself (example: van Fraassen's unit cube). This may well be right, but in practice, the principle of indifference often makes sense and can be used to "objectivize" Bayesian models by using diffuse priors and letting the data do the work. (We'll do a bit of this in §5. See MacKay 2003: 50-1 for a defense of this approach in applied contexts.)

There are also some Bayesians who believe that objective chances (e.g., propensities) exist in addition to credences, and are an object of knowledge. Lewis (1980) proposes the "Principal Principle": roughly, your subjective estimate of the probability of an event should correspond to your estimate of the objective chance of that event. In the extreme, if you believe with probability 1 that the objective chance of an event is p, your degree of belief in that event should also be p.

Many probabilistic models in recent linguistics and cognitive science self-consciously describe themselves as "Bayesian": see Chater, Tenenbaum & Yuille 2006; Griffiths, Kemp & Tenenbaum 2008; Tenenbaum et al. 2011 for discussion. The ideological affinity is clear, but for cognitive science the main interest is in understanding mental processes rather than classifying people as "rational" or "irrational". (A method known as "rational analysis" does plan an important role in Bayesian cognitive science, but as a theory-building method rather than an end in itself.) I don't know whether it's crucial for practitioners of Bayesian cognitive modeling to take a stand in the internecine struggles among Bayesians in philosophy, but there may well be some interesting philosophical commitments lurking in cognitive applications of Bayesian methods.

3 What is probability? Semantic features and four ways to derive them

We start with a simple intensional logic. Here are some assumptions and conventions:

- W is the set of possible worlds; roughly, all of the ways that the world could conceivably turn out to be — independent of whether we have information indicating that some of them aren't actually realistic possibilities. (Probability theorists usually call W the "sample space" and write it Ω.)

Technical note: In some cases it's convenient to pretend that W only contains as many worlds as there are relevant outcomes to the experiment we're analyzing or relevant answers to the question we're asking. For example, when thinking about a single toss of a die we might think of W as containing six worlds: w_1, where the die comes up 1; w_2, where it comes up 2; etc. This means we're ignoring variation between possible worlds that doesn't matter for the problem we're analyzing — our model of the die toss doesn't differentiate worlds according to whether it's sunny in Prague. Technically, then, we're not really dealing with a set of worlds but with a *partition* over the set of possible worlds, which is more or less coarse-grained depending on what we're analyzing. Being casual about this semantic distinction generally doesn't hurt anything, as long as we don't accidentally ignore differences between possibilities that really *are* relevant to the problem at hand.

- The meanings of English sentences are propositions, i.e. functions from possible worlds to truth-values. The meaning of *It is raining* is a function that takes a world w as an argument and returns 1 (true) if it's raining at w and 0 (false) otherwise.
  - We'll ignore context-sensitivity (important but mostly orthogonal); so we'll talk about the "proposition" that it is raining rather than the proposition that it is raining at time t in location l ...
- Each sentence ϕ is associated with a unique set of worlds: the set of worlds where ϕ is true. ϕ can also be associated with a function from worlds w to truth-values, returning 1 if ϕ is true at w and 0 otherwise. For notational convenience, I will use the term "proposition" and propositional variables ambiguously and let context clarify. So, ϕ represents either (a) a sentence which denotes some function from worlds to truth-values, (b) the function itself, or (c) the set of worlds of which the function is true. This allows us to move between notations as convenient, without spending a lot of time worrying about variables. (This is standard practice in formal semantics, but would probably horrify a lot of logicians.) Hopefully this won't cause any confusion — but if it does, please ask.
- Conventions about variables:
  - $w, u, v, w', u', v', \ldots$ are variables ranging over possible worlds.
  - $p, q, r, p', q', r', \ldots$ are variables ranging over ambiguously over atomic propositions or sentences that denote them.

  - $\phi, \psi, \chi, \phi', \psi', \chi', \ldots$ are variables ranging ambiguously over atomic or complex propositions or sentences that denote them.
  - $w_@$ is a distinguished variable representing the actual world.

- We'll assume that, for any two propositions/sentences ϕ and ψ that we can talk about, we can also talk about the sentences $\phi \wedge \psi$, $\phi \vee \psi$, $\neg\phi$, and $\neg\psi$ or equivalently the propositions $\phi \cap \psi$, $\phi \cup \psi$, $-\phi$, and $-\psi$. This shouldn't be very controversial, given that English (like any other natural language) allows us to put "It's not the case that" in front of any sentence and to join any two sentences by "and" or "or". (Technically, this corresponds to the assumption that the space of propositions is a (σ-)algebra. Saying it that way makes it sound like a less obvious choice than it is.)

- We can define **truth** of ϕ very simply as: ϕ is true at a world w if and only if $w \in \phi$. If we don't specify a world, we're implicitly talking about truth at the actual world; so ϕ is true (without further conditions) iff $w_@ \in \phi$. Consequently,

  - $\phi \wedge \psi$ is true iff both ϕ and ψ are true, i.e. iff $w_@ \in (\phi \cap \psi)$.
  - $\phi \vee \psi$ is true iff either ϕ is true, ψ is true, or both, i.e. iff $w_@ \in (\phi \cup \psi)$.
  - $\neg\phi$ is true iff ϕ is false, i.e. iff $w_@ \in -\phi$.

  (with the obvious modifications if we're talking about truth at worlds other than $w_@$.)

As a final note: in some cases I'll give definitions that work as intended only if the sample space/set of worlds W is finite. I'm not going to mention it explicitly every time I do this. If you take a class on probability theory or mathematical statistics, they'll give you more complicated definitions that allow you to deal with infinite W. This is important for many purposes, but worrying about infinite sets is hard and I don't think that it adds anything to conceptual understanding at this point, so we're not going to do it except when it's really necessary. If you later see these ideas somewhere and wonder why the math looks harder, this may be why.

3.1 Probability as measurement

The formalization of probability widely considered to be "standard" is due to Kolmogorov (1933). Here we think of probabilities as measures on propositions, more or less as heights are measures of objects' spatial extent in a vertical dimension, and temperatures are measures of heat energy. Keep in mind that the rules don't tell us what a measurement *means*, and so in principle are neutral between the philosophical interpretations that we've discussed.

First let's consider the simpler case of finite W. (Remember that we're being careless about whether probabilities attach to propositions or to sentences denoting propositions, with the result that complex sentences/propositions are formed equivalently with $\cup$ and $\vee$, or $\cap$ and $\wedge$, or $-$ and $\neg$.)

(2) **Def:** A **Finitely Additive Probability Space** is a triple $\langle W, \Phi, pr \rangle$, where
  a. W is a set of possible worlds;

b. $\Phi \subseteq \mathcal{P}(W)$ is an algebra of propositions (sets of worlds) containing W which is closed under union and complement;

c. $pr : \Phi \to [0,1]$ is a function from propositions to real numbers in the interval [0,1];

d. $pr(W) = 1$;

e. **Additivity**: If ϕ and ψ are in Φ and $\phi \cap \psi = \varnothing$, then $pr(\phi \cup \psi) = pr(\phi) + pr(\psi)$.

Exercise 1. Prove that $pr(\varnothing) = 0$.

Exercise 2. Suppose $pr(\phi) = d$. What is $pr(-\phi)$? Prove it.

Exercise 3. Suppose $pr(\phi) = .6$ and $pr(\psi) = .7$: e.g., there's a 60% chance that it will rain tomorrow and a 70% chance that the Heat will win the NBA finals. Why isn't $pr(\phi \cup \psi)$ equal to 1.3? Roughly what should it be?

Exercise 4. Using your reasoning from the previous exercise as a guide, can you figure out what $pr(p \cup q)$ is in the general case, when p and q may not be mutually exclusive? (Hint: how could you turn $p \cup q$ into something equivalent that you can apply rule (2e) to? It may be useful to draw a Venn diagram.)

Exercise 5. Can you derive from (2) a rule that tells us how to relate $pr(\phi)$ and $pr(\psi)$ to $pr(\phi \cap \psi)$? If so, what is it? If not, try reasoning about extreme cases; can you use $pr(\phi)$ and $pr(\psi)$ to place upper and lower bounds on $pr(\phi \cap \psi)$?

Exercise 6. Why should $pr(W)$ be 1? What would change if we were to require instead that pr maps propositions to $[0, 23]$, and $pr(W) = 23$? What if $pr(W)$ were required to be -23?

Exercise 7. Can you think of an intuitive justification for rule (2e) (additivity)? If not, try to think of an intuitive justification for the weaker rule of **positivity**: "If neither ϕ nor ψ has probability 0 and $\phi \cap \psi = \varnothing$, then $pr(\phi \cup \psi) > pr(\phi)$ and $pr(\phi \cup \psi) > pr(\psi)$."

When dealing with infinite sets of worlds, you need something a bit fancier to make the math work out right. This matters a lot, for example, when dealing with continuous sample spaces, i.e. situations where a variable can take on an uncountably infinite number of values. I'll present the axioms for completeness, though it's beyond beyond the scope of this course to discuss what the difference is and why it matters.

(3) **Def:** A **Countably Additive Probability Space.** is a triple $\langle W, \Phi, pr \rangle$ where

a. W is a set of possible worlds;

b. Φ is a σ-algebra (an algebra containing W which is closed under complement and countable union);

c. $pr : \Phi \to [0,1]$;

d. $pr(W) = 1$;

e. **Countable Additivity**: If $\{\phi_1, \phi_2, ...\}$ is a (possibly infinite) set of mutually exclusive propositions each of which is in Φ, then

$$pr\left(\bigcup_{i=1}^{\infty} \phi_i\right) = \sum_{i=1}^{\infty} pr(\phi_i)$$

In Kolmogorov's system, the logically basic notion is the "unconditional" probability of a proposition, $pr(\phi)$. In many contexts, however, we want to be able to talk about the "conditional" probability of ϕ given some other proposition ψ. This is defined as:

(4) **Def:** The **conditional probability** of ϕ given ψ, $pr(\phi|\psi)$, is defined as

$$pr(\phi|\psi) =_{df} \frac{pr(\phi \cap \psi)}{pr(\psi)}$$

Intuitively, the conditional probability of ϕ given ψ is the probability that we think ϕ would have if we were certain that ψ is true. (With apologies to the various philosophers who've pointed out that this gloss isn't quite right; it's still instructive, I think.) Another intuition is this: we temporarily ignore worlds in which q is false and make the minimal adjustments to make sure that we still have a valid probability measure, without altering the relative probabilities of any propositions. So the conditional probability of ϕ given ψ is just the probability that ϕ and ψ are both true — remember, we only want to look at worlds where ψ holds — "normalized" by dividing by the probability of ψ. Normalizing ensures that conditional probabilities behave like regular probabilities: e.g. $pr(\phi|\psi) + pr(\neg\phi|\psi) = 1$, even though in many cases $pr(\phi \wedge \psi) + pr(\neg\phi \wedge \psi) \neq 1$.

This is often known as the "ratio analysis" of conditional probability. As we'll see, conditional probability is taken to be the basic kind of probability in some other systems, so naturally these approaches will need to define it differently.

Kolmogorov's axiomatization of probability is simple and mathematically convenient, but has been criticized in various ways as being stipulative, uninsightful, or incorrect in its assumption that conditional probability is derived rather than basic. On the first two counts, at least, I think that this is a mistake: Kolmogorov's axioms are just mathematical definitions, and their value or lack of value is demonstrated by their usefulness/uselessness when applied to real problems. Indeed, the other derivations of probability that we'll consider can be seen not as competitors but as detailed arguments for the basic correctness of his system. (Jaynes (2003: 651-5) and Lassiter (2011: 75-6) suggest such an interpretation of the derivations that we'll see in §3.3 and §3.4, respectively.)

Finally, in preparation for §3.3, note the following property of (conditional) probability:

(5) **Product rule**. $pr(\phi \cap \psi) = pr(\phi) \times pr(\psi|\phi) = pr(\psi) \times pr(\phi|\psi)$.

Exercise 8. Derive the product rule from the ratio definition of conditional probability.
Exercise 9. Derive the conditional product rule, $pr(\phi \cap \psi|\chi) = pr(\phi|\chi) \times pr(\psi|\phi \cap \chi)$.

3.2 Probabilities as proportions

For those who interpret probabilities as the relative frequencies of actual events, the justification of the rules of probability is clear. For these theorists, the probability of ϕ is simply *defined* as the proportion of events in some reference class which satisfy ϕ, and the logic of proportions is guaranteed to obey the axioms of finitely additive probability. For instance, the probability that an American citizen is male is just the proportion of males among the U.S. citizenry.

Exercise 10. For each of the axioms in (2), explain why it is satisfied if we interpret probabilities as relative frequencies. Also explain why your unrestricted disjunction rule from exercise 4 is satisfied.

Even for non-frequentists, the correspondence between probabilities and proportions of events in (appropriately delineated) large samples is useful. We'll see a lot of this in later sections when we talk about sampling and simulation.

3.3 Plausibility: The Cox Axioms

A quite different way to derive probability is to start from qualitative assumptions about sound reasoning. Cox (1946) suggests such a derivation, elaborations of which are favored by many Bayesians. On this approach, probability is a generalization of deductive logic to uncertain reasoning, and deductive logic is the limiting case of probability theory when we restrict ourselves to reasoning about things that are either certain or impossible. (The presentation in this section follows closely Jaynes 2003: §2 and Van Horn 2003.)

We start with the intuitive concept of *plausibility*. Plausibility is a scalar concept: e.g., ϕ can be more or less plausible than ψ, or they can be equally plausible. Please don't assume that plausibility = probability — let's keep it intuitive and see what we have to assume in order to *derive* this equivalence.

The plausibility of a proposition is always relative to a state of information; if you had different evidence (and you made appropriate use of it, etc.), some propositions would be more plausible to you and others would be less plausible. So it doesn't really make sense on this conception to talk about the plausibility of a proposition in itself, since the proposition will have various plausibilities depending on the evidence that is available. When we talk about the plausibility of a proposition simpliciter, it is always implicitly relativized to some (logically consistent) state of information X which is clear from context: $plaus(\phi|X)$.

Some assumptions:

(6) a. The plausibility of any proposition is represented by a real number. Letting Φ represent the set of propositions that our language has the resources to talk about and $\mathcal{X}$ represent the possible information states X, $plaus : \Phi \to (\mathcal{X} \to \mathbb{R})$. We assume again that Φ is closed under union and complement (and therefore intersection).

b. There is a maximum plausibility $\top$. For all $\phi \in \Phi$, $plaus(\phi|X) \leq \top$.

c. If ϕ is a tautology, $plaus(\phi|X) = \top$.

d. If $plaus(\phi|X) < \top$ then $\neg\phi$ is consistent with X, i.e. $\neg\phi \wedge X$ is not a contradiction.

e. Logically equivalent propositions are equally plausible relative to any info. state X.

f. **Negation**: For some strictly decreasing $f : \mathbb{R} \to \mathbb{R}$, $plaus(\neg\phi|X) = f(plaus(\phi|X))$. In other words, if ϕ is more plausible than ψ then $\neg\phi$ is less plausible than $\neg\psi$. Likewise, ϕ is at least as plausible as ψ iff $\neg\phi$ is at most as plausible as $\neg\psi$.

Fact 1: $\forall \phi \in \Phi : \bot \leq plaus(\phi|X) \leq \top$ (plausibilities are bounded by $\bot$ and $\top$), where $\bot = f(\top)$.

- Proof: By (6b), $plaus(\phi|X) \leq \top$. By (6f) $plaus(\phi|X) = f(plaus(\neg\phi|X))$, which is greater than or equal to $f(\bot)$ since $plaus(\neg\phi|X) \leq \top$ and f is decreasing.

Fact 2: $\bot < \top$.

- Follows because A is non-empty and f is strictly decreasing.

Fact 3: $plaus(\phi|X) = f(f(plaus(\phi|X)))$.

- Proof: $\phi \equiv \neg\neg\phi$, and logically equivalent propositions have the same plausibility by assumption (6e). So $plaus(\phi|X) = plaus(\neg\neg\phi|X) = f(plaus(\neg\phi|X)) = f(f(plaus(\phi|X)))$.

Some further assumptions.

(7) a. **Richness**: For some nonempty dense $A \subseteq \mathbb{R}$, (a) both $\bot$ and $\top$ are in A, and (b) for every $x, y, z \in A$ there are a possible information state X and three atomic propositions p, q, r such that $plaus(p|X) = x$, $plaus(q|p, X) = y$, and $plaus(r|p, q, X) = z$. (This looks complicated, but as far as I know its only controversial feature is the density assumption.)

b. **Conjunction**: There is a function $g : A \times A \to A$, strictly increasing in both arguments, such that $plaus(\phi \wedge \psi|X) = g(plaus(\phi|\psi, X), plaus(\psi|X))$.

Clearly, g should depend on ϕ, ψ, and X in some way, but why the particular requirement on g given in (7b)? . It turns out that most of the other options have unacceptable consequences or are equivalent to this requirement, but there are still several options that can't be ruled out *a priori* (Van Horn 2003: 13-15). This one is the simplest, though. Jaynes (2003: §2) argues in favor of **Conjunction** that

> In order for $\phi \wedge \psi$ to be a true proposition, it is necessary that ψ is true. Thus the $plaus(\psi|X)$ should be involved. In addition, if ψ is true, it is further necessary that ϕ should be true; so $plaus(\phi|\psi \wedge X)$ is also needed. But if ψ is false, then of course $\phi \wedge \psi$ is false independently of whatever one knows about ϕ, as expressed by $plaus(\phi|\neg\psi \wedge X)$; if the robot reasons first about ψ, then the plausibility of ϕ will be relevant only if ψ is not true. Thus, if the robot has $plaus(\psi|X)$ and $plaus(\phi|\psi \wedge X)$ it will not need $plaus(\phi|X)$. That would tell it nothing about $\phi \wedge \psi$ that it did not have already. (Notation modified -DL)

Also important in the statement of **Conjunction** is the requirement that g be strictly increasing in both arguments. This seems intuitive: if ϕ becomes more plausible then $\phi \wedge \psi$ should presumably be more plausible as well, though not necessarily *as much* more plausible. The same goes for ψ. But we might also just require that $\phi \wedge \psi$ can't become *less* plausible when ϕ or ψ becomes more plausible. If we took this route we would leave room for a system where $plaus(\phi \wedge \psi|X) = \mathbf{min}(plaus(\phi|X), plaus(\psi|X))$ — a feature that some alternative representations of uncertainty do in fact have, such as fuzzy logic.

Exercise 11. Construct an argument for or against treating $plaus(\phi \wedge \psi|X)$ as equal to $\mathbf{min}(plaus(\phi|X), plaus(\psi|X))$. Give concrete examples of cases where this would give the right/wrong result. If you are arguing for this treatment, also give an example where allowing $plaus(\phi \wedge \psi|X)$ to be greater than the min of the two would give the *wrong* result.

We're now done assuming, and can move on to the consequences. The proof is somewhat intricate, and I don't want to get into the details here; the net result is that, no matter what *plaus* is, if it satisfies these assumptions then there is a one-to-one mapping from plausibilities to a continuous,

strictly increasing function p with the properties that, for any propositions ϕ and ψ and information state X,

(8) a. $p(\phi|X) = 0$ iff ϕ is known to be false given the information in X.
b. $p(\phi|X) = 1$ iff ϕ is known to be true given the information in X.
c. $0 \leq p(\phi|X) \leq 1$.
d. $p(\phi \wedge \psi|X) = p(\phi|X) \times p(\psi|\phi \wedge X)$.
e. $p(\neg\phi|X) = 1 - p(\phi|X)$.

Exercise 12. See if you can prove from (8) that Cox's assumptions derive the conditional version of the disjunction rule that we derived from Kolmogorov's axioms in exercise 4: $p(\phi \vee \psi|X) = p(\phi|X) + p(\psi|X) - p(\phi \wedge \psi|X)$. (Hint: Find something logically equivalent to $\phi \vee \psi$ that you can repeatedly apply (8d) and (8e) to.)

(8) plus the result of exercise 12 is enough to show clearly that p is a finitely additive probability measure according to the definition we saw in the last section! In other words, if you accept all of the requirements we've imposed on plausibilities, then you're committed to treating plausibilities (relative to an information state) as being isomorphic to conditional probability measures (conditioned on that information state, cf. (3) and (4)). Conversely, if you don't want to be committed to probabilistic reasoning as the unique rational way to deal with uncertainty, you'd better figure out which of the Cox assumptions you want to deny.

Note, however, that Cox's derivation does not give us countable additivity (3). Jaynes (2003) vigorously defends this feature of Cox's system, arguing that applications of probability which appear to require infinite sets are either unnecessary or can be reinterpreted as limiting cases of probabilities of finite sets. (This is a minority opinion, though.)

Various objections have been raised to Cox's derivation of probability.

- Is is obvious that degrees of plausibility should be represented as real numbers?
- More generally, are density and infinity of plausibility values (7a) reasonable assumptions? (For arguments against and for, see Halpern 1999, Van Horn 2003.)
- Using real-valued plausibilities begs the question of whether any two propositions are always comparable in plausibility. Is this intuitively obvious? Objections? Replies to objections?
- Frequency-minded probabilists have argued that it doesn't make sense to derive probability from plausibilities; plausibility is a psychological concept, and so just has the wrong subject matter. In other words, if you're don't already have Bayesian inclinations, the force of Cox's arguments is unclear.

If you find the idea that probability should be thought of as a way to assign plausibilities to propositions, and you don't mind assuming that degrees of plausibility are infinite and always comparable, Cox's theorem is an powerful argument in support of the conclusion that a reasonable system of ranking propositions in terms of their probability must follow the rules of the probability calculus, or be isomorphic to a system that does.

3.4 Linguistic derivation

For the linguists' sake, I want to mention briefly a quite different way of getting to probability stemming from recent work by Yalcin (2010); Lassiter (2010, 2011). The idea is that the mathematics of probability is already discernible in the structure of epistemic modality in English, and in particular the meanings of the epistemic adjectives *likely* and *probable*. If so, a knowledge of probability must form part of our knowledge of the semantics of the English language. (And, I imagine, other languages as well.)

To start, note that *likely* and *probable* are **gradable**. ϕ can be *very likely* or *somewhat likely* or *more likely than* ψ, just as Sam can be *very tall* or *somewhat tall* or *taller than Stan*. Standard semantics for gradable adjectives like *tall* and *full* treats them as relating individuals to points on a scale such as $(-\infty, \infty)$ or $[0,1]$ (e.g. Kennedy & McNally 2005; Kennedy 2007). Similarly, we presumably want *likely* and *probable* to relate propositions to points on a scale.

For argument's sake, grant me that this scale is $[0,1]$. (This can be justified linguistically, but I don't want to go into it here.) The question is then what other properties these expressions have. Well, we know from studying other gradable adjectives that some of them are additive (for non-overlapping objects) and some are not: *tall* and *heavy* vs. *hot* and *dangerous*. Are *likely* and *probable* associated with additive measures? If they are, then we're most of the way to a probability scale, with a minimum of 0, a maximum of 1, and an additive measure.

Here's an argument that they are. Many theories of epistemic modality don't even give truth-conditions for epistemic comparisons. The most widely-accepted semantics for epistemic modals in linguistics — Kratzer's (1991) — does better, but it also validates the following inference pattern:

(9) a. ϕ is as likely as ψ.
b. ϕ is as likely as χ.
c. $\therefore$ ϕ is as likely as $(\psi \vee \chi)$.

Imagine a lottery with 1 million tickets. Sam, Bill, and Sue buy two tickets each, and no one else buys more than two tickets. The lottery is fair, and only one ticket will be chosen as the winner. (10) is clearly true, and equivalent to (11).

(10) Sam is as likely to win the lottery as anyone else is.

(11) $\forall x$: Sam is as likely to win the lottery as x is.

We can use (9) and (10)/(11) to prove (12).

(12) Sam is as likely to win the lottery as he is not to win.

Exercise 13. Prove that (12) follows from (9) and (10).

Since (12) is clearly false in the situation described, (9) can't be a valid inference. So we want a semantics for *likely* (and *probable*) that doesn't validate (9). What kind of measure should we assign to them?

Exercise 14. Prove, by giving a counter-model, that (9) is not valid if *likely*'s scale is equivalent to finitely additive probability.

Exercise 15. Think of a weaker condition than additivity that would also make it possible to

avoid validating this inference.

3.5 Probability and rational choice

Another influential Bayesian argument for the (unique) rationality of probability is associated with Ramsey (1926) and much following literature on epistemology, probability, and rational choice. These are called **Dutch book arguments**, and they go roughly like this.

Suppose that an agent's willingness to take a bet on ϕ (e.g., whether the Heat will win the NBA finals) depends on the agent's relative degrees of belief in ϕ and $\neg\phi$. Call these $bel(\phi)$ and $bel(\neg\phi)$. In particular, imagine a bet that pays \$1 if ϕ happens and nothing otherwise. (We could make the stakes higher without affecting the reasoning.) We assume that the agent considers \$$d$ a fair price for a bet on ϕ if and only if $bel(\phi) = d$. A good bet is any bet which costs at most as much as the fair price for that bet. For instance, an agent with $bel(\phi) = bel(\neg\phi) = .5$ should be willing to pay up to 50 cents for a bet on ϕ. An agent with $bel(\phi) = .9$ should be willing to pay up to 90 cents, and so on.

Dutch book arguments suggest that, given this set-up, an agent whose degrees of belief fail to conform to the probability calculus can always be taken for a ride. For instance, suppose that the agent's degrees of belief fail to add up to 1, e.g. $bel(\phi) = bel(\neg\phi) = .6$. Then the agent will pay up to 60 cents for a \$1 bet on ϕ and up to 60 cents for a \$1 bet on $\neg\phi$. A clever bookie, detecting this weakness, will sell our agent bets on *both* ϕ and $\neg\phi$ for 60 cents each. But since only one of these can happen, the agent will pay \$1.20 but will earn \$1 no matter what, for a guaranteed loss of 20 cents. Similar arguments can be constructed to justify other features of probability. So, given these assumptions about the way that degrees of belief influence betting behavior, it would be irrational for anyone not to have their degrees of belief arranged in a way that follows the rules of the probability calculus.

The relationship between probability and rational decision is important and fascinating, with a huge literature spanning many fields, including a very healthy philosophical literature on Dutch books alone. Getting into these topics in greater detail would take us too far astray, though. Classic references include Savage 1954 and the entertaining, readable, and highly instructive Jeffrey 1965. The following articles should also whet your appetite and give some further references to investigate.

- http://plato.stanford.edu/entries/epistemology-bayesian/
- http://plato.stanford.edu/entries/decision-causal/
- http://plato.stanford.edu/entries/game-theory/

4 Probability and random variables: Basic math and simulation

The different philosophies of probability that we've seen are, mercifully, in agreement on almost all of the basic mathematics of probability. The main points of divergence are

- Whether probabilities are countably or only finitely additive;
- Whether conditional probability or unconditional probability is the fundamental concept.

Here we can safely skirt over these points of disagreement and make use of finitely additive probability, repeated here from (2). For those who think that conditional probability is fundamental, think of the unconditional probability measures in what follows as implicitly conditioned on some fixed body of information or worldly circumstances.

4.1 Reasoning with propositions

(13) Reminder: a **Finitely Additive Probability Space** is a triple $\langle W, \Phi, pr \rangle$, where

a. W is a set of possible worlds;

b. $\Phi \subseteq \mathcal{P}(W)$ is an algebra of propositions (sets of worlds) containing W which is closed under union and complement;

c. $pr : \Phi \to [0,1]$ is a function from propositions to real numbers in the interval [0,1];

d. $pr(W) = 1$;

e. If ϕ and ψ are in Φ and $\phi \cap \psi = \varnothing$, then $pr(\phi \cup \psi) = pr(\phi) + pr(\psi)$.

In an earlier exercise we used (13) to prove the product rule and the conditional product rule:

(14) a. **Product rule**. $pr(\phi \wedge \psi) = pr(\phi) \times pr(\psi|\phi) = pr(\psi) \times pr(\phi|\psi)$.

b. **Conditional PR**. $pr(\phi \wedge \psi|\chi) = pr(\phi|\chi) \times pr(\psi|\phi \wedge \chi) = pr(\psi|\chi) \times pr(\phi|\psi \wedge \chi)$.

We also derived rules for manipulating negations and disjunctions:

$$\begin{aligned} pr(\phi) &= 1 - pr(\neg\phi) \\ pr(\phi \vee \psi) &= pr(\phi) + pr(\psi) - pr(\phi \wedge \psi) \end{aligned}$$

Time for our first simulation! Open R and a new R source file, or download and open the code at http://www.stanford.edu/~danlass/NASSLLI-R-code.R and run it from within R. The first thing you should do is run the following line from the top of the code file, or else type into the prompt:

R code

```
source("http://www.stanford.edu/~danlass/NASSLLI-R-functions.R")
```

This will load some simple functions that we'll use below for counting, checking equalities, etc.

R has excellent (pseudo-)random number generating facilities that we'll make use of. The simplest case is `runif(1,0,1)`, which generates a floating point number between 0 and 1. Likewise `runif(5,0,1)` will generate 5 such numbers.

R code

```
> runif(1,0,1)
   0.5580357
> runif(5,0,1)
   0.5038063 0.5804765 0.8397822 0.7587819 0.2585851
```

Suppose we know (never mind how) that ϕ is true with probability p. The following function uses R's `runif` to *sample* from a distribution equivalent to the distribution on ϕ, returning either `TRUE` or `FALSE`. We can think of sampling from this distribution as flipping a coin which is biased to give heads with probability p, thus the name `flip`.

R code

```
flip = function(p) {
   if (runif(1,0,1) < p) {
      return(TRUE)
   } else {
      return(FALSE)
   }
}
```

After loading this function, type `flip(.8)` into the console a couple of times. It should return `TRUE` most of the time, but occasionally it will return `FALSE`. If we run this function many times, it will return `TRUE` about 80% of the time. This is because of the **Law of Large Numbers**, which we'll talk about more when we discuss random variables below.

If we want to take a lot of samples from `flip(p)` at once, we could use a `for`-loop and store the values in a vector, as in the `flip.n.slow` function.

R code

```
flip.n.slow = function(p,n) {
   vec = rep(-1, n)
   for (i in 1:n) {
      if (runif(1,0,1) < p) {
         vec[i] = TRUE
      } else {
         vec[i] = FALSE
      }
   }
   return(vec)
}
```

This has the further advantage that we can ask R to calculate the proportion of `TRUE`s in the sample:

R code

```
> n.samples = 1000
> samp = flip.n.slow(.8, n.samples)
> howmany(samp, eq(TRUE))
   778
```

```
> prop(samp, eq(TRUE))
   .778
```

When you run this, you'll probably get a different precise number, but it should be close to 800 true samples and proportion .8, as mine was. (Since R coerces `TRUE` to 1 and `FALSE` to 0, we could also have gotten the proportion of true samples by asking for `mean(samples)`, but that would be cheating since we haven't defined means yet.)

Sampling many times from the distribution on ϕ gives us an approximation to the true probability. This may help clarify why $pr(\phi)$ must be equal to $1 - pr(\neg\phi)$. If we're approximating $pr(\phi)$ by the number of samples in which ϕ is true divided by the total number of samples n, then of course our approximation of $pr(\neg\phi)$ should be the number of samples in which $\neg\phi$ is true divided by the total number n. Since every sample is either true or false, the approximate value of $pr(\neg\phi)$ must be n minus the number of samples where ϕ is true divided by n, i.e. 1 minus the approximation to $pr(\phi)$.

Exercise 16. Explain in similar terms why additivity must hold for mutually exclusive ϕ and ψ, and why $pr(\phi \vee \psi) \neq pr(\phi) + pr(\psi)$ when $pr(\phi \wedge \psi)$ is non-zero. Write down the formula for finding the approximation to $pr(\phi \vee \psi)$ in the general case, assuming that we have n samples. (Extra credit: how could additivity hold accidentally in a sample? What can we do to guard against this?)

It's clear what `flip.n.slow` is doing, but it has a distinct disadvantage: it's slow. To see this, try increasing `n.samples` from 1000 to 100,000,000 (or don't actually, you'll be waiting for a looooong time). The reason for this has to do with the internal workings of R, which is not optimized for `for`-loops. We can avoid this by using R's ability to generate a vector of random numbers all at once and then compare the whole vector to p rather than doing it 1 item at a time. This accomplishes the same thing as `flip.n.slow`, but much more quickly for large `n`.

R code

```
flip.n = function(p,n) {
   return(runif(n,0,1) < p)
}
```

R code

```
> prop(flip.n(.8,1000), eq(TRUE))
   .819
```

Assuming our random number generator works well, this should return a vector of length 1000 with approximately 800 `TRUE`s and 200 `FALSE`s. Run this a couple of times and observe how the values change; the proportion is usually very close to .8, but the difference varies a bit. In fact, it can be informative to run this simulation a bunch of times and look at how the return values are distributed:

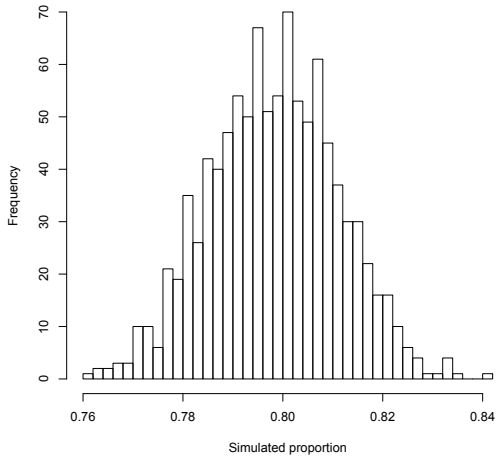
R code

```
> n.sims = 1000
> sim.props = rep(-1, n.sims) # make a vector to store the sim results
> for (i in 1:n.sims) {
+     sim.props[i] = prop(flip.n(.8,1000), eq(TRUE))
+ }
```

Here's what I got.[2]

R code

```
> summary(sim.props)
    Min.  1st Qu.  Median  Mean  3rd Qu.  Max.
  0.7600 0.7900   0.7990 0.7987 0.8080 0.8420
> hist(sim.props, xlab="Simulated proportion", main="", breaks=50)
```



Let's think now about distributions with multiple propositions that may interact in interesting ways.

(15) **Def: Joint distribution**. A *joint distribution* over n propositions is a specification of the probability of all 2^n possible combinations of truth-values. For example, a joint distribution over ϕ and ψ will specify $pr(\phi \wedge \psi)$, $pr(\neg\phi \wedge \psi)$, $pr(\phi \wedge \neg\psi)$, and $pr(\neg\phi \wedge \neg\psi)$.

In general, if we consider n logically independent propositions there are 2^n possible combinations of truth-values. The worst-case scenario is that we need to specify $2^n - 1$ probabilities. (Why not all 2^n of them?) If some of the propositions are probabilistically independent of others (cf. (17) below), we can make do with fewer numbers.

2 Note that the distribution is approximately bell-shaped, i.e. Gaussian/normal. This illustrates an important result about large samples from random variables, the Central Limit Theorem.

(16) **Def: Marginal probability**. Suppose we know $pr(\phi)$, $pr(\psi|\phi)$ and $pr(\psi|\neg\phi)$. Then we can find the *marginal probability* of ψ as a weighted average of the conditional probability of ψ given each possible value of ϕ.

Exercise 17. Using the ratio definition of conditional probability, derive a formula for the marginal probability of ψ from the three formulas in (16).

To illustrate, consider a survey of 1,000 students at a university. 200 of the students in the survey like classical music, and the rest do not. Of the students that like classical music, 160 like opera as well. Of the ones that do not like classical music, only 80 like opera. This gives us:

	Like classical	Don't like classical	Marginal
Like opera	160	80	240
Don't like opera	40	720	760
Marginal	200	800	1000

Exercise 18. What is the probability that a student in this sample likes opera but not classical? What is the marginal probability of a student's liking opera? Check that your formula from the last exercise agrees on the marginal probability.

Suppose we wanted to take these values as input for a simulation and use it to guess at the joint distribution over liking classical music and opera the next time we survey 1,000 (different) students. Presumably we don't expect to find that *exactly* the same proportion of students will be fans of each kind of music, but at the moment the data we've gathered is our best guess about future behavior.

R code

```
> sample.size = 1000
> p.classical = .2
> p.opera.given.classical = .8
> p.opera.given.no.classical = .1
> classical.sim = flip.n(p.classical, sample.size)
> opera.sim = rep(-1, sample.size)
> for (i in 1:sample.size) {
+    if (classical.sim[i] == TRUE) {
+       opera.sim[i] = flip(p.opera.given.classical)
+    } else {
+       opera.sim[i] = flip(p.opera.given.no.classical)
+    }
+ }
```

Note that we're representing individuals by an index `i` and using this correspondence to make the way we generated samples for `opera.sim` conditional on the output of our sampling process for `classical.sim`, via a conditional statement with `if...else`.

Exercise 19. Suppose we had instead computed `opera.sim` without making reference to `classical.sim`, using `flip.n` and the marginal probability of liking opera (.24). Intuitively, why

would this be a mistake? How would the simulation's predictions about the joint distribution over classical- and opera-liking differ?

(17) **Definition: Independence**. There are three ways to define independence, all equivalent on the ratio definition of conditional probability. ϕ and ψ are *independent* iff any/all of the following hold:

a. $pr(\phi|\psi) = pr(\phi)$

b. $pr(\psi|\phi) = pr(\psi)$

c. $pr(\phi \wedge \psi) = pr(\phi) \times pr(\psi)$

Independence is a very important concept in probability theory. Intuitively, if ϕ and ψ are independent then learning about one will not influence the probability of the other. This means that I can ignore ϕ when reasoning about ψ, and vice versa. Practically speaking, this can lead to important computational advantages. Independence has also been argued to be an important organizing feature of probabilistic reasoning for humans and other intelligent agents (Pearl 1988, 2000). In Pearl's example, everyone assumes that the price of beans in China and the traffic in L.A. are independent: if you ask someone to make a guess about one they'll never stop to consider the other, because there's no way that the answer to one of these questions could be informative about the other. If people didn't make the assumption that most things are independent of most other things, probabilistic reasoning would be extremely difficult. We would have to check the probabilities of a huge number of propositions in order to make any inference (increasing exponentially with the number of propositions we're considering).

Exercise 20. Assuming $pr(\phi) = .2$ and $pr(\psi) = .8$, describe how we could operationalize independence in a simulation and check that it has the properties in (17).

Exercise 21. If we add a third variable χ, what are the possible (in)dependence relations between the three? What would each look like in a data set? in a simulation?

Exercise 22. How did we encode dependence and independence in earlier simulation examples and exercises? Which led to simpler models?

For the next definition, recall that a *partition* of a set $\mathcal{A}$ is a set of disjoint subsets of $\mathcal{A}$ whose union is $\mathcal{A}$.

(18) **The rule of total probability**

a. If $\{\phi_1, ..., \phi_n\}$ is a partition of $\mathcal{A} \subseteq W$, then $\sum_{i=1}^{n} pr(\phi_i) = pr(\mathcal{A})$.

b. Special case: If $\{\phi_1 ... \phi_n\}$ is a partition of W, then $\sum_{i=1}^{n} pr(\phi_i) = 1$.

Finally, we can derive a result which — although mathematically trivial according to the ratio definition of conditional probability — is considered by Bayesians to be of the most useful results in probability theory.

(19) **Bayes' Rule**. $pr(\phi|\psi) = \dfrac{pr(\psi|\phi) \times pr(\phi)}{pr(\psi)} = \dfrac{\text{likelihood} \times \text{prior}}{\text{normalizing constant}}$

Exercise 23. Use the ratio definition of conditional probability to prove Bayes' rule.

In the Bayesian literature, you'll often see Bayes' rule given using hypothesis-talk instead of proposition-talk, along with an explicit declaration of the prior, likelihood, and hypothesis space. Setting $\mathcal{H} = h_1, ...h_n$ to be the set of possible hypotheses and E to be the evidence we've received:

$$pr(h_i|E) = \frac{pr(E|h_i) \times pr(h_i)}{pr(E)}$$

At first glance this looks unhelpful, since we need to know the prior probability of the evidence E that we've received, and there's often no obvious way to estimate this. But fortunately the rule of total probability helps us out here: if $\mathcal{H}$ really is a partition of W, then we can find the probability of E by calculating the joint probability of h_j and E for each j, which we can then convert into something that we may know how to estimate.

Exercise 24. Show that $pr(E) = \sum_{j=1}^{n} \left[pr(E|h_j) \times pr(h_j)\right]$ if $\mathcal{H}$ is a partition of W with n elements $\{h_1, ...h_n\}$. (Hint: use $\mathcal{H}$ and E to form a new, smaller partition.)

This result gives us a more usable form of Bayes' rule, which depends only on our assumption that $\mathcal{H}$ exhausts the possible hypotheses that could explain E. Calculating $pr(h_i|E)$ with this formula also requires us to be able to estimate priors and likelihoods for each possible hypothesis.

$$pr(h_i|E) = \frac{pr(E|h_i) \times pr(h_i)}{\sum_{j=1}^{n} \left(pr(E|h_j) \times pr(h_j)\right)}$$

Ex. 8

Bayes' rule in Medical Diagnosis #1. "To illustrate Bayes's rule in action, suppose we observe John coughing (d), and we consider three hypotheses as explanations: John has h_1, a cold; h_2, lung disease; or h_3, heartburn. Intuitively only h_1 seems compelling. Bayes's rule explains why. The likelihood favors h_1 and h_2 over h_3: only colds and lung disease cause coughing and thus elevate the probability of the data above baseline. The prior, in contrast, favors h_1 and h_3 over h_2: Colds and heartburn are much more common than lung disease. Bayes's rule weighs hypotheses according to the product of priors and likelihoods and so yields only explanations like h_1 that score highly on both terms. ... (Tenenbaum et al. 2011)

Exercise 25. In this example from Tenenbaum et al. 2011, it seems unlikely that the three hypotheses considered really do exhaust the possible explanations. Does this invalidate their reasoning? Why or why not?

Ex. 9

Bayes' rule in Medical diagnosis #2. A particular disease affects 300,000 people in the U.S., or about 1 in 1,000. There is a very reliable test for the disease: on average, if we test 100 people that have the disease, 99 will get a positive result; and if we test 100 people that do not have the disease, 99 will get a negative result.

Exercise 26. Use Bayes' rule to calculate the probability that a randomly chosen individual with a positive test result has the disease. Check this against your answer to the second question

("How worried should I be?") from the first time we saw this example. Does your answer to the earlier question make sense given your answer to this exercise? If not, what's going on?

Ex. 10 **Random variables warmup (the ubiquitous urn).** An urn contains 5 balls of identical size, shape, and texture. Three of them are red and two are green. I shake the urn so that position of the balls is unpredictable, and then select three balls, one after the other. (This is called "sampling without replacement".) I label them "1,2,3" so as not to forget what order I drew them in. Let ϕ be the proposition that that the first ball I pick is red; ψ be the proposition that the second is red; and χ be the proposition that the third ball is red.

Exercise 27. What is $pr(\phi)$?
Exercise 28. What is $pr(\psi|\phi)$? What about $pr(\psi|\neg\phi)$?
Exercise 29. What is $pr(\psi)$? Don't try to intuit it; reason by cases, thinking about values of ϕ.
Exercise 30. What is the probability that none of the balls will be red? One? Two? Three?
Exercise 31. I put all the balls back in the urn and start again. This time, each time I draw a ball I write down its color, put it back in the urn, and shake again. (This is called "sampling with replacement".) Now what is the probability that none of three balls will be red? One? Two? Three?

4.2 Random variables

(20) **Def: random variable**. A *random variable* $X : W \to \mathbb{R}$ is a function from possible worlds to real numbers.

Note that propositions can be thought of as a simple kind of random variable. We're treating propositions equivalently as sets of worlds or as the characteristic functions of such sets. On the latter conception, propositions are functions from worlds to $\{0, 1\}$, so they fit the definition.

Ex. 11 **Aside on random variables and the semantics of questions.** A proposition partitions W into two sets: the worlds where the proposition is true and the worlds where it is false. Similarly, every random variable is naturally associated with a partition on W: for any random variable X and any $v \in V_X$, there is a set of worlds $\{w \mid X(w) = v\}$. For instance, in the urn example, let X be a function mapping worlds to the number of red balls that I draw in that world. The corresponding partition divides W into four sets, the worlds where I pick 0, 1, 2, or 3 red balls. The probability that $X(w_@) = v$ is the same as the probability that the actual world is in the corresponding cell of the partition on W.

I mention this because it suggests a connection between probability talk and the semantics of questions. The definition of random variables in probability theory is closely related to the partition semantics of questions due to Groenendijk & Stokhof (1984) and developed in various directions by people doing Alternative Semantics and Inquisitive Semantics as well as

question-based models of discourse pragmatics (cf. Roberts' and Groenendijk & Roelefsen's courses).

Asking about the probability of a proposition ϕ is like asking the polar question "Is it true that ϕ?". There are two possible answers ("yes" and "no") each with some probability of being true, just as there are two cells in the partition induced by a polar question in the semantic treatment. Asking about the probabilities of possible values of $X(w_@)$ is like asking for the probability of each possible answer to the *wh*-question "How many red balls will Dan draw?" The difference is that in addition to a partition we also have a probability distribution over the cells of the partition. So the concept of a random variable is just a straightforward upgrade of familiar concepts from intensional semantics for natural language.

For some really cool connections between probability models and the semantics and pragmatics of questions, check out van Rooij 2003, 2004.

Most of the complicated mathematics in probability theory comes in when we start worrying about random variables, especially continuous ones. Here we'll concentrate on discrete random variables, i.e. ones whose range is a countable (often finite) subset of $\mathbb{R}$. This is because the math is simpler and they're sufficient to illustrate the basic concepts of random variables, sampling, and inference. When you look at more advanced material in probability you'll see a lot of inscrutable-looking formulas, but don't fear: it's mostly derived in a pretty straightforward fashion from what we'll now discuss, with integrals replacing summations and some other stuff thrown in to deal with special problems involving infinite sets.

Ex. 12

Urns in RV-speak. Again we have an urn with three red balls and two green ones. We sample three balls with replacement, shaking the urn between draws so that the position of the balls is unpredictable.

Previously we defined propositions ϕ = "The first ball drawn is red", ψ = "The second ball drawn is red", and χ = "The third ball drawn is red". We can rephrase the urn problem from the last section as a question about a random variable $X : W \to \{0,1,2,3\}$ which maps a world w to the number of red balls that I draw in w.

Exercise 32. Define the possible values of the random variable X in terms of the propositions ϕ,ψ, and χ. Which notation is easier to work with? What would happen if we had drawn 5 balls instead of 3, and introduced two more propositions to stand for the outcome that the fourth and fifth draws return red?

Exercise 33. Which notation is more expressive (i.e., allows us to define a finer-grained partition on W)? Exactly what information are we giving up when we ask the coarser-grained question?

(21) **Convention**: instead of writing $pr(X(w_@) = x)$, where x is some real number, I'll write $pr(X = x)$.

Exercise 34. Returning to the urn problem, for each possible value of x of X, find $pr(X = x)$.

Exercise 35. Generalize your solution to the last problem to a rule for finding the probability that n balls will be red in m draws in a sampling-with-replacement setup, given some probability p that a given ball will be red.

In the characterization of sampling with replacement in the urn problem, we had to specify that the urn is shaken each time we replace the ball drawn. If we didn't shake the urn, our choice on one draw might affect our choice on the next draw because the ball is on top, or because we remember where we put it down and are subtly drawn to (or away from) that location, etc. What we were trying to ensure by shaking the urn after each draw by was that each draw was **independent** of all other draws.

(22) **Def: Independence of random variables**. Random variables X and Y are independent if and only if, for all real numbers x and y, the propositions $X = x$ and $Y = y$ are independent, i.e. if $pr(X = x \wedge Y = y) = pr(X = x) \times pr(Y = y)$.

Independent random variables are variables where learning about one tells you nothing about the other, like the price of beans in China and the amount of traffic in Los Angeles. Dependent random variables are those where learning about one would allow you to make a better guess about the other, like learning about someone's height and learning about the same person's weight.

(23) **Def: random vector**. Let $\mathbf{X} = [X_1, ..., X_n]$ be a sequence of random variables. We call this a *random vector*. Sampling from $\mathbf{X}$ will return a vector $\mathbf{x} = [x_1, ..., x_n]$, where for each i and j the probability that $X_i = x_j$ is given by the distribution on X_i.

Note that the definition of independence for random variables implies that, if all of the random variables in $\mathbf{X}$ are independent, then

$$pr(X_1 = x_1 \wedge ... \wedge X_n = x_n) = \prod_{i=1}^{n} pr(X_i = x_i)$$

In the urn example, in addition to shaking the urn ensure independence of draws, we replaced the ball after each draw in the urn example in order to ensure that each draw in the sequence is **identically distributed**, i.e. has the same probability of returning a red ball.

(24) **Def: independent and identically distributed (i.i.d.)** A random vector $\mathbf{X} = [X_1, ...X_n]$ is *i.i.d* if and only if, for all i, j, X_i and X_j are independent; and for all $y \in V_{X_i}$, $pr(X_i = y) = pr(X_j = y)$.

Many statistical techniques assume that samples are drawn from i.i.d. random vectors, and practitioners have to do a considerable amount of work to ensure that this assumption is satisfied. If it isn't, the statistical conclusions are suspect.

Exercise 36. Think of a practical data-gathering situation in which samples might be independent but not identically distributed.

Exercise 37. Think of a situation in which samples might be identically distributed but not independent.

Exercise 38. Think of a situation in which neither property would hold.

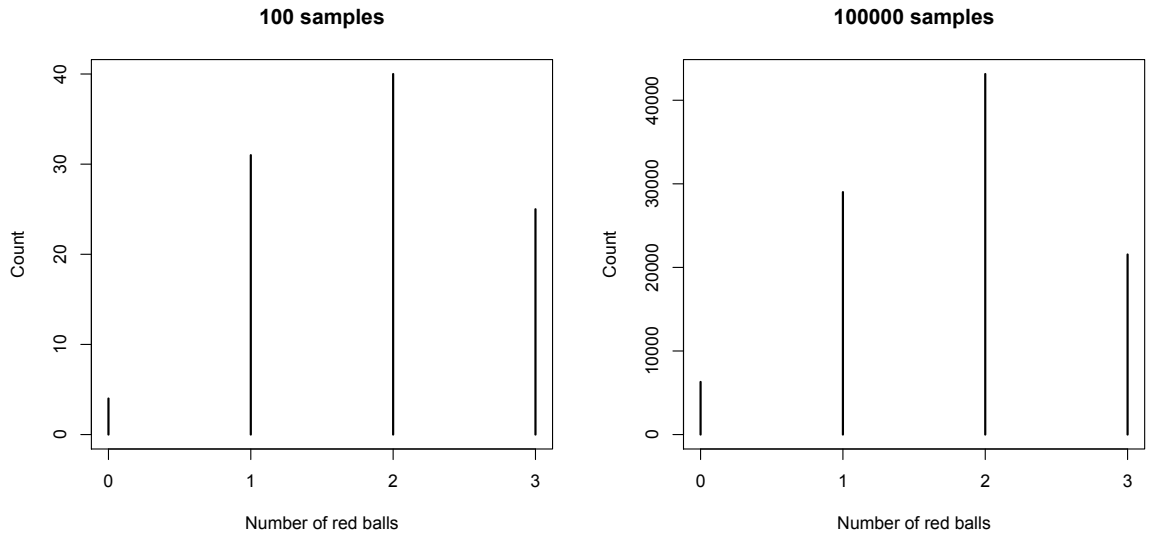
With these concepts in hand, let's do a simulation of the urn example to solidify intuitions and to check that our answers to earlier exercises were correct.

R code

```
urn.model = function(n.sims) {
   draws.per.sim = 3
   p.red = .6
   urn.results = rep(-1, n.sims)
   for (i in 1:n.sims) {
      draws = flip.n(p.red, draws.per.sim)
      num.red = howmany(draws, eq(TRUE))
      urn.results[i] = num.red
   }
   return(urn.results)
}

urn.100.samples = urn.model(100)
table(urn.100.samples)/100
   0    1    2    3
   0.04 0.31 0.40 0.25
plot(table(urn.100.samples), type='h', main="100 samples",xlab="Number of red
balls", ylab="Count")
# How do the results compare to your answer from ex.  34?

# What happens to the approximation if we increase the number of simulations?
urn.100000.samples = urn.model(100000)
table(urn.100000.samples)/100000
   0       1       2       3
   0.06300 0.29013 0.43140 0.21547
```



What we're doing here is really a roundabout way of sampling from a family of distributions called the *binomial*.

(25) **Def: Binomial distribution.** Suppose that we sample from an i.i.d. random vector of length n, where each sample returns 1 with probability p and 0 otherwise. This is the *binomial(n,p)* distribution. For each $x \in \{0,...,n\}$, the probability of getting exactly x 1's is equal to

$$\binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$$

(This was the solution to exercise 35, by the way.) The usual way to introduce the binomial is in terms of an experiment which is either a success or a failure, with probability p of being a success. If you repeat the experiment n times and the trials are i.i.d., then the distribution of successes and failures in the results has a binomial(n,p) distribution.

(26) **Def: Expectation/Mean**. The *expectation* or *mean* of a random variable X is the average of the possible values, weighted by their probability. For a random variable with n possible values $x_1,...,x_n$, this is

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i \times pr(X = x_i)$$

Sometimes instead of $\mathbb{E}(X)$ we write μ_X.

Exercise 39. Show that the expectation of a proposition is its probability. (Hint: expand the definition of expectation, undoing the abbreviation "$X = x_i$" defined in (21).)

Exercise 40. What is the expectation of a binomial(n,p) random variable?

(27) **Def: Variance.** The *variance* of a distribution is a measure of how spread out it is — of how far we can expect sample values to be from the mean. It's defined by

$$var(X) = \mathbb{E}((X - \mu_X)^2) = \mathbb{E}(X^2) - \mu_X^2$$

The *standard deviation* is the square root of the variance: $sd(X) = \sqrt{var(X)}$.

(28) **Def: Sample mean.** Let $\mathbf{x} = [x_1, ..., x_n]$ be a vector of samples from i.i.d. random vector $\mathbf{X}$. Then the *sample mean* of $\mathbf{x}$ is written $\bar{\mathbf{x}}$ and defined as

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Exercise 41. `mean` is the R function that calculates the sample mean of a vector. Type `mean(urn.100000.samples)` into the R console and see what it returns. Explain why this is the right result intuitively, and then compare it to the true mean that you get by applying the definition of expectation to the known probabilities from the urn model.

Ex. 13

Population distributions and sampling distributions. What's the average number of televisions in a household in United States? To find out for the exact value, we'd have to ask one person from each household in the U.S. how many TVs they have, and then average the results. If we could do this, the sample mean would of course be the same as the true mean.

But most of the time our desire to estimate such values precisely is tempered by our desire not to spend all of our money and the rest of our lives getting an answer. (Plus, the answer would probably change while we're conducting our huge survey.) For most purposes, an answer that is close to the true value is good enough. One way surveys like this are often done is to generate random telephone numbers and call the number to ask whoever answers. On the assumption that this procedure generates i.i.d. samples, if we ask enough people how many TVs they have, we can use the *sample distribution* to help us estimate the *population distribution*. For instance, imagine we call 10,000 people and find that 500 have no TV, 4000 have 1 TV, 3000 have 2 TVs, 2000 have 3 TVs, and the rest have 4. Then our best guess for the average number of TVs in a U.S. household is

$$.05 \times 0 + .4 \times 1 + .3 \times 2 + .2 \times 3 + .05 \times 4 = 1.8$$

Even though we certainly don't expect any particular household to have 1.8 televisions, these results suggest that the expected number of televisions in a U.S. household is about 1.8.

Exercise 42. Why might dialing random telephone numbers not be enough for us to generate an i.i.d. sample?

Exercise 43. If a vector of samples $\mathbf{x}$ is i.i.d., the expected value of the sample mean $\bar{\mathbf{x}}$ is equal to the expectation of the random variable μ_X from which it was drawn: $\mathbb{E}(\bar{\mathbf{x}}) = \mu_X$. Thinking about the survey example, explain in intuitive terms why this should be so.

Exercise 44. Calculate the sample variance and standard deviation in this survey.

Exercise 45. Suppose, instead of 10,000 people, we had gotten this sample distribution in a survey of only 20 people. Why might the sample variance not be a reliable estimate of the true variance in this case?

Using the sample mean to estimate the population mean seems intuitive, but we haven't officially shown that the sample mean of a big i.i.d. sample should be informative about a random variable whose expected value is unknown. At least for the case of means, there's an important result that tells us that we can rely on large i.i.d. samples to give us good estimates of the expectation of a random variable.

(29) **Weak law of large numbers.** Let $\mathbf{x} = \{x_1, ..., x_n\}$ be a vector of samples from i.i.d. random vector $\mathbf{X} = [X_1, ...X_n]$. Then as $n \to \infty$, $\bar{\mathbf{x}} \to \mathbb{E}(X_i)$ for any $X_i \in \mathbf{X}$.

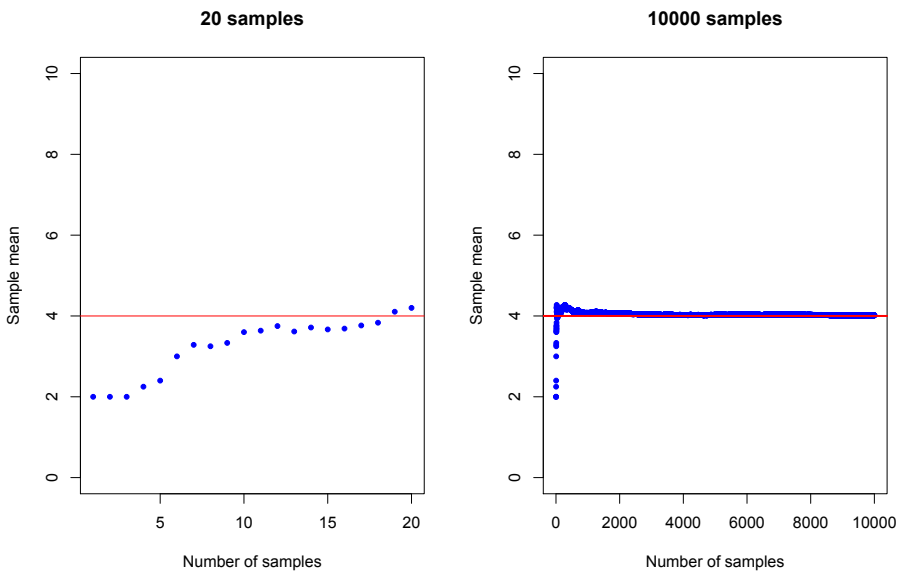
Instead of proving it, let's do a sanity check by simulating it. We'll generate a lot of samples from a distribution for which we know the true value (because we specified it): the binomial(10,.4). Recall that the expectation of a binomial(n, p) distribution is $n * p$, so the weak law of large numbers leads us to expect a mean of 4 once n is large enough. To verify this, each time we take a sample we'll compute the mean of all the samples we've taken so far, and at the end we'll plot the way the sample mean changes as n increases.

(Note: now that we've explicitly introduced the binomial distribution it would be better and quicker to do this using R's `rbinom` function. Type `?rbinom` in the console to see how it works. I'll keep using `flip.n` and `for`-loops, but only for continuity.)

R code

```
true.proportion = .4
n.samples = 10000
n.trials.per.sample = 10
binom.results = rep(-1, n.samples)
cumulative.mean = rep(-1, n.samples)
for (i in 1:n.samples) {
   samp = flip.n(true.proportion, n.trials.per.sample)
   binom.results[i] = howmany(samp, eq(TRUE))
   cumulative.mean[i] = mean(binom.results[1:i])
}
par(mfrow=c(1,2))   # tell R to plot in 2 panes, aligned horizontally
# plot cumulative mean of first 20 samples
plot(1:20, cumulative.mean[1:20], main="20 samples", pch=20, col="blue",
+ ylim=c(0,10), xlab="Number of samples", ylab="Sample mean")
abline(h=4, col="red", lwd=2)
# plot cumulative mean of all 1000 samples
plot(1:n.samples, cumulative.mean, main="10000 samples", pch=20, col="blue",
+ ylim=c(0,10), xlab="Number of samples", ylab="Sample mean")
abline(h=4, col="red", lwd=2)
```

Here's what I got:



Notice that, in the early parts of the sequence, the sample mean was off by as much as 50% of the true value. By the time we get to 200 samples, though, we've converged to the true proportion of .4. Type `cumulative.mean[9900:10000]` to see how little variation there is in the mean value once n is large. The law of large numbers tells us that this property will hold for any distribution, not just the binomial.

5 Statistical Inference

Depending on your philosophical commitments, the crucial question of statistical inference is either

> Suppose we have some samples that are known to come from a common distribution, but we don't know what the distribution is. How can we infer it? —

or the more general question

> Suppose we have some new information. How should we incorporate it into our existing beliefs? —

If you are of the Bayesian persuasion, the latter question subsumes the former as a special case. Frequentist statistics is exclusively concerned with the first question, and uses a variety of special-purpose tools to answer it. Bayesian statistics uses an all-purpose inferential tool, Bayes' rule, to answer both. This means that Bayesian statistics is in principle applicable to a much broader class of problems. Bayesian statistics also makes it possible to bring to bear prior knowledge in making statistical estimates, which is sometimes useful in practical data analysis and is absolutely critical if we're interested in cognitive modeling. However, under certain conditions the answers that the two methods give are the same.

We already talked a little bit about statistical inference in the context of estimating means by looking at samples. Now we'll make the assumptions more explicit and introduce the basic concepts

associated with the two schools. Remember, our focus here isn't on applied statistics, so we're not going to talk about t-tests, ANOVA, or anything like that. However, this material should still be helpful if you want to do practical data analysis in the future, because you'll be in a better position to evaluate the meaning of statistical tests and judge whether they are indeed applicable to a data set with a certain form that was gathered in a certain way.

5.1 Frequentist statistics: Basic concepts

A crucial fact about frequentist interpretations of probability is that the concept of the "probability of a single event" either doesn't make any sense, or is trivial (0 or 1). If you accept this, it has important consequences for the practice of statistics. In frequentist statistics, when we draw samples from a population — say, asking randomly chosen people how many TVs they have, or flipping a coin — we usually assume that the distribution is governed by one or more *fixed parameters*. In the coin-flipping example, for instance, we might suppose that the distribution on heads and tails in a reference class is governed by a fixed, but unknown, probability p of the coin coming up heads.

The "*fixed*" part of "*fixed parameters*" is really important. The true value of the parameter is a single event — a fixed fact about the world — and not a random variable. So it doesn't make sense to look at our data and speculate about the probability that the true value is .4 or .5 or .6. The true value just is. When we analyze data with our frequentist hats on, we have to be careful not to say things like "Given that we've seen 18 heads in 20 flips, it's very unlikely that the coin is fair". Instead, we have to say "If this is in fact a fair coin, we've just witnessed an extremely improbable event" — or equivalently, "Either this coin isn't fair, or something very unlikely just happened." (I might make this mistake from time to time. Point it out if I do, since — while we're being frequentists — saying things like that can only lead to confusion.)

That said, we do have some random variables to work with, such as the sample mean and variance. We might want to use the proportion of heads in the sample, $\hat{p}$, as our current best estimate of the fixed but unknown parameter p. But the interpretation of the estimate is still subtle: $\hat{p}$ isn't the "most likely value of the parameter" given the data; its probability is either 0 if $\hat{p} \neq p$, or 1 if $\hat{p} = p$.

One of the most commonly used methods in applied statistics is called **null hypothesis testing** (NHT). In its simplest form, the idea is roughly this. Pick a hypothesis, H_0, which you're trying to show false. It's usually wise to pick a H_0 that is plausible, or else your results won't be interesting even if you succeed in rejecting H_0. (In other words, it's possible to get "significant" results simply by setting up a straw-man H_0.) The alternative hypothesis, H_1, is simply the negation of H_0. In many cases, H_0 will be the hypothesis that there is no association between two variables, or that there is no difference between two groups.

For example, suppose I want to know whether height and income are related. The obvious choice of H_0 is that they are not: the distribution of incomes is the same for all heights. H_1 would be that the distribution of incomes is not the same. (This is a "2-sided hypothesis": data indicating either that taller people make more money or that shorter people do would be enough to justify rejecting H_0.) We would be justified in rejecting this H_0 if our data showed a statistical relationship between height and income that is very unlikely to have occurred by chance in a sample of this size,

assuming that there is fact no relationship. Or suppose I want to test a new treatment for insomnia. I find some insomniacs and divide them into two groups, one which receives the new treatment and one which does not (the control group). H_0 here would be that the new treatment is not any better than no treatment. H_1 would be that patients in the treatment group sleep better. (This is a "1-sided hypothesis".)

The logic of NHT requires that we specify in advance how unlikely the data must be according to a null hypothesis before we are justified in rejecting the null. This parameter is called the **level of significance** and usually abbreviated α. The standard setting of α is .05. Before actually analyzing any data, we use α to set up a **critical value** — a threshold such that all values more extreme that that will fall into the **region of rejection**. If H_0 is true, the probability that the data will fall into the region of rejection is at most α. Results that do fall into the region of rejection are **statistically significant**.

Now we analyze the data, and we reject H_0 if and only if the data fall into the region of rejection that we specified before starting our data analysis. If we don't reject H_0, we don't accept it either: rather, we *fail to reject* it. This is really important. In fact, given the way that the logic of NHT works, it's *impossible* to prove that the null hypothesis is true. Null hypotheses are something that you can only reject, or fail to reject. If the data do not allow us to reject the null hypothesis, this is emphatically *not* evidence that the null hypothesis is true. (This feature of NHT has been criticized and led some to adopt Bayesian methods, cf. Gallistel 2009.)

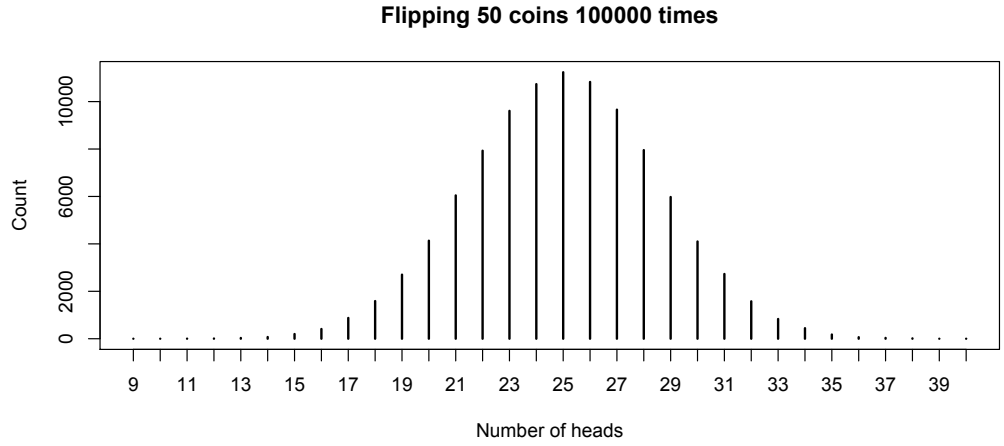
Statistics books and software make it easy to find significance levels for hypotheses that fall into common distributions like the binomial, but it is instructive to find the answers ourselves by doing a simulation. Simulating such values is also a useful skill to develop for use in situations in which you want to build statistical models that R doesn't have built-in capacities for.

So, suppose we're planning to flip a coin 50 times. H_0 is that the coin is fair, and H_1 is that the coin is not fair. This is a two-sided hypothesis, so the region of rejection should be divided into two parts: an area where the number of heads is improbably high if H_0 is true, and an area where the number is improbably low if H_0 is true. Each area should have probability .025 under H_0.

To find the region of rejection, let's flip many fair coins 50 times each, and find out what distribution of outcomes to expect. This time, we'll speed things up by using R's built-in function for flipping coins, `rbinom`. For instance, the command `rbinom(10,50,.5)` will return a vector of length 10, where each element of the vector is the number of heads in 50 flips of a fair coin. This is like repeating `flip.n(50,.5)` 10 times with a `for`-loop, as we did above, but faster and easier to type.

R code

```
n.sims = 100000
sim.results = rbinom(n.sims, 50, .5)
summary(sim.results)
plot(table(sim.results), type="h", main="Flipping 50 coins 100000
+ times",xlab="Number of heads", ylab="Count")
```



(Note the bell shape: the Central Limit Theorem strikes again.) We can now inspect these simulated coin flips to see what the rejection thresholds are. First we'll sort the data, and then we'll check to see what the values at the 2.5th and 97.5th percentiles are.

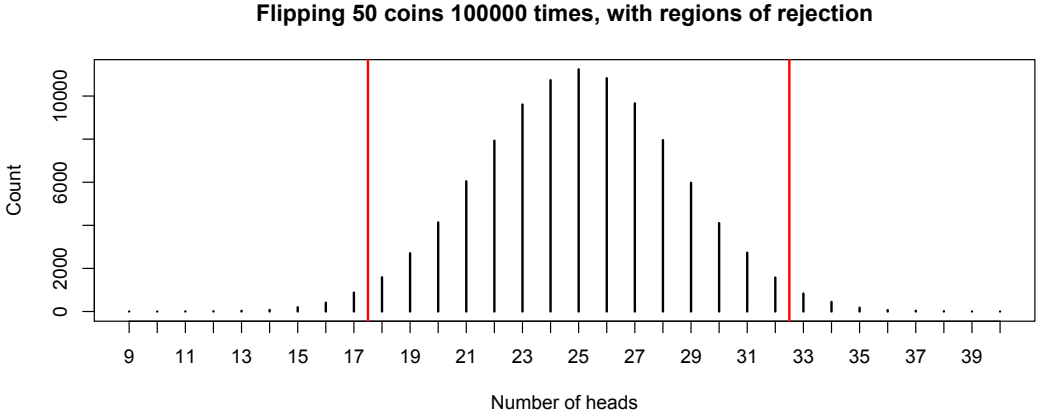
R code

```
sorted = sort(sim.results)
sorted[.025*n.sims]
   18
sorted[.975*n.sims]
   32
```

Let's add lines indicating the borders between the non-significant and significant regions:

R code

```
plot(table(sim.results), type="h", main="Flipping 50 coins 100000 times, with
+ regions of rejection", xlab="Number of heads", ylab="Count")
abline(v=c(18,32), col="red", lwd=2)
```



There's actually an easier way to do this for simulated values, for your future convenience:

R code
```
quantile(sim.results, c(.025, .975))
   18   32
```

And indeed an even easier way to find the rejection thresholds which is specific to binomials and doesn't require us to do a simulation.

R code
```
qbinom(p=c(.025,.975), size=50, prob=.5)
   18   32
```

I think that it's helpful to do the simulations, though, even when R has built-in functions that compute the values for you. This is for two reasons. First, it allows us to assure ourselves that the we understand the statistical model we're using. If these values had not agreed, we'd know that something was wrong with our model, or else the R commands don't do what we think they do. Second, when you're dealing with more complicated models or ones you've designed yourself, functions to compute these statistics may not exist, so you need to know how to compute these values without help.

The ***p*-value** of some data is just the probability that you would observe the data that you in fact observed *or something more extreme*, on the assumption that the null hypothesis is true. This is why lower p-values are "better". If $p < .001$, the chance of getting a value so extreme in a sample as large as your data set is less than 1 in a thousand if H_0 is true. In this case, H_1 is looking pretty good.

Exercise 46. Suppose we have two coins c_1 and c_2, and H_0 = "c_i is fair" for each. We flip both coins and get back 20 heads for c_1 and 36 heads for c_2. What are the p-values in each case? Find out by inspecting the simulation results, using the command `which(sorted == 20)` and `which(sorted == 36)`. With each command you'll get back a vector of indices in the sorted simulation results, representing the positions in the sorted results vector which contain the values 20 or 36 respectively. Explain what the indices mean.

Exercise 47. Check your answer to the previous question using the R commands `pbinom(q=20, size=50, prob=.5)` and `pbinom(q=36, size=50, prob=.5)`. How do these values compare to your estimate from the last question? If there are divergences, can you explain them?

There are two different kinds of errors that you could make in NHT:

(30) a. **Type 1 error**: You reject H_0 when it's actually true.
b. **Type 2 error**: You fail to reject H_0 when it's actually false.

The choice of α is, in principle, meant to be governed by the relative practical importance of avoiding type 1 vs. type 2 errors.

Exercise 48. Think of a situation where it would be so important to avoid type 1 errors that it would be reasonable to lower α, say to .01 or .001.

Ex. 14

Statistical significance ≠ practical importance. It's easy — for example in reading popular science reporting — to get the impression that if a study shows a significant effect of some factor then the factor must be important. For example, if a well-publicized study of a large number of adult males showed that eating red meat was associated with a significantly increased risk of heart disease, doubtless many men would think that eating red meat made them much more likely to get heart disease, and would stop. This interpretation is utterly inappropriate. The mere fact that a result is statistically significant tells you nothing at all about the **effect size**. If an effect is real but very small, you can easily make it statistically significant by gathering a huge amount of data. Before I stopped eating red meat, I'd want to know what the estimated increase in risk is — e.g., whether the increase is from 10% to 11% or from 10% to 30%.

In addition to null hypothesis testing, frequentist statistical techniques often make use of **parameter estimation** and associated **confidence intervals**. Here is a much-too-quick introduction. When we flip a coin with unknown parameter p, our estimate of the parameter is written $\hat{p}$. There are various ways to form an estimate, but the most commonly used is **maximum likelihood estimation**. First, we define the likelihood function for a parameter θ, assuming that $\mathbf{x} = [x_1, ..., x_n]$ are samples from an i.i.d. random vector.

(31) **Def: likelihood function**. $\mathcal{L}(\theta) = \prod_{i=i}^{n} pr(x_i; \theta)$

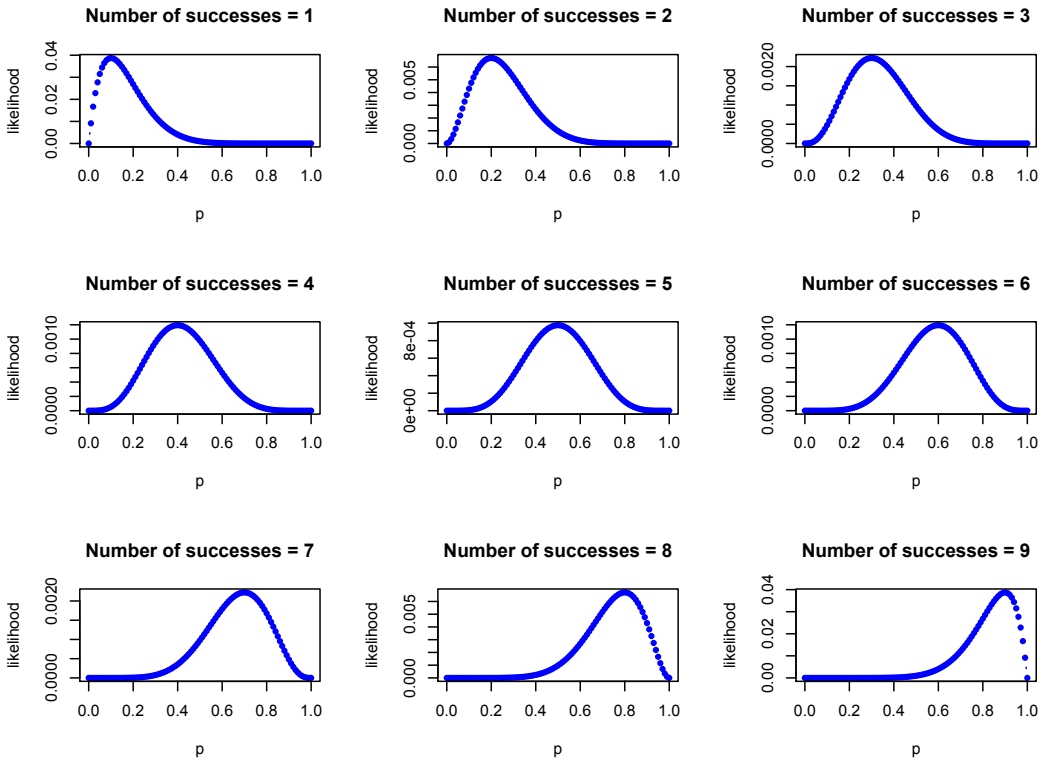
Note that in frequentist statistics we write "$pr(x_i; \theta)$", with a semicolon, rather than ("$pr(X_i|\theta)$") in order to make it clear that θ is a fixed parameter. It is not a random variable that we could condition on, so the use of the bar would be inappropriate.

If we have data drawn from a binomial distribution, the maximum likelihood estimate $\hat{p}$ of the parameter p is the proportion of successes (heads). To convince yourself of this, note that the likelihood of m successes in n independent trials under parameter p is $p^m \times (1-p)^{n-m}$. Suppose we're flipping a coin 10 times. For each possible number of successes, we can search over a grid of possible values of p spaced at .01 to check for ourselves that the maximum likelihood estimate is indeed the number of successes divided by 10.

R code

```
p=seq(from=0, to=1, by=.01)
n.successes = seq(from=1, to=9, by=1)
n.flips = 10
res=matrix(-1, nrow=length(p), ncol=length(n.successes))
#rows are values of p, columns are numbers of successes
for (i in 1:length(p)) {
   for (j in 1:length(n.successes)) {
      lik = p[i]^(n.successes[j]) * (1-p[i])^(n.flips - n.successes[j])
      res[i,j] = lik
   }
```

```
}
par(mfrow=c(3,3))
for (i in 1:9) {
   plot(p, res[,i], xlab="p", ylab="likelihood", main=paste("Number of
   successes =", i), pch=20, type='b', col="blue")
}
```



It's often a good idea to construct a *confidence interval* (CI) around a parameter estimate. Usually we use 95% CIs. The correct frequentist interpretation of a confidence interval is *not* that we can be 95% certain that the true value of the parameter falls into some range. Rather, the interpretation is this: if we conduct many experiments and construct 95% confidence intervals with each experiment, then in 95% of our experiments the true value of the unknown parameter will fall within the interval that we constructed for that experiment.

Let's flip a single fair coin 20 times and see what $\hat{p}$ and the 95% CI are, and how they compare to the fixed parameter $p = .5$. Now that we know how to use the `qbinom` function to find quantiles of the binomial(n, p) distribution, we can use it to find out what the 95% CI is given the sample that we have. Here's one way we could do it. (N.B. there are various ways to calculate CIs for the binomial, and this is certainly not the best; in particular, if `n.flips` were larger a normal approximation would be better. It's enough to illustrate the idea, though.)

R code

```
n.flips = 20
one.sample = rbinom(n=1, size=n.flips, prob=.5)
p.hat = one.sample/n.flips
p.hat
   0.55
sample.ci = qbinom(p=c(.025,.975), size=n.flips, prob=p.hat)
p.hat.ci = sample.ci/n.flips
p.hat.ci    0.35   0.75
```

The 95% CI estimated from my sample was $[.35, .75]$. This includes the true value of p, .5.

Exercise 49. Modify the example code to sample 10,000 times from a binomial(20,.5) distribution. Compute the 95% confidence interval for the estimator $\hat{p}$ with each sample. How many of the sample CIs include the true parameter value of .5? (Hint: it may help to use a matrix to store the values, as in `res = matrix(-1, nrow=10000, ncol=2)`. The command for copying the length 2 vector `p.hat.ci` into row `i` of matrix `res` is `res[i,] = p.hat.ci`.)

Exercise 50. Was your result from the previous exercise equal to the expected 5% of CIs for which the true value is not included in the interval? If not, what does your result tell us about the method of estimating CIs that we're using?

Further reading: For an easy but illuminating entry into frequentist statistical methods, see Hacking (2001: §15-19). Wasserman (2004) provides an excellent survey at an intermediate mathematical level. Cohen (1996) is a good textbook on frequentist statistical methods in psychology. Baayen (2008) and Gries (2009) are introductory texts geared toward linguistic analysis which make extensive use of R.

Discussion question: Many of us, I take it, are interested in language and reasoning. What are some ways that we could use frequentist statistical techniques to inform our understanding of these topics?

Optional topic (time permitting): Parametric models aren't obligatory in frequentist statistics; bootstrapping is a cool and useful non-parametric technique. R makes it easy to find bootstrap estimates of distributions, even wonky ones.

5.2 Bayesian statistics: Basic concepts

Philosophically, Bayesian statistics is very different from frequentist statistics — naturally, as different as the Bayesian and frequentist interpretations of probability. Bayesian methods apply in a wider range of situations, and are more flexible. However, in cases in which a data set can be analyzed in both ways you will frequently get the same answer. We have to be careful not to overstate the differences, then; but they are substantial. Here are a few:

- Bayesian statistics is built around a single all-purpose inferential mechanism, conditionalization (implemented via Bayes' rule).
- There is no difficulty in applying the same methods to estimating probabilities for repeatable

and non-repeatable events.

  - This means that we can deal with a much broader class of inference problems.
  - It's also crucial if you want to use probability to do cognitive modeling (including reasoning and language understanding).

- Parameters are treated as random variables, and parameter estimation as an ordinary inference problem.
- In Bayesian statistics it's possible to incorporate prior information into a model.
  - This makes the method considerably more flexible, which can be invaluable for some purposes, but has also been the target of criticism.
  - It's also crucial for cognitive modeling (reasoning & language), since real agents always approach inference problems with some prior knowledge.
- Bayesian methods make it possible to show that a null hypothesis is probably correct, rather than merely "failing to reject" it.
- Bayesian methods also extend to model comparison, which is treated as just another inference problem.

Finally, Bayesian thinking encourages us — instead of trying to assimilate data to some known distribution whenever possible — to create our own models of how some data was generated and the "invert" the model via Bayes' rule to make inferences about the likely values of model parameters given the data. They're also great for building and making inferences with hierarchical models, where the values of parameters can influence the values of other parameters. (See Noah Goodman's course for a lot more about this.)

One of the main criticisms that have been leveled at Bayesian methods is their "subjectivity", the fact that statisticians with different priors could in principle come to different conclusions. There is some truth in this criticism, but there are methods for managing it when it is desirable to, which we'll talk about a bit below. (And, of course, for cognitive modeling purposes this feature is desirable, since different people do draw different inferences from the same data depending on their prior beliefs!)

Remember the statement of Bayes' rule from §4:

$$pr(h|E) = \frac{pr(E|h) \times pr(h)}{pr(E)} = \frac{pr(E|h) \times pr(h)}{\sum_{h'} (pr(h'|E) \times pr(h'))} = \frac{\text{likelihood} \times \text{prior}}{\text{normalizing constant}}$$

In many cases, the normalizing constant will drop out of an equation, e.g. if we're only interested in comparing the relative plausibilities of two hypotheses. In this case we can just write

$$\frac{pr(h_1|E)}{pr(h_2|E)} = \frac{pr(E|h_1) \times pr(h_1)}{pr(E|h_2) \times pr(h_2)}$$

Often Bayesians give the proportional version even when there's only a single hypothesis under discussion —

$$pr(h|E) \propto pr(E|h) \times pr(h) = \text{likelihood} \times \text{prior}$$

— because the normalizing constant is the same for all hypotheses. Computing it is trivial once you've found the likelihood × prior for each hypothesis: you just sum over them.

Whenever you want to set up an inference problem using Bayesian methods, you have to specify three components:

- Hypothesis space
- Prior
- Likelihood function

Once you have these three components, you can do inference by conditionalizing the prior on the new evidence E using Bayes' rule.

At a high level, this is the whole theory! The devil is in the details, though — it's often non-trivial to specify these components. Let's return to our coin-flipping example, and see how we should think about it with our Bayesian hats on now.

Ex. 15 **Fair and weighted coins**. Forgetting for a moment about statistical theory and just thinking about common sense: suppose someone you don't know takes out a coin which you haven't ever seen and flips it 10 times. It comes up heads 4 of the 10 times.

Exercise 51. As we saw in the last subsection, the frequentist maximum likelihood estimate given this data is $\hat{p} = .4$. Is this a reasonable estimate? Why or why not?

Bayesian inference for estimating coin weight

What we want to know: the coin weight P.
What we already know: the number of flips N and the number of heads X.

Bayes' rule tells us

$$pr(P = p|X = x, N = n) = \frac{pr(X = x|N = n, P = p) \times pr(P = p)}{pr(X = x)} = \frac{pr(X = x|N = n, P = p) \times pr(P = p)}{\sum_{p'} [pr(X = x|N = n, P = p') \times pr(P = p')]}$$

Note that N, X, and P are all random variables, so we can do inference about any or all of them as needed. Since X and N are known, we have

$$pr(p|X = 4, N = 10) = \frac{pr(X = 4|N = 10, p) \times pr(p)}{\sum_{p'} [pr(X = 4|N = 10, p') \times pr(p')]}$$

What we're doing here can be thought of as "inference by inversion". We have a model of a sampling process with some parameters that generated the data, and we run the model in reverse to infer likely values of the parameters given that the process actually generated the data that it did.

Hypothesis space: Possible coin weights are in the interval $[0,1]$, so P can take any of these values.

Likelihood: The natural Bayesian approach to the problem of estimating the coin weight is to use a binomial model, just as the frequentist model did. So, given some coin weight p, the probability of getting x heads in n i.i.d. flips is given by

$$p(X = x|N = n, P = p) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!(n-x)!}{x!} p^x (1-p)^{n-x}$$

which is, remember, just a way of adding up the probabilities of all the different ways that n flips could give you x heads.

Prior: Depending on how you answered exercise 51, you might have two inclinations. The assumption that most people naturally make, I think, is that *most coins are fair*. You really have to go out of your way to bias a coin, and frankly it's kind of a weird activity to be involved in, without much hope of profit. Making a double-headed or double-tailed coin is easier, but still, it's not really all that common. Non-statisticians typically consider it perfectly reasonable to bring this prior information to bear in drawing inferences about the case at hand. For Bayesians, this is perfectly legitimate information for a statistician to use as well.

If you're attracted to this reasoning, you could use a prior which favors fair coins:

> **Prior #1: Fair coins dominate.** The probability that a coin is fair is high, say .95. The probability that it is double-headed or double-tailed is low, say .02 each. The probability that it has any other weight is just .01. That tiny quantity is further divided by being distributed among the possible weights other than 0, .5, and 1.

If you aren't attracted to the reasoning, or if you have philosophical qualms about letting prior beliefs influence your inferences, that's OK too:

> **Prior #2: Flat/uninformative/maxent.** If you don't have reason to think that most coins are fair, or if you decline to make use of this information, you can assign a *flat* or *uninformative* prior: all weights are equally likely. (Cf. Keynes' "principle of indifference", §2.)

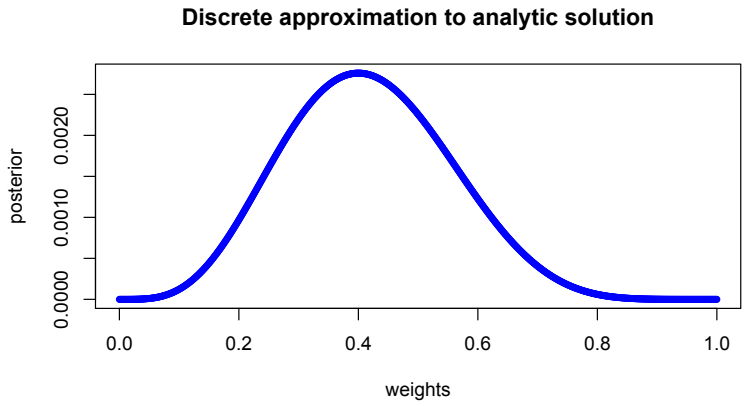
Suppose first that we adopt the flat prior. Inference is simply finding the posterior distribution on coin weights, given the observation that 4 of 10 flips came up heads. We can solve the problem in various ways. In this case, it's possible to solve it analytically (though for many problems this won't be possible). With a flat prior, $pr(p)$ is the same for every possible value of p; call this value c. We have

$$pr(p|X = 4, N = 10) = \frac{pr(X = 4|N = 10, p) \times c}{\sum_{p'} [pr(X = 4|N = 10, p') \times c]} = \frac{c \times pr(X = 4|N = 10, p)}{c \times \sum_{p'} pr(X = 4|N = 10, p')}$$

c cancels out, so the posterior weight of p is proportional to the likelihood of the observed data given p.

$$pr(p|X = 4, N = 10) \propto pr(X = 4|N = 10, p) = \binom{10}{4} p^4 (1-p)^6$$

Plotting, using a very fine discrete approximation to the hypothesis space:

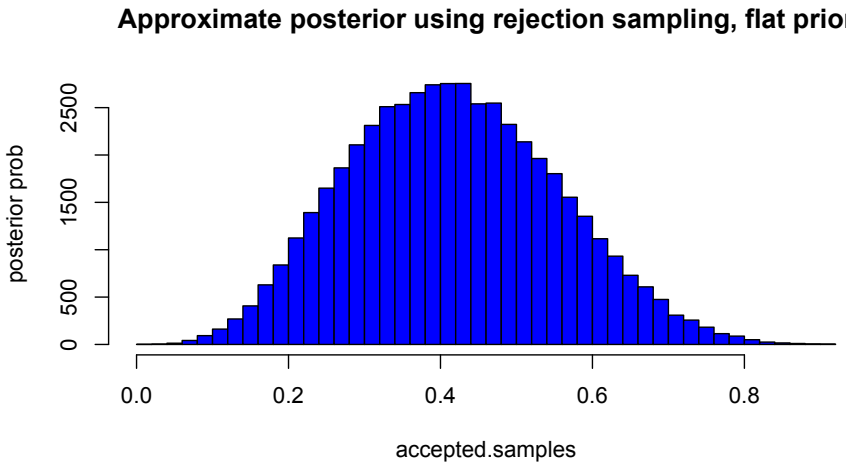


Exercise 52. Write some R code that will generate the graph you see.

Let's do a simulation to check our answer and explore an alternative way of approximating the posterior. As before, we'll simplify things by using a discrete approximation to the continuous hypothesis space; so instead of all weights in $[0,1]$, we'll only consider $0,.01,.02,....99,1$. The method we'll use this time is **rejection sampling**. Rejection sampling is just sampling from the prior and then discarding samples that don't satisfy the conditioning expression. The proportion of accepted samples taken from h is approximately equal to the conditional probability of h given E.

R code

```
weights = seq(from=0, to=1, by=.01)
accepted.samples = c()
n.samples = 50000
while (length(accepted.samples) < n.samples) {
   sample.weight = runif(1,0,1)
   sim = rbinom(n=1, size=10, prob=sample.weight)
   if (sim==4) accepted.samples = c(accepted.samples, sample.weight)
}
hist(accepted.samples, breaks=50, col="blue", main="Approximate posterior using
rejection sampling, flat prior", ylab="posterior prob")
```



Exercise 53. The graphs suggest that our two methods of finding the Bayesian posterior are doing the same thing. (Whew!) What might be surprising is the similarity between both of them and the corresponding graph (from a few pages ago) of the frequentist likelihood statistic when the number of successes is 4 out of 10. Can you explain why the frequentist and Bayesian parameter estimates are so similar here? What would we need to do to our model in order to decouple them?

Exercise 54. Use `quantile` to get an estimate of the Bayesian 95% CI.

Exercise 55. Re-run the simulation, assuming that this time 9 of 10 flips were heads. What are the estimated weight and CI?

Rejection sampling is frequently very slow, especially when the prior probability of the conditioning proposition is low. (Why is there a connection?) There are much better sampling methods which do the same thing more cleverly, but this is good enough for our purposes.

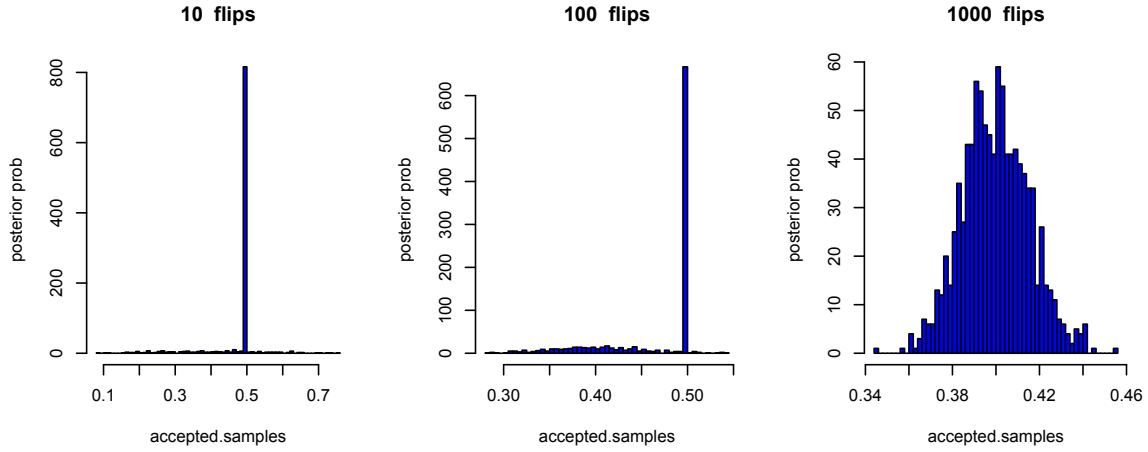
But keep in mind how deeply counter-intuitive this all is. After seeing 4 out of 10 flips of a coin come up heads, my best estimate of the coin's weight wouldn't be .4; it would still be .5, and I would assume that this is ordinary variation in sampling. However, things would be very different if I had seen 4000 out of 10,000 flips come up heads; suddenly it seems really unlikely that this is just sampling noise. If we're interested in modeling the inferences that people actually make, then, assuming a flat prior may be a bad idea. But can capture these intuitions in a Bayesian model by adopting prior #1.

This time, instead of sampling a coin weight at random from $[0,1]$, we'll honor the intuition that most coins are fair and that biased coins are usually double-sided. Let's also see how the posterior is sensitive to the number of flips, keeping the proportion of observed heads (.4) constant.

R code

```
flip.ns = c(10, 100, 1000)
par(mfrow=c(1,3))
for (i in 1:3) {
   n.flips = flip.ns[i]
   accepted.samples = c()
   n.samples = 10000
```

```
   while (length(accepted.samples) < n.samples) {
      coin.type = sample(c("fair", "double", "uniform bias"), 1)
      if (coin.type == "fair") {sample.weight = .5 }
      else if (coin.type == "double") {sample.weight = sample(c(0,1),1)}
      else {sample.weight = runif(1,0,1)}
      sim = rbinom(n=1, size=n.flips, prob=sample.weight)
      if (sim==.4*n.flips) accepted.samples = c(accepted.samples,
sample.weight)
   }
   hist(accepted.samples, breaks=50, col="blue", main=paste(n.flips,"flips"),
ylab="posterior prob")
}
```



Even with a very strong prior in favor of fair coins, the evidence overwhelms the prior as the number of flips increases and the chance of getting 40% heads by chance decreases. With 400 heads out of 1000 flips, the posterior even has the eponymous bell shape centered around .4, and weight .5 — which was heavily favored by the prior — doesn't even make it into the simulation results. This behavior fits the intuitions we (or at least I) expressed earlier quite well.

Note also that, despite having much higher prior weight than any arbitrarily biased coin, double-headed and double-tailed coins have zero posterior weight after we've observed 4 heads and 6 tails. This is as it should be.

Exercise 56. Why, practically speaking, shouldn't you try to answer the question in this way for n much larger than 10,000?

Bayesians get a lot of flack for needing priors in their model. But as we just saw, using an uninformative prior a Bayesian posterior can mimic frequentist likelihood. In this case, the extra degree of freedom just isn't doing anything. On the other hand, with a strongly biased prior, the evidence overwhelms the prior preference for fair coins as the number of flips increases. This is as

it should be; but note that NHT behaves in a similar way, in that the null can't be rejected if the true parameter value is close to it, unless the sample size is large.

One of the distinct advantages of Bayesian methods is the ease and intuitiveness of writing down **hierarchical models**. These are models where the value of one parameter may influence the value of others, and we have to estimate them jointly. For example, suppose you're going to meet someone, and you don't know their ethnic background. When you meet them, you'll be in a better position to guess their ethnic background, and this will in turn enable you to make more informed guesses about properties of their family members, such as hair color and languages spoken. This is true even though there is variation in hair color and language among most ethnic groups; it's just that there is a strong tendency for such traits to co-vary among members of families, and learning about one is informative even if not determinate about the others.

Since this is hard to model, let's go back to a hokey coin-flipping example. (Sorry.)

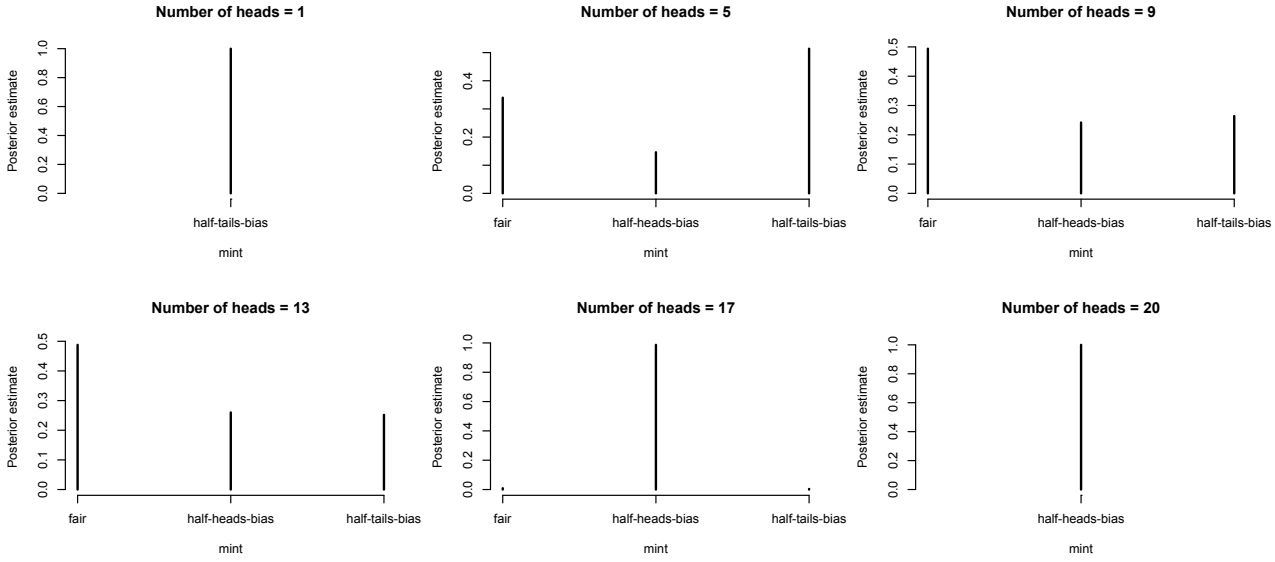
Ex. 16

The three mints. There are three mints: one makes only fair coins; one makes half fair coins and half coins with a bias of .75; and a third makes half fair coins and half with a bias of .25. You've got a coin, and you have no idea what mint it comes from. You flip it 20 times and get some number of heads n. How does your guess about which mint it came from depend on n?

R code

```
heads.ns = c(1,5,9,13,17,20)
par(mfrow=c(2,3))
for (i in 1:9) {
   n.flips = 20
   accepted.samples = c()
   n.samples = 1000
   observed.heads = heads.ns[i]
   while (length(accepted.samples) < n.samples) {
      mint = sample(c("fair", "half-heads-bias", "half-tails-bias"), 1)
      if (mint == "fair") {
         sample.weight = .5
      } else if (mint == "half-heads-bias") {
         sample.weight = sample(c(.5, .9), 1)
      } else {
         sample.weight = sample(c(.5, .1), 1)
      }
      sim = rbinom(n=1, size=n.flips, prob=sample.weight)
      if (sim==observed.heads) accepted.samples = c(accepted.samples, mint)
}
plot(table(accepted.samples)/n.samples, xlab="mint", ylab="Posterior estimate",
main=paste(observed.heads, "Heads"))
```

```
}
```



This is a type of inference that people are generally very good at, and about which Bayesian models make intuitively correct predictions. Hierarchical models are very useful for both applied data analysis and cognitive modeling. (Noah Goodman's course will have much more on this topic.) Here's a famous example.

Ex. 17 **Wet grass**. You wake up one morning and observe that the grass is wet. Two hypotheses suggest themselves: either it rained, or someone left the sprinkler on. (Pearl 1988)

Exercise 57. Intuitively, how likely is it that it rained? That the sprinkler was left on? How do these compare to the likelihood of these events on an arbitrary day, when you don't know whether the grass is wet?

Exercise 58. Suppose you learn that it rained. How likely is it now that the sprinkler was left on? How does this compare to the probability that this proposition would have if you didn't know whether the grass was wet? (This is called **explaining away**.)

Exercise 59. Design a simulation in R, using rejection sampling, that displays the qualitative behavior suggested by your answers to exercises 57 and 58.

Discussion question: What are some ways that we could use Bayesian statistics to inform our understanding of language and reasoning?

Further reading: Kruschke (2012) is an entertaining and thorough introduction to conceptual, mathematical, and computational aspects of Bayesian data analysis. Gelman & Hill (2007) is

an excellent intermediate-level text focusing on practical aspects of regression and hierarchical modeling and borrowing from both frequentist and Bayesian schools. Gelman, Carlin, Stern & Rubin (2004) is the Bible, but is more challenging. There are several good, brief surveys of the main ideas behind Bayesian cognitive models, including Chater et al. 2006 and Tenenbaum et al. 2011. Griffiths et al. (2008) is more detailed.

References

Baayen, R.H. 2008. *Analyzing linguistic data: A practical introduction to statistics using r*. Cambridge Univ Press.

Butler, Joseph. 1736. *The analogy of religion, natural and revealed, to the constitution and course of nature: to which are added, two brief dissertations: on personal identity, and on the nature of virtue; and fifteen sermons*.

Carnap, R. 1950. *Logical foundations of probability*. University of Chicago Press.

Chater, Nick, Joshua B. Tenenbaum & Alan Yuille. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences* 10(7). 287–291. doi:10.1016/j.tics.2006.05.007.

Cohen, B. 1996. *Explaining psychological statistics*. Thomson Brooks/Cole Publishing.

Cox, R.T. 1946. Probability, frequency and reasonable expectation. *American journal of physics* 14(1). 1–13. http://algomagic.org/ProbabilityFrequencyReasonableExpectation.pdf.

de Finetti, Bruno. 1937. Foresight: Its logical laws, its subjective sources, 53–118. Reprinted in H. E. Kyburg and H. E. Smokler, H. E., editors,(eds.), *Studies in subjective probability*. Krieger.

Gallistel, CR. 2009. The importance of proving the null. *Psychological Review* 116(2). 439.

Gelman, A. & J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern & Donald B. Rubin. 2004. *Bayesian data analysis*. Chapman and Hall/CRC.

Gigerenzer, Gerd. 1991. How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology* 2(1). 83–115. doi:10.1080/14792779143000033.

Gries, S.T. 2009. *Statistics for linguistics with R: A practical introduction*. Walter de Gruyter.

Griffiths, Thomas L., Charles Kemp & Joshua B. Tenenbaum. 2008. Bayesian models of cognition. In R. Sun (ed.), *Cambridge handbook of computational psychology*, 59–100. Cambridge University Press.

Groenendijk, Jeroen & Martin Stokhof. 1984. *Studies in the Semantics of Questions and the Pragmatics of Answers*: University of Amsterdam dissertation.

Hacking, I. 2001. *An introduction to probability and inductive logic*. Cambridge University Press.

Halpern, Joseph Y. 1999. Cox's theorem revisited. *Journal of Artificial Intelligence Research* 11. 429–435.

Jaynes, E.T.. 2003. *Probability theory: The logic of science*. Cambridge University Press. http://omega.albany.edu:8008/JaynesBook.html.

Jeffrey, Richard C. 1965. *The logic of decision*. University of Chicago Press.

Jeffrey, Richard C. 2004. *Subjective probability: The real thing*. Cambridge University Press. http://www.princeton.edu/~bayesway/Book*.pdf.

Kennedy, Chris. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30(1). 1–45.

Kennedy, Chris & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2). 345–381.

Keynes, John Maynard. 1921. *A Treatise on Probability*. Macmillan.

Kolmogorov, Andrey. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer.

Kratzer, Angelika. 1991. Modality. In von Stechow & Wunderlich (eds.), *Semantics: An international handbook of contemporary research*, de Gruyter.

Kruschke, John. 2012. *Doing bayesian data analysis: A tutorial introduction with r and bugs*. Academic Press.

Laplace, Pierre. 1814. Essai philosophique sur les probabilités.

Lassiter, Daniel. 2010. Gradable epistemic modals, probability, and scale structure. In Li & Lutz (eds.), *Semantics and Linguistic Theory (SALT) 20*, 197–215. Ithaca, NY: CLC Publications.

Lassiter, Daniel. 2011. *Measurement and Modality: The Scalar Basis of Modal Semantics*: New York University dissertation.

Lewis, D. 1980. A subjectivist's guide to objective chance 263–293.

MacKay, David J.C. 2003. *Information theory, inference, and learning algorithms*. Cambridge University Press. http://www.inference.phy.cam.ac.uk/itprnn/book.pdf.

Mellor, D.H. 2005. *Probability: A philosophical introduction*. Routledge.

von Mises, R. 1957. *Probability, statistics, and truth*. Allen and Unwin.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, Judea. 2000. *Causality: Models, reasoning and inference*. Cambridge University Press.

Popper, Karl R. 1959. The propensity interpretation of probability. *The British journal for the philosophy of science* 10(37). 25–42.

Ramsey, F.P. 1926. Truth and probability. In *The foundations of mathematics and other logical essays*, 156–198.

van Rooij, Robert. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy* 26(6). 727–763. doi:10.1023/B:LING.0000004548.98658.8f.

van Rooij, Robert. 2004. Utility, informativity and protocols. *Journal of Philosophical Logic* 33(4). 389–419. doi:10.1023/B:LOGI.0000036830.62877.ee.

Savage, Leonard J. 1954. *The Foundations of Statistics*. Wiley.

Tenenbaum, J.B. 1999. *A bayesian framework for concept learning*: MIT dissertation. http://dspace.mit.edu/bitstream/handle/1721.1/16714/42471842.pdf.

Tenenbaum, J.B., C. Kemp, T.L. Griffiths & N.D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022). 1279. http://www.cogsci.northwestern.edu/speakers/2011-2012/tenenbaumEtAl_2011-HowToGrowAMind.pdf.

Van Horn, K.S. 2003. Constructing a logic of plausible inference: A guide to Cox's theorem. *International Journal of Approximate Reasoning* 34(1). 3–24. http://ksvanhorn.com/bayes/Papers/rcox.pdf.

Wasserman, L. 2004. *All of statistics: A concise course in statistical inference*. Springer Verlag.

Williamson, Jon. 2009. *In defence of objective bayesianism*. Oxford.

Yalcin, Seth. 2010. Probability Operators. *Philosophy Compass* 5(11). 916–937.