Theory of Statistics

Contents

1 Overview

Typically, statistical inference uses the following structure:

We observe a realization of random variables on a sample space $\mathcal{X}$,

$$X(s) = (X_1(s), \cdots, X_n(s))$$

where each of the X_i has the same distribution. The random variable may be independent in the case of *sampling with replacement* or more generally *exchangeable* as in the case of *sampling without replacement* from a finite population. The aim of the inference is to say something about which distribution it is.

We usually restrict the allowable distributions to be from some class $\mathcal{P}$. If these distributions can be indexed by a set $\Omega \subset R^d$, then $\mathcal{P}$ is called a *parametric family.* We generally set up this indexing so that the parameterization is *identifiable.* i.e., the mapping from Ω to $\mathcal{P}$ is one to one.

For a parameter choice $\theta \in \Omega$, we denote the distribution of the observations by P_θ and the expectation by E_θ.

Often, the distributions of X are absolutely continuous with respect to some reference measure ν on $\mathcal{X}$ for each value of the parameter. Thus, X has a density $f_{X|\Theta}(x|\theta)$ and

$$P_\theta\{X \in B\} = \int_B f_{X|\Theta}(x|\theta)\ \nu(dx).$$

The most typical choices for $(\mathcal{X}, \nu)$ are

1. $\mathcal{X}$ is a subset of R^k and ν is Lebesgue measure. Thus, $\int_B f_{X|\Theta}(x|\theta)\ \nu(dx) = \int_B f_{X|\Theta}(x|\theta)\ dx$.

2. $\mathcal{X}$ is a subset of Z^k and ν is counting measure. Thus, $\int_B f_{X|\Theta}(x|\theta)\ \nu(dx) = \sum_{x\in B} f_{X|\Theta}(x|\theta)$.

For an observation $X(s) = x$, we can consider the density as a function, L, of θ. $L(\theta) = f_{X|\Theta}(x|\theta)$ is called the *likelihood function.*

1.1 Classical Statistics

Suppose we are interested in deciding if the parameter Θ lies in one portion Ω_H of the parameter space. We can then set a *hypothesis*

$$H : \Theta \in \Omega_H,$$

versus the *alternative hypothesis*,

$$A : \Theta \notin \Omega_H.$$

A simple *test* of this hypothesis would be to choose a *rejection region* $\mathcal{R}$, and a decision function d and reject H if $d(x) \in \mathcal{R}$. The *power function*

$$\beta(\theta) = P_\theta\{d(X) \in \mathcal{R}\}$$

gives, for each value of the parameter θ, the probabilty that the hyothesis is rejected.

Example. Suppose that, under the probability P_θ, X consists of n independent $N(\theta, 1)$ random variables. The usual two-sided α test of

$$H : \Theta = \theta_0 \qquad \text{versus} \qquad A : \Theta \neq \theta_0$$

is to reject H if $\bar{X} \in \mathcal{R}$,

$$\mathcal{R} = (\theta_0 - \frac{1}{\sqrt{n}}\Phi^{-1}(\frac{\alpha}{2}), \theta_0 + \frac{1}{\sqrt{n}}\Phi^{-1}(1 - \frac{\alpha}{2}))^c.$$

Φ is the cumulative distribution function of a standard normal random variable.

An *estimator* ϕ of $g(\theta)$ is *unbiased* if

$$E_\theta[\phi(X)] = g(\theta) \qquad \text{for all} \quad \theta \in \Omega.$$

An *estimator* ϕ of θ is a *maximum likelihood estimator (MLE)* if

$$\sup_{\theta\in\Omega} L(\theta) = L(\phi(x)) \qquad \text{for all} \quad x \in \mathcal{X}.$$

An estimator ψ of $g(\theta)$ is a *maximum likelihood estimator* if $\psi = g \circ \phi$, where ϕ is defined above.

Exercise. $\bar{x}$ is both an unbiased and a maximum likelihood estimate of the parameter for independent $N(\theta, 1)$ random variables.

1.2 Bayesian Statistics

In Bayesian statistics, (X, Θ) is a random variable with state space $\mathcal{X} \times \Omega$. The distribution, μ of Θ on Ω is called the *prior distribution.* Thus, the prior distribution and $\{P_\theta : \theta \in \Omega\}$ determine the joint distribution of (X, Θ).

$$Pr\{(X, \Theta) \in B\} = \int\int I_B(x, \theta)\mu_{X|\Theta}(dx|\theta)\mu_\Theta(d\theta).$$

Here, $\mu_{X|\Theta}(\cdot|\theta)$ is the distribution of X under P_θ.

Consider the case in which μ_Θ has density f_Θ and $\mu_{X|\Theta}(\cdot|\theta)$ has density $f_{X|\Theta}$ with respect to Lebesgue measure, then

$$Pr\{(X, \Theta) \in B\} = \int\int I_B(x, \theta)f_{X|\Theta}(x|\theta)f_\Theta(\theta)\ dx\ d\theta.$$

After observing $X = x$, one constructs the conditional density of Θ given $X = x$ using *Bayes' theorem.*

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x|t)f_\Theta(t)\ dt}.$$

This is called the *posterior distribution.*

Example. Suppose that given $\Theta = \theta$, X consists of n conditionally independent $N(\theta, 1)$ random variables. In addition, suppose that Θ is $N(\theta_0, 1/\lambda)$. The likelihood function is

$$\begin{aligned} f_{X|\Theta}(x|\theta) &= (2\pi)^{-n/2}\exp\big(-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2\big) \\ &= (2\pi)^{-n/2}\exp\big(-\frac{n}{2}(\theta - \bar{x})^2 - \frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2\big). \end{aligned}$$

and the prior density is

$$f_\Theta(\theta) = \sqrt{\frac{\lambda}{2\pi}} \exp(-\frac{\lambda}{2}(\theta - \theta_0)^2).$$

The numerator for posterior density has the form

$$k(x) \exp(-\frac{1}{2}(n(\theta - \bar{x})^2 + \lambda(\theta - \theta_0)^2)) = \tilde{k}(x) \exp(-\frac{n+\lambda}{2}(\theta - \theta_1(x))^2).$$

where $\theta_1(x) = (\lambda\theta_0 + n\bar{x})/(\lambda + n)$. Thus, the posterior distribution is $N(\theta_1(x), 1/(\lambda + n))$. If n is small, then $\theta_1(x)$ is near θ_0 and if n is large, $\theta_1(x)$ is near $\bar{x}$.

Inference is based on the posterior distribution. In the example above,

$$\int \theta f_{(\Theta|X)}(\theta|x)\ d\theta = \theta_1(x)$$

is an estimate for Θ.

2 Probability Theory

2.1 σ-fields and Measures

We begin with a definition.

Definition. A nonempty collection $\mathcal{A}$ of subsets of a set S is called a *field* if

1. $S \in \mathcal{A}$.
2. $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$.
3. $A_1, A_2 \in \mathcal{A}$ implies $A_1 \cup A_2 \in \mathcal{A}$.
   If, in addition,
4. $\{A_n : n = 1, 2, \cdots\} \subset \mathcal{A}$ implies $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$,
   then $\mathcal{A}$ is called a σ-field.

The pair $(S, \mathcal{A})$ is called a *measurable space.*

Exercise. An arbitrary intersection of σ-fields is a σ-field. The power set of S is a σ-field.

Let $\mathcal{C}$ be any collection of subsets. Then, $\sigma(\mathcal{C})$ will denote the smallest sigma field containing $\mathcal{C}$. By the exercise above, this is the (non-empty) intersection of all σ-fields containing $\mathcal{C}$.

Examples.

1. For a single set A, $\sigma(A) = \{\emptyset, A, A^c, S\}$
2. If $\mathcal{C}$ is a σ-field, then $\sigma(\mathcal{C}) = \mathcal{C}$
3. If $S \subset R^d$, or, more generally, S is a topological space, and $\mathcal{C}$ is the set of the open sets in S, then $\sigma(\mathcal{C})$ is called the *Borel σ-field* and denoted $\mathcal{B}(S)$

Because we often look at sequences of random variables, we will often consider the product space,

$$S = \prod_{\lambda \in \Lambda} S_\lambda$$

If, for each λ, $\mathcal{A}_\lambda$ is a σ-field on S_λ, then the *product σ-field* is the smallest σ-field that contains all sets of the form $\prod_{\lambda \in \Lambda} A_\lambda$, where $A_\lambda \in \mathcal{A}_\lambda$ for all λ and $A_\lambda = S_\lambda$ for all but finitely many λ .

Proposition. The Borel σ-field $\mathcal{B}(R^d)$ of R^d is the same as the product σ-field of k copies of $\mathcal{B}(R^1)$.

Definition. Let $(S, \mathcal{A})$ be a measurable space. A function $\mu : \mathcal{A} \to [0, \infty]$ is called a *measure* if

1. $\mu(\emptyset) = 0$.
2. (Additivity) If $A \cap B = \emptyset$ then $\mu(A \cup B) = \mu(A) + \mu(B)$.
3. (Continuity) If $A_1 \subset A_2 \subset \cdots$, and $A = \cup_{n=1}^{\infty} A_n$, then $\mu(A) = \lim_{n\to\infty} \mu(A_n)$.
   If in addition,

4. (Normalization) $\mu(S) = 1$, μ is called a *probability.*

The triple $(S, \mathcal{A}, \mu)$ is called a *measure space* or a *probability space* in the case that μ is a probability. In this situation, an element in S is called an *outcome* or *realization* and a member of $\mathcal{A}$ is called an *event.*

A measure μ is called *σ-finite* if can we can find $\{A_n; n \geq 1\} \in \mathcal{A}$, so that $S = \cup_{n=1}^{\infty} A_n$ and $\mu(A_n) < \infty$ for each n.

Examples.

1. (Counting measure, ν) For $A \in \mathcal{A}$, $\nu(A)$ is the number of elements in A. Thus, $\nu(A) = \infty$ if A has infinitely many elements.

2. (Lebesgue measure m on $(R^1, \mathcal{B}(R^1))$ For the open interval (a, b), set $m(a, b) = b-a$. Lebesgue measure generalizes the notion of length. There is a maximum σ-field which is smaller than the power set in which this measure can be defined.

3. (Product measure) Let $\{(S_i, \mathcal{A}_i, \nu_i); 1 \leq i \leq k\}$ be k σ-finite measure spaces. Then the product measure $\nu_1 \times \cdots \times \nu_k$ is the unique measure on $\sigma(\mathcal{A}_1 \times \cdots \times \mathcal{A}_n)$ such that

$$\nu_1 \times \cdots \times \nu_k(A_1 \times \cdots \times A_k) = \nu_1(A_1) \cdots \nu(A_k) \quad \text{for all} \quad A_i \in \mathcal{A}_i, i = 1, \ldots k.$$

Lebesgue measure on R^k is the product measure of k copies of Lebesgue measure on R^1.

We say A *occurs almost everywhere* (A *a.e.*) if $\mu(A^c) = 0$. If μ is a probability, we say A *occurs almost surely* (A *a.s.*). If $f = g$ a.e., then we say that g is a *version* of f.

2.2 Measurable Functions and Integration

Let $f : (S, \mathcal{A}) \to (T, \mathcal{B})$ be a function between two measurable spaces. We say that f is *measureable* if

$$f^{-1}(B) \in \mathcal{A} \qquad \text{for every} \qquad B \in \mathcal{B}.$$

If the measure on $(S, \mathcal{A})$ is a probability, f is called a *random variable.* We typically use capital letters near the end of the alphabet for random variables.

Exercise. The collection

$$\{f^{-1}(B) : B \in \mathcal{B}\}$$

is a σ-field, denoted $\sigma(f)$. Thus, f is measurable if and only if $\sigma(f) \subset \mathcal{A}$.

If μ is a measure on $(S, \mathcal{A})$, then f *induces* a measure ν on $(T, \mathcal{B})$ by $\nu(B) = \mu(f^{-1}(B))$ for $B \in \mathcal{B}$.

Examples.

1. Let A be a measurable set. The *indicator function* for A, $I_A(s)$ equals 1 if $s \in A$, and 0 is $s \notin A$.

2. A simple function e take on a finite number of distinct values, $e(x) = \sum_{i=1}^n a_i I_{A_i}(x)$, $A_1, \cdots, A_n \in \mathcal{A}$, and $a_1, \cdots, a_n \in S$. Call this class of functions $\mathcal{E}$.

3. $\mathcal{A}$ and $\mathcal{B}$ are Borel σ-fields and f is a continuous function.

For a simple function e define the *integral of* e *with respect to the measure* μ as

$$\int e \; d\mu = \sum_{i=1}^{n} a_i \mu(A_i).$$

You can check that the value of $\int e \; d\mu$ does not depend on the choice for the representation of e. By convention $0 \times \infty = 0$.

For f a non-negative measurable function, define the *integral of* f *with respect to the measure* μ as

$$\int_S f(s) \; \mu(ds) = \int f \; d\mu = \sup\{\int e \; d\mu : e \in \mathcal{E}, e \leq f\}.$$

Again, you can check that the integral of a simple function is the same under either definition.

For general functions, denote the positive part of f, $f^+(s) = \max\{f(s), 0\}$ and the negative part of f by $f^-(s) = -\min\{f(s), 0\}$. Thus, $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

If f is a real valued measurable function, then define the *integral of* f *with respect to the measure* μ as

$$\int f(s) \; \mu(ds) = \int f^+(s) \; \mu(ds) - \int f^-(s) \; \mu(ds).$$

provided at least one of the integrals on the right is finite. If $\int |f| \; du < \infty$, then we say that f is *integrable.*

We typically write $\int_A f(s) \; \mu(ds) = \int I_A(s) f(s) \; \mu(ds)$.

If the underlying measure is a probability, then we typically call the integral, the *expectation* or the *expected value* and write,

$$EX = \int X \; d\mu.$$

Exercise. If $f = g$ a.e., then $\int f \; d\mu = \int g \; d\mu$.

Examples.

1. If μ is counting measure on S, then $\int f \; d\mu = \sum_{s \in S} f(s)$.
2. If μ is Lebesgue measure and f is Riemann integrable, then $\int f d\mu = \int f \; dx$, the Riemann integral.

The integral is a positive linear functional, i.e.

1. $\int f \; d\mu \geq 0$ whenever f is non-negative and measurable.
2. $\int (af + bg) \; d\mu = a \int f \; d\mu + \int g \; d\mu$ for real numbers a, b and integrable functions f, g.

Exercise. Any non-negative real valued measurable function is the increasing limit of measurable functions, e.g.

$$f_n(s) = \sum_{i=1}^{n2^n} \frac{i-1}{2^n} I_{\{\frac{i-1}{2^n} < f \le \frac{i}{2^n}\}}(s) + nI_{\{f>n\}}$$

Exercise. If $\{f_n : n \ge 1\}$ is a sequence of measurable functions, then $f(s) = \liminf_{n\to\infty} f_n(s)$ is measurable.

Theorem. (Fatou's Lemma) Let $\{f_n : n \ge 1\}$ be a sequence of non-negative measurable functions. Then

$$\int \liminf_{n\to\infty} f_n(s)\ \mu(ds) \le \liminf_{n\to\infty} \int f_n(s)\ \mu(ds).$$

Theorem. (Monotone Convergence) Let $\{f_n : n \ge 1\}$ be an increasing sequence of non-negative measurable functions. Then

$$\int \lim_{n\to\infty} f_n(s)\ \mu(ds) = \lim_{n\to\infty} \int f_n(s)\ \mu(ds).$$

Example. Let X be a non-negative random variable with *cumulative distribution function* $F_X(x) = Pr\{X \le x\}$. Set $X_n(s) = \sum_{i=1}^{n2^n} \frac{i-1}{2^n} I_{\{\frac{i-1}{2^n} < X \le \frac{i}{2^n}\}}(s)$. Then by the monotone convergence theorem and the definition of the Riemann-Stieltjes integral

$$\begin{aligned} EX &= \lim_{n\to\infty} EX_n \\ &= \lim_{n\to\infty} \sum_{i=1}^{n2^n} \frac{i-1}{2^n} Pr\{\frac{i-1}{2^n} < X \le \frac{i}{2^n}\} \\ &= \lim_{n\to\infty} \sum_{i=1}^{n2^n} \frac{i-1}{2^n} (F_X(\frac{i}{2^n}) - F_X(\frac{i-1}{2^n})) \\ &= \int_0^\infty x\ dF_X(x) \end{aligned}$$

This can be generalized.

Theorem. Let $f : (S_1, \mathcal{A}_1) \to (S_2, \mathcal{A}_2)$. For a measure μ_1 on S_1, define the induced measure $\mu_2(A) = \mu_1(f^{-1}(A))$. Then, if $g : S_2 \to R$,

$$\int g(s_2)\ \mu_2(ds_2) = \int g(f(s_1))\ \mu_1(ds_1).$$

To prove this, use the "standard machine".

1. Show that the identity holds for indicator functions.
2. Show, by the linearity of the integral, that the identity holds for simple functions.
3. Show, by the monotone convergence theorem, that the identity holds for non-negative functions.
4. Show, by decomposing a function into its positive and negative parts, that it holds for integrable functions.

To compute integrals based on product measures, we use Fubini's theorem.

Theorem. Let $(S_i, \mathcal{A}_i, \mu_i), i = 1, 2$ be two σ-finite measures. If $f : S_1 \times S_2 \to R$ is integrable with respect to $\mu_1 \times \mu_2$, then

$$\int f(s_1, s_2)\ \mu_1 \times \mu_2(ds_1 \times ds_2) = \int [\int f(s_1, s_2)\ \mu_1(ds_1)]\mu_2(ds_2) = \int [\int f(s_1, s_2)\ \mu_2(ds_2)]\mu_1(ds_1).$$

Use the "standard machine" to prove this. Begin with indicators of sets of the form $A_1 \times A_2$. (This requires knowing arbitrary measurable sets can be approximated by a finite union of rectangles.) The identity for non-negative functions is known as Tonelli's theorem.

Exercises.

1. Let $\{f_k : k \geq 1\}$ be a sequence of non-negative measurable functions. Then
$$\int \sum_{k=1}^{\infty} f_k(s)\ \mu(dx) = \sum_{k=1}^{\infty} \int f_k(s)\ \mu(dx).$$
2. Let f be a non-negative measurable function, then
$$\nu(A) = \int_A f(x)\ \mu(dx)$$
is a measure.

Note that
$$\mu(A) = 0 \qquad \text{implies} \qquad \nu(A) = 0.$$

Whenever this holds for any two measures defined on the same measure space, we say that ν is *absolutely continuous* with respect to μ. This is denoted by

$$\nu << \mu.$$

This exercise above has as its converse the following theorem.

Theorem. (Radon-Nikodym) Let $\mu_i, i = 1, 2$, be two measures on $(S, \mathcal{A})$ such that $\mu_2 << \mu_1$, and μ_1 is σ-finite. Then there exists an extended real valued function $f : S \to [0, \infty]$ such that for any $A \in \mathcal{A}$,

$$\mu_2(A) = \int_A f(s)\ \mu_1(ds).$$

The function f, called the *Radon-Nikodym derivative* of μ_2 with respect to μ_1, is unique a.e. This derivative is sometimes denoted

$$\frac{d\mu_2}{d\mu_1}(s).$$

We also have the following calculus identities.

1. (Substitution) $\int g\ d\mu_2 = \int g \frac{d\mu_2}{d\mu_1}\ d\mu_1$.
2. If $\lambda_i,\ i = 1, 2$, are measures, $\lambda_i << \mu_1$, then $\lambda_1 + \lambda_2 << \mu_1$ and
$$\frac{d(\lambda_1 + \lambda_2)}{d\mu_1} = \frac{d\lambda_1}{d\mu_1} + \frac{d\lambda_2}{d\mu_1}. \qquad \text{a.e. } \mu_1$$
3. (Chain Rule) If μ_3 is a measure, μ_2 is σ-finite, and $\mu_3 << \mu_2$, then
$$\frac{d\mu_3}{d\mu_1} = \frac{d\mu_3}{d\mu_2}\frac{d\mu_2}{d\mu_1} \qquad \text{a.e. } \mu_1$$
In particular, if $\mu_1 << \mu_2$, then
$$\frac{d\mu_1}{d\mu_2} = (\frac{d\mu_2}{d\mu_1})^{-1} \qquad \text{a.e. } \mu_1 \text{ or } \mu_2.$$
4. Let $\nu_i,\ i = 1, 2$, be two measures on $(T, \mathcal{B})$ such that $\nu_2 << \nu_1$, and ν_1 is σ-finite. Then
$$\mu_2 \times \nu_2 << \mu_1 \times \nu_1,$$
and
$$\frac{d(\mu_2 \times \nu_2)}{d(\mu_1 \times \nu_1)} = \frac{d\mu_2}{d\mu_1}\frac{d\nu_2}{d\nu_1}.$$

In the study of sufficient statistics, we will need the following theorem.

Theorem. (Halmos and Savage) Let μ be a σ-finite measure on $(S, \mathcal{A})$. Let $\mathcal{N}$ be a collection of nontrivial measures on $(S, \mathcal{A})$ such that $\nu << \mu$ for all $\nu \in \mathcal{N}$. Then, there exists a sequence of non-negative numbers $\{c_i : i \geq 1\}$, $\sum_{i=1}^{\infty} c_i = 1$ and a sequence of elements $\{\nu_i : i \geq 1\} \subset \mathcal{N}$ such that

$$\nu << \sum_{i=1}^{\infty} c_i\nu_i \qquad \text{for every} \qquad \nu \in \mathcal{N}.$$

Proof. For μ finite, set $\lambda = \mu$.

For μ infinite, pick a countable partition, $\{S_i : i \geq 1\}$ of S such that $0 < \mu(S_i) < \infty$. Set

$$\lambda(B) = \sum_{i=1}^{\infty} \frac{\mu(B \cap S_i)}{2^i \mu(S_i)} \quad \text{for} \quad B \in \mathcal{A}.$$

Thus, λ is a finite measure and $\nu << \lambda$ for all $\nu \in \mathcal{N}$. Define

$$\mathcal{Q} = \{\sum_{i=1}^{\infty} a_i \nu_i : \sum_{i=1}^{\infty} a_i = 1, a_i \geq 0, \nu_i \in \mathcal{N}\}.$$

If $\beta \in \mathcal{Q}$, then $\beta << \lambda$. Now set

$$\mathcal{D} = \{C \in \mathcal{A} : \lambda\{x \in C : dQ/d\lambda(x) = 0\} = 0, \text{and } Q(C) > 0, \text{for some } Q \in \mathcal{Q}\}.$$

Claim 1. $\mathcal{D} \neq \emptyset$.

Set $C = \{x : d\nu/d\lambda(x) > 0\}$ and $Q = \nu$.

Then, $\{x \in C : d\nu/d\lambda(x) = 0\} = \emptyset$, and $Q(C) = \nu(C) = \nu(S) > 0$. Thus, $C \in \mathcal{D}$. Note that

$$\sup_{C \in \mathcal{D}} \lambda(C) = c \leq \lambda(S) < \infty.$$

Thus, choose a sequence $\{C_i : i \geq 1\} \subset \mathcal{D}$ so that $c = \lim_{i \to \infty} \lambda(C_i)$.

Because $C_i \in \mathcal{D}$, we can find Q_i so that

$$\lambda\{x \in C_i : dQ_i/d\lambda(x) = 0\} = 0, \quad \text{and } Q_i(C_i) > 0.$$

Define $C_0 = \cup_{i=1}^{\infty} C_i$ and $Q_0 = \sum_{i=1}^{\infty} 2^{-i} Q_i \in \mathcal{Q}$. Then, $Q_0(C_0) > 0$, $dQ_0/d\lambda = \sum_{i=1}^{\infty} 2^{-i} dQ_i/d\lambda$ and

$$\{x \in C_0 : \frac{dQ_0}{d\lambda}(x) = 0\} \subset \cup_{i=1}^{\infty}\{x \in C_i : \frac{dQ_i}{d\lambda}(x) = 0\}.$$

Thus, $C_0 \in \mathcal{D}$ and $\lambda(C_0) = c$.

Claim 2. $\nu << Q_0$ for all $\nu \in \mathcal{N}$.

We show that for $\nu \in \mathcal{N}$, $Q_0(A) = 0$ implies $\nu(A) = 0$.

Define $C = \{x : d\nu/d\lambda(x) > 0\}$ and write

$$\nu(A) = \nu(A \cap C_0) + \nu(A \cap C_0^c \cap C^c) + \nu(A \cap C_0^c \cap C).$$

Now, $Q_0(A \cap C_0) = 0$ and $dQ_0/d\lambda > 0$ on C_0 imples that $\lambda(A \cap C_0) = 0$ and because $\nu << \lambda$, $\nu(A \cap C_0) = 0$.

Because $d\nu/d\lambda = 0$ on C^c, $\nu(A \cap C_0^c \cap C^c) = 0$.

Finally, for $D = A \cap C_0^c \cap C$, suppose $\nu(D) > 0$. Because $\nu << \lambda$, $\lambda(D) > 0$.

$D \subset C$ implies $d\nu/d\lambda(x) > 0$ on D, and

$$\lambda\{x \in D : d\nu/d\lambda(x) = 0\} = \lambda(\emptyset) = 0.$$

Thus, $D \in \mathcal{D}$ and hence $C \cup D \in \mathcal{D}$ However, $C_0 \cap D = \emptyset$, thus $\lambda(C_0 \cup D) > \lambda(C_0)$, contradicting the definition of c. Thus, $\nu(D) = 0$.

2.3 Conditional Expectation.

We will begin with a general definition of conditional expectation and show that this agrees with more elementary definitions.

Definition. Let Y be an integrable random variable on $(S, \mathcal{A}, \mu)$ and let $\mathcal{C}$ be a sub-σ-field of $\mathcal{A}$. The *conditional expectation* of Y given $\mathcal{C}$, denoted $E[Y|\mathcal{C}]$ is the a.s. unique random variable satisfying the following two conditions.

1. $E[Y|\mathcal{C}]$ is $\mathcal{C}$-measurable.
2. $E[E[Y|\mathcal{C}]I_A] = E[YI_A]$ for any $A \in \mathcal{C}$.

Thus, $E[Y|\mathcal{C}]$ is essentially the only random variable that uses the information provided by $\mathcal{C}$ and gives the same averages as Y on events that are in $\mathcal{C}$. The uniqueness is provided by the Radon-Nikodym theorem. For Y positive, define the measure

$$\nu(C) = E[YI_C] \qquad \text{for} \qquad C \in \mathcal{C}.$$

Then ν is absolutely continuous with respect to the underlying probability restricted to $\mathcal{C}$. Set $E[Y|\mathcal{C}]$ equal to the Radon-Nikodym derivative $d\nu/dP|_{\mathcal{C}}$.

For $B \in \mathcal{A}$, the *conditional probability* of B given $\mathcal{C}$ is defined to be $P(B|\mathcal{C}) = E[I_B|\mathcal{C}]$.

Exercises. Let $C \in \mathcal{A}$, then $P(B|\sigma(C)) = P(B|C)I_C + P(B|C^c)I_{C^c}$.

If $\mathcal{C} = \sigma\{C_1, \cdots, C_n\}$, the σ-field generated by a finite partition, then

$$P(A|\mathcal{C}) = \sum_{i=1}^{n} P(A|C_i)I_{C_i}.$$

Theorem. (Bayes Formula) Let $C \in \mathcal{C}$, then

$$P(C|A) = \frac{E[P(A|\mathcal{C})I_C]}{E[P(A|\mathcal{C})]}.$$

Proof. $E[P(A|\mathcal{C})I_C] = E[I_A I_C] = P(A \cap C)$ and $E[P(A|\mathcal{C})] = P(A)$.

Exercise. Show the Bayes formula for a finite partition $\{C_1, \cdots, C_n\}$ is

$$P(C_j|A) = \frac{P(A|C_j)P(C_j)}{\sum_{i=1}^n P(A|C_i)P(C_i)}.$$

If $\mathcal{C} = \sigma(X)$, then we usually write $E[Y|\mathcal{C}] = E[Y|X]$. For these circumstances, we have the following theorem which can be proved using the standard machine.

Theorem. Let X be a random variable and let Z is a measurable function on $(S, \sigma(X))$ if and only if there exists a measurable function h on the range of X so that $Z = h(X)$.

Thus, $E[g(Y)|X] = h(X)$. If X is discrete, taking on the values $x_1, x_2, \cdots$, then by property 2,

$$\begin{aligned} E[g(Y)I_{\{X=x_i\}}] &= E[E[g(Y)|X]I_{\{X=x_i\}}] \\ &= E[h(X)I_{\{X=x_i\}}] \\ &= h(x_i)P\{X = x_i\}. \end{aligned}$$

Thus,

$$h(x_i) = \frac{E[YI_{\{X=x_i\}}]}{P\{X = x_i\}}.$$

If, in addition, Y is discrete and the pair (X, Y) has joint mass function $f_{(X,Y)}(x, y)$. Then,

$$E[g(Y)I_{\{X=x_i\}}] = \sum_j g(y_j)f_{(X,Y)}(x_i, y_j).$$

Typically, we write $h(x) = E[g(Y)|X = x]$, and for $f_X(x) = P\{X = x\}$, we have

$$E[g(Y)|X = x] = \sum_j \frac{g(y_j)f_{(X,Y)}(x, y_j)}{f_X(x)} = \sum_j g(y_j)f_{Y|X}(y_j|x).$$

We now look to extend the definition of $E[g(X, Y)|X = x]$. Let ν and λ be σ-finite measures and consider the case in which (X, Y) has a density $f_{(X,Y)}$ with respect to $\nu \times \lambda$. Then the marginal density $f_X(x) = \int f_{(X,Y)}(x, y)\ \lambda(dy)$ and the conditional density

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}.$$

if $f_X(x) > 0$ and 0 if $f_X(x) = 0$. Set

$$h(x) = \int g(x, y)f_{Y|X}(y|x)\ \lambda(dy).$$

Claim. If $E|g(X, Y)| < \infty$, then $E[g(X, Y)|X] = h(X)$

By the theorem above, $h(X)$ is $\sigma(X)$ measurable.

A typical element of $\sigma(X)$ is $\{X \in B\}$ for some Borel set B. We must show that

$$E[h(X)I_B(X)] = E[g(X, Y)I_B(X)].$$

Thus,

$$\begin{aligned} E[h(X)I_B(X)] &= \int I_B(x)h(x)f_X(x)\ \nu(dx) \\ &= \int I_B(x)(\int g(x, y)f_{Y|X}(y|x)\ \lambda(dy))f_X(x)\ \nu(dx) \\ &= \int\int g(x, y)I_B(x)f_{(X,Y)}(x, y)\ \lambda(dy)\nu(dx) \\ &= E[g(X, Y)I_B(X)]. \end{aligned}$$

Theorem. (Bayes) Suppose that X has a parametric family of distributions $\mathcal{P}_0$ of distributions with parameter space Ω. Suppose that $P_\theta << \nu$ for all $\theta \in \Omega$, and let $f_{X|\Theta}(x|\theta)$ be the conditional density of X with respect to ν given that $\Theta = \theta$. Let μ_Θ be the prior distribution of Θ and let $\mu_{\Theta|X}(\cdot|x)$ denote the conditional distribution of Θ given $X = x$.

1. Then $\mu_{\Theta|X}(\cdot|x) << \mu_\Theta$, a.s. with respect to the marginal of X with Radon-Nikodym derivative
$$\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)}{\int_\Omega f_{X|\Theta}(x|t)\mu_\Theta(dt)}.$$
for those x for which the denominator is neither 0 nor infinite.

2. If $\mu_\Theta << \lambda$ for some σ-finite measure λ, and $d\mu_\Theta/d\lambda = f_\Theta$, then
$$\frac{d\mu_{\Theta|X}}{d\lambda}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x|t)f_\Theta(t)\lambda(dt)}.$$

Proof. Statement 2 follows from statement 1 by the chain rule.
To show 1, write $m(x) = \int_\Omega f_{X|\Theta}(x|t)\ \mu_\Theta(dt)$. Then
$$\int_{\mathcal{X}} m(x)\ \nu(dx) = \int_{\mathcal{X}}\int_\Omega f_{X|\Theta}(x|t)\ \mu_\Theta(dt)\ \nu(dx) = \int_\Omega\int_{\mathcal{X}} f_{X|\Theta}(x|t)\ \ \nu(dx)\mu_\Theta(dt) = 1.$$
Thus, $m < \infty$ a.e. ν.

Choose $x \in \mathcal{X}$ with $m(x) < \infty$ and define
$$P(B|x) = \frac{1}{m(x)}\int_B f_{X|\Theta}(x|t)\ \mu_\Theta(dt)$$
for B in the σ-field on Ω. Note that $P(\cdot|x)$ is a probability on Ω a.e. ν. Thus, it remains to show that
$$P(B|x) = \mu_{\Theta|X}(B|x).$$

By Fubini's theorem, $P(B|\cdot)$ is a measurable function of x. If μ is the joint distribution of (X, Θ), then for any measurable subset A of $\mathcal{X}$.

$$\begin{aligned}
E[I_B(\Theta)I_A(X)] &= \int_{A\times\Omega} I_B(\theta)\ \mu(dx\times d\theta) = \int_A\int_B f_{X|\Theta}(x|\theta)\ \mu_\Theta(d\theta)\ \nu(dx) \\
&= \int_A [\int_B \frac{f_{X|\Theta}(x|\theta)}{m(x)}\ \mu_\Theta(d\theta)][\int_\Omega f_{X|\Theta}(x|t)\ \mu_\Theta(dt)]\ \nu(dx) \\
&= \int_\Omega\int_A [\int_B \frac{f_{X|\Theta}(x|\theta)}{m(x)}\ \mu_\Theta(d\theta)] f_{X|\Theta}(x|t)\ \nu(dx)\mu_\Theta(dt) \\
&= \int_{A\times\Omega} P(B|x)\ \mu(dx\times d\theta) = E[P(B|X)I_A(X)]
\end{aligned}$$

We now summarize the properties of conditional expectation.

Theorem. Let $Y, Y_1, Y_2, \cdots$ have finite absolute mean on $(S, \mathcal{A}, \mu)$ and let $a_1, a_2 \in R$. In addition, let $\mathcal{B}$ and $\mathcal{C}$ be σ-algebras contained in $\mathcal{A}$. Then

(a) If Z is any version of $E[Y|\mathcal{C}]$, then $EZ = EY$. ($E[E[Y|\mathcal{C}]] = EY$).

(b) If Y is $\mathcal{C}$ measurable, then $E[Y|\mathcal{C}] = Y$, a.s.

(c) **(Linearity)** $E[a_1Y_1 + a_2Y_2|\mathcal{C}] = a_1E[Y_1|\mathcal{C}] + a_2E[Y_2|\mathcal{C}]$, a.s.

(d) **(Positivity)** If $Y \geq 0$, then $E[Y|\mathcal{C}] \geq 0$.

(e) **(cMON)** If $Y_n \uparrow Y$, then $E[Y_n|\mathcal{C}] \uparrow E[Y|\mathcal{C}]$, a.s.

(f) **(cFATOU)** If $Y_n \geq 0$, then $E[\liminf_{n\to\infty} Y_n|\mathcal{C}] \leq \liminf_{n\to\infty} E[Y_n|\mathcal{C}]$.

(g) **(cDOM)** If $\lim_{n\to\infty} Y_n(s) = Y(s)$, a.s., if $|Y_n(s)| \leq V(s)$ for all n, and if $EV < \infty$, then

$$\lim_{n\to\infty} E[Y_n|\mathcal{C}] = E[Y|\mathcal{C}],$$

almost surely.

(h) **(cJENSEN)** If $c : \mathbf{R} \to \mathbf{R}$ is convex, and $E|c(Y)| < \infty$, then

$$E[c(Y)|\mathcal{C}] \geq c(E[Y|\mathcal{C}]),$$

almost surely. In particular,

(i) **(Contraction)** $||E[Y|\mathcal{C}]||_p \leq ||Y||_p$ for $p \geq 1$.

(j) **(Tower Property)** If $\mathcal{B} \subset \mathcal{C}$, then

$$E[E[Y|\mathcal{C}]|\mathcal{B}] = E[Y|\mathcal{B}],$$

almost surely.

(k) **(cCONSTANTS)** If Z is $\mathcal{C}$-measurable, then

$$E[ZY|\mathcal{C}] = ZE[Y|\mathcal{C}]$$

holds almost surely whenever any one of the following conditions hold:

(i) Z is bounded.

(ii) $E|Y|^p < \infty$ and $E|Z|^q < \infty$. $\frac{1}{p} + \frac{1}{q} = 1, p > 1$.

(iii) $Y, Z \geq 0, EY < \infty$, and $E[YZ] < \infty$.

(l) **(Role of Independence)** If $\mathcal{B}$ is independent of $\sigma(\sigma(Y), \mathcal{C})$, then

$$E[Y|\sigma(\mathcal{C}, \mathcal{B})] = E[Y|\mathcal{C}],$$

almost surely. In particular,

(m) if Y is independent of $\mathcal{B}$, then $E[Y|\mathcal{B}] = EY$.

Exercise. Let $\mathcal{C}$ be a sub-σ-field and let $EY^2 < \infty$, then

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y|\mathcal{C})] + \mathrm{Var}(E[Y|\mathcal{C}]).$$

Definition. On a probability space $(S, \mathcal{A}, Pr)$, let $\mathcal{C}_1, \mathcal{C}_2, \mathcal{B}$ be sub-σ-fields of $\mathcal{A}$. Then the σ-fields $\mathcal{C}_1$ and $\mathcal{C}_2$ are *conditionally independent* given $\mathcal{B}$ if

$$Pr(A_1 \cap A_2|\mathcal{B}) = Pr(A_1|\mathcal{B})Pr(A_2|\mathcal{B}),$$

for $A_i \in \mathcal{C}_i$, $i = 1, 2$.

Proposition. Let $\mathcal{B}, \mathcal{C}$ and $\mathcal{D}$ be σ-fields. Then

(a) If $\mathcal{B}$ and $\mathcal{C}$ are conditionally independent given $\mathcal{D}$, then $\mathcal{B}$ and $\sigma(\mathcal{C}, \mathcal{D})$ are conditionally independent given $\mathcal{D}$.

(b) Let $\mathcal{B}_1$ and $\mathcal{C}_1$ be sub-σ-fields of $\mathcal{B}$ and $\mathcal{C}$, respectively. Suppose that $\mathcal{B}$ and $\mathcal{C}$ are independent. Then $\mathcal{B}$ and $\mathcal{C}$ are conditionally independent given $\sigma(\mathcal{B}_1, \mathcal{C}_1)$

(c) Let $\mathcal{B} \subset \mathcal{C}$. Then $\mathcal{C}$ and $\mathcal{D}$ are conditionally independent given $\mathcal{B}$ if and only if, for every $D \in \mathcal{D}$, $Pr(D|\mathcal{B}) = Pr(D|\mathcal{C})$.

2.4 Modes of Convergence

Let $X, X_1, X_2, \cdots$ be a sequence of random variables taking values in a metric space $\mathcal{X}$ with metric d.

1. We say that X_n *converges to* X *almost surely* $(X_n \to^{a.s.} X)$ if

$$\lim_{n\to\infty} X_n = X \qquad \text{a.s..}$$

2. We say that X_n *converges to* X *in probability* $(X_n \to^{P} X)$ if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} Pr\{d(X_n, X) > \epsilon\} = 0.$$

3. We say that X_n *converges to* X *in distribution* $(X_n \to^{\mathcal{D}} X)$ if, for every bounded continuous $f : \mathcal{X} \to R$.

$$\lim_{n\to\infty} Ef(X_n) = Ef(X).$$

4. We say that X_n *converges to* X *in* L^p, $p > 0$, $(X_n \to^{L^p} X)$ if,

$$\lim_{n\to\infty} E[d(X_n, X)^p] = 0.$$

Note that $X_n \to^{a.s.} X$ or $X_n \to^{L^p} X$ implies $X_n \to^{P} X$ which in turn implies $X_n \to^{\mathcal{D}} X$. If $X_n \to^{\mathcal{D}} c$, then $X_n \to^{\mathcal{P}} c$

Exercise. Let $X_n \to X$ under one of the first three modes of convergence given above and let g be a continuous, then $g(X_n)$ converges to $g(X)$ under that same mode of convergence.

The converse of these statements requires an additional concept.

Definition. A collection of real-valued random variables $\{X_\gamma : \gamma \in \Gamma\}$ is *uniformly integrable* if

1. $\sup_\gamma E|X_\gamma| < \infty$, and

2. for every $\epsilon > 0$, there exists a $\delta > 0$ such that for every γ,

$$P(A) < \delta \qquad \text{implies} \qquad |E[X_\gamma I_A]| < \epsilon.$$

Theorem. The following are equivalent:

1. $\{X_\gamma : \gamma \in \Gamma\}$ is uniformly integrable.
2. $\lim_{n\to\infty} \sup_{\gamma\in\Gamma} E[|X_\gamma| I_{\{|X_\gamma|>n\}}] = 0$.
3. $\lim_{n\to\infty} \sup_{\gamma\in\Gamma} E[|X_\gamma| - \min\{n, |X_\gamma|\}] = 0$.
4. There exists an increasing convex function $c : [0,\infty) \to R$ such that $\lim_{x\to\infty} c(x)/x = \infty$, and

$$\sup_{\gamma\in\Gamma} E[c(|X_\gamma|)] < \infty.$$

Theorem. If $X_n \to^{\mathcal{D}} X$ and $\{X_n; n \geq 1\}$ is uniformly integrable, then $\lim_{n\to\infty} EX_n = EX$.
Conversely, if the X_n are integrable, $X_n \to^{\mathcal{D}} X$ and $\lim_{n\to\infty} E|X_n| = E|X|$, then $\{X_n; n \geq 1\}$ is uniformly integrable.

The following will be useful in establishing the "delta method".

Theorem. (Slutsky) Let $X, X_1, X_2, \cdots, Y_1, Y_2, \cdots$ be random variables and let $c \in R$. Suppose that $X_n \to^{\mathcal{D}} X$ and $Y_n \to^P c$. Then

1. $X_n + Y_n \to^{\mathcal{D}} X + c$.
2. $Y_n X_n \to^{\mathcal{D}} cX$.
3. $X_n/Y_n \to^{\mathcal{D}} X/c$ provided $c \neq 0$.

Convergence in distribution depends only on the laws P_n of X_n. Thus, we can write

$$\lim_{n\to\infty} \int f \, dP_n = \int f \, dP,$$

where P is the law for X. Under these circumstances, we say P_n *converges weakly to* P and write $P_n \to^W P$.

Theorem. (Portmanteau) The following statements are equivalent for a sequence $\{P_n : n \geq 1\}$ of probability measures on a metric space.

1. $P_n \to^W P$.
2. $\lim_{n\to\infty} \int f \, dP_n = \int f \, dP$.
3. $\limsup_{n\to\infty} P_n(F) \leq P(F)$ for all closed sets F.
4. $\liminf_{n\to\infty} P_n(G) \geq P(G)$ for all open sets G.

5. $\lim_{n\to\infty} P_n(A) = P(A)$ for all P-continuity sets, i.e. sets A so that $P(\partial A) = 0$.

Suppose P_n and P are absolutely continuous with respect to some measure ν and write $f_n = dP_n/d\nu$, $f = dP/d\nu$. If $f_n \to f$ a.e. ν, then by Fatou's lemma,

$$\liminf_{n\to\infty} P_n(G) = \liminf_{n\to\infty} \int_G f_n \, d\nu \geq \int_G f \, d\nu = P(G).$$

Thus, convergence of the densities implies convergence of the distributions.

On a separable metric space, weak convergence is a metric convergence based on the Prohorov metric, defined by

$$\rho(P,Q) = \inf\{\epsilon > 0 : P(F) \leq Q(F^\epsilon) + \epsilon \text{ for all closed sets } F\}.$$

Here $F^\epsilon = \{x : d(x,y) < \epsilon \text{ for some } y \in F\}$. The statements on weak convergence are equivalent to

$$\lim_{n\to\infty} \rho(P_n, P) = 0.$$

For probability distributions on R^d, we can associate to P_n and P its cumulative distribution functions F_n and F, and characteristic functions $\phi_n(s) = \int e^{i\langle s,x\rangle} \, dP_n$ and $\phi(s) = \int e^{i\langle s,x\rangle} \, dP$. Then the statements above are equivalent to

1. $\lim_{n\to\infty} F_n(x) = F(x)$ for all continuity points x of F.

2. (Lévy-Cramér continuity theorem) $\lim_{n\to\infty} \phi_n(s) = \phi(s)$ for all s.

3. (Cramér-Wold device) $\langle c, X_n\rangle \to^{\mathcal{D}} \langle c, X\rangle$ for all $c \in R^d$

In R^d, the Prohorov metric is equivalent to the Lévy metric,

$$\rho_L(F,G) = \inf\{\epsilon > 0 : F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon\},$$

for cumulative distribution functions F and G.

2.5 Limit Theorems

Theorem. Let $X_1, X_2, \cdots$ and Y be R^d random variables satisfying

$$a_n(X_n - c) \to^{\mathcal{D}} Y,$$

where $\{a_n : n \geq 1\} \subset R, c \subset R^d$ with $\lim_{n\to\infty} a_n = \infty$. Let $g : R^d \to R$. Then,

(i) If g is differentiable at c, then

$$a_n(g(X_n) - g(c)) \to^{\mathcal{D}} \nabla g(c) Y^T.$$

(ii) If $g \in C^m$ in a neighborhood of c, with all k-th order derivatives vanishing at c, $1 \leq k < m$. Then,

$$a_n^m(g(X_n) - g(c)) \to^{\mathcal{D}} \sum_{i_1,\cdots,i_m} \frac{\partial^m g}{\partial x_{i_1} \cdots \partial x_{i_m}}(c) Y_{i_1} \cdots Y_{i_m}.$$

Proof. (i). By Slutsky's theorem, $X_n - c \to^{\mathcal{D}} 0$ and hence $X_n - c \to^P 0$.
Let

$$Z_n = a_n(g(X_n) - g(c)) - a_n \nabla g(c)(X_n - c)^T.$$

If we show that $Z_n \to^P 0$, then the hypothesis, and Slutsky's theorem imply the theorem.

Because g is differentiable at c, given $\eta > 0$, we can find $\delta_\eta > 0$ so that

$$|g(x) - g(c) - \nabla g(c)(x - c)^T| \leq \eta |x - c|$$

whenever $|x - c| \leq \delta_\eta$. Fix ϵ, and note that

$$\begin{aligned} Pr\{|Z_n| \geq \epsilon\} &\leq Pr\{|Z_n| \geq \epsilon, |X_n - c| \geq \delta_\eta\} + Pr\{|Z_n| \geq \epsilon, |X_n - c| \leq \delta_\eta\} \\ &\leq Pr\{|X_n - c| \geq \delta_\eta\} + Pr\{a_n |X_n - c| \geq \epsilon/\eta\}. \end{aligned}$$

The first term goes to zero as $n \to \infty$. By the Portmanteau Theorem,

$$\limsup_{n\to\infty} Pr\{|Z_n| \geq \epsilon\} \leq \limsup_{n\to\infty} Pr\{a_n|X_n - c| \geq \epsilon/\eta\} \leq Pr\{|Y| \geq \epsilon/\eta\}.$$

Because η is arbitrary, the theorem is complete.

Part (ii) is similar and is left as an exercise.

Corollary. (Delta method) Assume the conditions of the theorem above. If Y has a $N(0, \Sigma)$ distribution, then

$$a_n(g(X_n) - g(c)) \to^{\mathcal{D}} W.$$

W is $N(0, \nabla g(c) \Sigma \nabla g(c)^T)$.

Examples. If $\sqrt{n}(X_n - c) \to^{\mathcal{D}} Z$, Z is a standard normal, then for $c \neq 0$

$$\sqrt{n}(X_n^2 - c^2) \to^{\mathcal{D}} W.$$

W is $N(0, 4c^2)$. For $c = 0$,

$$\sqrt{n} X_n^2 \to^{\mathcal{D}} \chi_1^2,$$

a chi-square random variable with one degree of freedom.

The most general *central limit theorem* for independent random variables is due to Lindeberg.

Theorem. Let $\{X_{nj} : j = 1, \cdots, k_n\}$ be independent mean zero random variables with $k_n \to \infty$ as $n \to \infty$ and $0 < \sigma_n = \mathrm{Var}(\sum_{j=1}^{k_n} X_{nj}) < \infty$, and

If, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \sum_{j=1}^{k_n} E[X_{nj}^2 I_{\{|X_{nj}|>\epsilon\sigma_n\}}] = 0,$$

then

$$\frac{1}{\sigma_n} \sum_{j=1}^{k_n} X_{nj}$$

converges in distribution to a standard normal random variable.

Exercises.

1. Show that the Liapounov condition below implies the Lindeberg condition.

$$\lim_{n\to\infty} \frac{1}{\sigma_n^{2+\delta}} \sum_{j=1}^{k_n} E[|X_{nj}|^{2+\delta}] = 0,$$

for some $\delta > 0$.

2. Let $\{X_i : i \geq 1\}$ be Bernoulli random variables. $EX_i = p_i$. If $\sigma_n^2 = \sum_{i=1}^n p_i(1-p_i) \to \infty$ as $n \to \infty$, then

$$\frac{1}{\sigma_n} \sum_{i=1}^n (X_i - p_i) \to^{\mathcal{D}} Z$$

as $n \to \infty$ with Z a standard normal random variable.

3. Let $\{X_i : i \geq 1\}$ be i.i.d R^d-valued random variables with mean μ and variance $\Sigma = \mathrm{Var}(X_1)$, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \to^{\mathcal{D}} W,$$

with W a $N(0, \Sigma)$ random variable.

4. In Exercise 2, let $d = 1$, $\bar{X}_n$ and S_{n-1}, be respectively, the sample mean and standard deviation of the first n observations, then

$$\frac{\bar{X}_n - \mu}{S_{n-1}/\sqrt{n}} \to^{\mathcal{D}} Z$$

as $n \to \infty$ with Z a standard normal random variable.

2.6 Exponential and Location-Scale families

Defnition. A parametric family of distributions $\{P_\theta : \theta \in \Omega\}$ dominated by a σ-finite measure ν on $(\mathcal{X}, \mathcal{B})$ is called an *exponential family* if

$$f_{X|\Theta}(x|\theta) = \frac{dP_\theta}{d\nu}(\theta) = c(\theta)h(x)\exp\{\sum_{i=1}^k \pi_i(\theta)t_i(x)\} = c(\theta)h(x)e^{\langle \pi(\theta), t(x)\rangle}.$$

In particular, there exists a dominating measure (e.g. $d\lambda = h\ d\nu$) such that $f_{X|\Theta}(x|\theta) > 0$ for all $(x, \theta) \in \text{supp}(\lambda) \times \Omega$. We can use this, for example, to show that the choice of P_θ, a $U(0, \theta)$-random variable, does not give an exponential family.

This choice for the representation is not unique. The transformation $\tilde{\pi}(\theta) = \pi(\theta)D$ and $\tilde{t}(x) = t(x)(D^T)^{-1}$ for a nonsingular matrix D gives another representation. Also, if $\tilde{\nu} << \nu$ then we can use $\tilde{\nu}$ as the dominating measure.

Note that

$$c(\theta) = (\int_{\mathcal{X}} h(x)e^{\langle \pi(x), t(x) \rangle}\ d\nu(x))^{-1}.$$

For $j = 1, 2$, let X_j be independent exponential families with parameter set Ω_j, and dominating measure ν_j, then the pair (X_1, X_2) is an an exponential family with parameter set (Ω_1, Ω_2) and dominating measure $\nu_1 \times \nu_2$.

Consider the reparameterization $\pi = \pi(\theta)$, then

$$\tilde{f}_{X|\Pi}(x|\pi) = \tilde{c}(\pi)h(x)e^{\langle \pi, t(x) \rangle}.$$

We call the vector $\Pi = \pi(\Theta)$ the *natural parameter* and

$$\Gamma = \{\pi \in R^k : \int_{\mathcal{X}} h(x)e^{\langle \pi, t(x) \rangle}\nu(dx) < \infty\}$$

is called the natural parameter space.

Examples.

1. Let $\{P_\theta : \theta \in (0, 1)\}$ be $Bin(n, \theta)$ and let ν be counting measure on $\{0, 1, \cdots, n\}$, then

$$f_{X|\Theta}(x|\theta) = (1-\theta)^n \binom{n}{x} \exp\{x \log \frac{\theta}{1-\theta}\}.$$

Thus,

$$c(\theta) = (1-\theta)^n, \quad h(x) = \binom{n}{x}, \quad t(x) = x, \quad \pi(\theta) = \log \frac{\theta}{1-\theta}.$$

The natural parameter is $\pi = \log \frac{\theta}{1-\theta}$. We can recover θ via $\theta = e^\pi/(1 + e^\pi)$.

2. Let $\{P_\theta : \theta = (\mu, \sigma) \in R \times R_+\}$ be the distribution of $X = (X_1, \cdots, X_n)$, an i.i.d sequence of $N(\mu, \sigma^2)$ random variables. Let ν be n-dimensional Lebesgue measure. Then,

$$\begin{aligned} f_{X|\Theta}(x|\theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\} \\ &= \frac{1}{(2\pi)^{n/2}} \sigma^{-n} \exp\{\frac{n\mu^2}{2\sigma^2}\} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2} n\bar{x}\}. \end{aligned}$$

Thus,

$$c(\theta) = \sigma^{-n}\exp\{\frac{n\mu^2}{2\sigma^2}\}, \quad h(x) = \frac{1}{(2\pi)^{n/2}}, \quad t(x) = (n\bar{x}, \sum_{i=1}^{n} x_i^2), \quad \pi(x) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}).$$

An exponential family of distributions is called *degenerate* if the image of the natural parameterization is almost surely contained in a lower dimensional hyperplane of R^k. In other words, there is a vector α and a scalar r so that $P_\theta\{\langle, \alpha X\rangle = r\} = 1$ for all θ.

For X distributed $Mult_k(\theta_1, \cdots, \theta_k), \Omega = \{\theta \in R^k : \theta_i > 0, \sum_{i=1}^{k} \theta_i = 1\}$. Then,

$$c(\theta) = 1, \quad h(x) = \binom{n}{x_1, \cdots, x_k}, \quad t(x) = x$$

The natural parameter is $\pi(\theta) = (\log(\theta_1), \cdots, \log(\theta_k))$.

This family is degenerate because $Pr\{\langle 1, X\rangle = n\}$. We can choose the reparameterization

$$\tilde{\pi}(\theta) = (\log(\theta_1/\theta_k), \cdots, \log(\theta_{k-1}/\theta_k))$$

with

$$\tilde{t}(x) = (x_1, \cdots, x_{k-1}), \quad \tilde{c}(\theta) = \theta_k^n.$$

This gives a nondegenerate exponential family.

Formulas for exponential families often require this non-degeneracy. Thus, we will often make a linear transformation to remove this degeneracy.

Definition. (Location-scale families.) Let Pr be a probability measure on $(R^d, \mathcal{B}(R^d))$ and let $\mathcal{M}_d$ be the collection of all $d \times d$ symmetric positive definite matrices. Set

$$Pr_{(\mu,\Sigma)}(B) = Pr((B-\mu)\Sigma^{-1/2}), \qquad B \in \mathcal{B}(R^d).$$

Then, the collection

$$\{Pr_{(\mu,\Sigma)}(B) : \mu \in R^d, \Sigma \in \mathcal{M}_d\}$$

is called a *location-scale family.*

A *location family* is obtained by restricting the choice of matrices to the identity matrix. A *scale family* is obtained by restricting the choice of vectors to $\mu = 0$. The normal distributions on R^d form a location-scale family. The uniform distributions on $(0, \theta)$ form a location family.

3 Sufficiency

We begin with some notation.

Denote the underlying probability space by $(S, \mathcal{A}, \mu)$. We will often write $Pr(A)$ for $\mu(A)$.

We refer to a measurable mapping $X : (S, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ as the *data.* $\mathcal{X}$ is called the sample sapce.

We begin with a parametric family of distributions $\mathcal{P}_0$ of $\mathcal{X}$. The *parameter* Θ is a mapping from the *parameter space* (Ω, τ) to $\mathcal{P}_0$. Preferably, this mapping will have good continuity properties.

The distribution of X under the image of θ is denoted by

$$P'_\theta\{X \in B\} = Pr\{X \in B | \Theta = \theta\} = P_\theta(B), \qquad B \in \mathcal{B}.$$

Thus, if Θ has distribution μ_Θ,

$$Pr\{X \in B, \Theta \in D\} = \int_D P_\theta(B)\ \mu_\Theta(d\theta).$$

For example, let $\theta = (\alpha, \beta), \alpha > 0, \beta > 0$. Under the distribution determined by θ, X is a sequence of n independent $Beta(\alpha, \beta)$-random variables. Pick $B \in \mathcal{B}([0,1]^n)$. Then

$$P_\theta(B) = \int_B \prod_{i=1}^n \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1}\ dx.$$

Let $\mathcal{C}$ be a sigma field on $\mathcal{T}$ that contains singletons, then $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{C})$ is a *statistic* and $T(X)$ is called a *random quantity.* We write

$$P_{\theta,T}(C) = P'_\theta\{T(X) \in C\} = P'_\theta\{T \in C\}.$$

3.1 Basic Notions

Definition. Let $\mathcal{P}_0$ be a parametric family of distributions on $(\mathcal{X}, \mathcal{B})$. Let (Ω, τ) be the parameter space and let $\Theta : \Omega \to \mathcal{P}_0$ be the parameter. Let $T : \mathcal{X} \to \mathcal{T}$ be a statistic.

1. T is *sufficient in the classical sense* if the conditional distribution of X given $T(X)$ does not depend on θ.

2. T is *sufficient in the Bayesian sense* if, for every prior μ_Θ, there exists versions of the posteriors $\mu_{\Theta|X}$ and $\mu_{\Theta|T}$ such that for every $B \in \tau$,

   $$\mu_{\Theta|X}(B|x) = \mu_{\Theta|T}(B|T(x)) \qquad \text{a.s. } \mu_X,$$

   where μ_X is the marginal distribution of X.

If T is sufficient in the classical sense, then there exists a function $r : \mathcal{B} \times \mathcal{T} \to [0,1]$ such that

1. $r(\cdot|t)$ is a probability on $\mathcal{X}$ for every $t \in \mathcal{T}$.

2. $r(B|\cdot)$ is measurable on $\mathcal{T}$ for every $B \in \mathcal{B}$.

3. For every $\theta \in \Omega$ and every $B \in \mathcal{B}$,

$$P_\theta\{B|T = t\} = r(B|t) \qquad \text{a.e. } P_{\theta,T}.$$

Thus, given $T = t$, one can generate the conditional ditribution of X without any knowledge of the parameter θ.

Example. Let X be n independent $Ber(\theta)$ random variables and set $T(x) = \sum_{i=1}^n x_i$. Then

$$P'_\theta\{X = x|T(X) = t\} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}.$$

Thus, T is sufficient in the classical sense and $r(\cdot|t)$ has a uniform distribution.

By Bayes' formula, the Radon-Nikodym derivative,

$$\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(x|\theta) = \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\int \psi^{\sum_{i=1}^n x_i}(1-\psi)^{n-\sum_{i=1}^n x_i}\ \mu_\Theta(d\psi)}.$$

Given $\Theta = \theta$, $T(X)$ is $Bin(n, \theta)$, thus

$$\frac{d\mu_{\Theta|T}}{d\mu_\Theta}(x|t) = \frac{\binom{n}{t}\theta^t(1-\theta)^{n-t}}{\int \binom{n}{t}\psi^t(1-\psi)^{n-t}\ \mu_\Theta(d\psi)}.$$

Thus, T is sufficient in the Bayesian sense.

Checking the two basic properties of conditional expectation, we have the following lemma.

Lemma. A statistic T is sufficient in the Bayesian sense if and only if, for every prior μ_Θ, there exists a version of the posterior distribution given X, $\mu_{\Theta|X}$ such that for all $B \in \tau$, $\mu_{\theta|X}(B|\cdot)$ is $\sigma(T)$-measurable, i.e. $\mu_{\theta|X}(B|x)$ is a function of T.

Exercise. If $X = (X_1, \cdots, X_n)$ are exchangeable, then the order statistics $(X_{(1)}, \cdot, X_{(n)})$ are sufficient.

The two notions of sufficiency are quite similar as the following theorem demonstrates.

Theorem.

1. If T is sufficient in the classical sense, then T is sufficient in the Bayesian sense.

2. Let T be sufficient in the Bayesian sense. If $P_\theta << \nu$ for all θ and some σ finite measure ν, then T is sufficient in the classical sense.

Proof. Let r be as desribed above and let μ_Θ be a prior for Θ. Then,

$$\mu_{X|T}(B|T=t) = \int_\Omega P_\theta(B|T=t)\ \mu_{\Theta|T}(d\theta|t) = \int_\Omega r(B|t)\ \mu_{\Theta|T}(d\theta|t) = r(B|t) = \mu_{X|T,\Theta}(B|T=t,\Theta=\theta),$$

where $\mu_{\Theta|T}$ is the posterior distribution of Θ given T. Thus, the conditional distribution of X given (Θ, T) is the conditional distribution of X given T. Consequently, X and Θ are conditionally independent given T and

$$\mu_{\Theta|T,X} = \mu_{\Theta|T}.$$

Because T is a function of X, we always have

$$\mu_{\Theta|T,X} = \mu_{\Theta|X}$$

and T is sufficient in the Bayesian sense.

Part 2 requires the following lemma.

Lemma. let ν be a σ-finite measure dominating P_θ for each $\theta \in \Omega$. If T is sufficient in the Bayesian sense, then there exists a probability measure ν^* such that $P_\theta << \nu^* << \nu$ for all θ and

$$f_{X|\Theta}(x|\theta) = \frac{dP_\theta}{d\nu^*}(x) = h(\theta, T(x))$$

for some measurable function $h : \Omega \times \mathcal{T} \to R$.

Proof. We can choose $\Omega_\nu = \{\theta_i : i \geq 1\}$ and $\{c_i : i \geq 1\}$ so that $c_i \geq 0, \sum_{i=1}^\infty c_i = 1$ and for every $\theta \in \Omega$,

$$P_\theta << \nu^* = \sum_{i=1}^\infty c_i P_{\theta_i}.$$

For $\theta \in \Omega \backslash \Omega_\nu$ specify the prior distribution

$$Pr\{\Theta = \theta\} = \frac{1}{2}, \qquad Pr\{\Theta = \theta_i\} = \frac{c_i}{2}.$$

Then,

$$\begin{aligned} Pr\{\Theta = \theta | X = x\} &= \frac{Pr\{X = x|\Theta = \theta\} Pr\{\Theta = \theta\}}{Pr\{X = x\}} \\ &= \frac{f_{X|\Theta}(x|\theta)}{f_{X|\Theta}(x|\theta) + \sum_{i=1}^\infty c_i f_{X|\Theta}(x|\theta_i)} \\ &= (1 + \frac{\sum_{i=1}^\infty c_i f_{X|\Theta}(x|\theta_i)}{f_{X|\Theta}(x|\theta)})^{-1}. \end{aligned}$$

Because T is sufficient in the Bayesian sense, $Pr\{\Theta = \theta | X = x\}$ is, for each θ, a function of $T(x)$. Thus, we write

$$h(\theta, T(x)) = \frac{f_{X|\Theta}(x|\theta)}{\sum_{i=1}^\infty c_i f_{X|\Theta}(x|\theta_i)} = \frac{dP_\theta}{d\nu}(x) / \frac{d\nu^*}{d\nu}(x) = \frac{dP_\theta}{d\nu}(x).$$

If $\theta \in \Omega_\nu$, use the prior $Pr\{\Theta = \theta_i\} = c_i$ and repeat the steps above.

Proof. (Part 2 of the theorem) Write $\tilde{r}(B|t) = \nu^*(X \in B|T = t)$ and set $\nu_T^*(C) = \nu^*\{T \in C\}$. Then

$$\nu^*(T^{-1}(C) \cap B) = \int I_C(T(x))I_B(x)\ \nu^*(dx) = \int_C \tilde{r}(B|t)\ \nu_T^*(dt).$$

Thus by the standard machine, for every integrable $g : \mathcal{T} \to R$,

$$\int g(T(x))I_B(x)\ \nu^*(dx) = \int g(t)\tilde{r}(B|t)\ \nu_T^*(dt).$$

Claim. For all θ, $\tilde{r}(B|t) = P_\theta\{X \in B|T = t\}$
Note that $P_\theta\{X \in B|T = t\}$ is characterized by satisfying

1. It is a function $m : \mathcal{T} \to [0, 1]$.
2. $E_\theta[I_B(X)I_C(T(X))] = E_\theta[m(T(X))I_C(T(X))]$.

Clearly, $\tilde{r}(B|t)$ satisfies 1. By the lemma,

$$\frac{dP_{\theta,T}}{d\nu_T^*}(t) = h(\theta, t).$$

$(P_{\theta,T}(B) = P_\theta\{T(X) \in B\} = \int I_B(T(x))h(\theta, T(x))\ \nu^*(dx) = \int I_B(t)h(\theta, t)\ \nu_T^*(dt))$
Thus,

$$\begin{aligned}\int_C \tilde{r}(B|t)\ P_{\theta,T}(dt) &= \textstyle\int I_C(t)\tilde{r}(B|t)h(\theta, t)\ \nu_T^*(dt) &&= \int I_B(x)I_C(T(x))h(\theta, T(x))\ \nu^*(dx)\\ &= \textstyle\int I_B(x)I_C(T(x))\ P_\theta(dx) &&= E_\theta[I_B(X)I_C(T(X))].\end{aligned}$$

This gives the claim.

3.2 Fisher-Neyman Factorization Theorem

If all the conditional distributions are absolutely continuous with respect to a single σ-finite measure, then the two senses of sufficiency agree. In this circumstance, the Fisher-Neyman factorization theorem gives a simple characterization of sufficiency.

Theorem. Let $\{P_\theta : \theta \in \Omega\}$ be a parametric family such that $P_\theta << \nu$ for all θ. Write

$$\frac{dP_\theta}{d\nu}(x) = f_{X|\Theta}(x|\theta).$$

Then T is sufficient if and only if there exists functions m_1 and m_2 such that

$$f_{X|\Theta}(x|\theta) = m_1(x)m_2(T(x), \theta).$$

Proof. By the theorem above, it suffices to prove this theorem for sufficiency in the Bayesian sense.

Write $f_{X|\Theta}(x|\theta) = m_1(x)m_2(T(x),\theta)$ and let μ_Θ be a prior for Θ. Bayes' theorem states that the posterior distribution of Θ is absolutely continuous with respect to the prior with Radon-Nikodym derivative

$$\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta|x) = \frac{m_1(x)m_2(T(x),\theta)}{\int_\Omega m_1(x)m_2(T(x),\psi)\ \mu_\Theta(d\psi)} = \frac{m_2(T(x),\theta)}{\int_\Omega m_2(T(x),\psi)\ \mu_\Theta(d\psi)}.$$

This is a function of $T(x)$. Thus, T is sufficient.

Now, assume that T is sufficient. Then there exists a measure ν^* such that

1. $P_\theta << \nu^*$ for all θ.
2. $dP_\theta/d\nu^*(x) = h(\theta, T(x))$ for some measurable h.
3. $\nu^* << \nu$.

Therefore,

$$f_{X|\Theta}(x|\theta) = \frac{dP_\theta}{d\nu^*}(x)\frac{d\nu^*}{d\nu}(x) = h(\theta, T(x))\frac{d\nu^*}{d\nu}(x).$$

Now set $m_1 = d\nu^*/d\nu$ and $m_2 = h$.

Example. (truncation families) Let ϕ be a non-negative Borel function on $(R, \mathcal{B}(R))$ such that for any α and β, we have $\int_\alpha^\beta \phi(x)\ dx < \infty$. Use the parameter $\theta = (\alpha, \beta)$, $\Omega = \{(\alpha,\beta) \in R^2 : \alpha < \beta\}$, and set the densities with respect to Lebesgue measure

$$f_{X|\Theta}(x|\theta) = c(\theta)\phi(x)I_{(\alpha,\beta)}(x)$$

with $c(\theta) = (\int_\alpha^\beta \phi(x)\ dx)^{-1}$.

The joint density function of n independent identically distributed random variables from this truncation family is

$$\begin{aligned}\prod_{i=1}^n f_{X|\Theta}(x_i|\theta) &= c(\theta)^n \prod_{i=1}^n I_{(\alpha,\beta)}(x_i) \prod_{i=1}^n \phi(x_i). \\ &= c(\theta)^n I_{(\alpha,\infty)}(x_{(1)})I_{(-\infty,\beta)}(x_{(n)}) \prod_{i=1}^n \phi(x_i)\end{aligned}$$

Write this as $m_1(x)m_2(T(x),\theta)$ with $m_1(x) = \prod_{i=1}^n \phi(x_i)$. and $m_2(T(x),\theta) = c(\theta)^n I_{(\alpha,\infty)}(x_{(1)})I_{(-\infty,\beta)}(x_{(n)})$, $T(x) = (x_{(1)}, x_{(n)})$. Thus, the minimum and maximum are sufficient statistics.

Lemma. Assume the conditions of the factorization theorem and assume that $T : \mathcal{X} \to \mathcal{T}$ is sufficient. Then there exists a measure $\nu_\mathcal{T}$ on $(\mathcal{T}, \mathcal{C})$ such that $P_{\theta,T} << \nu_\mathcal{T}$ and $dP_{\theta,T}/d\nu_\mathcal{T} = m_2(t,\theta)$.

Proof. Define ν^* are before, then $P_\theta << \nu^*$ for each θ and

$$\frac{dP_\theta}{d\nu^*}(x) = \frac{f_{X|\Theta}(x|\theta)}{\sum_{i=1}^\infty c_i f_{X|\Theta}(x|\theta_i)} = \frac{m_2(T(x),\theta)}{\sum_{i=1}^\infty c_i m_2(T(x),\theta_i)}.$$

Therefore,

$$\begin{aligned} P_{\theta,T}(B) &= \int_{T^{-1}(B)} \frac{dP_\theta}{d\nu^*}(x)\ \nu^*(dx) \\ &= \int_{T^{-1}(B)} \frac{m_2(T(x),\theta)}{\sum_{i=1}^\infty c_i m_2(T(x),\theta_i)}\ \nu^*(dx) \\ &= \int_B \frac{m_2(t,\theta)}{\sum_{i=1}^\infty c_i m_2(t,\theta_i)}\ d\nu_T^*(x) \end{aligned}$$

with $\nu_T^*(B) = \nu^*\{T \in B\}$. Now, set $d\nu_{\mathcal{T}}/d\nu_T^*(t) = (\sum_{i=1}^\infty c_i m_2(t,\theta_i))^{-1}$ to complete the proof.

Example. (exponential families) By the factorization theorem, for densities

$$f_{X|\Theta}(x|\theta) = \frac{dP_\theta}{d\nu}(x) = c(\theta)h(x)e^{\langle \pi(\theta), t(x)\rangle},$$

we have that t is a sufficient statistic, sometimes called the *natural sufficient statistic.*

Note that $t(X)$ is an exponential family. We will sometimes work in the *sufficient statistic space.* In this case, the parameter is the natural parameter which we now write as Ω. The reference measure is $\nu_{\mathcal{T}}$ described in the lemma above. The density is

$$f_{T|\Theta}(t|\theta) = \frac{dP_{T,\theta}}{d\nu_{\mathcal{T}}}(x) = c(\theta)e^{\langle \theta, t\rangle},$$

3.3 Regularity Properties of Exponential Families

Theorem. The natural parameter space Ω of an exponential family is convex and $1/c(\theta)$ is convex.

Proof. Working with the sufficient statistics space, write

$$\frac{1}{c(\theta)} = \int e^{\langle \theta, t\rangle} \nu_{\mathcal{T}}(dt).$$

Choose $\theta_1, \theta_2 \in \Omega$ and $\alpha \in (0,1)$. Then, by the convexity of the exponential, we have that

$$\begin{aligned} \frac{1}{c(\alpha\theta_1 + (1-\alpha)\theta_2)} &= \int e^{\langle (\alpha\theta_1 + (1-\alpha)\theta_2), t\rangle} \nu_{\mathcal{T}}(dt) \\ &= \int (\alpha e^{\langle \theta_1, t\rangle} + (1-\alpha)e^{\langle \theta_2, t\rangle}) \nu_{\mathcal{T}}(dt) \\ &= \frac{\alpha}{c(\theta_1)} + \frac{1-\alpha}{c(\theta_2)} < \infty. \end{aligned}$$

Moreover, if $\int |\phi(t)|e^{\langle \theta, t\rangle}\nu_{\mathcal{T}}(dt) < \infty$ for θ in the interior of the natural parameter space for $\phi : \mathcal{T} \to R$, then

$$f(z) = \int \phi(t)e^{\langle t, z\rangle}\nu_{\mathcal{T}}(dt)$$

is an analytic function of z in the region for which the real part of z is in the interior of the natural parameter space. Consequently,

$$\frac{\partial}{\partial z_i} f(z) = \int t_i \phi(t) e^{\langle z,t \rangle} \nu_T(dt)$$

Taking $\phi(\theta) = 1$, we have

$$E_\theta T_i = c(\theta) \frac{\partial}{\partial \theta_i} \frac{1}{c(\theta)} = -\frac{\partial}{\partial \theta_i} \log c(\theta).$$

More generally, if $\ell = \ell_1 + \cdots + \ell_k$,

$$E_\theta[\prod_{i=1}^k T_i^{\ell^i}] = c(\theta) \frac{\partial^\ell}{\partial \theta_1^{\ell_1} \cdots \partial \theta_k^{\ell_k}} \frac{1}{c(\theta)}.$$

For example,

$$\mathrm{Cov}(T_i, T_j) = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log c(\theta).$$

Examples.

1. The family $Exp(\psi)$ has densities $f_{X|\Psi}(x|\psi) = \psi e^{-\psi x}$ with respect to Lebesgue measure. Thus, the natural parameter, $\theta = -\psi \in (-\infty, 0)$, $c(\theta) = -1/\theta$, a convex function.

$$E_\theta T = \frac{\partial}{\partial \theta} \log(-\theta) = \frac{1}{-\theta}, \quad \mathrm{Var}(T) = \frac{\partial^2}{\partial \theta^2} \log(-\theta) = \frac{1}{\theta^2}.$$

2. For the sum of n-independent $N(\mu, \sigma^2)$ random variables, the natural parameter $\theta = (\mu/\sigma^2, -1/2\sigma^2)$, and the natural sufficient statistic $T(x) = (n\bar{x}, \sum_{i=1}^n x_i^2)$. Thus,

$$\log c(\theta) = \frac{n}{2} \log(-2\theta_2) + \frac{n}{4} \frac{\theta_1^2}{\theta_2},$$

$$E_\theta[n\bar{X}] = -\frac{\partial}{\partial \theta_1} \log c(\theta) = -\frac{n}{2} \frac{\theta_1}{\theta_2} = n\mu, \quad E_\theta[\sum_{i=1}^n X_i^2] = -\frac{\partial}{\partial \theta_2} \log c(\theta) = -\frac{n}{2\theta_2} + \frac{n}{4} \frac{\theta_1^2}{\theta_2^2} = n(\sigma^2 + \mu^2),$$

$$\mathrm{Cov}(n\bar{X}, \sum_{i=1}^n X_i^2) = -\frac{\partial}{\partial \theta_2} \frac{\partial}{\partial \theta_1} \log c(\theta) = -\frac{\partial}{\partial \theta_2} \frac{n}{2} \frac{\theta_1}{\theta_2} = \frac{n}{2} \frac{\theta_1}{\theta_2^2} = 2n\mu\sigma^2.$$

Definition. Let $\mathcal{P}_0$ be a family of distributions on $\mathcal{X}$. A second family of distributions $\mathcal{P}^*$ is called a *conjugate family* provided that any choice of prior $\mu_\Theta \in \mathcal{P}^*$ implies that the posterior $\mu_{\Theta|X} \in \mathcal{P}^*$.

Exercise. Let $X = (X_1, \cdots, X_n)$ be independent $Ber(\theta)$ random variables and set $T(X) = \sum_{i=1}^n X_i$. Then the beta family of distributions forms a natural conjugate pair.

This fact, in the case of integer parameters, is a consequence of the following theorem.

Theorem. Suppose, for any choice n of sample size, there exists a natural number k and a sufficient statistic T_n whose range in contained in R^k, functions $m_{1,n}$ and $m_{2,n}$ such that

$$f_{X_1,\cdots,X_n|\Theta}(x|\theta) = m_{1,n}(x_1,\cdots,x_n)m_{2,n}(T_n(x_1,\cdots,x_n),\theta).$$

In addition, assume that for all n and all $t \in \mathcal{T}$,

$$0 < c(t,n) = \int_\Omega m_{2,n}(t,\theta)\ \lambda(d\theta) < \infty.$$

Then the family

$$\mathcal{P}^* = \{\frac{m_{2,n}(t,\cdot)}{c(t,n)} : t \in \mathcal{T}, n = 1,2,\cdots\}$$

is a conjugate family.

We can apply the computational ideas above to computing posterior distributions.

Theorem. Let $X = (X_1,\cdots,X_n)$ be i.i.d. given $\Theta = \theta \in R^k$ with density $c(\theta)\exp(\langle\theta, T(x)\rangle)$. Choose $a > 0$ and $b \in R^k$ so that the prior for Θ is proportional to $c(\theta)^a \exp(\langle\theta, b\rangle)$. Write the predictive density of X, $f_X(x) = g(t_1,\cdots,t_k)$, where $t_i = \sum_{j=1}^n T_i(x_j)$. Set $\ell = \ell_1 + \cdots + \ell_k$, then

$$E[\prod_{i=1}^k \Theta_i^{\ell_i}|X = x] = \frac{1}{f_X(x)} \frac{\partial^\ell}{\partial t_1^{\ell_1}\cdots\partial t_k^{\ell_k}} g(t_1,\cdots,t_k).$$

3.4 Minimal and Complete Sufficiency

The entire data is always a sufficient statistic. Here we look for a sufficient statistics T that minimizes $\sigma(T)$.

Definition. A sufficient statistic $T : \mathcal{X} \to \mathcal{T}$ is called *minimal sufficient* if for every sufficient statistic $U : \mathcal{X} \to \mathcal{U}$, there is a measurable function $g : \mathcal{U} \to \mathcal{T}$ such that $T = g(U)$, a.s. P_θ for all θ.

Theorem. let ν be σ-finite and suppose that there exist versions of $dP_\theta/d\nu = f_{X|\Theta}(\cdot|\theta)$ for every θ and a measurable function $T : \mathcal{X} \to \mathcal{T}$ that is constant on the sets

$$\mathcal{D}(x) = \{y \in \mathcal{X} : f_{X|\Theta}(y|\theta) = f_{X|\Theta}(x|\theta)h(x,y) \text{ for all } \theta \text{ and some } h(x,y) > 0\},$$

then T is a minimal sufficient statistic.

Proof. Note that the sets $\mathcal{D}(x)$ partition $\mathcal{X}$. To show that T is sufficient, choose a prior μ_Θ. The density of the posterior using Bayes' theorem.

$$\begin{aligned}
\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta|x) &= \frac{f_{X|\Theta}(x|\theta)}{\int f_{X|\Theta}(x|\psi)\ \mu_\Theta(d\psi)} \\
&= \frac{h(x,y)f_{X|\Theta}(x|\theta)}{\int h(x,y)f_{X|\Theta}(x|\psi)\ \mu_\Theta(d\psi)} \\
&= \frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta|y)
\end{aligned}$$

provided $y \in \mathcal{D}(x)$. Hence, the posterior is a function of $T(x)$ and T is sufficient.

To prove that T is minimal, choose a sufficient statistic U. We will show that $U(y) = U(x)$ implies that $y \in \mathcal{D}(x)$ and hence that T is a function of U. Use the Fisher-Neyman factorization theorem to write

$$f_{X|\Theta}(x|\theta) = m_1(x)m_2(U(x), \theta).$$

Because, for each θ, $m_1 > 0$ with P_θ-probability 1, we can choose a version of m_1 that is positive for all x. If $U(x) = U(y)$, then

$$f_{X|\Theta}(x|\theta) = f_{X|\Theta}(y|\theta)\frac{m_1(y)}{m_2(x)}$$

for all θ. Thus, we can choose $h(x, y) = m_1(y)/m_2(x)$ to place $y \in \mathcal{D}(x)$.

Examples.

1. Let $X_1, \cdots, X_n$ be independent $Ber(\theta)$ random variables. Then

   $$f_{X|\Theta}(x|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

   So, the ratio

   $$\frac{f_{X|\Theta}(y|\theta)}{f_{X|\Theta}(x|\theta)} = (\frac{\theta}{1-\theta})^{\sum_{i=1}^n y_i - \sum_{i=1}^n x_i}.$$

   Thus, $h(x, y) = 1$ and $\mathcal{D}(x) = \{y \in \{0,1\}^n : \sum_{i=1}^n y_i = \sum_{i=1}^n x_i\}$. This gives that $T(x) = \sum_{i=1}^n x_i$ is a minimal sufficient statistic.

2. Let $X_1, \cdots, X_n$ be independent $U(0, \theta)$ random variables. Then

   $$f_{X|\Theta}(x|\theta) = \theta^{-n} I_{[0,\theta]}(\max_i x_i).$$

   Now suppose that, for all θ

   $$\theta^{-n} I_{[0,\theta]}(\max_i x_i) = h(x, y)\theta^{-n} I_{[0,\theta]}(\max_i y_i).$$

   Then $\max_i x_i = \max_i y_i$ and $h(x, y) = 1$ and $\mathcal{D}(x) = \{y \in R_+^n : \max_i x_i = \max_i y_i\}$. Consequently, $T(x) = \max_i x_i$ is a minimal sufficient statistic.

Definition. A statistic T is *(boundedly) complete* if, for every (bounded) measurable function g,

$$E_\theta[g(T)] = 0 \text{ for all } \theta \in \Omega \text{ implies } g(T) = 0, \text{ a.s. } P_\theta \text{ for all } \theta.$$

Examples.

1. Let $X_1, \cdots, X_n$ be independent $Ber(\theta)$ and let T denote the sum. Choose g so that $E_\theta[g(T)] = 0$ for all θ. Then

$$\begin{aligned} 0 = E_\theta[g(T)] &= \sum_{i=1}^{n} g(i)\binom{n}{i}\theta^i(1-\theta)^{n-i} \\ &= (1-\theta)^n \sum_{i=1}^{n} g(i)\binom{n}{i}(\frac{\theta}{1-\theta})^i \end{aligned}$$

   This polynomial in $\theta/(1-\theta)$ must have each of its coefficients equal to zero. Thus, $g(i) = 0$ for all i in the range of T. Hence, T is complete.

2. Let $X_1, \cdots, X_n$ be independent $U(0, \theta)$ and let T denote the maximum. Choose g so that $E_\theta[g(T)] = 0$ for all θ. Then

$$0 = E_\theta[g(T)] = \int_0^\theta g(t)nt^{n-1}\theta^n \ dt.$$

   Thus, the integrand and hence $g(t) = 0$ a.e. and T is complete.

Theorem. If the natural parameter space Ω of an exponential family contains an open set in R^k, then the natural sufficient statistic is complete.

Proof. Let $T(X)$ have density $c(\theta)\exp\langle\theta, t\rangle$ with respect to a measure ν_T. Let g be a function such that

$$0 = E_\theta[g(T)] = \int g(t)c(\theta)\exp\langle\theta, t\rangle \ \nu_T(dt).$$

Thus, for each θ

$$\int g^+(t)\exp\langle\theta, t\rangle \ \nu_T(dt) = \int g^-(t)\exp\langle\theta, t\rangle \ \nu_T(dt).$$

Fix θ_0 in the interior of the natural parameter space and let $Z(\theta_0)$ be the common value of the integrals above. Define two probability measures

$$\begin{aligned} P^+(A) &= \frac{1}{Z(\theta_0)}\int g^+(t)\exp\langle\theta, t\rangle \ \nu_T(dt) \\ P^-(A) &= \frac{1}{Z(\theta_0)}\int g^-(t)\exp\langle\theta, t\rangle \ \nu_T(dt). \end{aligned}$$

Now the equality above can be written

$$\int \exp\langle(\theta-\theta_0), t\rangle \ P^+(dt) = \int \exp\langle(\theta-\theta_0), t\rangle \ P^-(dt).$$

Thus, the Laplace transforms of P^+ and P^- agree on an open set, and hence $P^+ = P^-$. Consequently, $g^+ = g^-$ a.s. ν_T and $P_\theta\{g(T) = 0\} = 1$.

Thus, the sufficient statistics from normal, exponential, Poisson, Beta, and binomial distributions are complete.

Theorem. (Bahadur) If U is a finite dimensional boundedly complete and sufficient statistic, then it is minimal sufficient.

Proof. Let T be another sufficient statistic. Write $U = (U_1, \cdots, U_k)$ and set $V_i(u) = (1 + \exp(u_i))^{-1}$. Thus, V is a bounded, one-to-one function of u. Define

$$H_i(t) = E_\theta[V_i(U)|T = t], \quad L_i(u) = E_\theta[H_i(T)|U = u].$$

Because U and T are sufficient, these conditional means do not depend on θ. Because V is bounded, so are H and L. Note that by the tower property,

$$E_\theta[V_i(U)] = E_\theta[E_\theta[V_i(U)|T]] = E_\theta[H_i(T)] = E_\theta[E_\theta[H_i(T)|U]] = E_\theta[L_i(U)].$$

Thus, $E_\theta[V_i(U) - L_i(U)] = 0$ for all θ. Use the fact that U is boundedly complete to see that $P_\theta\{V_i(U) = L_i(U)\} = 1$ for all θ.

Now, use the conditional variance formula.

$$\mathrm{Var}_\theta(L_i(U)) = E_\theta[\mathrm{Var}_\theta(L_i(U))|T] + \mathrm{Var}_\theta(H_i(T)), \qquad \mathrm{Var}_\theta(H_i(T)) = E_\theta[\mathrm{Var}_\theta(H_i(T))|U] + \mathrm{Var}_\theta(L_i(U)).$$

Add these equations and simplify to obtain

$$0 = E_\theta[\mathrm{Var}_\theta(L_i(U))|T] + E_\theta[\mathrm{Var}_\theta(H_i(U))|T].$$

Because conditional variances are non-negative, we have that $0 = \mathrm{Var}_\theta(L_i(U)|T) = \mathrm{Var}_\theta(V_i(U)|T)$, a.s. P_θ. Thus, $V_i(U) = E_\theta[V_i(U)|T] = H_i(T)$ or $U_i = V_i^{-1}(H_i(T))$, a.s. P_θ, and U is a function of T. Consequently, U is minimal sufficient.

3.5 Ancillarity

Definition. A statistic U is called *ancillary* if the conditional distribution of U is independent of Θ.

Examples.

1. Let X_1, X_2 be independent $N(\theta, 1)$, then $X_2 - X_1$ is $N(0, 2)$.
2. Let $X_1, \cdots, X_n$ be independent observations from a location family, then $X_{(n)} - X_{(1)}$ is ancillary.
3. Let $X_1, \cdots, X_n$ be independent observations from a scale family, then any function of the random variables $X_1/X_n, \cdots, X_{n-1}/X_n$ is ancillary.

Sometimes a minimal sufficient statistic contains a coordinate that is ancillary. For example, for n i.i.d. observations $X_1, \cdots, X_n$, from a location family take $T = (T_1, T_2) = (\max_i X_i, \max_i X_i - \min_i X_i)$. Then T is minimal sufficient and T_2 is ancillary. In these types of situations, T_1 is called *conditionally sufficient given* T_2.

Theorem. **(Basu)** Suppose that T is a boundedly complete sufficient statistic and U is ancillary. Then U and T are independent given $\Theta = \theta$ and are marginally independent irrespective of the prior used.

Proof. Let A be a measurable subset of the range of U. Because U is ancillary, $P'_\theta\{U \in A\}$ is constant. Because T is sufficient, $P'_\theta\{U \in A|T\}$ is a function of T independent of θ. Note that

$$E_\theta[P'_\theta\{U \in A|T\} - P'_\theta\{U \in A\}] = 0.$$

Use the fact that T is boundedly complete to conclude that

$$P'_\theta\{U \in A|T\} = P'_\theta\{U \in A\} \quad \text{a.s } P_\theta.$$

Now, let B be a measurable subset of the range of T, then

$$\begin{aligned} P'_\theta\{T \in B, U \in A\} &= E_\theta[E_\theta[I_B(T(X))I_A(U(X))|T(X)]] \\ &= E_\theta[I_B(T(X))E_\theta[I_A(U(X))|T(X)]] \\ &= E_\theta[I_B(T(X))E_\theta[I_A(U(X))|T(X)]] \\ &= E_\theta[I_B(T(X))P'_\theta\{U \in A\}] \\ &= P'_\theta\{T \in B\}P'_\theta\{U \in A\} \end{aligned}$$

Let μ_Θ be a prior for Θ, then

$$\begin{aligned} Pr\{U(X) \in A, T(X) \in B\} &= \int_\Omega \int_B Pr\{U(X) \in A|T(X) = t\}\ P_{\Theta,T}(dt)\mu_\Theta(d\theta) \\ &= \int_\Omega Pr\{U(X) \in A\}\ P_{\Theta,T}(B)\mu_\Theta(d\theta) = Pr\{U(X) \in A\}Pr\{T(X) \in B\} \end{aligned}$$

Examples.

1. Let $X_1, \cdots, X_n$ be independent $N(\theta, 1)$. Then $\bar{X}$ is complete and $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is ancillary. Thus, they are independent.

2. Let $X_1, \cdots, X_n$ be independent $N(\mu, \sigma^2)$ Then, $(\bar{X}, S)$ is a complete sufficient statistic. Let

$$U = (\frac{X_1 - \bar{X}}{S}, \cdots, \frac{X_n - \bar{X}}{S}).$$

   Then U is ancillary and independent of $(\bar{X}, S)$. The distribution of U is uniform on a sphere of radius 1 in an $n-1$ dimensional hyperplane.

3. Conditionin on an ancillary statistic can sometimes give a more precise estimate. The following example is due to Basu.

   Let $\Theta = (\Theta_1, \cdots, \Theta_N)$, ($N$ known) and select labels $i_1, \cdots, i_n$ at random with replacement from $1, \cdots, N$, $n \leq N$. Set $X = (X_1, \cdots, X_n) = (\Theta_{i_1}, \cdots, \Theta_{i_n})$. Thus for all compatible values of x,

$f_X|\Theta(x|\theta) = 1/N^n$. Let M be the number of distinct labels drawn. Then, M is ancillary. Let $(X_1^*, \cdots, X_M^*)$ be the distinct values.

One possible estimate of the population average is $\bar{X}^* = \sum_{i=1}^M X_i/M$. For this, we have conditional variance

$$\mathrm{Var}(\bar{X}^*|\Theta = \theta, M = m) = \frac{N-m}{N-1}\frac{\sigma}{m},$$

where $\sigma^2 = \sum_{i=1}^N (\theta_i - \bar{\theta})^2/N$.

This is a better measure of the variance of $\bar{X}^*$ than the marginal variance for the case $n = 3$. In this case, the distribution of M is

$$f_M(m) = \frac{\binom{2}{m-1}(N)_{m-1}}{N^2} \qquad \text{for } m = 1, 2, 3,$$

and 0 otherwise. Because $E[\bar{X}^*|\Theta = \theta, M = m] = \bar{\theta}$ for all θ, we have that

$$\mathrm{Var}(\bar{X}^*|\Theta = \theta) = E[\frac{N-M}{N-1}\frac{\sigma^2}{M}|\Theta = \theta] = \frac{\sigma^2}{N^2}(1 + (N-2) + \frac{(N-2)(N-3)}{3}) = \frac{\sigma^2}{3}\frac{N^2 - 3N + 3}{N^2}.$$

4 Information

4.1 Fisher Information

Let Θ be a k dimensional parameter and let X have density $f_{X|\Theta}(x|\theta)$ with respect to ν. The following will be called the *Fisher Information (FI) regularity conditions.*

1. $0 = \partial f_{X|\Theta}(x|\theta)/\partial\theta$ exists for all θ, ν a.e.
2. For each $i = 1, \cdots, k$,
$$\frac{\partial}{\partial\theta_i}\int f_{X|\Theta}(x|\theta)\ \nu(dx) = \int \frac{\partial}{\partial\theta_i} f_{X|\Theta}(x|\theta)\ \nu(dx).$$
3. The set $C = \{x : f_{X|\Theta}(x|\theta) > 0\}$ is the same for all θ.

Definition Assume the FI regularity conditions above. Then the matrix $\mathcal{I}_X(\theta)$ with elements

$$I_{X,i,j}(\theta) = \mathrm{Cov}_\theta(\frac{\partial}{\partial\theta_i}\log f_{X|\Theta}(x|\theta), \frac{\partial}{\partial\theta_j}\log f_{X|\Theta}(x|\theta))$$

is called the *Fisher information matrix* about Θ based on X.

The random vector with coordinates $\nabla_\Theta \log f_{X|\Theta}(X|\theta)$ is called the *score function.*

For a statistic T, the *conditional score function* $\nabla_\Theta \log f_{X|T,\Theta}(X|t,\theta)$.

$\mathcal{I}_{X|T}(\theta|t)$, the *conditional Fisher information* given $T = t$, is the conditional covariance matrix of the conditional score function.

Under the FI regularity conditions, the mean of the score function is zero. To see this, note that

$$\begin{aligned} E[\frac{\partial}{\partial\theta_i} f_{X|\Theta}(X|\theta)] &= \int (\frac{\partial}{\partial\theta_i}\log f_{X|\Theta}(x|\theta)) f_{X|\Theta}(x|\theta)\ \nu(dx) \\ &= \int \frac{\partial}{\partial\theta_i} f_{X|\Theta}(x|\theta)\ \nu(dx) = 0. \end{aligned}$$

If differentiation is permitted twice under the integral sign, then

$$0 = \int \frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{X|\Theta}(x|\theta)\ \nu(dx) = E_\theta[\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{X|\Theta}(X|\theta)}{f_{X|\Theta}(X|\theta)}].$$

Because

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{x|\Theta}(X|\theta) = \frac{(\frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{x|\Theta}(x|\theta)) f_{X|\Theta}(x|\theta) - (\frac{\partial}{\partial\theta_i} f_{X|\Theta}(x|\theta))\frac{\partial}{\partial\theta_j} f_{X|\Theta}(x|\theta))}{f_{X|\Theta}(X|\theta)^2},$$

we have that

$$E[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{X|\Theta}(X|\theta)] = 0 - \mathrm{Cov}_\theta(\frac{\partial}{\partial\theta_i}\log f_{X|\Theta}(X|\theta), \frac{\partial}{\partial\theta_j}\log f_{X|\Theta}(X|\theta)) = -I_{X,i,j}(\theta).$$

In the case of exponential families using the natural parameter,

$$f_{X|\Theta}(x|\theta) = c(\theta)e^{\langle\theta,T(x)\rangle}.$$

Thus,

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{X|\Theta}(x|\theta) = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\log c(\theta).$$

In the case that $X_1, \cdots, X_n$ are i.i.d given $\Theta = \theta$, then

$$\log f_{X|\Theta}(X|\theta) = \sum_{i=1}^n \log f_{X_i|\Theta}(X_i|\theta).$$

Consequently,

$$\mathcal{I}_X(\theta) = n\mathcal{I}_{X_1}(\theta).$$

Examples.

1. Let $\theta = (\mu, \sigma)$, and let X be $N(\mu, \sigma^2)$. Then,

$$\log f_{X|\Theta}(x|\theta) = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(x-\mu)^2,$$

and

$$\begin{array}{rlr} \frac{\partial}{\partial\mu}\log f_{X|\Theta}(x|\theta) &= \frac{x-\mu}{\sigma^2} & \mathrm{Var}_\theta(\frac{\partial}{\partial\mu}\log f_{X|\Theta}(X|\theta)) = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2} \\ \frac{\partial}{\partial\sigma}\log f_{X|\Theta}(x|\theta) &= -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3} & \mathrm{Var}_\theta(\frac{\partial}{\partial\sigma}\log f_{X|\Theta}(X|\theta)) = \frac{2}{\sigma^2} \\ \frac{\partial}{\partial\sigma}\frac{\partial}{\partial\mu}\log f_{X|\Theta}(x|\theta) &= -2\frac{x-\mu}{\sigma^3} & \mathrm{Cov}_\theta(\frac{\partial}{\partial\mu}\log f_{X|\Theta}(X|\theta), \frac{\partial}{\partial\sigma}\log f_{X|\Theta}(X|\theta)) = 0. \end{array}$$

This give the information matrix

$$\mathcal{I}(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.$$

2. The $\Gamma(\alpha, \beta)$ family is an exponential family whose density with respect to Lebesgue measure is

$$f_{X|A,B}(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{x} \exp(\alpha \log x - \beta x).$$

Thus, $T(x) = (\log x, x)$ is a natural sufficient statistic. $\theta = (\theta_1, \theta_2) = (\alpha, -\beta)$ is the natural parameter. To compute the information matrix, note that

$$\log c(\theta) = \theta_1 \log(-\theta_2) - \log \Gamma(\theta_1),$$

$$\frac{\partial}{\partial \theta_1} \log c(\theta) = \log(-\theta_2) - \frac{\partial}{\partial \theta_1} \log \Gamma(\theta_1), \qquad \frac{\partial}{\partial \theta_2} \log c(\theta) = -\frac{\theta_1}{\theta_2},$$

$$\frac{\partial^2}{\partial \theta_1^2} \log c(\theta) = -\frac{\partial^2}{\partial \theta_1^2} \log \Gamma(\theta_1), \quad \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log c(\theta) = -\frac{1}{\theta_2}, \quad \frac{\partial^2}{\partial \theta_2^2} \log c(\theta) = \frac{\theta_1}{\theta_2^2}.$$

Thus the information matrix $\mathcal{I}_X(\alpha, \beta)$ is

$$\begin{pmatrix} \frac{\partial^2}{\partial \alpha^2} \log \Gamma(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}$$

Theorem. Let $Y = g(X)$ and suppose that $P_\theta << \nu_X$ for all θ. Then $\mathcal{I}_X(\theta) - \mathcal{I}_Y(\theta)$ is positive semidefinite. This difference of matrices is the 0 matrix if and only if Y is sufficient.

Proof. Define $Q_\theta(C) = P'_\theta\{(X,Y) \in C\}$ and $\nu(C) = \nu_X\{x : (x, g(s)) \in C\}$. Note that

$$\int h(x,y)\ \nu(dx \times dy) = \int h(x, g(x))\ \nu_X(dx).$$

Thus,

$$Q_\theta(C) = \int I_C(x, g(x)) f_{X|\Theta}(x|\theta)\ \nu_X(dx) = \int I_C(x,y) f_{X|\Theta}(x|\theta)\ \nu(dx \times dy),$$

and, consequently, $Q_\theta << \nu$ with Radon-Nikodym derivative $f_{X,Y|\Theta}(x,y|\theta) = f_{X|\Theta}(x|\theta)$. Because,

$$f_{X|Y,\Theta}(x|y,\theta) = \frac{f_{X,Y|\Theta}(x,y|\theta)}{f_{Y|\Theta}(y|\theta)},$$

we have

$$f_{X,Y|\Theta}(x,y|\theta) = f_{X|\Theta}(x|\theta) = f_{Y|\Theta}(y|\theta) f_{X|Y,\Theta}(x|y,\theta).$$

or

$$\frac{\partial}{\partial \theta_i} \log f_{X|\Theta}(x|\theta) = \frac{\partial}{\partial \theta_i} \log f_{Y|\Theta}(y|\theta) + \frac{\partial}{\partial \theta_i} \log f_{X|Y,\Theta}(x|y,\theta).$$

a.s. Q_θ for all θ.

Claim The two terms on the right are uncorrelated.

$$\begin{aligned}
& \mathrm{Cov}_\theta(\frac{\partial}{\partial\theta_i}\log f_{Y|\Theta}(Y|\theta), \frac{\partial}{\partial\theta_j}\log f_{X|Y,\Theta}(X|Y,\theta)) \\
= \quad & E_\theta[\frac{\partial}{\partial\theta_i}\log f_{Y|\Theta}(Y|\theta)\frac{\partial}{\partial\theta_j}\log f_{X|Y,\Theta}(X|Y,\theta)] \\
= \quad & E_\theta[E_\theta[\frac{\partial}{\partial\theta_i}\log f_{Y|\Theta}(Y|\theta)\frac{\partial}{\partial\theta_j}\log f_{X|Y,\Theta}(X|Y,\theta)|Y]] \\
= \quad & E_\theta[\frac{\partial}{\partial\theta_i}\log f_{Y|\Theta}(Y|\theta)E_\theta[\frac{\partial}{\partial\theta_j}\log f_{X|Y,\Theta}(X|Y,\theta)|Y]] \\
= \quad & E_\theta[\frac{\partial}{\partial\theta_i}\log f_{Y|\Theta}(Y|\theta)(0)] = 0.
\end{aligned}$$

Thus, by the conditional variance formula

$$\mathcal{I}_X(\theta) = \mathcal{I}_X(\theta) + E_\theta\mathcal{I}_{X,Y}(\theta|Y).$$

Thus, $\mathcal{I}_X(\theta) - \mathcal{I}_X(\theta)$ is positive semidefinite. The last term is zero if and only if the conditional score function is zero a.s. Q_θ. This happens if and only if $f_{X|Y,\Theta}(x|y,\theta)$ is constant in θ, i.e. Y is sufficient.

Let $H = h(\Theta)$ be a one-to-one reparameterization, and let $\mathcal{I}_X^*$ be the Fisher information matrix with respect to this new parameterization. Then, by the chain rule,

$$\mathcal{I}_X^*(\eta) = \Delta(\eta)\mathcal{I}_X(h^{-1}(\eta))\Delta(\eta)^T,$$

where $\Delta(\eta)$ is a matrix with ij-th entry $\partial h_j^{-1}(\eta)/\partial\eta_i$.

4.2 Kullback-Leibler Information

Definition. Let P and Q be probability measures on the same space. Let p and q be their respective densities with respect to some measure ν. The *Kullback-Leibler information* in X is defined as

$$\mathcal{I}_X(P;Q) = \int \log\frac{p(x)}{q(x)}p(x)\ \nu(dx).$$

In the case of parametric families, let $\theta, \psi \in \Omega$ The Kullback-Leibler information is then

$$\mathcal{I}_X(\theta;\psi) = E_\theta[\log\frac{f_{X|\Theta}(X|\theta)}{f_{X|\Theta}(X|\psi)}].$$

For a statistic T, let p_t and q_t denote the conditional densities for P and Q given $T = t$ with respect to some measure ν_t. Then the *conditional Kullback-Leibler information* is

$$\mathcal{I}_{X|T}(P;Q) = \int \log\frac{p_t(x)}{q_t(x)}p_t(x)\ \nu_t(dx).$$

Examples.

1. If X is $N(\theta, 1)$, then

   $$\log \frac{f_{X|\Theta}(x|\theta)}{f_{X|\Theta}(x|\psi)} = \frac{1}{2}((x-\psi)^2 - (x-\theta)^2) = \frac{1}{2}(\theta-\psi)(2x-\psi-\theta).$$

   Thus $\mathcal{I}_X(\theta;\psi) = \mathcal{I}_X(\psi;\theta) = (\theta-\psi)^2$.

2. If X is $Ber(\theta)$, then

   $$\log \frac{f_{X|\Theta}(x|\theta)}{f_{X|\Theta}(x|\psi)} = x\log\frac{\theta}{\psi} + (1-x)\log\frac{1-\theta}{1-\psi}.$$

   Thus,

   $$\mathcal{I}_X(\theta;\psi) = \theta\log\frac{\theta}{\psi} + (1-\theta)\log\frac{1-\theta}{1-\psi}.$$

   Here, $\mathcal{I}_X(\theta;\psi) \neq \mathcal{I}_X(\psi;\theta)$.

Some properties of Kullback-Leibler information are readily verifiable.

1. From Jensen's inequality, $\mathcal{I}_X(P;Q) \geq 0$ and equals 0 if and only if $P = Q$.
2. $\mathcal{I}_{X|T}(P;Q) \geq 0$ a.s. P_T and equals 0 if and only if $p_t = q_t$ a.s. P.
3. If X and Y are conditionally indepedent given Θ, then $\mathcal{I}_{X,Y}(\theta;\psi) = \mathcal{I}_X(\theta;\psi) + \mathcal{I}_Y(\theta;\psi)$

We have the following theorem in analogy to Fisher information.

Theorem. Let $Y = g(X)$, then $\mathcal{I}_X(\theta;\psi) \geq \mathcal{I}_Y(\theta;\psi)$ with equality if and only if Y is sufficient.

Proof. Using the same setup as before, we have

$$\begin{aligned}
\mathcal{I}_X(\theta;\psi) &= E_\theta[\log \frac{f_{X|\Theta}(X|\theta)}{f_{X|\Theta}(X|\psi)}] \\
&= E_\theta[\log \frac{f_{Y|\Theta}(Y|\theta)}{f_{Y|\Theta}(Y|\psi)}] + E_\theta[\log \frac{f_{X|Y,\Theta}(X|Y,\theta)}{f_{X|Y,\Theta}(X|Y,\psi)}] \\
&= \mathcal{I}_Y(\theta;\psi) + E_\theta[\mathcal{I}_{X|Y}(\theta;\psi|Y)] \geq \mathcal{I}_Y(\theta;\psi).
\end{aligned}$$

To obtain equality, we use Jensen's inequality to conclude that

$$f_{X|Y,\Theta}(X|Y,\theta) = f_{X|Y,\Theta}(X|Y,\psi), \text{ a.s. } P_\theta$$

.

Using the same ideas, we have the following.

Theorem. Both Fisher and Kullback-Leibler information is the mean of the conditional information given an ancillary statistic U.

Proof. Let P_θ have density $f_{X|\Theta}$ with respect to a measure ν. For $u = U(x)$, we can write

$$f_{X|\Theta}(x|\theta) = f_U(u) f_{X|U,\Theta}(x|u,\theta)$$

because U does not depend on Θ.

If the FI regularity conditions hold,

$$\frac{\partial}{\partial\theta_i} f_{X|\Theta}(x|\theta) = \frac{\partial}{\partial\theta_i} f_{X|U,\Theta}(x|u,\theta).$$

Because the mean of the conditional score function is 0 a.s.,

$$\mathcal{I}_X(\theta) = E_\theta \mathcal{I}_{X|U}(\theta|U).$$

Similarly, for the Kullback-Leibler information,

$$\frac{f_{X|\Theta}(x|\theta)}{f_{X|\Theta}(x|\psi)} = \frac{f_{X|U,\Theta}(x|u,\theta)}{f_{X|U,\Theta}(x|u,\psi)}$$

and

$$\mathcal{I}_X(\theta;\psi) = E\mathcal{I}_{X|U}(\theta;\psi|U).$$

Some advantages of the Kullback-Leibler information are

1. It is not affected by the parameterization.
2. It has no smoothness conditions.

We have the following connection under the appropriate regularity conditions.

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathcal{I}_X(\theta_0;\theta)|_{\theta=\theta_0} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int \log\frac{f_{X|\Theta}(X|\theta_0)}{f_{X|\Theta}(X|\theta)} f_{X|\Theta}(X|\theta_0)\ \nu(dx)|_{\theta=\theta_0}$$

$$= -\int \frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{X|\Theta}(X|\theta_0) f_{X|\Theta}(X|\theta_0)\ \nu(dx) = \mathcal{I}_{X,i,j}$$

Examples.

1. For X a $Ber(\theta)$ random variable, and $\delta > 0$.

$$\frac{\partial^2}{\partial\psi}\mathcal{I}_X(\theta;\psi)|_{\theta=\psi} = (\frac{\theta}{\psi^2} + \frac{1-\theta}{(1-\psi)^2})|_{\theta=\psi} = \frac{1}{\theta(1-\theta)} = \mathcal{I}_X(\theta).$$

2. Let X be $U(0,\theta)$, then

$$\begin{aligned}
\mathcal{I}_X(\theta,\theta+\delta) &= \int_0^\theta \log(\frac{\theta+\delta}{\theta})\frac{1}{\theta}\,dx = \log(1+\frac{\delta}{\theta}) \\
\mathcal{I}_X(\theta+\delta.\theta) &= \int_0^\theta \log(\frac{\theta}{\theta+\delta})\frac{1}{\theta+\delta}\,dx + \int_\theta^{\theta+\delta} \infty\frac{1}{\theta+\delta}\,dx = \infty.
\end{aligned}$$

In words, from observations of a $U(0,\theta)$ distribution, we have some information to distinguish it from a $U(0,\theta+\delta)$ distribution. On the other hand, any observation of $X > \theta$ eliminates the $U(0,\theta)$ distribution. This is infinite information in the Kullback-Leibler sense.

Example. This example shows how the Kullback-Leibler information appears in the theory of large deviations.

Let $X_1,\cdots,X_n$ be independent $Ber(\psi)$ random variables. Choose $0<\psi<\theta<1$. Then, by Chebyshev's inequality, we have for $\alpha>0$,

$$P_\psi\{\frac{1}{n}\sum_{i=1}^n X_i \geq \theta\} \leq \frac{E[e^{\alpha\sum_{i=1}^n X_i}]}{e^{\alpha n\theta}} = \frac{(\psi e^\alpha + (1-\psi))^n}{e^{\alpha n\theta}}.$$

or

$$\frac{1}{n}\log P_\psi\{\frac{1}{n}\sum_{i=1} nX_i \geq \theta\} \leq \log(\psi e^\alpha + (1-\psi)) - \alpha\theta$$

The right side has a minimum value when α satisfies

$$\frac{\psi e^\alpha}{\psi e^\alpha + (1-\psi)} = \theta,$$

$$e^\alpha = \frac{(1-\psi)\theta}{\psi(1-\theta)} \quad \text{or} \quad \alpha = \log\frac{(1-\psi)\theta}{\psi(1-\theta)}.$$

Thus, this minimum value is

$$\begin{aligned}
\log(\frac{(1-\psi)\theta}{1-\theta} + (1-\psi)) - \theta\log\frac{(1-\psi)\theta}{\psi(1-\theta)} &= \log(\frac{(1-\psi)}{1-\theta}) - \theta\log\frac{(1-\psi)\theta}{\psi(1-\theta)} \\
&= -(\theta\log\frac{\theta}{\psi} + (1-\theta)\log\frac{1-\theta}{1-\psi}) \\
&= -\mathcal{I}_X(\psi:\theta)
\end{aligned}$$

In summary,

$$P\{\frac{1}{n}\sum_{i=1}^n X_i \geq \theta\} \leq \exp(-n\mathcal{I}_X(\psi;\theta)).$$

In words, The probability that a ψ coin can perform better than a θ coin is exponential small with power in the exponent equal to negative the number of coin tosses times the Kullback-Leibler information.

5 Statistical Decision Theory

5.1 Framework

A *statistical decision* is an *action* that we take after we make an observation $X : S \to \mathcal{X}$ from $P_\theta \in \mathcal{P}$. Call A the set of allowable actions or the *action space*. Assume that A is a measurable space with σ-field $\mathcal{F}$.

A *randomized decision rule* is a mapping from the sample space to probability measures on the action space $\delta : \mathcal{X} \to \mathcal{P}(A, \mathcal{F})$ so that $\delta(\cdot)(B)$ is measurable for every $B \in \mathcal{F}$. In other words, δ is a regular conditional distribution on A given X. A *nonrandomized decision rule* is one in which δ is a point mass. Denote this mapping by $\delta : \mathcal{X} \to A$.

Let $V : S \to \mathcal{V}$ be measurable. The criterion for assessing a nonrandomized decision rule is a loss function $L : \mathcal{V} \times A \to R$. For a randomized decision rule, we use

$$L(v, \delta(x)) = \int_A L(v, a)\ \delta(x)(da).$$

Example. Let n be an even integer and let $X_1, \cdots, X_n$ be independent $Ber(\theta)$ random variables. Let the parameter space $\Omega = \{1/3, 2/3\}$ and the action space $A = \{a_0, a_1\}$. Set $V = \Theta$ and take the loss function

$$L(v, a) = \begin{cases} 0 & \text{if } (v = \frac{1}{3} \text{ and } a = a_0) \text{ or } (v = \frac{2}{3} \text{ and } a = a_1), \\ 1 & \text{otherwise.} \end{cases}$$

and randomized decision rule

$$\delta(x) = \begin{cases} \text{probability 1 on } a_0 & \text{if } \sum_{i=1} x_i < \frac{n}{2}, \\ \text{probability 1 on } a_1 & \text{if } \sum_{i=1} x_i > \frac{n}{2}, \\ \text{probability } \frac{1}{2} \text{ on each} & \text{if } \sum_{i=1} x_i = \frac{n}{2}. \end{cases}$$

5.2 Classical Decision Theory

Define the *risk function* as the mean of the loss function.

$$R(\theta, \delta) = E_\theta[L(V, \delta(X))] = \int_{\mathcal{X}} \int_{\mathcal{V}} L(v, \delta(x))\ P_{\theta,V}(dv) P_\theta(dx).$$

where $P_{\theta,V}(D) = P_\theta\{V \in D\}$.

The most common choices for V is Θ. In this case,

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x))\ P_\theta(dx).$$

This suggests that we define

$$L(\theta, a) = \int_{\mathcal{V}} L(v, a)\ P_{\theta,V}(dv).$$

Then, the formula for R holds for all choices of V.

Exercise. Here is a decision theoretic basic for the standard measures of center. Let δ be a nonrandomized decision function with finite range.

1. If $L(\theta, \delta) = I_{\{\delta(x) \neq \phi_0(\theta)\}}$, take $\phi_0(\theta)$ to be the *mode* of $\delta(X)$ to minimize risk.
2. If $L(\theta, \delta) = |\delta(x) - \phi_1(\theta)|$, take $\phi_1(\theta)$ to be the *median* of $\delta(X)$ to minimize risk.
3. If $L(\theta, \delta) = (\delta(x) - \phi_2(\theta))^2$, take $\phi_2(\theta)$ to be the *mean* to minimize risk. This minimum risk is $\mathrm{Var}(\delta(X))$.

A decision rule with small loss is preferred. If a decision rule δ_* in a class of allowable decision rules $\mathcal{D}$ minimzes risk for any choice of $\theta \in \Omega$, then we say that δ_* is $\mathcal{D}$-optimal.

Sufficient statistics play an important role in classical decision theory.

Theorem. Let δ_0 be a randomized decision rule and let T be a sufficient statistic. Then, there exists a rule δ_1 that is a function of the sufficient statistic and has the same risk function.

Proof. For $C \in \mathcal{F}$, define

$$\delta_1(t)(C) = E_\theta[\delta_0(X)(C)|T(X) = t].$$

Because T is sufficient, the expectation does not depend on θ. By the standard machine, for any integrable $h : A \to R$,

$$E_\theta[\int h(a)\ \delta_0(X)(da)|T = t] = \int h(a)\ \delta_1(t)(da).$$

Taking $h(a) = L(\theta, a)$, we have

$$\begin{aligned} R(\theta, \delta_0) &= E_\theta[L(\theta, \delta_0(X))] \\ &= E_\theta[E_\theta[L(\theta, \delta_0(X))|T(X)]] \\ &= E_\theta[L(\theta, \delta_1(T(X)))] = R(\theta, \delta_1) \end{aligned}$$

Note that even if δ_0 is nonrandomized, then δ_1 will be randomized if T is not one-to-one.

Theorem. Suppose that $A \subset R^m$ is convex and that for all $\theta \in \Omega$, $L(\theta, a)$ is a convex function of a. Let δ be a randomized rule and set

$$F = \{x \in \mathcal{X} : \int_A |a|\ \delta(x)(da) < \infty\}.$$

Consider the nonrandomized decision rule

$$\delta_0(x) = \int_A a\ \delta(x)(da),$$

for $x \in F$. Then $L(\theta, \delta_0(x)) \leq L(\theta, \delta(x))$ for all $x \in F$ and $\theta \in \Omega$.

Proof. Because A is convex, $\delta_0(x) \in A$ for all $x \in F$. By Jensen's inequality,

$$L(\theta, \delta_0(x)) = L(\theta, \int_A a\ \delta(x)(da)) \leq \int_A L(\theta, a)\ \delta(x)(da) = L(\theta, \delta(x)).$$

Thus, if $F = \mathcal{X}$, the nonrandomized rule obtain from averaging the randomized rule cannot have larger loss.

Theorem. (Rao-Blackwell) Suppose that $A \subset R^m$ is convex and that for all $\theta \in \Omega$, $L(\theta, a)$ is a convex function of a. Suppose that T us sufficient and δ_0 is a nonrandomized decision rule with $E_\theta[||\delta_0(X)||] < \infty$. Define

$$\delta_1(t) = E_\theta[\delta_0(X)|T = t],$$

Then, for all θ,

$$R(\theta, \delta_1) \leq R(\theta, \delta_0).$$

Proof. Using the conditional form of Jensen's inequality, we have

$$\begin{aligned} R(\theta, \delta_0) &= E_\theta[L(\theta, \delta_0(X))] = E_\theta[E_\theta[L(\theta, \delta_0(X))|T(X)]] \\ &\geq E_\theta[L(\theta, E_\theta[\delta_0(X)|T(X)])] = E_\theta[L(\theta, \delta_1(X))] = R(\theta, \delta_1) \end{aligned}$$

Example. Let $X = (X_1, \cdots, X_n)$ be independent $N(\theta, 1)$ random variables. Set $A = [0, 1]$ and fix $c \in R$. For loss function

$$L(\theta, a) = (a - \Phi(c - \theta))^2,$$

a naïve decision rule is

$$\delta_0(X) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, c]}(X_i).$$

However, $T(X) = \bar{X}$ is sufficient and δ_0 is not a function of $T(X)$.

As the Rao-Blackwell theorem suggests, we compute

$$E_\theta[\delta_0(X)|T(X)] = \frac{1}{n} \sum_{i=1}^{n} E_\theta[I_{(-\infty, c]}(X_i)|T(X)] = P_\theta\{X_1 \leq x|T(X)\} = \Phi(\frac{c - T(X)}{\sqrt{(n-1)/n}})$$

because X_1 given $T(X) = t$ is $N(t, (n-1)/n)$

5.3 Bayesian Decision Theory

In the Bayesian paradigm, we might begin by computing the *posterior risk*

$$r(\delta|x) = \int_{\mathcal{V}} L(v, \delta(x))\ \mu_{V|X}(dv|x).$$

A rule, δ_0, is called a *formal Bayes rule* if,

1. for every x, the risk $r(\delta_0|x)$ is finite, and
2. $r(\delta_0|x) \leq r(\delta|x)$ for every rule δ.

Example. Let $A = \Omega$ and $L(\theta, a) = (\theta - a)^2$. Then

$$\int_\Omega L(\theta, a)\mu_{\Theta|X}(d\theta|x) = a^2 + 2aE[\Theta|X = x] + E[\Theta^2|X = x].$$

If Θ has finite variance, then $\delta_0(x) = E[\Theta|X = x]$ minimizes the expression above and thus is a formal Bayes rule.

If no decision rule exists that minimizes risk for all θ, one alternative is to choose a probability measure η on Ω and minimize the *Bayes risk*,

$$r(\eta, \delta) = \int_\Omega R(\theta, \delta)\ \eta(d\theta).$$

Each δ that minimizes $r(\eta, \delta)$ is called a *Bayes rule.* If the measure η has infinite mass, then a rule that minimizes the integral above is called a *generalized Bayes rule* and η is called an improper prior.

If the loss is nonnegative and if $P_\theta << \nu$ for all θ, with density $f_{X|\Theta}$, then by Tonelli's theorem,

$$\begin{aligned} r(\eta, \delta) &= \int_\Omega R(\theta, \delta)\ \eta(d\theta) \\ &= \int_\Omega \int_{\mathcal{X}} L(\theta, \delta(x)) f_{X|\Theta}(x|\theta)\ \nu(dx)\eta(d\theta) \\ &= \int_{\mathcal{X}} \int_\Omega L(\theta, \delta(x)) f_{X|\Theta}(x|\theta)\ \eta(d\theta)\nu(dx) \\ &= \int_{\mathcal{X}} r(\delta|x)\ \mu_X(dx), \end{aligned}$$

where $\mu_X(B) = \int_\Omega P_\theta(B)\ \eta(d\theta)$ the the marginal for X. In this circumstance, Bayes rules and formal Bayes rules are the same a.s. μ_X.

5.4 Admissibility

The previous results give us circumstances in which one decision rule is at least as good as another. This leads to the following definition.

Definition. A decision rule δ is *inadmissible* if there exits a decision rule δ_0 such that $R(\theta, \delta_0) \leq R(\theta, \delta)$ with strict inequality for at least one value of θ. The decision δ_0 is said to *dominate* δ.

If no rule dominates δ, we say that δ is *admissible.*

Let λ be a measure on (Ω, τ) and let δ be a decision rule. If

$$R(\theta, \delta_0) \leq R(\theta, \delta) \quad \text{a.e. } \lambda \text{ implies } R(\theta, \delta_0) = R(\theta, \delta) \quad \text{a.e. } \lambda.$$

Then δ is λ-admissible.

Theorem. Suppose that λ is a probability and that δ is a Bayes rule with respect to λ. Then δ is λ-admissible.

Proof. Let δ be a Bayes rule with respect to λ and let δ_0 be a decision rule. Then

$$\int_\Omega (R(\theta,\delta) - R(\theta,\delta_0))\ \lambda(d\theta) \leq 0.$$

If $R(\theta,\delta_0) \leq R(\theta,\delta)$ a.s. λ, then the integrand is nonnegative a.s. λ. Because the integral is nonpositive, the integrand is 0 a.s. λ and δ is λ-admissible.

A variety of results apply restrictions on λ so that λ-admissibility implies admissibility.

Theorem. Let Ω be discrete. If a probability λ has $\lambda\{\theta\} > 0$ for all $\theta \in \Omega$, and if δ is a Bayes rule with respect to λ, then δ is admissible.

Proof. Suppose that δ_0 dominates δ. Then, $R(\theta,\delta_0) \leq R(\theta,\delta)$ for all θ with strict inequality for at least one value of θ. Consequently,

$$r(\lambda,\delta_0) = \sum_\theta \lambda\{\theta\}R(\theta,\delta_0) < \sum_\theta \lambda\{\theta\}R(\theta,\delta) = r(\lambda,\delta),$$

and δ is not a Bayes rule.

The following are examples of sufficient conditions for admissibility.

1. Every Bayes rule with respect to a prior λ has the same risk function. In particular, a unique Bayes rule is admissible.

2. The parameter set $\Omega \subset R^k$ is contained in the closure of its interior, λ is absolutely continuous with respect to Lebesgue measure. For all δ having finite risk, $R(\theta,\delta)$ is continuous in θ. δ_0 is λ-admissible with finite risk.

3. A is a convex subset of R^m, $\{P_\theta : \theta \in \Omega\}$ are mutually absolutely continuous, $L(\theta,a)$ is strictly convex in a for all δ, and δ_0 is λ-admissible.

Examples.

1. Let $X_1,\cdots,X_n$ be independent $U(0,\theta)$ and set $Y = \max\{X_1,\cdots,X_n\}$. Choose loss function $L(\theta,a) = (\theta - a)^2$. Then

$$R(\theta,\delta) = \frac{n}{\theta^n}\int_0^\theta (\theta - \delta(y))^2 y^{n-1}\ dy = \theta^2 - \frac{2n}{\theta^{n-1}}\int_0^\theta \delta(y)y^{n-1}\ dy + \frac{n}{\theta^n}\int_0^\theta \delta(y)^2 y^{n-1}\ dy.$$

Choose a rule δ with finite risk function. Then $R(\cdot,\delta)$ is continuous. Let $\lambda \in \mathcal{P}(0,\infty)$ have strictly positive density $\ell(\theta)$ with respect to Lebesgue measure. Then the formal Bayes rule with respect to λ is admissible.

2. Let $X_1, \cdots, X_n$ be independent $Ber(\theta)$. The action space $A = [0,1]$. The lost is $L(\theta, a) = (\theta - a)^2/(\theta(1-\theta))$. Define $Y = \sum_{i=1}^n X_i$ and select Lebesgue measure to be the prior on $[0,1]$. Then the posterior, given $X = x$ is $Beta(y+1, n-y+1)$ where $y = \sum_{i=1}^n x_i$. Then

$$E[L(\Theta, a)|X = x] = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \int_0^1 (\theta - a)^2 \theta^{y-1}(1-\theta)^{n-y-1}\, d\theta.$$

This has a minimum at $a = y/n$, for all x and all n. Consequently, $\delta(x) = \sum_{i=1}^n x_i/n$ is a Bayes rule with respect to Lebesgue measure and hence is admissible.

3. Let X have an exponential family of distributions with natural parameter $\Omega \subset R$. Let $A = \Omega$ and $L(a, \theta) = (\theta - a)^2$. Note that all P_θ are mutually absolutely continuous. Take the prior λ to be point mass at $\theta_0 \in \Omega$. Then $\delta_{\theta_0}(c) = \theta_0$ is the formal Bayes rule with respect to λ and so is λ-admissible. By the theorem above, it is also admissible.

4. Returning to Example 1, note that the P_θ are not mutually absolutely continuous and that $\delta_{\theta_0}(c) = \theta_0$ is not admissible. To verify this last statement, take $\delta'_{\theta_0} = \max\{y, \theta_0\}$. Then

$$R(\theta, \delta_{\theta_0}) < R(\theta, \delta'_{\theta_0}) \text{ for all } \theta > \theta_0 \quad \text{and} \quad R(\theta, \delta_{\theta_0}) = R(\theta, \delta'_{\theta_0}) \text{ for all } \theta \leq \theta_0.$$

We also have some results that allow us to use admissibility in one circumstance to imply admissibility in other circumstances.

Theorem. Suppose that $\Theta = (\Theta_1, \Theta_2)$. For each choice of $\tilde{\theta}_2$ for Θ_2, define $\Omega_{\tilde{\theta}_2} = \{(\theta_1, \theta_2) : \theta_2 = \tilde{\theta}_2\}$. Assume, for each $\tilde{\theta}_2$, δ is admissible on $\Omega_{\tilde{\theta}_2}$, then it is admissible on Ω.

Proof. Suppose that δ is inadmissible on Ω. Then there exist δ_0 such that

$$R(\theta, \delta_0) \leq R(\theta, \delta) \text{ for all } \theta \in \Omega \text{ and } R(\tilde{\theta}, \delta_0) < R(\tilde{\theta}, \delta) \text{ for some } \tilde{\theta} \in \Omega.$$

Writing $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$ yields a contradiction to the admissibility of δ on $\Omega_{\tilde{\theta}_2}$.

Theorem. Let $\Omega \subset R^k$ be open and assume that $c(\theta) > 0$ for all θ and d is a real vauled function of θ. Then δ is admissible with loss $L(\theta, a)$ if and only if δ is admissible with loss $c(\theta)L(\theta, a) + d(\theta)$.

Example. Returning to the example above for Bernoulli random variables, the theorem above tells us that $\delta(x) = \sum_{i=1}^n x_i/n$ is admissible for a quadratic loss function.

Theorem. Let δ be a decision rule. Let $\{\lambda_n : n \geq 1\}$ be a sequence of measures on Ω such that a generalized Bayes rule δ_n with respect to λ_n exists for every n with

$$r(\lambda_n, \delta_n) = \int_\Omega R(\theta, \delta_n)\ \lambda_n(d\theta), \quad \lim_{n\to\infty} r(\lambda_n, \delta) - r(\lambda_n, \delta_n) = 0.$$

Furthermore, assume that one of the following conditions hold.

1. $\{P_\theta : \theta \in \Omega\}$ are mutually absolutely continuous; A is convex; $L(\theta, \cdot)$ is strictly convex for all θ, and there exits a constant c, a measurable set C, and a measure λ so that

$$\lambda_n << \lambda, \quad \frac{d\lambda_n}{d\lambda}(\theta) \geq c \text{ for } \theta \in C \text{ with } \lambda(C) > 0.$$

2. Ω is contained in the closure of its interior, for every open set $G \subset \Omega$, there exists $c > 0$ such that $\lambda_n(G) \geq c$ for all n, and the risk function is continuous in θ for every decision rule.

Then δ is admissible.

Proof. We will show that δ inadmissible implies that the limit condition above fails to hold.

With this in mind, choose δ_0 so that $R(\theta, \delta_0) \leq R(\theta, \delta)$ with strict inequality for θ_0.

Using the first condition, set $\tilde{\delta} = (\delta + \delta_0)/2$. Then

$$L(\theta, \tilde{\delta}(x)) < (L(\theta, \delta(x)) + L(\theta, \delta_0(x)))/2.$$

for all θ and all x with $\delta(x) \neq \delta_0(x)$. Because $P'_{\theta_0}\{\tilde{\delta}(X) = \delta(X)\} < 1$ and $\{P_\theta : \theta \in \Omega\}$ are mutually absolutely continuous, we have $P'_\theta\{\tilde{\delta}(X) = \delta(X)\} < 1$ for all θ. Consequently, $R(\theta, \tilde{\delta}) < R(\theta, \delta)$ for all θ. For each n

$$\begin{aligned} r(\lambda_n, \delta) - r(\lambda_n, \delta_n) &\geq r(\lambda_n, \delta) - r(\lambda_n, \tilde{\delta}) \geq \int_C (R(\theta, \delta) - R(\theta, \tilde{\delta}))\ \lambda_n(d\theta) \\ &\geq c \int_C (R(\theta, \delta) - R(\theta, \tilde{\delta}))\ \lambda(d\theta) > 0. \end{aligned}$$

This contradicts the hypothesis.

Using the second condition, there exists $\epsilon > 0$ and an open set $G \subset \Omega$ such that $R(\theta, \delta_0) < R(\theta, \delta) - \epsilon$ for all $\theta \in G$. Note that for each n,

$$r(\lambda_n, \delta) - r(\lambda_n, \delta_n) \geq r(\lambda_n, \delta) - r(\lambda_n, \delta_0) \geq \int_G (R(\theta, \delta) - R(\theta, \delta_0))\ \lambda_n(d\theta) \geq \epsilon \lambda_n(G) \geq \epsilon c,$$

again contradicting the hypothesis.

Example. Let $\theta = (\mu, \sigma)$ and suppose that $X_1, \cdots, X_n$ be independent $N(\mu, \sigma^2)$. Choose the loss function $L(\theta, a) = (\mu - a)^2$.

Claim. $\delta(x) = \bar{x}$ is admissible

$R(\theta, \delta) = \sigma^2$. For each value σ_0, we will show that δ is admissible for the parameter space $\Omega_0 = \{(\mu, \sigma_0) : \mu \in R\}$. Let λ_n be the measure equal to $\sqrt{n}$ times $N(0, \sigma_0^2 n)$. Check that

1. The generalized Bayes rule with respect to λ_n is $\delta_n(x) = nx/(n + 1)$.

2. $r(\lambda_n, \delta_n) = n^{3/2}\sigma_0^2/(n + 1)$

3. $r(\lambda_n, \delta) = n^{1/2}\sigma_0^2$

4. $\lim_{n\to\infty}(n^{1/2}\sigma_0^2 - n^{3/2}\sigma_0^2/(n + 1)) = \lim_{n\to\infty} n^{1/2}\sigma_0^2/(n + 1) = 0.$

5. The densities of λ_n with respect to Lebesgue measure increase for each value of μ. Thus, $\lambda_n(G) > \lambda_1(G)$ for any nonempty open $G \subset R$, and $d\lambda_n/d\lambda_1(\mu) \geq 1$.

Thus, the theorem above applies using condition 1 and $\delta(x) = \bar{x}$ is admissible.

5.5 James-Stein Estimators

Theorem. Consider $X = (X_1, \cdots, X_n)$, n independent random variables with X_i having a $N(\mu_i, 1)$ distribution. Let $A = \Omega = R^n = \{(\mu_1, \cdots, \mu_n) : \mu_i \in R\}$ and let the loss function be $L(\mu, a) = \sum_{i=1}^n (\mu_i - a_i)^2$. Then $\delta(x) = x$ is inadmissible with dominating rule

$$\delta_1(x) = \delta(x)[1 - \frac{n-2}{\sum_{i=1}^n x_i^2}].$$

To motivate this estimator, suppose that M has distribution $N_n(\mu_0, \tau I)$. Then, the Bayes estimate for M is

$$\mu_0 + (X - \mu_0)\frac{\tau^2}{\tau^2 + 1}.$$

The marginal distribution of X is $N_n(\mu_0, (1 + \tau^2)I)$. Thus, we could estimate $1 + \tau^2$ by $\sum_{j=1}^n (X_j - \mu_{0,j})^2/c$ for some choice of c. Consequently, a choice for estimating

$$\frac{\tau^2}{\tau^2 + 1} = 1 - \frac{1}{\tau^2 + 1}$$

is

$$1 - \frac{c}{\sum_{j=1}^n (X_j - \mu_{0,j})^2}.$$

Giving an estimate $\delta_1(x)$ in the case $\mu = 0$ using the choice $c = n - 2$.

The proof requires some lemmas.

Lemma. (Stein) Let $g \in C^1(R, R)$ and let X have a $N(\mu, 1)$ distribution. Assume that $E[|g'(X)|] < \infty$. Then $E[g'(X)] = \mathrm{Cov}(X, g(X))$.

Proof. Let ϕ be the standard normal density function. Because $\phi'(x) = x\phi(x)$, we have

$$\phi(x - \mu) = \int_x^\infty (z - \mu)\phi(z - \mu)\ dz = -\int_{-\infty}^x (z - \mu)\phi(z - \mu)\ dz.$$

Therefore,

$$E[g'(X)] = \int_{-\infty}^\infty g'(x)\phi(x - \mu)\ dx = \int_{-\infty}^0 g'(x)\phi(x - \mu)\ dx + \int_0^\infty g'(x)\phi(x - \mu)\ dx$$

$$\begin{aligned}
&= -\int_{-\infty}^0 g'(x)\int_{-\infty}^x (z - \mu)\phi(z - \mu)\ dzdx + \int_0^\infty g'(x)\int_x^\infty (z - \mu)\phi(z - \mu)\ dzdx \\
&= -\int_{-\infty}^0 (z - \mu)\phi(z - \mu)\int_z^0 g'(x)\ dxdz + \int_0^\infty (z - \mu)\phi(z - \mu)\int_0^z g'(x)\ dxdz \\
&= \int_{-\infty}^0 (z - \mu)\phi(z - \mu)(g(z) - g(0))\ dz + \int_0^\infty (z - \mu)\phi(z - \mu)(g(z) - g(0))\ dz \\
&= \int_{-\infty}^\infty (z - \mu)\phi(z - \mu)g(z)\ dz = \mathrm{Cov}(X, g(X)).
\end{aligned}$$

Lemma. Let $g \in C^1(R^n, R^n)$ and let X have a $N_n(\mu, I)$ distribution. For each i, define

$$h_i(y) = E_\mu[g_i(X_1, \cdots, X_{i-1}, y, X_{i+1} \cdots, X_n)]$$

and assume that

$$h_i'(y) = E_\mu[\frac{d}{dy} g_i(X_1, \cdots, X_{i-1}, y, X_{i+1} \cdots, X_n)]$$

and that $E_\mu[h'(X_i)|] < \infty$. Then

$$E_\mu||X + g(X) - \mu||^2 = n + E_\mu[||g(X)||^2 + 2\nabla \cdot g(X)].$$

Proof.

$$\begin{aligned} E_\mu||X + g(X) - \mu||^2 &= E_\mu||X - \mu||^2 + E_\mu[||g(X)||^2] + 2E_\mu\langle (X - \mu), g(X)\rangle \\ &= n + E_\mu[||g(X)||^2] + 2E_\mu \sum_{i=1}^{n} ((X_i - \mu_i) g_i(X)). \end{aligned}$$

Note that, by the previous lemma

$$E_\mu[(X_i - \mu_i) g_i(X)] = E_\mu[(X_i - \mu_i) h_i(X_i)] = \mathrm{Cov}_\mu(X_i, h_i(X_i)) = E_\mu[h_i'(X_i)] = E_\mu[\frac{\partial}{\partial x_i} g_i(X)].$$

Proof. (Jones-Stein estimator) Set

$$g_i(x) = -\frac{x_i(n-2)}{\sum_{j=1} n x_j^2}.$$

To see that the lemma above applies, note that, for $x \neq 0$, $\partial^2 g / \partial x_i^2$ is bounded in a neighborhood of x. Moreover,

$$\begin{aligned} E_\mu[|h_i'(X_i)|] &\leq (n-2) \int_{R^n} \frac{|\sum_{j=1}^n x_j^2 - 2x_i^2|}{(\sum_{j=1}^n x_j^2)^2} f_{X|M}(x|\mu)\ dx \\ &\leq (n-2) \int_{R^n} \frac{3}{\sum_{j=1}^n x_j^2} f_{X|M}(x|\mu)\ dx = (n-2)\frac{3n}{n-2} = 3n \end{aligned}$$

Thus, the risk function

$$\begin{aligned} R(\mu, \delta_1) &= n + E_\mu[||g(X)|| + 2\nabla \cdot g(X)] \\ &= n + (n-2)^2 E_\mu[\frac{\sum_{i=1}^n X_i^2}{(\sum_{j=1}^n X_j^2)^2} - 2\frac{\sum_{i=1}^n X_i^2}{(\sum_{j=1}^n X_j^2)^2}] \\ &= n - (n-2)^2 E_\mu[\frac{1}{\sum_{j=1}^n X_j^2}] < n = R(\mu, \delta_0) \end{aligned}$$

remark. The Jones-Stein estimator is also not admissible. It is dominated by

$$\delta^+(x) = x(1 - \min\{1, \frac{n-2}{(\sum_{j=1}^n x_j^2)^2}\}.$$

5.6 Minimax Procedures

Definition A rule δ_0 is called *minimax* if,

$$\sup_{\theta\in\Omega} R(\theta, \delta_0) = \inf_{\delta} \sup_{\theta\in\Omega} R(\theta, \delta)$$

Proposition. If δ has constant risk and is admissible, then it is minimax.

Theorem. Let δ_0 be the Bayes rule for λ. If $\lambda(\Omega_0) = 1$, where

$$\Omega_0 = \{\theta : R(\theta, \delta_0) = \sup_{\psi\in\Omega} R(\psi, \delta_0)\} = 1,$$

then δ_0 is minimax.

Proof. Let δ be any other rule, then

$$\sup_{\psi\in\Omega} R(\psi, \delta_0) = \int_\Omega R(\theta, \delta_0) I_{\Omega_0}\ \lambda(d\theta) \le \int_\Omega R(\theta, \delta) I_{\Omega_0}\ \lambda(d\theta) \le \sup_{\psi\in\Omega} R(\psi, \delta).$$

Note that if δ_0 is the unique Bayes rule, then the second inequality in the line above should be replaced by strict inequality and, therefore, δ_0 is the unique minimax rule.

Example.

1. For X a $N(\mu, \sigma^2)$ random variable, $\delta(x)$ is admissible with loss $L(\theta, a) = (\mu - a)^2$ and hence loss $\tilde{L}(\theta, a) = (\mu - a)^2/\sigma^2$. The risk function for $\tilde{L}$ is constant $\tilde{R}(\theta, \delta) = 1$ and δ is minimax.

2. For $X_1, \cdots, X_n$ are independent $Ber(\theta)$. The Bayes rule with respect to a $Beta(\alpha, \beta)$ is

   $$\delta(x) = \frac{\alpha + \sum_{i=1}^n x_i}{\alpha + \beta + n},$$

   with risk

   $$R(\theta, \delta) = \frac{n\theta(1-\theta) + (\alpha - \alpha\theta - \beta\theta)^2}{(\alpha + \beta + n)^2}.$$

   Check that $R(\theta, \delta)$ is constant if and only if $\alpha = \beta = \sqrt{n}/2$ which leads to the unique minimax estimator

   $$\delta_0(x) = \frac{\sqrt{n}/2 + \sum_{i=1}^n x_i}{\sqrt{n} + n}.$$

   The risk $R(\theta, \delta_0) = 1/(4(1 + \sqrt{n})^2)$.

Theorem. Let $\{\lambda_n : n \geq 1\}$ be a sequence of probability measures on Ω and let δ_n be the Bayes rule with respect to λ_n. Suppose that $\lim_{n\to\infty} r(\lambda_n, \delta_n) = c < \infty$. If a rule δ_0 satisfies $R(\theta, \delta_0) \leq c$ for all θ, then δ_0 is minimax.

Proof. Assume that δ_0 is not minimax. Then we can find a rule δ and a number $\epsilon > 0$ such that

$$\sup_{\psi} R(\psi, \delta) \leq \sup_{\psi\in\Omega} R(\psi, \delta_0) - \epsilon \leq c - \epsilon.$$

Choose N_0 so that $r(\lambda_n, \delta_n) \geq c - \epsilon/2$ for all $n \geq N_0$. Then, for such n,

$$r(\lambda_n, \delta) = \int R(\theta, \delta)\ \lambda_n(d\theta) \leq (c - \epsilon) \int \lambda_n(d\theta) < c - \epsilon/2 \leq r(\lambda_n, \delta_n)$$

and thus δ_n is not a Bayes rule with respect to λ_n.

Example. Let the independent observations X_i be $N(\mu_i, 1)$, $i = 1, \cdots, m$. Set $\delta(x) = x$, $A = R^m$, and $L(\mu, a) = \sum_{i=1}^n (\mu_i - a_i)^2$. $\delta_n(x) = nx/(n + 1)$ is a Bayes rule for λ_n, the law of a $N_m(0, nI)$ with Bayes risk $r(\lambda_n, \delta_n) = mn/(n + 1)$. Thus,

$$\lim_{n\to\infty} r(\lambda_n, \delta_n) = m = R(\mu, \delta).$$

Thus, δ, which is not admissible, is minimax.

A prior distribution λ_0 for Θ is called *least favorable* if

$$\inf_{\delta} r(\lambda_0, \delta) = \sup_{\lambda} \inf_{\delta} r(\lambda, \delta).$$

Theorem. If δ_0 is a Bayes rule with respect to λ_0 and $R(\theta, \delta_0) \leq r(\lambda_0, \delta_0)$ for all θ, then δ_0 is minimax and λ_0 is least favoriable.

Proof. For any rule $\tilde{\delta}$ and prior $\tilde{\lambda}$, $\inf_\delta r(\tilde{\lambda}, \delta) \leq r(\tilde{\lambda}, \tilde{\delta}) \leq \sup_\lambda r(\lambda, \tilde{\delta})$. Thus,

$$\begin{aligned}
\inf_{\delta} \sup_{\theta} R(\theta, \delta) &\leq \sup_{\theta} R(\theta, \delta_0) \leq r(\lambda_0, \delta_0) = \inf_{\delta} r(\lambda_0, \delta) \\
&\leq \sup_{\lambda} \inf_{\delta} r(\lambda, \delta) \leq \inf_{\delta} \sup_{\lambda} r(\lambda, \delta) \leq \inf_{\delta} \sup_{\theta} R(\theta, \delta).
\end{aligned}$$

Thus, the inequalities above are all equalities.

Example. For $X_1, \cdots, X_n$ are independent $Ber(\theta)$ and quadratic loss function, we saw that the minimax rule was a Bayes rule with respect to $Beta(\sqrt{n}/2, \sqrt{n}/2)$ with constant risk function. Thus, this prior distribution is least favorable.

Definition. Suppose that $\Omega = \{\theta_1, \cdots, \theta_k\}$. Then the *risk set*

$$R = \{z \in R^k : z_i = R(\theta_i, \delta) \text{ for some decision rule } \delta \text{ and } i = 1, \cdots, k\}.$$

The *lower boundary* of a set $C \subset R^k$ is

$$\partial_L(C) = \{z \in \bar{C} : x_i \leq z_i \text{ for all } i \text{ and } x_i < z_i \text{ for some } i \text{ implies } x \notin C\}.$$

The set C is *closed from below* if $\partial_L(C) \subset C$.

Note that the risk set is convex. Interior points correspond to randomized decision rules.

Theorem. (Minimax Theorem) Suppose that the loss function is bounded below and that Ω is finite. Then

$$\sup_\lambda \inf_\delta r(\lambda, \delta) = \inf_\delta \sup_\theta R(\theta, \delta).$$

In addition, there exists a least favorable distribution λ_0. If the risk set is closed from below, then there is a minimax rule that is a Bayes rule with respect to λ_0.

We will use the following lemmas in the proof.

Lemma. For Ω finite, the loss function is bounded below if and only if the risk function is bounded below.

Proof. If the loss function is bounded below, then its expectation is also bounded below.

Because Ω is finite, if the loss function is unbounded below, then there exist $\theta_0 \in \Omega$ and a sequence of actions a_n so that $L(\theta_0, a_n) < -n$. Now take $\delta_n(x) = a_n$ to see that the risk set is unbounded below.

Lemma. If $C \subset R^k$ is bounded below, them $\partial_L(C) \neq \emptyset$.

Proof. Clearly $\partial_L(C) = \partial_L(\bar{C})$. Set

$$c_1 = \inf\{z_1 : z \in \bar{C}\},$$

and

$$c_j = \inf\{z_j : z_i = c_i, i = 1, \cdots, j-1\}.$$

Then $(c_1, \cdots, c_k) \in \partial_L(C)$.

Lemma. If their exists a minimax rule for a loss function that is bounded below, then there is a point on $\partial_L(R)$ whose maximum coordinate value is the same as the minimax risk.

Proof. Let $z \in R^k$ be the risk function for a minimax rule and set

$$s = \max\{z_1, \cdots, z_k\}$$

be the minimax risk. Define

$$C = R \cap \{x \in R^k : x_i \leq s \text{ for all } i\}.$$

Because the loss function is bounded below, so is R and hence C. Therefore, $\partial_L(C) \neq \emptyset$. Clearly, $\partial_L(C) \subset \partial_L(R)$ and each point in C is the risk function of a minimax rule.

Proof. (Minimax Theorem) For each real s define the closed convex set $C_s = (-\infty, s]^k$ and set $s_0 = \inf\{s : C_s \cap R \neq \emptyset\}$.

Claim. There is a least favorable decision.

Note that

$$s_0 = \inf_\delta \sup_\theta R(\theta, \delta).$$

Because the interior of C_{s_0} is convex and does not intersect R, the separating hyperplane theorem guarantees a vector v and a constant c such that

$$\langle v, z\rangle \ \geq c \text{ for all } z \in R \text{ and } \langle v, z\rangle \ < c \text{ for all } z \in \text{int}(C_{s_0}).$$

It is easy to see that each coordinate of v is non-negative. Normalize v so that its sum is 1. Define a probability λ_0 on $\Omega = \{\theta_1 \cdots, \theta_k\}$ with with respective masses

$$\lambda_0(\theta_i) = v_i.$$

Use the fact that $\{s_0, \cdots, s_0\}$ is in the closure of the interior of C_{s_0} to obtain

$$c \geq s_0 \sum_{j=1}^{n} v_j = s_0.$$

Therefore,

$$\inf_\delta r(\lambda_0, \delta) = \inf_{z \in R} \langle v, z\rangle \geq c \geq s_0 = \inf_\delta \sup_\theta R(\theta, \delta) \geq \inf_\delta r(\lambda_0, \delta),$$

and λ_0 is a least favorable distribution.

Note that for $s > s_0$, $\bar{R} \cap C_s$ is closed, bounded and non-empty. Let $\{s_n : n \geq 1\}$ decrease to s_0. Then by the finite interection property, $\bar{R} \cap C_{s_0} \neq \emptyset$. The elements in this set are risks functions for minimax rules. By a lemma above, $\partial(\bar{R} \cap C_s) \neq \emptyset$. Because R is closed from below, $\partial(\bar{R} \cap C_s) \subset R$, we have a point in R that is the risk function for a minimax rule.

Finally, note that $R(\theta, \delta) \leq s_0$ for all θ implies that $r(\lambda, \delta) \leq s_0 = \inf_\delta r(\lambda_0, \delta)$.

5.7 Complete Classes

Definition. A class of decision rules $\mathcal{C}$ is *complete* if for every $\delta \notin \mathcal{C}$, there exists $\delta_0 \in \mathcal{C}$ that dominates δ. A *minimal complete class* contains no proper complete class.

Certainly, a complete class contains all admissible rules. If a minimal complete class exists, then it consits of exactly the admissible rules.

Theorem. (Neyman-Pearson fundamental lemma) Let $\Omega = A = \{0, 1\}$. The loss function

$$\begin{array}{ll} L(0,0) = 0, & L(0,1) = k_0 > 0, \\ L(1,0) = k_1 > 0, & L(1,1) = 0. \end{array}$$

Set $\nu = P_0 + P_1$ and $f_i = dP_i/d\nu$ and let δ be a decision rule. Define the *test function* $\phi(x) = \delta(x)\{1\}$ corresponding to δ. Let $\mathcal{C}$ denote the class of all rules with test functions of the following forms:

For each $k > 0$ and each function $\gamma : \mathcal{X} \to [0, 1]$,

$$\phi_{k,\gamma}(x) = \begin{cases} 1 & : \text{ if } f_1(x) > kf_0(x), \\ \gamma(x) & : \text{ if } f_1(x) = kf_0(x), \\ 0 & : \text{ if } f_1(x) < kf_0(x). \end{cases}$$

For $k = 0$,

$$\phi_0(x) = \begin{cases} 1 & : \text{ if } f_1(x) > 0, \\ 0 & : \text{ if } f_1(x) = 0. \end{cases}$$

For $k = \infty$,

$$\phi_\infty(x) = \begin{cases} 1 & : \text{ if } f_0(x) = 0, \\ 0 & : \text{ if } f_0(x) > 0. \end{cases}$$

Then $\mathcal{C}$ is a minimal complete class.

The Neyman-Pearson lemma uses the $\phi_{k,\gamma}$ to construct a likelihood ratio test

$$\frac{f_1(x)}{f_0(x)}$$

and a test level k. If the probability of hitting the level k is positive, then the function γ is a randomization rule needed to resolve ties.

The proof proceeds in choosing a level α for $R(0, \delta)$, the probability of a type 1 error, and finds a test function from the list above that yields a decision rule δ^* that matches the type one error and has a lower type two error, $R(1, \delta^*)$.

Proof. Append to $\mathcal{C}$ all rules having test functions of the form $\phi_{0,\gamma}$. Call this new collection $\mathcal{C}'$.

Claim. The rules in $\mathcal{C}'\backslash\mathcal{C}$ are inadmissible.

Choose a rule $\delta \in \mathcal{C}'\backslash\mathcal{C}$. Then δ has test function $\phi_{0,\gamma}$ for some γ such that $P_0\{\gamma(X) > 0, f_1(X) = 0\} > 0$. Let δ_0 be the rule whose test function is ϕ_0. Note that $f_1(x) = 0$ whenever $\phi_{0,\gamma}(x) \neq \phi_0(x)$,

$$\begin{aligned} R(1, \delta) &= E_1[L(1, \delta(X))] = L(1, 0)E_1[\delta(X)\{0\}] \\ &= k_1 E_1[1 - \phi_{0,\gamma}(X)] = k_1(1 - \int \phi_{0,\gamma}(x) f_1(x)\ \nu(dx)) \\ &= k_1(1 - \int \phi_0(x) f_1(x)\ \nu(dx)) = R(1, \delta_0) \end{aligned}$$

Also,

$$\begin{aligned} R(0, \delta) &= E_0[L(0, \delta(X))] = L(0, 1)E_1[\delta(X)\{1\}] \\ &= k_0 E_0[\phi_{0,\gamma}(X)] = k_0(E_0[\gamma(X) I_{\{f_1(X)=0\}}] + E_0[1 I_{\{f_1(X)>0\}}]) \\ &= k_0 E_0[\gamma(X) I_{\{f_1(X)=0\}}] + R(0, \delta_0) > R(0, \delta_0). \end{aligned}$$

Thus, δ_0 dominates δ and δ is not admissible.

To show that $\mathcal{C}'$ is a complete class, choose a rule $\delta \notin \mathcal{C}'$ and let ϕ be the corresponding test function. Set

$$\alpha = R(0, \delta) = \int k_0 \phi(x) f_0(x)\ \nu(dx).$$

We find a rule $\delta^* \in \mathcal{C}'$ such $R(0, \delta^*) = \alpha$ and $R(1, \delta^*) < R(1, \delta)$. We do this by selecting an appropriate choice for k^* and γ^* for the δ^* test function ϕ_{k^*,γ^*}.

To this end, set

$$g(k) = k_0 P_0\{f_1(X) \geq k f_0(X)\} = \int_{\{f_1 \geq k f_0\}} k_0 f_0(x)\ \nu(dx).$$

Note that

1. g is a decreasing function.
2. $\lim_{k\to\infty} g(k) = 0$.
3. $g(0) = k_0 \geq \alpha$.
4. By the monotone convergence theorem, g is left continuous.
5. By the dominated convergence theorem, g has right limit

$$g(k+) = k_0 P_0\{f_1(X) > k f_0(X)\}.$$

6. If $\gamma(x) = 1$ for all x, then $g(k) = R(0, \delta^*)$.
7. If $\gamma(x) = 0$ for all x, then $g(k+) = R(0, \delta^*)$.

Set $k^* = \inf\{k : g(k) \leq \alpha\}$. Because g decreases to zero, if $\alpha > 0$, then $k^* < \infty$. To choose γ^*, we consider three cases and check that $R(0, \delta) = R(0, \delta^*)$.

Case 1. $\alpha = 0$, $k^* < \infty$. Choose $\gamma^* = 0$. Then

$$R(0, \delta^*) = k_0 E_0[\phi_{k^*,\gamma^*}(X)] = g(k^*+) = 0 = \alpha.$$

Case 2. $\alpha = 0$, $k^* = \infty$.

$$R(0, \delta^*) = k_0 E_0[\phi_\infty(X)] = \int k_0 \phi_\infty(x) f_0(x)\ \nu(dx) = 0 = \alpha.$$

Case 3. $\alpha > 0$, $k^* < \infty$.

Note that

$$k_0 P_0\{f_1(X) = k^* f_0(X)\} = g(k^*) - g(k^*+).$$

For those x which satisfy $f_1(x) = k^* f_0(x)$, define

$$\gamma^*(x) = \frac{\alpha - g(k^*+)}{g(k^*) - g(k^*+)}.$$

Then,

$$\begin{aligned} R(0, \delta^*) &= k_0 \int \phi_{k^*,\gamma^*}(x) f_0(x)\ \nu(dx) = g(k^*+) + k_0 \int_{\{f_1(x) = k^* f_0(x)\}} \frac{\alpha - g(k^*+)}{g(k^*) - g(k^*+)} f_0(x)\ \nu(dx) \\ &= g(k^*+) + k_0 \frac{\alpha - g(k^*+)}{g(k^*) - g(k^*+)} P_0\{f_1(X) = k^* f_0(X)\} = \alpha. \end{aligned}$$

We now verify that $R(1,\delta^*) < R(1,\delta)$ in two cases.

Case1. $k^* < \infty$

Define

$$h(x) = (\phi_{k^*,\gamma^*}(x) - \phi(x))(f_1(x) - k^* f_0(x)).$$

Here, we have

1. $1 = \phi_{k^*,\gamma^*}(x) \geq \phi(x)$ for all x satisfying $f_1(x) - k^* f_0(x)) > 0$.
2. $0 = \phi_{k^*,\gamma^*}(x) \leq \phi(x)$ for all x satisfying $f_1(x) - k^* f_0(x) < 0$.

Because ϕ is not one of the $\phi_{k,\gamma}$, there exists a set B such that $\nu(B) > 0$ and $h > 0$ on B.

$$\begin{aligned} 0 &< \int_B h(x)\ \nu(dx) \leq \int h(x)\ \nu(dx) \\ &= \int (\phi_{k^*,\gamma^*}(x) - \phi(x)) f_1(x)\ \nu(dx) - \int (\phi_{k^*,\gamma^*}(x) - \phi(x)) k^* f_0(x)\ \nu(dx) \\ &= \int ((1-\phi(x)) - (1 - \phi_{k^*,\gamma^*}(x))) f_1(x)\ \nu(dx) + \frac{k^*}{k_0}(\alpha - \alpha) \\ &= \frac{1}{k_1}(R(1,\delta) - R(1,\delta^*)). \end{aligned}$$

Case2. $k^* = \infty$

In this case, $0 = \alpha = R(0,\delta)$ and hence $\phi(x) = 0$ for almost all x for which $f_0(x) > 0$. Because ϕ and ϕ_∞ differ on a set of ν positive measure,

$$\int_{\{f_0=0\}} (\phi_\infty(x) - \phi(x)) f_1(x)\ \nu(dx) = \int_{\{f_0=0\}} (1-\phi(x)) f_1(x)\ \nu(dx) > 0.$$

Consequently,

$$R(1,\delta) = k_1 P_0\{f_0(X) > 0\} + k_1 \int_{\{f_0=0\}} (1-\phi(x)) f_1(x)\ \nu(dx) > k_1 P_0\{f_0(X) > 0\} = R(1,\delta^*).$$

This gives that $\mathcal{C}'$ is complete. Check that no element of $\mathcal{C}$ dominates any other element of $\mathcal{C}$, thus $\mathcal{C}$ is minimal complete.

All of the rules above are Bayes rules which assign positive probability to both parameter values.

Example.

1. Let $\theta_1 > \theta_0$ and let f_i have a $N(\theta_i, 1)$ density. Then, for any k,

$$\frac{f_1(x)}{f_0(x)} > k \text{ if and only if } x > \frac{\theta_1 + \theta_0}{2} + \frac{\log k}{\theta_1 - \theta_0}.$$

2. Let $1 > \theta_1 > \theta_0 > 0$ and let f_i have a $Bin(n, \theta_i)$ density. Then, for any k,

$$\frac{f_1(x)}{f_0(x)} > k \text{ if and only if } x > \frac{n\log(\frac{1-\theta_0}{1-\theta_1}) + \log k}{\log(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)})}.$$

3. Let ν be Lebesgue measure on $[0, n]$ plus counting measure on $\{0, 1, \cdots, n\}$. Consider the distributions $Bin(n, p)$ and $U(0, n)$ with respective densities f_0 and f_1. Note that

$$f_1(x) = \sum_{i=1}^{n} \frac{1}{n} I_{(i-1,i)}(x).$$

Then

$$\frac{f_1(x)}{f_0(x)} = \begin{cases} \infty & : \text{ if } 0 < x < n, x \text{ not an integer,} \\ 0 & : \text{ if } x = 0, 1, \cdots, n. \\ \text{undefined} & : \text{ otherwise} \end{cases}$$

The only admissible rule is to take the binomial distribution if and only if x is an integer.

6 Hypothesis Testing

We begin with a function

$$V : S \to \mathcal{V}.$$

In classical statistics, the choices for V are functions of the parameter Θ.

Consider $\{\mathcal{V}_H, \mathcal{V}_A\}$, a partition of $\mathcal{V}$. We can write a *hypothesis* and a corresponding *alternative* by

$$H : V \in \mathcal{V}_H \quad \text{versus} \quad A : V \in \mathcal{V}_A.$$

A decision problem is called *hypothesis testing* if

1. The action space $A = \{0, 1\}$.
2. The loss function L satisfies

$$\begin{array}{ll} L(v,1) \geq L(v,0) & \text{for } v \in \mathcal{V}_H \\ L(v,1) \leq L(v,0) & \text{for } v \in \mathcal{V}_A \end{array}$$

The action $a = 1$ is called *rejecting the hypothesis.* Rejecting the hypothesis when it is true is called a *type I error.*

The action $a = 0$ is called *failing to reject the hypothesis.* Failing to reject the hypothesis when it is false is called a *type II error.*

We can take $L(v,0) = 0, L(v,1) = c$ for $v \in \mathcal{V}_H$ and $L(v,0) = 1, L(v,1) = 0$ for $v \in \mathcal{V}_A$ and keep the ranking of the risk function. Call this a $0 - 1 - c$ *loss function*

A randomized decision rule in this setting can be described by a test function $\phi : \mathcal{X} \to [0,1]$ by

$$\phi(x) = \delta(x)\{1\}.$$

Suppose $V = \Theta$. Then,

1. The *power function* of a test $\beta_\phi(\theta) = E_\theta[\phi(X)]$.
2. The *operating characteric curve* $\rho_\phi = 1 - \beta_\phi$.
3. The *size* of ϕ is $\sup_{\theta \in \Omega_H} \beta_\phi(\theta)$.
4. A test is called *level* α if its size is at most α.
5. The *base* of ϕ is $\inf_{\theta \in \Omega_A} \beta_\phi(\theta)$.
6. A test is called *floor* γ if its base is at most γ.
7. A hypothesis (alternative) is *simple* if Ω_H (Ω_A) is a singleton set. Otherwise, it is called *composite.*

This sets up the following duality between the hypothesis and its alternative.

hypothesis	alternative
test function ϕ	test function $\psi = 1 - \phi$
power	operating characteristic
level	base
size α	floor γ

To further highlight this duality, note that for a $0 - 1 - c$ loss function,

$$R(\theta, \delta) = \begin{cases} c\beta_\phi(\theta) & \text{if } \theta \in \Omega_H \\ 1 - \beta_\phi(\theta) & \text{if } \theta \in \Omega_A \end{cases}$$

If we let $\Omega'_H = \Omega_A$ and $\Omega'_A = \Omega_H$ and set the test function to be $\phi' = 1 - \phi$. Then, for c times a $0 - 1 - 1/c$ loss function and $\delta' = 1 - \delta$

$$R'(\theta, \delta') = \begin{cases} \beta_{\phi'}(\theta) & \text{if } \theta \in \Omega'_H \\ c(1 - \beta_{\phi'}(\theta)) & \text{if } \theta \in \Omega'_A \end{cases},$$

and $R(\theta, \delta) = R'(\theta, \delta')$

Definition. A level α test ϕ is *uniformly most powerful (UMP)* level α if for every level α test ψ

$$\beta_\psi(\theta) \leq \beta_\phi(\theta) \text{ for all } \theta \in \Omega_A.$$

A floor γ test ϕ is *uniformly most cautious (UMC)* level γ if for every floor γ test ψ

$$\beta_\psi(\theta) \geq \beta_\phi(\theta) \text{ for all } \theta \in \Omega_H.$$

Note that if T is a sufficient statistic, then

$$E_\theta[\phi(X)|T(X)]$$

has the same power function as ϕ. Thus, in choosing UMP and UMC tests, we can confine ourselves to functions of sufficient statistics.

6.1 Simple Hypotheses and Alternatives

Throughout this section, $\Omega = \{\theta_0, \theta_1\}$, we consider the hypothesis

$$H : \Theta = \theta_0 \quad \text{versus} \quad A : \Theta = \theta_1$$

and write type I error $\alpha_0 = \beta_\phi(\theta_0)$ and type II error $\alpha_1 = 1 - \beta_\phi(\theta_1)$. The risk set is a subset of $[0, 1]^2$ that is closed, convex, symmetric about the point $(1/2, 1/2)$ (Use the decision rule $1 - \phi$.), and contains the portion of the line $\alpha_1 = 1 - \alpha_0$ lying in the unit square. (Use a completely randomized decision rule without reference to the data.)

In terms of hypothesis testing, the Neyman-Pearson lemma states that all of the decisions rules in $\mathcal{C}$ lead to most powerful and most cautious tests of their respective levels and floors.

Note that

$$\beta_{\phi_\infty}(\theta_0) = E_{\theta_0}[\phi_\infty(X)] = P_{\theta_0}\{f_0(X) = 0\} = 0.$$

The test ϕ_∞ size 0 and never rejects H. On the other hand, ϕ_0 has the largest possible size

$$\beta_\phi(\theta_0) = P_{\theta_0}\{f_1(X) > 0\}.$$

for an admissible test.

Lemma. Assume for $i = 0, 1$ that $P_{\theta_i} << \nu$ has density f_i with respect to ν. Set

$$B_k = \{x : f_1(x) = kf_0(x)\}.$$

and suppose that $P_{\theta_i}(B_k) = 0$ for all $k \in [0, \infty]$ and $i = 0, 1$. Let ϕ be a test of the form

$$\phi = I_{\{f_1 > kf_0\}}.$$

If ψ is any test satisfying $\beta_\psi(\theta_0) = \beta_\phi(\theta_0)$, then, either

$$\psi = \phi \text{ a.s. } P_{\theta_i}, i = 0, 1 \quad \text{or} \quad \beta_\psi(\theta_1) > \beta_\psi(\theta_1).$$

In this circumstance, most powerful tests are essentially unique.

Lemma. If ϕ is a MP level α test, then either $\beta_\phi(\theta_1) = 1$ or $\beta_\phi(\theta_0) = \alpha$.

Proof. To prove the contrapositive, assume that $\beta_\phi(\theta_1) < 1$ and $\beta_\phi(\theta_0) < \alpha$. Define, for $c \geq 0$,

$$g(c, x) = \min\{c, 1 - \phi(x)\},$$

and

$$h_i(c) = E_{\theta_i}[g(c, X)].$$

Note that g is bounded. Because g is continuous and non-decreasing in c, so is h_i. Check that $h_0(0) = 0$ and $h_0(1) = 1 - \beta_\phi(\theta_0)$. Thus, by the intermediate value theorem, there exists $\tilde{c} > 0$ so that $h_0(\tilde{c}) = \alpha - \beta_\phi(\theta_0)$. Define a new test function $\phi'(x) = \phi(x) + g(\tilde{c}, x)$. Consequently, $\beta_{\phi'}(\theta_0) = \beta_\phi(\theta_0) + \alpha - \beta_\phi(\theta_0) = \alpha$.

Note that

$$P_{\theta_1}\{\phi(X) < 1\} = 1 - P_{\theta_1}\{\phi(X) = 1\} \geq 1 - E_{\theta_1}[\phi(X)] = 1 - \beta_\phi(\theta_1) > 0.$$

On the set $\phi < 1$, $\phi' > \phi$. Thus,

$$\beta_{\phi'}(\theta_1) > \beta_\phi(\theta_1)$$

and ϕ is not most powerful.

In other words, a test that is MP level α must have size α unless all tests with size α are inadmissible.

Remarks.

1. If a test ϕ corresponds to the point $(\alpha_0, \alpha_1) \in (0, 1)^2$, then ϕ is MC floor $1 - \alpha_1$ if and only if it is MP level α_0.

2. If ϕ is MP level α, then $1 - \phi$ has the smallest power at θ_1 among all tests with size at least $1 - \alpha$.

3. If ϕ_1 is a level α_1 test of the form of the Neyman-Pearson lemma and if ϕ_2 is a level α_2 test of that form with $\alpha_1 < \alpha_2$, then $\beta_{\phi_1}(\theta_1) < \beta_{\phi_2}(\theta_1)$

6.2 One-sided Tests

We now examine hypotheses of the form

$$H : \Theta \leq \theta_0 \quad \text{versus} \quad A : \Theta > \theta_0$$

or

$$H : \Theta \geq \theta_0 \quad \text{versus} \quad A : \Theta < \theta_0.$$

Definition. Suppose that $\Omega \subset R$ and that

$$\frac{dP_\theta}{d\nu}(x) = f_{X|\Theta}(x|\theta)$$

for some measure ν. Then the parametric family is said to have a *monotone likelihood ratio (MLR)* in T, a real valued statistic, if whenever $\theta_1 < \theta_2$, the ratio

$$\frac{f_{X|\Theta}(x|\theta_2)}{f_{X|\Theta}(x|\theta_1)}$$

is a monotone function of $T(x)$ a.e. $P_{\theta_1} + P_{\theta_2}$. We use *increasing MLR* and *decreasing MLR* according to the properties of the ratio above. If $T(x) = x$, we will drop its designation.

Examples.

1. If X is $Cau(\theta, 1)$, then
$$\frac{f_{X|\Theta}(x|\theta_2)}{f_{X|\Theta}(x|\theta_1)} = \frac{\pi(1 + (x - \theta_1)^2)}{\pi(1 + (x - \theta_2)^2)}.$$
This is not monotone in x.

2. For X a $U(0, \theta)$ random variable,
$$\frac{f_{X|\Theta}(x|\theta_2)}{f_{X|\Theta}(x|\theta_1)} = \begin{cases} \text{undefined} & \text{if } x \leq 0, \\ \frac{\theta_1}{\theta_2} & \text{if } 0 < x < \theta_1, \\ \infty & \text{if } \theta_1 \leq x \leq \theta_2 \\ \text{undefined} & \text{if } x \geq \theta_2. \end{cases}$$
This is MLR. The undefined regions have $P_{\theta_1} + P_{\theta_2}$ measure 0.

3. If X has a one parameter exponential with natural parameter θ, then
$$\frac{f_{X|\Theta}(x|\theta_2)}{f_{X|\Theta}(x|\theta_1)} = \frac{c(\theta_2)}{c(\theta_1)} \exp((\theta_2 - \theta_1)T(x))$$
is increasing in $T(x)$ for all $\theta_1 < \theta_2$.

4. Uniform family, $U(\theta, \theta + 1)$.

5. Hypergeometric family, $Hyp(N, \theta, k)$.

Lemma. Suppose that $\Omega \subset R$ and that the parametric family $\{P_\theta : \theta \in \Omega\}$ is increasing MLR in T. If ψ is nondecreasing as a function of T, then

$$g(\theta) = E_\theta[\psi(T(X))]$$

is nondecreasing as a function of θ.

Proof. Let $\theta_1 < \theta_2$,

$$A = \{x : f_{X|\Theta}(x|\theta_1) > f_{X|\Theta}(x|\theta_2)\}, \quad a = \sup_{x \in A} \psi(T(x)),$$

$$B = \{x : f_{X|\Theta}(x|\theta_1) < f_{X|\Theta}(x|\theta_2)\}, \quad b = \inf_{x \in B} \psi(T(x)).$$

Then, on A, the likelihood ratio is less than 1. On B, the likelihood ratio is greater than 1. Thus, $b \geq a$.

$$\begin{aligned} g(\theta_2) - g(\theta_1) &= \int \psi(T(x))(f_{X|\Theta}(x|\theta_2) - f_{X|\Theta}(x|\theta_1))\nu(dx) \\ &\geq \int_A a(f_{X|\Theta}(x|\theta_2) - f_{X|\Theta}(x|\theta_1))\nu(dx) + \int_B b(f_{X|\Theta}(x|\theta_2) - f_{X|\Theta}(x|\theta_1))\nu(dx) \\ &= (b-a)\int_B (f_{X|\Theta}(x|\theta_2) - f_{X|\Theta}(x|\theta_1))\nu(dx) > 0. \end{aligned}$$

because $\nu(B) > 0$.

Theorem. Suppose that $\{P_\theta : \theta \in \Omega\}$ is a parametric family with increasing MLR in T, and consider tests of the form

$$\phi(x) = \begin{cases} 0 & \text{if } T(x) < t_0, \\ \gamma & \text{if } T(x) = t_0, \\ 1 & \text{if } T(x) > t_0 \end{cases}$$

Then,

1. ϕ has a nondecreasing power function.
2. Each such test for each θ_0 is UMP of its size for testing

$$H : \Theta \leq \theta_0 \quad \text{versus} \quad A : \Theta > \theta_0$$

3. For $\alpha \in [0, 1]$ and each $\theta_0 \in \Omega$, there exits $t_0 \in [-\infty, +\infty]$ and $\gamma \in [0, 1]$ such that the test ϕ is UMP level α.

Proof. The first statement follows from the previous lemma.
Let $\theta_0 < \theta_1$ and consider the simple hypothesis

$$\tilde{H} : \Theta = \theta_0 \quad \text{versus} \quad \tilde{A} : \Theta = \theta_1$$

Because $\{P_\theta : \theta \in \Omega\}$ has increasing MLR in T, the UMP test in the Neyman-Pearson lemma is the same as the test given above as long as γ and t_0 satisfy $\beta_\phi(\theta_0) = \alpha$.

Also, the same test is used for each of the simple hypotheses, and thus it is UMP for testing H against A.

Note that for exponential families, if the natural parameter π is an increasing function of θ, then the theorem above holds.

Fix θ_0, and choose ϕ_α of the form above so that $\beta_{\phi_\alpha}(\phi_0) = \alpha$. If the $\{P_\theta : \theta \in \Omega\}$ are mutually absolutely continuous, then $\beta_{\phi_\alpha}(\phi_1)$ is continuous in α. To see this, pick a level α. If $P_\theta\{t_0\} > 0$ and $\gamma \in (0,1)$, then small changes in α will result in small changes in γ and hence small changes in $\beta_{\phi_\alpha}(\phi_1)$. $P_\theta\{t_0\} = 0$, then small changes in α will result in small changes in t_0 and hence small changes in $\beta_{\phi_\alpha}(\phi_1)$. The remaining cases are similar.

By reversing inequalities throughout, we obtain UMP tests for

$$H : \Theta \geq \theta_0 \quad \text{versus} \quad A : \Theta < \theta_0$$

Examples.

1. Let $X_1, \cdots, X_n$ be independent $N(\mu, \sigma_0^2)$. Assume that σ_0^2 is known, and consider the hypothesis

$$H : M \leq \mu_0 \quad \text{versus} \quad A : M > \mu_0.$$

Take $T(x) = \bar{x}$, then $\bar{X}$ is $N(\mu, \sigma_0^2/n)$ and a UMP test is $\phi(x) = I_{(\bar{x}_\alpha,\infty)}(\bar{X})$ where

$$\bar{x}_\alpha = \sigma_0 \Phi^{-1}(1-\alpha)/\sqrt{n} + \mu_0.$$

In other words, let

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}$$

and reject H if $Z > z_\alpha$ where $z_\alpha = \Phi^{-1}(1-\alpha)$. The power function

$$\begin{aligned} \beta_\phi(\mu) &= P_\mu\{\bar{X} > \frac{\sigma_0 z_\alpha}{\sqrt{n}} + \mu_0\} \\ &= P_\mu\{\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} > z_\alpha + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\} \\ &= 1 - \Phi(z_\alpha + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}) \end{aligned}$$

2. Let $X_1, \cdots, X_n$ be independent $Ber(\theta)$ random variables. Then $T(x) = \sum_{i=1}^n X_i$ is the natural sufficient statistic and $\pi(\theta) = \log(\theta/(1-\theta))$, the natural parameter, is an increasing function. Thus, a UMP test of

$$H : \Theta \leq \theta_0 \quad \text{versus} \quad A : \Theta > \theta_0$$

has the form above.

3. Let $X_1, \cdots, X_n$ be independent $Pois(\theta)$ random variables. Then $T(x) = \sum_{i=1}^n X_i$ is the natural sufficient statistic and $\pi(\theta) = \log\theta$, the natural parameter, is an increasing function.

4. Let $X_1, \cdots, X_n$ be independent $U(0,\theta)$ random variables. Then $T(x) = \max_{1\leq i\leq n} x_i$ is a sufficient statistic. $T(X)$ has density
$$f_{T|\Theta}(t|\theta) = n\theta^{-n}t^{n-1}I_{(0,\theta)}(t)$$
with respect to Lebesgue measure. The UMP test of
$$H : \Theta \leq \theta_0 \quad \text{versus} \quad A : \Theta > \theta_0$$
is nonrandomized with t_0 determined by
$$\alpha = \beta_\phi(\theta_0) = \frac{n}{\theta_0^n}\int_{t_0}^{\theta_0} t^{n-1}\,dt = 1 - \frac{t_0^n}{\theta_0^n},$$
or $t_0 = \theta_0(1-\alpha)^{1/n}$. The power function
$$\beta_\phi(\theta) = \frac{n}{\theta_0^n}\int_{t_0}^{\theta} t^{n-1}\,dt = 1 - \frac{t_0^n}{\theta^n} = 1 - \frac{\theta_0^n(1-\alpha)}{\theta^n}.$$
A second UMP test is
$$\tilde{\phi}(x) = \begin{cases} 1 & \text{if } T(X) > \theta_0 \\ \alpha & \text{if } T(X) \leq \theta_0. \end{cases}$$

6.3 Two-sided Tests

Let $\theta_1 < \theta_2$. The following situations are called *two-sided hypotheses.*

$$H : \Theta \leq \theta_1 \text{ or } \Theta \geq \theta_2 \quad \text{versus} \quad A : \theta_1 < \Theta < \theta_2,$$

$$H : \theta_1 \leq \Theta \leq \theta_2 \quad \text{versus} \quad A : \Theta < \theta_1 \text{ or } \Theta > \theta_2,$$

$$H : \Theta = \theta_0 \quad \text{versus} \quad A : \Theta \neq \theta_0.$$

The first case has a *two sided hypothesis.* The second and third has a *two sided alternative.*

We will focus on the case of a one parameter exponential family.

Theorem. (Lagrange multipliers) Let $f, g_1, \cdots, g_n$ be real valued functions and let $\lambda_1, \cdots, \lambda_n$ be real numbers. If ξ_0 minimizes
$$f(\xi) + \sum_{i=1}^n \lambda_i g_i(\xi)$$
and satisfies $g_i(\xi_0) = c_i,\ i = 1, \cdots, n$, then ξ_0 minimizes f subject to $g_i(\xi_0) \leq c_i$ for each $\lambda_i > 0$ and $g_i(\xi_0) \geq c_i$ for each $\lambda_i < 0$.

Proof. Suppose that there exists $\tilde{\xi}$ such that $f(\tilde{\xi}) < f(\xi_0)$ with g_i satisfying the conditions above at $\tilde{\xi}$. Then ξ_0 does not minimize $f(\xi) + \sum_{i=1}^n \lambda_i g_i(\xi)$.

Lemma. Let $p_0, \cdots p_n \in L^1(\nu)$ have positive norm and let

$$\phi_0(x) = \begin{cases} 1 & \text{if } p_0(x) > \sum_{i=1}^n k_i p_i(x), \\ \gamma(x) & \text{if } p_0(x) = \sum_{i=1}^n k_i p_i(x), \\ 0 & \text{if } p_0(x) < \sum_{i=1}^n k_i p_i(x), \end{cases}$$

where $0 \le \gamma(x) \le 1$ and the k_i are real numbers. Then ϕ_0 minimizes

$$\int (1 - \phi(x)) p_0(x) \ \nu(dx)$$

subject to the range of ϕ in $[0,1]$,

$$\int \phi(x) p_j(x) \ \nu(dx) \le \int \phi_0(x) p_j(x) \ \nu(dx) \quad \text{whenever } k_j > 0,$$

$$\int \phi(x) p_j(x) \ \nu(dx) \ge \int \phi_0(x) p_j(x) \ \nu(dx) \quad \text{whenever } k_j < 0.$$

Proof. Choose ϕ with range in [0,1] satisfying the inequality constraints above. Clearly,

$$\phi(x) \le \phi_0(x) \quad \text{whenever} \quad p_0(x) - \sum_{i=1}^n k_i p_i(x) > 0,$$

$$\phi(x) \ge \phi_0(x) \quad \text{whenever} \quad p_0(x) - \sum_{i=1}^n k_i(x) < 0.$$

Thus,

$$\int (\phi(x) - \phi_0(x))(p_0(x) - \sum_{i=1}^n k_i p_i(x)) \ \nu(dx) \le 0.$$

or

$$\int (1 - \phi_0(x)) p_0(x) \ \nu(dx) + \sum_{i=1}^n k_i \int \phi_0(x) p_i(x)) \ \nu(dx)$$

$$\le \int (1 - \phi(x)) p_0(x) \ \nu(dx) + \sum_{i=1}^n k_i \int \phi(x) p_i(x)) \ \nu(dx).$$

Let ξ be a measurable function from $\mathcal{X}$ to $[0,1]$, and let

$$f(\xi) = \int (1 - \xi(x)) p_0(x) \ \nu(dx), \ g_i(\xi) = \int \xi(x) p_i(x) \ \nu(dx).$$

for $i = 1, \cdots, n.$ we see that ϕ_0 minimizes

$$f(\xi) + \sum_{i=1}^n k_i g_i(\xi).$$

Thus, ϕ_0 minimizes $f(\xi)$ subject to the constraints.

Lemma. Assume $\{P_\theta : \theta \in \Omega\}$ has an increasing monotone likelihood ratio in T. Pick $\theta_1 < \theta_2$ and for a test ϕ, define $\alpha_i = \beta_\phi(\theta_i),\ i = 1, 2$. Then, there exists a test of the form

$$\psi(x) = \begin{cases} 1 & \text{if } t_1 < T(x) < t_2, \\ \gamma_i & \text{if } T(x) = t_i, \\ 0 & \text{if } t_i > T(x) \text{ or } t_2 < T(x), \end{cases}$$

with $t_1 \le t_2$ such that $\beta_\psi(\theta_i) = \alpha_i$, $i = 1, 2$.

Proof. Let ϕ_α be the UMP α-level test of

$$H : \Theta \le \theta_1 \quad \text{versus} \quad A : \Theta > \theta_1.$$

For each $\tilde\alpha \in [0, 1 - \alpha_1]$, set

$$\tilde\phi_{\tilde\alpha}(x) = \phi_{\alpha_1+\tilde\alpha}(x) - \phi_{\tilde\alpha}(x).$$

Note that for all x, $\tilde\phi_\alpha(x) \in [0, 1]$, i.e., $\tilde\phi_\alpha$ is a test. Because the form of test ϕ_α, $\tilde\phi_{\tilde\alpha}$ has the form of the test above (with t_1 or t_2 possibly infinite.) In addition,

$$\beta_{\tilde\phi_\alpha}(\theta_1) = (\alpha_1 + \tilde\alpha) - \tilde\alpha = \alpha_1.$$

$\tilde\phi_0 = \phi_{\alpha_1}$ is the MP and $\tilde\phi_{1-\alpha_1} = 1 - \phi_{1-\alpha_1}$ is the least powerful level α_1 test of

$$\tilde H : \Theta = \theta_1 \quad \text{versus} \quad \tilde A : \Theta = \theta_2.$$

ϕ is also a level α_1 test of $\tilde H$ versus $\tilde A$, we have that

$$\beta_{\tilde\phi_{1-\alpha_1}}(\theta_2) \le \alpha_2 \le \beta_{\tilde\phi_0}(\theta_2).$$

Now use the continuity of $\alpha \to \beta_{\phi_\alpha}(\theta_2)$, to find $\tilde\alpha$ so that $\beta_{\tilde\phi_{\tilde\alpha}}(\theta_2) = \alpha_2$.

Theorem. Let $\{P_\theta : \theta \in \Omega\}$ be an exponential family in its natural parameter. If $\Omega_H = (-\infty, \theta_1] \cup [\theta_2, +\infty)$ and $\Omega_A = (\theta_1, \theta_2)$, $\theta_1 < \theta_2$, then a test of the form

$$\phi_0(x) = \begin{cases} 1 & \text{if } t_1 < T(x) < t_2, \\ \gamma_i & \text{if } T(x) = t_i, \\ 0 & \text{if } t_i > T(x) \text{ or } t_2 < T(x), \end{cases}$$

with $t_1 \le t_2$ minimizes $\beta_\phi(\theta)$ for all $\theta < \theta_1$ and for all $\theta > \theta_1$, and it maximizes $\beta_\phi(\theta)$ for all $\theta_1 < \theta < \theta_2$ subject to $\beta_\phi(\theta_i) = \alpha_i = \beta_{\phi_0}(\theta_i)$ for $i = 1, 2$. Moreover, if $t_1, t_2, \gamma_1, \gamma_2$ are chosen so that $\alpha_1 = \alpha_2 = \alpha$, then ϕ_0 is UMP level α.

Proof. Given α_1, α_2, choose $t_1, t_2, \gamma_1, \gamma_2$ as determined above. Choose ν so that $f_{T|\theta}(t|\theta) = c(\theta)\exp(\theta t)$. Pick $\theta_0 \in \Omega$ and define

$$p_i(t) = c(\theta_i)\exp(\theta_i t),$$

for $i = 0, 1, 2$.

Set $b_i = \theta_i - \theta_0$, $i = 1, 2$, and consider the function

$$d(t) = a_1 \exp(b_1 t) + a_2 \exp(b_2 t).$$

Case I. $\theta_1 < \theta_0 < \theta_2$

Solve for a_1 and a_2 in the pair of linear equations

$$d(t_1) = 1, \quad d(t_2) = 1$$

and note that the solutions $\tilde{a}_1$ and $\tilde{a}_2$ are both positive.

To verify this, check that d is monotone if $\tilde{a}_1\tilde{a}_2 < 0$ and negative if both $\tilde{a}_1 < 0$ and $\tilde{a}_2 < 0$.

Apply the lemma with $k_i = \tilde{a}_i c(\theta_0)/c(\theta_i)$. Note that minimizing $\int(1 - \phi(x))p_0(x)\ \nu(dx)$ is the same as maximizing $\beta_\phi(\theta_0)$ and that the constraints are

$$\beta_\phi(\theta_i) \le \beta_{\phi_0}(\theta_i)$$

The test that achieves this maximum has

$$\phi(x) = 1 \quad \text{if} \quad c(\theta_0)\exp(\theta_0 t) > k_1 c(\theta_1)\exp(\theta_1 t) + k_2 c(\theta_2)\exp(\theta_2 t)$$

or

$$\phi(x) = 1 \quad \text{if} \quad 1 > \tilde{a}_1 \exp(b_1 t) + \tilde{a}_2 \exp(b_2 t).$$

Because $\tilde{a}_1$ and $\tilde{a}_2$ are both positive, the inequality holds if $t_1 < t < t_2$, and thus $\phi = \phi_0$.

Case II. $\theta_0 < \theta_1$

To minimize $\beta_\phi(\theta_0)$, we modify the lemma, reversing the roles of 0 and 1 and replacing *minimum* with *maximum.* Now the function $d(t)$ is strictly monotone if a_1 and a_2 have the same sign. If $a_1 < 0 < a_2$, then

$$\lim_{t\to-\infty} d(t) = 0, \quad \lim_{t\to\infty} d(t) = \infty$$

and equals 1 for a single value of t. Thus, we have $\tilde{a}_1 > 0 > \tilde{a}_2$ in the solution to

$$d(t_1) = 1, \quad d(t_2) = 1.$$

As before, set $k_i = \tilde{a}_i c(\theta_0)/c(\theta_i)$ and the argument continues as above. A similar argument works for the case $\theta_1 > \theta_2$.

Choose t_1, t_2, γ_1, and γ_2 in ϕ_0 so that $\alpha_1 = \alpha_2 = \alpha$ and consider the trivial test $\phi_\alpha(x) = \alpha$ for all x. Then by the optimality properties of ϕ_0,

$$\beta_{\phi_0}(\theta) \le \alpha \text{ for every } \theta \in \Omega_H.$$

Consequently, ϕ_0 has level α and maximizes the power for each $\theta \in \Omega_A$ subject to the constraints $\beta_\phi(\theta_i) \le \alpha$ for $i = 1, 2$, and thus is UMP.

Examples.

1. Let $X_1, \cdots, X_n$ be independent $N(\theta, 1)$. The UMP test of

$$H : \Theta \leq \theta_1 \text{ or } \Theta \geq \theta_2 \quad \text{versus} \quad A : \theta_1 < \Theta < \theta_2$$

is

$$\phi_0(x) = I_{(\bar{x}_1, \bar{x}_2)}(\bar{x}).$$

The values of $\bar{x}_1$ and $\bar{x}_2$ are determined by

$$\Phi(\sqrt{n}(\bar{x}_2 - \theta_1)) - \Phi(\sqrt{n}(\bar{x}_1 - \theta_1)) = \alpha \text{ and } \Phi(\sqrt{n}(\bar{x}_2 - \theta_2)) - \Phi(\sqrt{n}(\bar{x}_1 - \theta_2)) = \alpha.$$

2. Let Y be $Exp(\theta)$ and let $X = -Y$. Thus, θ is the natural parameter. Set $\Omega_H = (0, 1] \cup [2, \infty)$ and $\Omega_A = (1, 2)$. If $\alpha = 0.1$, then we must solve the equations

$$e^{t_2} - e^{t_1} = 0.1 \quad \text{and} \quad e^{2t_2} - e^{2t_1} = 0.1$$

Setting $a = e^{t_2}$ and $b = e^{t_1}$, we see that these equations become $a - b = 0.1$ and $a^2 - b^2 = 0.1$. We have solutions $t_1 = \log 0.45$ and $t_2 = \log 0.55$. Thus, we reject H if

$$-\log 0.55 < Y < -\log 0.45.$$

3. Suppose X is $Bin(n, p)$ and consider the hypothesis

$$H : P \leq \frac{1}{4} \text{ or } P \geq \frac{3}{4} \quad \text{versus} \quad A : \frac{1}{4} < \Theta < \frac{3}{4}.$$

Then $\theta = \log(p/(1 - p))$ is the natural parameter and $\Omega_H = (-\infty, -\log 3] \cup [\log 3, \infty)$. For $n = 10$, the UMP $\alpha = 0.1$ test has $t_1 = 4, t_2 = 6$ with $\gamma_1 = \gamma_2 = 0.2565$.

4. Let $X_1, \cdots, X_n$ be independent $N(\theta, 1)$. Consider the test

$$H : \Theta = \theta_0 \quad \text{versus} \quad A : \Theta \neq \theta_0.$$

For a test level α and a parameter value $\theta_1 < \theta_0$, the test ϕ_1 that rejects H if $\bar{X} < -z_\alpha/\sqrt{n} + \theta_0$ is the unique test that has the highest power at θ_1. On the other hand, the test ϕ_2 that rejects H if $\bar{X} > z_\alpha/\sqrt{n} + \theta_0$ is also an α level test that has the highest power for $\theta_2 > \theta_0$. Using the z-score $Z = (\bar{X} - \theta)/(\sigma/\sqrt{n})$, we see

$$\beta_{\phi_1}(\theta_1) > P\{Z < -z_\alpha\} = P\{Z > z_\alpha\} > \beta_{\phi_2}(\theta_1).$$

ϕ_1 test has higher power at θ_1 than ϕ_2 and thus ϕ_2 is not UMP. Reverse the roles to see that the first test, and thus no test, is UMP.

6.4 Unbiased Tests

The following criterion will add an additional restriction on tests so that we can have optimal tests within a certain class of tests.

Definition. A test ϕ is said to be *unbiased* if for some level α,

$$\beta_\phi(\theta) \le \alpha,\ \theta \in \Omega_H \quad \text{and} \quad \beta_\phi(\theta) \ge \alpha,\ \theta \in \Omega_A.$$

A test of size α is call a *uniformly most powerful unbiased (UMPU)* test if it is UMP within the class of unbiased test of level α.

We can also define the dual concept of *unbiased floor* α and *uniformly most cautious unbiased (UMCU)* tests. Note that this restriction rules out many admissible tests.

We will call a test ϕ α-*similar* on G if

$$\beta_\phi(\theta) = \alpha \quad \text{for all } \theta \in G$$

and simply α-similar if $G = \bar{\Omega}_H \cap \bar{\Omega}_A$.

Lemma. Suppose that β_ϕ is continuous for every ϕ. If ϕ_0 is UMP among all α-similar tests and has level α, then ϕ_0 is UMPU level α.

Proof. Note that in this case an unbiased test is similar.

Because the test $\psi(\theta) = \alpha$ for all θ is α-similar, $\beta_{\phi_0}(\theta) \ge \alpha$ for $\theta \in \Omega_A$. Because ϕ_0 has level α, it is unbiased and consequently UMPU.

Definition. Let $G \subset \Omega$ be a subparameter space corresponding to a subfamily $\mathcal{Q}_0 \subset \mathcal{P}_0$ and let $\Psi : \mathcal{Q}_0 \to G$. If T is a sufficient statistic for Ψ, then a test ϕ has *Neyman structure* with respect to G and T if

$$E_\theta[\phi(X)|T(X) = t]$$

is constant a.s. $P_\theta,\ \theta \in G$.

If $\mathcal{Q}_0 = \{P_\theta : \theta \in \bar{\Omega}_H \cap \bar{\Omega}_A\}$, ϕ has Neyman structure, and

$$E_\theta[\phi(X)|T(X)] = \alpha,$$

then ϕ is α-similar. This always holds if $\phi(X)$ and $T(X)$ are independent.

Example. (t-test) Suppose that $X_1, \cdots, X_n$ are independent $N(\mu, \sigma^2)$ random variables. The usual α-level two-sided t-test of

$$H : M = \mu_0 \quad \text{versus} \quad A : M \ne \mu_0$$

is

$$\phi_0(x) = \begin{cases} 1 & \text{if } \frac{|\bar{x}-\mu_0|}{s/\sqrt{n}} > t_{n-1,\alpha/2}, \\ 0 & \text{otherwise.} \end{cases}$$

Here,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2,$$

and $t_{n-1,\alpha/2}$ is determined by $P\{T_{n-1} > t_{n-1,\alpha/2}\} = \alpha/2$, the right tail of a $t_{n-1}(0,1)$ distribution, and

$$\bar{\Omega}_H \cap \bar{\Omega}_A = \Omega_H = \{(\mu, \sigma) : \mu = \mu_0\}.$$

To check that ϕ is α-similar, note that the distribution of

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

is $t_{n-1}(0,1)$ for all σ. Hence, for all $\theta \in \Omega_H$.

$$\beta_\phi(\theta) = P_\theta\{\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > t_{n-1,\alpha/2}\} = \alpha.$$

A sufficient statistic for the subparameter Ω_H is

$$U(X) = \sum_{i=1}^{n}(X_i - \mu_0)^2 = (n-1)S^2 + n(\bar{X} - \mu_0)^2.$$

Write

$$W(X) = \frac{\bar{X} - \mu_0}{\sqrt{U(X)}} = \frac{\text{sign}(T)}{\sqrt{n}\sqrt{(n-1)/T^2 + 1}}.$$

Thus, W is a one-to-one function of T and $\phi(X)$ is a function of $W(X)$. Because the distribution of $(X_1 - \mu_0, \cdots, X_n - \mu_0)$ is spherically symmetric, the distribution of

$$(\frac{X_1 - \mu_0}{\sqrt{U(X)}}, \cdots, \frac{X_n - \mu_0}{\sqrt{U(X)}})$$

is uniform on the unit sphere and thus is independent of $U(X)$. Consequently, $W(X)$ is independent of $U(X)$ for all $\theta \in \Omega_H$ and thus ϕ has the Neyman structure relative to Ω_H and $U(X)$.

Lemma. Let T be a sufficient statistic for the subparameter space G. Then a necessary and sufficient condition for all tests similar on G to have the Neyman structure with respect to T is that T is boundedly complete.

Proof. First suppose that T is boundedly complete and let ϕ be an α similar test, then

$$E_\theta[E_\theta[\phi(X)|T(X)] - \alpha] = 0 \quad \text{for all } \theta \in G.$$

Because T is boundedly complete on G,

$$E_\theta[\phi(X)|T(X)] = \alpha \text{ a.s. } P_\theta \text{ for all } \theta \in G.$$

If T is not boundedly complete, there exists a non zero function $h, ||h|| \leq 1$, such that $E_\theta[h(T(X))] = 0$ for all $\theta \in G$. Define

$$\psi(x) = \alpha + ch(T(x)),$$

where $c = \min\{\alpha, 1-\alpha\}$. Then ψ is an α similar test. Because $E[\psi(X)|T(X)] = \psi(X)$ *psi* does not have Neyman structure with respect to T.

We will now look at the case of multiparameter exponential families $X = (X_1, \cdots, X_n)$ with natural parameter $\Theta = (\Theta_1, \cdots, \Theta_n)$. The hypothesis will consider the first parameter only. Thus, write $U = (X_2, \cdots, X_n)$ and $\Psi = (\Theta_2, \cdots, \Theta_n)$. Because the values of Ψ do not appear in the hypothesis tests, they are commonly called *nuisance parameters.*

Theorem. Using the notation above, suppose that (X_1, U) is a multiparameter exponential family.

1. For the hypothesis
$$H : \Theta_1 \le \theta_1^0 \quad \text{versus} \quad A : \Theta_1 > \theta_1^0,$$
a conditional UMP and a UMPU level α test is
$$\phi_0(x_1|u) = \begin{cases} 1 & \text{if } x_1 > d(u), \\ \gamma(u) & \text{if } x_1 = d(u), \\ 0 & \text{if } x_1 < d(u). \end{cases}$$
where d and γ are determined by
$$E_\theta[\phi_0(X_1|U)|U = u] = \alpha, \quad \theta_1 = \theta_1^0.$$

2. For the hypothesis
$$H : \Theta_1 \le \theta_1^1 \text{ or } \Theta_1 \ge \theta_1^2 \quad \text{versus} \quad A : \theta_1^1 < \Theta_1 < \theta_1^2,$$
a conditional UMP and a UMPU level α test is
$$\phi_0(x_1|u) = \begin{cases} 1 & \text{if } d_1(u) < x_1 < d_2(u), \\ \gamma_i(u) & \text{if } x_1 = d_i(u),\ i = 1, 2. \\ 0 & \text{if } x_1 < d_1(u) \text{ or } x > d_2(u). \end{cases}$$
where d and γ_i are determined by
$$E_\theta[\phi_0(X_1|U)|U = u] = \alpha, \quad \theta_1 = \theta_1^i,\ i = 1, 2.$$

3. For the hypothesis
$$H : \theta_1^1 \le \Theta_1 \le \theta_1^2 \quad \text{versus} \quad A : \Theta_1 < \theta_1^1 \text{ or } \Theta_1 > \theta_1^2$$
a conditional UMP and a UMPU level α test is
$$\phi_0(x_1|u) = \begin{cases} 1 & \text{if } x_1 < d_1(u) \text{ or } x > d_2(u), \\ \gamma_i(u) & \text{if } x_1 = d_i(u),\ i = 1, 2 \\ 0 & \text{if } d_1(u) < x_1 < d_2(u). \end{cases}$$
where d_i and γ_i are determined by
$$E_\theta[\phi_0(X_1|U)|U = u] = \alpha, \quad \theta_1 = \theta_1^i,\ i = 1, 2.$$

4. For testing the hypothesis
$$H : \Theta_1 = \theta_1^0 \quad \text{versus} \quad A : \Theta_1 \neq \theta_1^0,$$
a UMPU test of size α is has the form in part 3 with d_i and γ_i, $i = 1, 2$ determined by
$$E_\theta[\phi_0(X_1|U)|U = u] = \alpha, \text{ and } E_\theta[X\phi_0(X_1|U)] - \alpha E_\theta^0[X], \quad \theta_1 = \theta_1^0.$$

Proof. Because (X_1, U) is sufficient for θ, we need only consider tests that are functions of (X_1, U). For each of the hypotheses in 1-4,
$$\bar{\Omega}_H \cap \bar{\Omega}_A = \{\theta \in \Omega : \theta_1 = \theta_1^0\} \quad \text{or} \quad \bar{\Omega}_H \cap \bar{\Omega}_A = \{\theta \in \Omega : \theta_1 = \theta_1^i, i = 1, 2\}.$$
In all of these cucumstances U is boundedly complete and thus all test similar on $\bar{\Omega}_H \cap \bar{\Omega}_A$ have Neyman structure. The power of any test function is analytic, thus,the lemma above states that proving ϕ_0 is UMP among all similar tests establishes that it is UMPU.

The power function of any test ϕ,
$$\beta_\phi(\theta) = E_\theta[E_\theta[\phi(X_1|U)|U]].$$
Thus, for each fixed u and θ, we need only show that ϕ_0 maximizes
$$E_\theta[\phi(X_1|U)|U = u]$$
subject to the appropriate conditions. By the sufficiency of U, this expectation depends only on the first coordinate of θ.

Because the conditional law of X given $U = u$ is a one parameter exponential family, parts 1-3 follows from the previous results on UMP tests.

For part 4, any unbiased test ϕ must satisfy
$$E_\theta[\phi(X_1|U)|U = u] = \alpha, \quad \text{and} \quad \frac{\partial}{\partial\theta_1} E_\theta[\phi(X_1|U)] = 0, \ \theta \in \bar{\Omega}_H \cap \bar{\Omega}_A.$$
Differentiation under the integral sign yields
$$\begin{aligned}
\frac{\partial}{\partial\theta_1}\beta_\phi(\theta) &= \int \phi(x_1|u) \frac{\partial}{\partial\theta_1} \left(c(\theta)\exp(\theta_1 x_1 + \psi \cdot u)\right) \ \nu(dx) \\
&= \int \phi(x_1|u)(x_1 c(\theta) + \frac{\partial}{\partial\theta_1} c(\theta)) \exp(\theta_1 x + \psi \cdot u) \ \nu(dx) \\
&= E_\theta[X_1\phi(X_1|U)] + \beta_\phi(\theta)\frac{\partial c(\theta)/\partial\theta_1}{c(\theta)} \\
&= E_\theta[X_1\phi(X|U)] - \beta_\phi(\theta)E_\theta[X_1].
\end{aligned}$$

Use the fact that U is boundedly complete, $\partial\beta_\phi(\theta)/\partial\theta_1 = 0$, and that $\beta_\phi(\theta) = \alpha$ on $\bar{\Omega}_H \cap \bar{\Omega}_A$ to see that for every u
$$E_\theta[X_1\phi(X_1|U)|U = u] = \alpha E_\theta[X_1|U = u], \quad \theta \in \bar{\Omega}_H \cap \Omega_A.$$

Note that this condition on ϕ is equivalent to the condition on the derivative of the power function.

Now, consider the one parameter exponential family determined by conditioning with respect to $U = u$. We can write the density as

$$c_1(\theta_1, u) \exp(\theta_1 x_1).$$

Choose $\theta_1^1 \neq \theta_1^0$ and return to the Lagrange multiplier lemma with

$$p_0(x_1) = c_1(\theta_1^1, u) \exp(\theta_1^1 x_1), \quad p_1(x_1) = c_1(\theta_1^0, u) \exp(\theta_1^0 x_1), \quad p_1(x_1) = x_1 c_1(\theta_1^0, u) \exp(\theta_1^0 x_1)$$

to see that the test ϕ with the largest power at θ_1^1 has $\phi(x_1) = 1$ when

$$\exp(\theta_1^1 x_1) > k_1 \exp(\theta_1^0 x_1) + k_2 x_1 \exp(\theta_1^0 x_1),$$

or

$$\exp((\theta_1^1 - \theta_1^0) x_1) > k_1 + k_2 x_1.$$

Solutions to this have $\phi(x_1) = 1$ either in a semi-infinite interval or outside a bounded interval. The first option optimizes the power function only one side of the hypothesis. Thus, we need to take ϕ_0 according to the second option. Note that the same test optimizes the power function for all values of θ_1^1, and thus is UMPU.

Examples.

1. (t-test) As before, suppose that $X_1, \cdots, X_n$ are independent $N(\mu, \sigma^2)$ random variables. The density of $(\bar{X}, S^2)$ with respect to Lebesgue measure is

$$\begin{aligned} f_{\bar{X},S^2|M,\Sigma}(\bar{x}, s^2|\mu, \sigma) &= \frac{\sqrt{n}(\frac{n-1}{2\sigma^2})^{(n-1)/2}}{\sqrt{2\pi}\Gamma(\frac{n-1}{2})\sigma} \exp\left(-\frac{1}{2\sigma^2}(n((\bar{x}-\mu_0)-(\mu-\mu_0))^2+(n-1)^2 s^2)\right) \\ &= c(\theta_1, \theta_2) h(v, u) \exp(\theta_1 v + \theta_2 u) \end{aligned}$$

where

$$\theta_1 = \frac{\mu - \mu_0}{\sigma^2}, \quad \theta_2 = -\frac{1}{\sigma^2},$$

$$v = n(\bar{x} - \mu_0), \quad u = n(\bar{x} - \mu_0)^2 + (n-1)s^2.$$

The theorem states that the UMPU test of

$$H : \Theta_1 = 0 \quad \text{versus} \quad A : \Theta_1 \neq 0$$

has the form of part 4. Thus ϕ_0 is 1 in a bounded interval. From the requirement that

$$E_\theta[\phi_0(V|U)|U] = \alpha E_\theta[V|U] = 0 \text{ for } \theta_1 = 0,$$

we have that this interval is symmetric about zero. Taking $d_1(u) = -c\sqrt{u}$ and $d_2(u) = c\sqrt{u}$ gives the classical 2-sided t-test. Because it satisfies the criteria in part 4, it is UMPU.

2. (Contrasts) In an n parameter exponential family with natural parameter Θ, let $\tilde{\Theta}_1 = \sum_{i=1}^n c_i \Theta_i$ with $c_1 \neq 0$. Let $Y_1 = X_1/c_1$. For $i = 2, \cdots, n$, set

$$\tilde{\Theta}_i = \Theta_i, \quad \text{and} \quad Y_i = \frac{X_i - c_i X_1}{c_1}.$$

With the parameter space $\tilde{\Omega}$ and observations Y, we proceed as before.

3. Let Y_i be independent $Pois(\lambda_i)$, $i = 1, 2$. To consider the hypothesis

$$H : \Lambda_1 = \Lambda_2 \quad \text{versus} \quad A : \Lambda_1 \leq \Lambda_2$$

write the probability density function of $Y = (Y_1, Y_2)$

$$\frac{\exp\left(-(\lambda_1 + \lambda_2)\right)}{y_1! y_2!} \exp\left(y_2 \log(\lambda_2/\lambda_1) + (y_1 + y_2) \log \lambda_2\right)$$

with respect to counting measure in Z_+^2. If we set $\theta_1 = \log(\lambda_2/\lambda_1)$ then the hypothesis becomes

$$H : \Theta_1 = 1 \quad \text{versus} \quad A : \Theta_1 \leq 1.$$

The theorem above applies taking

$$\theta_2 = \log \lambda_2, \quad X_1 = Y_2, \quad U = Y_1 + Y_2.$$

Use the fact that U is $Pois(\lambda_1 + \lambda_2)$ to see that the conditional distribution of X_1 given $U = u$ is $Bin(p, u)$ where

$$p = \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{e^{\theta_1}}{1 + e^{\theta_1}}.$$

4. Let X be $Bin(n, p)$ and $\theta = \log(p/(1-p))$. Thus, the density

$$f_{X|\Theta}(x|\theta) \binom{n}{x} (1 + e^{\theta})^{-n} e^{x\theta}.$$

For the hypothesis

$$H : \Theta = \theta_0 \quad \text{versus} \quad A : \Theta \neq \theta_0$$

the UMPU level α test is

$$\phi_0(x) = \begin{cases} 1 & \text{if } x < d_1 \text{ or } x > d_2, \\ \gamma_i & \text{if } x = d_i, \ i = 1, 2 \\ 0 & \text{if } d_1 < x < d_2. \end{cases}$$

where d_i and γ_i are determined by

$$E_{\theta_0}[\phi_0(X)] = \alpha, \text{ and } E_{\theta_0}[X\phi_0(X)] - \alpha E_{\theta_0}[X], \quad \theta_1 = \theta_1^0.$$

Once d_1 and d_2 have been determined, solving for γ_1 and γ_2 comes from a linear system. For $p_0 = 1/4$, we have $\theta_0 = -\log 3$. With $n = 10$ and $\alpha = 0.05$, we obtain

$$d_1 = 0, \quad d_2 = 5, \quad \gamma_1 = 0.52084, \quad \gamma_2 = 0.00918.$$

An equal tailed-test, having

$$d_1 = 0, \quad d_2 = 5, \quad \gamma_1 = 0.44394, \quad \gamma_2 = 0.00928.$$

is not UMPU. The probability for rejecting will be less than 0.05 for θ in some interval below θ_0.

5. Let Y_i be independent $Bin(p_i, n_i),\ i = 1, 2$. To consider the hypothesis

$$H : P_1 = P_2 \quad \text{versus} \quad A : P_1 \neq P_2$$

The probability density function with respect to counting measure is

$$\binom{n_1}{x_1}\binom{n_1}{x_2}(1-p_1)^{n_1}(1-p_2)^{n_2}\exp\left(x_2 \log\frac{p_2(1-p_1)}{p_1(1-p_2)} + (x_1+x_2)\log\frac{p_1}{1-p_1}\right).$$

If we set $\theta_1 = \log(p_2(1-p_1)/p_1(1-p_2))$ then the hypothesis becomes

$$H : \Theta_1 = 0 \quad \text{versus} \quad A : \Theta_1 \neq 0.$$

The theorem above applies taking

$$\theta_2 = \log\frac{p_1}{1-p_1}, \quad X_1 = Y_2, \quad U = Y_1 + Y_2.$$

Then, for $u = 0, 1, \cdots, n_1 + n_2$,

$$P_\theta\{X_1 = x | U = u\} = K_u(\theta)\binom{n_1}{u-x}\binom{n_2}{x}e^{\theta x}, \quad x = 0, 1\cdots, \min(u, n_2),\ x \geq u - n_1.$$

If $\theta = 0$, this is a hypergeometric distribution and $K_u(0) = \binom{n_1+n_2}{u}^{-1}$.

6. (2×2 contingency tables) Let A and B be two events and consider n independent trials whose data are summarized in the following table

	A	A^c	Total
B	Y_{11}	Y_{12}	n_1
B^c	Y_{12}	Y_{22}	n_1
Total	m_1	m_2	n

The distribution of the table entries is $Multi(n, p_{11}, p_{12}, p_{21}, p_{22})$ giving a probability density function with respect to counting measure is

$$\binom{n}{y_{11}, y_{12}, y_{21}, y_{22}}p_{22}^n \exp\left(y_{11}\log\frac{p_{11}}{p_{22}} + y_{12}\log\frac{p_{12}}{p_{22}} + y_{21}\log\frac{p_{21}}{p_{22}}\right).$$

We can use the theorem to derive UMPU tests for any parameter of the form

$$\tilde{\theta}_1 = c_{11} \log \frac{p_{11}}{p_{22}} + c_{12} \log \frac{p_{12}}{p_{22}} + c_{21} \log \frac{p_{21}}{p_{22}}.$$

Check that a test for independence of A and B follows from the hypothesis

$$H : \tilde{\Theta}_1 = 0 \quad \text{versus} \quad A : \tilde{\Theta}_1 \neq 0.$$

with the choices $c_{11} = -1$, and $c_{12} = c_{21} = 1$, take $X_1 = Y_{11}$ and $U = (Y_{11} + Y_{12}, Y_{11} + Y_{21})$. Compute

$$P_\theta\{X_1 = x_1 | U = (n_1, m_2)\} = K_{m_2}(\theta_1) \binom{n_1}{y} \binom{m_2}{m_2 - x_1} e^{\theta_1(m_2 - x_1)},$$

$x = 0, 1, \cdots, \min(n_1, m_2),\ m_2 x \leq n_2$.

The choice $\theta_1 = 0$ gives the hypergeometric distribution. This case is called *Fisher's exact test.*

7. (two-sample problems, inference for variance) Let $X_{i1}, \cdots, X_{in_i}$, $i = 1, 2$ be independent $N(\mu_i, \sigma_i^2)$. Then the joint density takes the form

$$c(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \exp\left(-\sum_{i=1}^{2} (\frac{1}{2\sigma_i^2} \sum_{j=1}^{n_i} x_{ij}^2 + \frac{n_i \mu_i}{\sigma_i^2} \bar{x}_i) \right),$$

where $\bar{x}_i$ is the sample mean of the i-th sample.

Consider the hypothesis

$$H : \Sigma_2^2/\Sigma_1^2 \leq \delta_0 \quad \text{versus} \quad A : \Sigma_2^2/\Sigma_1^2 > \delta_0.$$

For the theorem, take

$$\theta = (\frac{1}{2\delta_0 \sigma_1^2} - \frac{1}{2\sigma_2^2}, -\frac{1}{2\sigma_1^2}, \frac{n_1 \mu_1}{\sigma_1^2}, \frac{n_2 \mu_2}{\sigma_2^2}),$$

$$Y_1 = \sum_{j=1}^{n_2} X_{2j}^2, \qquad U = (\sum_{j=1}^{n_1} X_{1j}^2 + \frac{1}{\delta_0} \sum_{j=1}^{n_2} X_{2j}^2, \bar{X}_1, \bar{X}_2).$$

Consider the statistic

$$V = \frac{(n_2 - 1)S_2^2/\delta_0}{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2/\delta_0} = \frac{(Y_1 - n_2 U_3^2)/\delta_0}{U_1 - n_1 U_2^2 - n_2 U_3^2/\delta_0}.$$

S_i^2 is the sample variance based on the i-th sample.

If $\sigma_2^2 = \delta_0 \sigma_1^2$, i.e. $\theta_1 = 0$, then V is ancillary and, by Basu's theorem, V and U are independent. Note that V is increasing in Y_1 for each U. Thus, the UMPU α-level test of

$$H : \Theta_1 \leq 0 \quad \text{versus} \quad A : \Theta_1 > 0.$$

rejects H whenever

$$V > d_0, \qquad P_0\{V > d_0\} = \alpha.$$

Note that

$$V = \frac{(n_2-1)F}{(n_1-1)+(n_2-1)F} \quad \text{with } F = \frac{S_2^2/\delta_0}{S_1^2}.$$

Thus, V is an increasing function of F, which has the F_{n_2-1,n_1-1}-distribution. Consequently, the classical F test is UMPU.

8. (two-sample problem, inference for means) Consider the hypothesis

$$H : M_1 = M_2 \quad \text{versus} \quad A : M_1 \neq M_2.$$

or

$$H : M_1 \geq M_2 \quad \text{versus} \quad A : M_1 < M_2.$$

The situaion $\sigma_1^2 \neq \sigma_2^2$ is called the Fisher-Behrens problem. To utilize the theorem above, we consider the case $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

The joint density function is

$$c(\mu_1,\mu_2,\sigma^2)\exp\left(\frac{1}{\sigma^2}\sum_{i=1}^{2}\sum_{j=1}^{n_i} x_{ij}^2 + \frac{n_1\mu_1}{\sigma^2}\bar{x}_1 + \frac{n_2\mu_2}{\sigma^2}\bar{x}_2\right).$$

For the theorem take,

$$\theta = (\frac{\mu_2-\mu_1}{(1/n_1+1/n_2)\sigma^2}, \frac{n_1\mu_1+n_2\mu_2}{(n_1+n_2)\sigma^2}, -\frac{1}{2\sigma^2}),$$

$$Y_1 = \bar{X}_2 - \bar{X}_1, \qquad U = (n_1\bar{X}_1 + n_2\bar{X}_2, \sum_{i=1}^{2}\sum_{j=1}^{n_i} X_{ij}^2).$$

Consider the statistic

$$T = \frac{(\bar{X}_2-\bar{X}_1)/\sqrt{1/n_1+1/n_2}}{\sqrt{((n_1-1)S_1^2+(n_2-1)S_2^2)/(n_1+n_2-2)}}.$$

Again, by Basu's theorem, when $\theta = 0$, T and U are independent. Also, T, a function of (Y_1, U) is increasing in Y_1 for each value of U. Thus, the exist values t_1 and t_2 so that the test takes the form of part 4 of the theorem.

Upon division of the numerator and denominator of T by σ^2, the numberator is a standard normal random variable. The denominator is $\chi^2_{n_1+n_2-1}$. Because the numerator and denominator are independent, T has the $t_{n_1+n_2-2}$-distribution.

6.5 Equivariance

Definition. Let $\mathcal{P}_0$ be a parametric family with parameter space Ω and sample space $(\mathcal{X},\mathcal{B})$. Let G be a group of transformations on $\mathcal{X}$. We say that G *leaves $\mathcal{X}$ invariant* if for each $g \in G$ and each $\theta \in \Omega$, there exists $g^* \in G$ such that

$$P_\theta(B) = P_{\theta^*}(gB) \quad \text{for every } B \in \mathcal{B}.$$

If the parameterization is identifiable, then the choice of θ^* is unique. We indicate this by writing $\theta^* = \bar{g}\theta$. Note that

$$P'_\theta\{gX \in B\} = P'_\theta\{X \in g^{-1}B\} = P'_{\bar{g}\theta}\{X \in gg^{-1}B\} = P'_{\bar{g}\theta}\{X \in B\}.$$

We can easily see that if G leaves $\mathcal{P}_0$ invariant, then the transformation $\bar{g} : \Omega \to \Omega$ is one-to-one and onto. Moreover, the mapping

$$g \mapsto \bar{g}$$

is a group isomorphism from G to $\bar{G}$.

We call a loss function L *invariant under* G if for each $a \in A$, there exists a unique $a^* \in A$ such that

$$L(\bar{g}\theta, a^*) = L(\theta, a).$$

Denote a^* by $\tilde{g}a$. Then, the transformation $\tilde{g} : A \to A$ is one-to-one and onto. The mapping

$$g \to \tilde{g}$$

is a group homomorphism.

Example. Pick $b \in R^n$ and $c > 0$ and consider the transformation $g_{(b,c)}$ on $\mathcal{X} = R^n$ defined by

$$g_{(b,c)}(x) = b + cx.$$

This forms a transformation group with $g_{(b_1,c_1)} \circ g_{(b_2,c_2)} = g_{(c_1b_2+b_1,c_1c_2)}$. Thus, the group is not abelian. The identity is $g_{(0,1)}$. The inverse of $g_{(b,c)}$ is $g_{(-b/c,1/c)}$.

Suppose that $X_1, \cdots, X_n$ are independent $N(\mu, \sigma^2)$. Then

$$\bar{g}_{(b,c)}(\mu, \sigma) = (\frac{\mu - b}{c}, \frac{\sigma}{c}).$$

To check this,

$$\begin{aligned}
P_\theta(g_{(b,c)}B) &= \int_{cB+b} \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (z_i - \mu)^2\right) dz \\
&= \int_{cB} \frac{1}{(\sigma)\sqrt{2\pi})^n} \exp\left(-\frac{1}{\sigma^2}\sum_{i=1}^n (y_i - (\mu - b))^2\right) dy \\
&= \int_{B} \frac{1}{((\sigma/c)\sqrt{2\pi})^n} \exp\left(-\frac{1}{2(\sigma/c)^2}\sum_{i=1}^n (x_i - \frac{1}{c}(b - \mu)^2\right) dx = P_{\bar{g}(b,c)}\theta(B)
\end{aligned}$$

Definition. A decision problem is *invariant under* G is $\mathcal{P}_0$ and the loss function L is invariant. In such a case a randomized decision rule δ is *equivariant* if

$$\delta(gx)(\tilde{g}B) = \delta(x)(B) \quad \text{for all measurable } B \text{ and for all } x \in \mathcal{X}.$$

For a non-randomized rule, this becomes

$$\delta(gx) = \tilde{g}\delta(x) \quad \text{for all } g \in G \text{ for all } x \in \mathcal{X}.$$

Definition.

1. The test
$$H : \Theta \in \Omega_H \quad \text{versus} \quad A : \Theta \in \Omega_A$$
is *invariant under* G if both Ω_H and Ω_A are invariant under $\bar{G}$.
2. A statistic T is *invariant under* G if T is constant on orbits, i.e.,
$$T(gx) = T(x) \qquad \text{for all } x \in \mathcal{X} \text{ and } g \in G.$$
3. A test of size α is called *uniformly most powerful invariant (UMPI)* if it is UMP with the class of α level tests that are invariant under G.
4. A statistic M is *maximal invariant under* G if it is invarian N(t and
$$M(x_1) = M(x_2) \qquad \text{implies} \qquad x_1 = g(x_2) \text{ for some } g \in G.$$

Thus, a statistic T is invariant if and only if there is a function h such that $T = h \circ M$.

Proposition. Let H be a hypothesis test invariant under a transformation group G. If there exists a UMPI test of size α, then it is unbiased. If there also exists a UMPU test of size α that is invariant under G, then the two tests have the same power function. If either the UMPI test or the UMPU test is unique, then the two tests are equal.

Proof. We only need to prove that UMPI tests of size α are unbiased. This follows from the fact that the test

$$\phi(x) = \alpha \qquad \text{for all } x \in \mathcal{X}$$

is invariant under G.

In the past, we have reduced tests to functions of a sufficient statistic U. If a test $\psi(U)$ is UMP among all invariant tests depending on U, then we can not necessarily conclude that $\psi(U)$ is UMPI. The following results provides a condition under which it is sufficient to consider tests that are function of U.

Proposition. Let $(G, \mathcal{G}, \lambda)$, a measure space with a σ-finite λ, be a group of transformations on $(\mathcal{X}, \mathcal{B})$. Consider a hypothesis that is invariant under G.

Suppose that for any set $B \in \mathcal{B}$,

$$\{(x, g) : g(x) \in B\} \in \sigma(\mathcal{B} \times \mathcal{G}).$$

Further, assume that

$$\lambda(D) = 0 \quad \text{imples} \quad \lambda\{h \circ g : h \in D\} = 0 \text{ for all } g \in G.$$

Suppose that there exists sufficient statistic U with the property that U is constant on the orbits of G, i.e.,

$$U(x_1) = U(x_2) \quad \text{implies} \quad U(gx_1) = U(gx_2) \text{ for all } g \in G.$$

Consequently, G induces a group G_U of transformations on the range of U via

$$g_U(U(x)) = U(gx).$$

Then, for any test ϕ invariant under G, there exist a test that is a function of U invariant under G (and G_U) that has the same power function as ϕ.

Exercise. Let $X_1, \cdots, X_n$ be independent $N(\mu, \sigma^2)$. The test

$$H : \Sigma^2 \geq \sigma_0^2 \quad \text{versus} \quad A : \Sigma^2 < \sigma_0^2$$

is invariant under $G = \{g_c : c \in R\}$, $g_c x = x + c$. The sufficient statistic $(\bar{X}, S^2)$ satisfy the conditions above $G_U = \{h_c : c \in R\}$ and $h_c(u_1, u_2) = (u_1 + c, u_2)$. The maximal invariant under G_U is S^2.

Use the fact that $(n-1)S^2/\sigma_0^2$ has a χ^2_{n-1}-distribution when $\Sigma^2 = \sigma_0^2$ and the proposition above to see that a UMPI test is the χ^2 test. Note that this test coincides with the UMPU test.

Consider the following *general linear model*

$$Y_i = X_i \beta^T + \epsilon_i, \qquad i = 1, \cdots, n,$$

where

- Y_i is the i-th observation, often called *response.*
- β is a p-vector of unknown parameters, $p < n$, the number of parameters is less that the number of observations.
- X_i is the i-th value of the p-vector of explanatory variables, often called the *covariates*, and
- $\epsilon_1, \cdots, \epsilon_n$ are random errors.

Thus, the data are

$$(Y_1, X_1), \cdots, (Y_n, X_n).$$

Because the X_i are considered to be nonrandom, the analysis is conditioned on X_i. The ϵ_i is viewed as random measurement errors in measuring the (unknown) mean of Y_i. The interest is in the parameter β.

For normal models, let ϵ_i be independent $N(0, \sigma^2)$, σ^2 unknown. Thus Y is $N_n(\beta X^T, \sigma^2 I_n)$, X is a fixed $n \times p$ matrix of rank $r \leq p < n$.

Consider the hypothesis

$$H : \beta L^T = 0 \quad \text{versus} \quad A : \beta L^T \neq 0.$$

Here, L is an $s \times p$ matrix, $\text{rank}(L) = s \leq r$ and all rows of L are in the range of X.

Pick an orthogonal matrix Γ such that

$$(\gamma, 0) = \beta X^T \Gamma,$$

where γ is an r-vector and 0 is the $n-r$-vector of zeros and the hypothesis becomes

$$H : \gamma_i = 0, \text{for all } i = 1, \cdots, s \quad \text{versus} \quad A : \gamma_i \neq 0, \text{for some } i = 1, \cdots, s.$$

If $\tilde{Y} = Y\Gamma$, then $\tilde{Y}$ is $N((\gamma, 0), \sigma^2 I_n)$. Write $y = (y_1, y_2)$ and $y_1 = (y_{11}, y_{12})$ where y_1 is an r-vector and y_{11} is an s-vector. Consider the group

$$G = \{g_{\Lambda,b,c} : b \in R^{r-s}, c > 0, \Lambda \in O(s, R)\}.$$

with

$$g_{\Lambda,b,c}(y) = c(y_{11}\Lambda, y_{12} + b, y_2).$$

Then, the hypothesis is invariant under G.

By the proposition, we can restrict our attention to the sufficient statistic $(\tilde{Y}_1, ||\tilde{Y}_2||)$.

Claim. The statistic $M(\tilde{Y}) = ||\tilde{Y}_{11}||/||\tilde{Y}_2||$ is maximal invariant.

Clearly, $M(\tilde{Y})$ is invariant. Choose $u_i \in R^s\backslash\{0\}$, and $t_i > 0$, $i = 1, 2$. If $||u_1||/t_1 = ||u_2||/t_2$, then $t_1 = ct_2$ with $c = ||u_1||/||u_2||$. Because $u_1/||u_1||$ and $u_2/||u_2||$ are unit vectors, there exists an orthogonal matrix Λ such that $u_1/||u_1|| = u_2/||u_2||\Lambda$, and therefore $u_1 = cu_2\Lambda$

Thus, if $M(y^1) = M(y^2)$ for $y^1, y^2 \in R^n$, then

$$y^1_{11} = cy^2_{11}\Lambda \quad \text{and} \quad ||y^1_2|| = c||y^2_2|| \text{ for some } c > 0, \ \Lambda \in O(s, R).$$

Therefore,

$$y^1 = g_{\Lambda,b,c}(y^2), \quad \text{with } b = c^{-1}y^1_{12} - y^2_{12}.$$

Exercise. $W = M(\tilde{Y})^2(n-r)/s$ has the noncentral F-distribution $NCF(s, n-r, \theta)$ with $\theta = ||\gamma||^2/\sigma^2$. Write $f_{W|\Theta}(w|\theta)$ for the density of W with respect to Lebesgue measure. Then the ratio

$$f_{W|\Theta}(w|\theta)/f_{W|\Theta}(w|0)$$

is an increasing function of w for any given $\theta \neq 0$ Therefore, a UMPI α-level test of

$$H : \Theta = 0 \quad \text{versus} \quad A : \Theta = \theta_1.$$

rejects H at critical values of the $F_{s,n-r}$ distribution. Because this test is the same for each value of θ_1, it is also a UMPI test for

$$H : \Theta = 0 \quad \text{versus} \quad A : \Theta \neq 0.$$

An alternative to finding Γ directly proceeds as follows. Because $\tilde{Y} = Y\Gamma$, we have

$$E[\tilde{Y}] = E[Y]\Gamma \quad \text{and} \quad ||\tilde{Y}_1 - \gamma||^2 + ||\tilde{Y}_2||^2 = ||Y - \beta X^T||^2.$$

Therefore

$$\min_\gamma ||\tilde{Y}_1 - \gamma||^2 + ||\tilde{Y}_2||^2 = \min_\beta ||Y - \beta X^T||^2,$$

or

$$||\tilde{Y}_2||^2 = ||Y - \hat{\beta}X^T||^2,$$

where $\hat{\beta}$ is the least square estimator, i.e., any solution to $\beta X^T X = YX$. If the inverse matrix exists, then

$$\hat{\beta} = YX(X^T X)^{-1}.$$

Similarly,

$$||Y_{11}||^2 + ||\tilde{Y}_2||^2 = \min_{\beta:\beta L^T=0} ||Y - \beta X^T||^2,$$

Denote by $\hat{\beta}_H$ the value of β that leads to this minimum. Then

$$W = \frac{(||Y - \beta_H X^T||^2 - ||Y - \beta X^T||^2)/s}{||Y - \beta X^T||^2/(n-r)}.$$

Examples.

1. (One-way analysis of variance (ANOVA)). Let Y_{ij}, $j = 1, \cdots, n_i$, $i = 1, \cdots, m$, be independent $N(\mu_i, \sigma^2)$ random variables. Consider the hypothesis test

$$H : \mu_i = \cdots = \mu_m \quad \text{versus} \quad A : \mu_i \neq \mu_k \text{ for some } i \neq k.$$

Note that $(\bar{Y}_{1,\cdot}, \cdots \bar{Y}_{m,\cdot})$ is the least squares estimate of $(\mu_1, \cdots, \mu_m)$ where $\bar{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i$. The least squares estimate under H for the *grand mean* is

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{ij}, \quad n = \sum_{i=1}^{m} n_i.$$

The *sum of squares of the residuals*

$$SSR = ||Y - \beta X^T||^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2,$$

The *total sum of squares*

$$SST = ||Y - \beta_H X^T||^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2,$$

A little algebra shows that

$$SSA = SST - SSR = \sum_{i=1}^{m} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2.$$

Thus,

$$W = \frac{SSA/(m-1)}{SSR/(n-m)}.$$

2. (Two-way balanced analysis of variance) Let Y_{ijk}, $i = 1, \cdots, a$, $j = 1, \cdots, b$, $k = 1, \cdots, c$, be independent $N(\mu_{ij}, \sigma^2)$ random variables where

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad \sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0.$$

Typically we consider the following hypotheses:

$$H : \alpha_i = 0 \text{ for all } i \quad \text{versus} \quad A : \alpha_i \neq 0 \text{ for some } i.$$

$$H : \beta_j = 0 \text{ for all } i \quad \text{versus} \quad A : \beta_j \neq 0 \text{ for some } j.$$

$$H : \gamma_{ij} = 0 \text{ for all } i, j \quad \text{versus} \quad A : \gamma_{i,j} \neq 0 \text{ for some } i, j.$$

In applications,

- α_i's are the effects of factor A,
- β_j's are the effects of factor B,
- γ_{ij}'s are the effects of the interaction of factors A and B.

Using dot to indicate averaging over the indicated subscript, we have the following least squares estimates:

$$\hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot}, \quad \hat{\beta}_i = \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot}, \quad \hat{\gamma}_{ij} = (\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot}) - (\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot}).$$

Let

$$SSR = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} (Y_{ijk} - \bar{Y}_{ij\cdot})^2, \quad SSA = bc \sum_{i=1}^{a} \hat{\alpha}_i^2, \quad SSB = ac \sum_{j=1}^{b} \hat{\beta}_j^2, \quad SSC = c \sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij}^2.$$

Then the UMPI tests for the respective hypotheses above are

$$\frac{SSA/(a-1)}{SSR/((c-1)ab)}, \quad \frac{SSB/(b-1)}{SSR/((c-1)ab)}, \quad \frac{SSC/((a-1)(b-1))}{SSR/((c-1)ab)}.$$

6.6 The Bayesian Approach

The Bayesian solution to a hypothesis-testing problem with a $0-1-c$ loss function is straightfoward. The posterior risk from choosing the action $a = 1$ is

$$cP\{v \in \mathcal{V}_H | X = x\},$$

and posterior risk from choosing the action $a = 0$ is

$$P\{v \in \mathcal{V}_A | X = x\}.$$

Thus, the optimal decision is to choose $a = 1$ if

$$cP\{v \in \mathcal{V}_H | X = x\} < P\{v \in \mathcal{V}_A | X = x\},$$

or

$$P\{v \in \mathcal{V}_H | X = x\} < \frac{1}{1+c}.$$

In classical hypothesis testing with $0-1-c$ loss function, we have

$$\begin{aligned} L(v, \delta(x)) &= cI_{\mathcal{V}_H}(v)\phi_(x) + I_{\mathcal{V}_A}(v)(1-\phi(x)) \\ L(\theta, \delta(x)) &= cP_{\theta,V}(\mathcal{V}_H)\phi_(x) + (1 - P_{\theta,V}(\mathcal{V}_H)(1-\phi(x))) \\ &= \phi(x)((c+1)P_{\theta,V}(\mathcal{V}_H) - 1) + 1 - P_{\theta,V}(\mathcal{V}_H) \\ R(\theta, \phi) &= \beta_\phi(\theta)((c+1)P_{\theta,V}(\mathcal{V}_H) - 1) + 1 - P_{\theta,V}(\mathcal{V}_H) \end{aligned}$$

Now define

$$\Omega_H = \{\theta : P_{\theta,V}(\mathcal{V}_H) \geq \frac{1}{1+c}\}, \qquad \Omega_A = \Omega_H^c,$$

$$e(\theta) = \begin{cases} 1 - P_{\theta,V}(\mathcal{V}_H) & \text{if } \theta \in \Omega_H, \\ -P_{\theta,V}(\mathcal{V}_H) & \text{if } \theta \in \Omega_A. \end{cases}$$

$$d(\theta) = |(c+1)P_{\theta,V}(\mathcal{V}_H) - 1|.$$

The risk function above is exactly equal to $e(\theta)$ plus $d(\theta)$ time the risk function from a $0-1$ loss for the hypothesis

$$H : \Theta \in \Omega_H \quad \text{versus} \quad A : \Theta \in \Omega_A.$$

If the power function is continuous at that θ such that

$$P_{\theta,V}(\mathcal{V}_H) = \frac{1}{1+c},$$

In replacing a test concerning an observable V with a test concerning a distribution of V given Θ, the predictive test function problem has been converted into a classical hypotheis testing problem.

Example. If X is $N(\theta, 1)$ and $\Theta_H = \{0\}$, then conjugate priors for Θ of the from $N(\theta_0, 1/\lambda_0)$ assign both prior and posterior probabilities to Ω_H.

Thus, in considering the hypothesis

$$H : \Theta = \theta_0 \quad \text{versus} \quad A : \Theta \neq \theta_0,$$

we must choose a prior that assigns a positive probability p_0 to $\{\theta_0\}$. Let λ be a prior distribution on $\Omega\backslash\{\theta_0\}$ for the conditional prior given $\Theta \neq \theta_0$. Assume that $P_\theta << \nu$ for all θ, then the joint density with respect to $\nu \times (\delta_{\theta_0} + \lambda)$ is

$$f_{X,\Theta}(x,\theta) = \begin{cases} p_0 f_{X|\Theta}(x|\theta) & \text{if } \theta = \theta_0, \\ (1-p_0) f_{X|\Theta}(x|\theta) & \text{if } \theta \neq \theta_0 \end{cases}$$

The marginal density of the data is

$$f_X(x) = p_0 f_{X|\Theta}(x|\theta_0) + (1-p_0)\int_\Omega f_{X|\Theta}(x|\theta)\ \lambda(d\theta).$$

The posterior distribution of Θ has density with respect to the sum of $\lambda + \delta_{\theta_0}$ is

$$f_{\Theta|X}(\theta|x) = \begin{cases} p_1 & \text{if } \theta = \theta_0, \\ (1-p_1)\frac{f_{X|\Theta}(x|\theta)}{\int_\Omega f_{X|\Theta}(\theta|x)\ \lambda(\theta)} & \text{if } \theta \neq \theta_0. \end{cases}$$

Here,

$$p_1 = \frac{f_{X|\Theta}(x|\theta)}{f_X(x)}.$$

the posterior probability of $\Theta = \theta_0$. To obtain the second term, note that

$$1 - p_1 = \frac{(1-p_0)\int_\Omega f_{X|\Theta}(x|\theta)\ \lambda(d\theta)}{f_X(x)} \quad \text{or} \quad \frac{1-p_0}{f_x(x)} = \frac{1-p_1}{\int_\Omega f_{X|\Theta}(x|\theta)\ \lambda(d\theta)}.$$

Thus,

$$\frac{p_1}{1-p_1} = \frac{p_0}{1-p_0}\frac{f_{X|\Theta}(x|\theta)}{\int_\Omega f_{X|\Theta}(x|\theta)\ \lambda(d\theta)}.$$

In words, we say that the posterior odds equals the prior odds times the *Bayes factor*.

Example. If X is $N(\theta, 1)$ and $\Omega_H = \{\theta_0\}$, then the Bayes factor is

$$\frac{e^{-\theta_0^2/2}}{\int_\Omega e^{x(\theta-\theta_0)-\theta^2/2)}\ \lambda(d\theta)}.$$

If $\lambda(-\infty, \theta_0) > 0$, and $\lambda(\theta_0, \infty) > 0$, then the denominator in the Bayes factor is a convex function of x and has limit ∞ as $x \to \pm\infty$. Thus, we will reject H if x falls outside some bounded interval.

The global lower bound on the Bayes factor,

$$\frac{f_{X|\Theta}(x|\theta)}{\sup_{\theta\neq\theta_0} f_{X|\Theta}(x|\theta)},$$

is closely related to the likelihood ratio test statistic.

In the example above, the distribution that minimizes the Bayes factor is $\lambda = \delta_x$, giving the lower bound

$$\exp(-(x-\theta_0)^2/2).$$

In the case that $x = \theta_0 + z_{0.025}$ the Bayes factor is 0.1465. Thus, rejection in the classical setting at $\alpha = 0.05$ corresponds to reducing the odds against the hypothesis by a factor of 7.

The Bayes factor for λ, a $N(\theta_0, \tau^2)$, is

$$\sqrt{1+\tau^2}\exp\left(-\frac{(x-\theta_0)^2\tau^2}{2(1+\tau^2)}\right).$$

The smallest Bayes factor occurs with value

$$\tau = \begin{cases} (x-\theta_0)^2 - 1 & \text{if } |x-\theta_0| > 1, \\ 0 & \text{otherwise.} \end{cases}$$

with value 1 if $|x - \theta_0| \leq 1$, and

$$|x - \theta_0| \exp\left(-(x - \theta_0)^2 + 1)/2\right),$$

if $|x - \theta_0| > 1$. At $x = \theta_0 + z_{0.025}$, this minimizing factor is 0.4734.

Example. Consider a $0 - 1 - c$ loss function and let X be $Pois(\theta)$ and let Θ have a $\Gamma(a, b)$ prior. Then the posterior, given $X = x$ is $\Gamma(a + x, b + 1)$. For fixed second parameter, the Γ distributions increase stochastically in the first parameter. For the hypothesis,

$$H : \Theta \leq 1 \quad \text{versus} \quad A : \Theta > 1.$$

$$P\{\Theta \leq 1 | X = x\} < \frac{1}{1+c}$$

if and only if $x \geq x_0$ for some x_0.

For the improper prior with $a = 0$ and $b = 0$ (corresponding to the density $d\theta/\theta$) and with $c = 19$, the value of x_0 is 4. This is the same as the UMP level $\alpha = 0.05$ test except for the randomization at $x = 3$.

Example. Consider a $0 - 1 - c$ loss function and let Y be $Exp(\theta)$. Let $X = -Y$ so that θ is the natural parameter. For the hypothesis

$$H : \Theta \leq 1 \text{ or } \Theta \geq 2 \quad \text{versus} \quad A : 1 < \Theta < 2,$$

use the improper prior having density $1/\theta$ with respect to Lebesgue measure. Then, given $Y = y$, the posterior distribution of Θ is $Exp(y)$. To find the formal Bayes rule, note that the posterior probability that H is true is

$$1 - e^{-y} + e^{-2y}.$$

Setting this equal to $1/(1 + c)$ we may obtain zero, one, or two solutions. In the cases of 0 or 1 solution, the formal Bayes always accepts H. For the case of 2 solutions, c_1 and c_2,

$$1 - e^{-c_1} + e^{-2c_1} = 1 - e^{-c_2} + e^{-2c_2} = \frac{1}{1+c},$$

or

$$e^{-c_1} - e^{-c_2} = e^{-2c_1} - e^{-2c_2}.$$

This equates the power function at $\theta = 1$ with the power function at $\theta = 2$. If α is the common value, then the test is UMP level α.

Example. Let X be $N(\mu, 1)$ and consider two different hypotheses:

$$H_1 : M \leq -0.5 \text{ or } M \geq 0.5 \quad \text{versus} \quad A_1 : -0.5 < M < 0.5,$$

and

$$H_2 : M \leq -0.7 \text{ or } M \geq 0.51 \quad \text{versus} \quad A_2 : -0.7 < M < 0.51.$$

A UMP level $\alpha = 0.05$ test of H_1 rejects H_1 if

$$-0.071 < X < 0.071.$$

A UMP level $\alpha = 0.05$ test of H_2 rejects H_2 if

$$-0.167 < X < -0.017.$$

Because $\Omega_{H_2} \subset \Omega_{H_1}$, then one may argue that rejection of H_1 should *a fortiori* imply the rejection of H_2. However, if

$$-0.017 < X < 0.071,$$

then we would reject H_1 at the $\alpha = 0, 05$ level but accept H_2 at the same level.

This lack of coherence cannot happen in the Bayesian approach using levels for posterior probabilities. For example, suppose that we use a improper Lebesgue prior for M, then the posterior probability for M is $N(x, 1)$. The $\alpha = 0.05$ test for H_1 rejects if the posterior probability of H_1 is less than 0.618. The posterior probability of H_2 is less than 0.618 whenver $x \in (-0.72, 0.535)$. Note that this contains the rejection region of H_2.

Example. Let X be $N(\theta, 1)$ and consider the hypothesis

$$H : |\Theta - \theta_0| \leq \delta \quad \text{versus} \quad A : |\Theta - \theta_0| > \delta$$

Suppose that the prior is $N(\theta_0, \tau^2)$. Then the posterior distribution of Θ given $X = x$ is

$$N(\theta_1, \frac{\tau^2}{1+\tau^2}), \quad \theta_1 = \frac{\theta_0 + x\tau^2}{1+\tau^2}.$$

If we use a $0 - 1 - c$ loss function, the Bayes rule is to reject H if the posterior probability

$$\int_{\theta_0-\delta}^{\theta_0+\delta} \sqrt{\frac{1+1/\tau^2}{2\pi}} \exp\left(-\frac{1}{2}(1+\frac{1}{\tau^2})(\theta-\theta_1)^2\right)\, d\theta = \int_{\theta_0-\theta_1-\delta}^{\theta_0-\theta_1+\delta} \sqrt{\frac{1+1\tau^2}{2\pi}} \exp(-\frac{1}{2}(1+\frac{1}{\tau^2})\theta^2)\, d\theta$$

is low. This integral is a decreasing function of

$$|\theta_0 - \theta_1| = |\theta_0 - \frac{\theta_0 + x\tau^2}{1+\tau^2}| = \frac{\tau^2}{1+\tau^2}|\theta_0 - x|.$$

Thus, the Bayes rule is to reject H if

$$|\theta_0 - x| > d$$

for some d. This has the same form as the UMPU test.

Alternatively, suppose that

$$P\{\Theta = \theta_0\} = p_0 > 0,$$

and, conditioned on $\Theta = \theta_0$, Θ is $N(\theta_0, \tau^2)$. Then the Bayes factor is

$$\sqrt{1+\tau^2} \exp\left(-\frac{(x-\theta_0)^2\tau^2}{2(1+\tau^2)}\right).$$

Again, the Bayes rule has the same form as the UMPU test.

If, instead we choose the conditional prior Θ is $N(\tilde{\theta}, \tau^2)$ whenever $\Theta \neq \theta_0$. This gives rise to a test that rejects H whenever

$$|((1-\tau^2)\theta_0 - \tilde{\theta})/\tau^2 - x| > d.$$

This test is admissible, but it is not UMPU if $\tilde{\theta} \neq \theta_0$. Consequently, the class of UMPU tests is not complete.

Example. Let X be $Bin(n, p)$. Then $\theta = \log(p/(1-p))$ is the natural parameter. If we choose the conditional prior for P to be $Beta(\alpha, \beta)$, then the Bayes factor for $\Omega_H = \{p_0\}$ is

$$\frac{p_0^x(1-p_0)^{n-x}\prod_{i=0}^{n-1}(\alpha+\beta+i)}{\prod_{i=0}^{x-1}(\alpha+i)\prod_{j=0}^{n-x-1}(\beta+j)}.$$

7 Estimation

7.1 Point Estimation

Definition. Let Ω be a parameter space and let $g : \Omega \to G$ be a measurable function. A measurable function

$$\phi : \mathcal{X} \to G' \quad G \subset G',$$

is called a *(point) estimator* of $g(\theta)$. An estimator is called *unbiased* if

$$E_\theta[\phi(X)] = g(\theta).$$

If G' is a vector space, then the *bias*

$$b_\phi(\theta) = E_\theta[\phi(X)] - g(\theta).$$

Example.

1. For $X_1, \cdots, X_n$ independent $N(\mu, \sigma)$ random varibles,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator of μ.

2. Let X be an $Exp(\theta)$ random variable. For ϕ to be an unbiased estimator of θ, then

$$\theta = E_\theta[\phi(X)] = \int_0^\infty \phi(x)\theta e^{-\theta x}\ dx,$$

for all θ. Dividing by θ and differentiating the Riemann integral with respect to θ yields

$$0 = \int_0^\infty x\phi(x)\theta e^{-\theta x}\ dx = \frac{1}{\theta}E_\theta[X\phi(X)].$$

Because X is a complete sufficient statistic

$$\phi(X) = 0 \text{ a.s. } P_\theta \text{ for all } \theta.$$

This contradicts the assumption that ϕ is unbiased and consequently θ has no unbiased estimators.

Using a quadratic loss function, and assuming that G' is a subset of R, the risk function of an estimator ϕ is

$$R(\theta, \phi) = E_\theta[(g(\theta) - \phi(X))^2] = b_\phi(\theta)^2 + \text{Var}_\theta \phi(X).$$

This suggests the following criterion for unbiased estimators.

Definition. An unbiased estimator ϕ is *uniformly minimum variance unbiased estimator (UMVUE)* if $\phi(X)$ has finite variance for all θ and, for every unbiased estimator ψ,

$$\text{Var}_\theta \phi(X) \leq \text{Var}_\theta \psi(X) \text{ for all } \theta.$$

Theorem. (Lehmann-Scheffé) Let T be a complete sufficient statistic. Then all unbiased estimators of $g(\theta)$ that are functions of T alone are equal a.s. P_θ for all $\theta \in \Omega$.

If an unbiased estimator is a function of a complete sufficient statistic, then it is UMVUE.

Proof. Let $\phi_1(T)$ and $\phi_2(T)$ be two unbiased estimators of $g(\theta)$, then

$$E_\theta[\phi_1(T) - \phi_2(T)] = 0 \quad \text{for all } \theta.$$

Because T is complete,

$$\phi_1(T) = \phi_2(T), \text{ a.s. } P_\theta.$$

If $\phi(X)$ is an unbiased estimator with finite variance then so is $\tilde{\phi}(T) = E_\phi[\phi(X)|T]$. The conditional variance formula states that

$$\text{Var}_\theta(\tilde{\phi}(T)) \leq \text{Var}_\theta(\phi(X)).$$

and $\tilde{\phi}(T)$ is UMVUE.

Example. Let $X_1, \cdots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Then $(\bar{X}, S^2)$ is a complete sufficent statistic. The components are unbiased estimators of μ and σ^2 respectively. Thus, they are UMVUE.

Define

$$\mathcal{U} = \{U : E_\theta[U(X)] = 0, \text{ for all } \theta\}.$$

If δ_0 is an unbiased estimator of $g(\theta)$, then every unbiased estimator of $g(\theta)$ has the form $\delta_0 + U$, for some $U \in \mathcal{U}$.

Theorem. An estimator δ is UMVUE of $E_\theta[\delta(X)]$ if and only if, for every $U \in \mathcal{U}$, $\text{Cov}_\theta(\delta(X), U(X)) = 0$.

Proof. (sufficiency) Let $\delta_1(X)$ be an unbiased estimator of $E_\theta[\delta(X)]$. Then, there exists $U \in \mathcal{U}$ so that $\delta_1 = \delta + U$. Because $\text{Cov}_\theta(\delta(X), U(X)) = 0$,

$$\text{Var}_\theta(\delta_1(X)) = \text{Var}_\theta(\delta(X)) + \text{Var}_\theta(U(X)) \geq \text{Var}_\theta(\delta(X)),$$

and $\delta(X)$ is UMVUE.

(necessity) For $\lambda \in R$ define the unbiased estimator

$$\delta_\lambda = \delta + \lambda U.$$

Then

$$\text{Var}_\theta(\delta(X)) \leq \text{Var}_\theta(\delta_\lambda(X)) = \text{Var}_\theta(\delta(X)) + 2\lambda \text{Cov}_\theta(\delta(X), U(X)) + \lambda^2 \text{Var}_\theta(U(X)),$$

or

$$\lambda^2 \mathrm{Var}_\theta(U(X)) \geq -2\lambda \mathrm{Cov}_\theta(\delta(X), U(X)).$$

This holds for all λ and θ if and only if $\mathrm{Cov}_\theta(\delta(X), U(X)) = 0$.

Example. Let $Y_1, Y_2, \cdots$ be independent $Ber(\theta)$ random variables. Set

$$X = \begin{cases} 1 & \text{if } Y_1 = 1 \\ \text{number of trials before 2nd failure} & \text{otherwise.} \end{cases}$$

Suppose that we observe X. Then

$$f_{X|\Theta}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ \theta^{x-2}(1-\theta)^2 & \text{if } x = 2, 3, \cdots. \end{cases}$$

(Note that the failures occur on the first and last trial.)

Define the estimator

$$\delta_0(x) = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{if } x = 2, 3, \cdots. \end{cases}$$

Then, δ_0 is an unbiased estimator of Θ. To try to find an UMVUE estimator, set

$$\begin{aligned} 0 &= E_\theta[U(X)] \\ &= U(1)\theta + \sum_{x=2}^{\infty} \theta^{x-2}(1-\theta)^2 U(x) \\ &= U(2) + \sum_{k=1}^{\infty} \theta^k (U(k) - 2U(k+1) + U(k+2)) \end{aligned}$$

Then $U(2) = 0$ and $U(k) = (2-k)U(1)$ for all $k \geq 3$. Thus, we can characterize $\mathcal{U}$ according to its value at $t = -U(1)$. Thus,

$$\mathcal{U} = \{U_t : U_t(x) = (x-2)t, \text{ for all } x\}.$$

Consequently, the unbiased estimators of Θ are

$$\delta_t(x) = \delta_0(x) + (x-2)t.$$

The choice that is UMVUE must have 0 covariance with every $U_s \in \mathcal{U}$. Thus, for all s and θ,

$$0 = \sum_{x=1}^{\infty} f_{X|\Theta}(x|\theta)\delta_t(x)U_s(x) = \theta(-s)(1-t) + \sum_{x=2}^{\infty} \theta^{x-2}(1-\theta)^2 ts(x-2)^2.$$

or

$$\sum_{x=2}^{\infty} ts\theta^{x-2}(x-2)^2 = s(1-t)\frac{\theta}{(1-\theta)^2} = s(1-t)\sum_{k=1}^{\infty} k\theta^k.$$

These two power series must be equal term by term. Thus,

$$s(1-t)k = tsk^2 \quad \text{or} \quad 1 - t = tk.$$

Thus, there is no UMVUE.

Given θ_0, there is a locally minimum variance unbiased estimator.

Theorem. (Cramér-Rao lower bound) Suppose that the three FI regularity conditions hold and let $\mathcal{I}_X(\theta)$ be the Fisher information. Suppose that $\mathcal{I}_{\mathcal{X}}(\theta) > 0$, for all θ. Let $\phi(X)$ be a real valued statistic satisfying $E|\phi(X)| < \infty$. for all θ and $\int \phi(x) f_{X|\Theta}(x|\theta)\ \nu(dx)$ can be differentiated under the integral sign. Then,

$$\mathrm{Var}_\theta(\phi(X)) \geq \frac{(\frac{d}{d\theta}E_\theta\phi(X))^2}{\mathcal{I}_X(\theta)}.$$

Proof. Define

$$B = \{x : \frac{\partial f_{X|\Theta}(x|\theta)}{\partial\theta} \text{ fails to exist for some } \theta\} \quad \text{and} \quad C = \{x : f_{X|\Theta}(x|\theta) > 0\}.$$

Then $\nu(B) = 0$ and C is independent of θ. Let $D = C \cap B^c$, then, for all θ,

$$1 = P_\theta(D) = \int_D \phi(x) f_{X|\Theta}(x|\theta)\ \nu(dx).$$

Taking the derivative, we obtain

$$0 = \int_D \frac{\partial f_{X|\Theta}(x|\theta)/\partial\theta}{f_{X|\Theta}(x|\theta)} f_{X|\Theta}(x|\theta)\ \nu(dx) = E_\theta[\frac{\partial}{\partial\theta}\log f_{X|\Theta}(X|\theta)].$$

Using this fact, we also have

$$\begin{aligned}\frac{d}{d\theta}E_\theta[\phi(X)] &= \int_D \phi(x)\frac{\partial}{\partial\theta} f_{X|\Theta}(x|\theta)\ \nu(d\theta) \\ &= E_\theta[\phi(X)\frac{\partial}{\partial\theta}\log f_{X|\Theta}(X|\theta)] = E_\theta[(\phi(X) - E_\theta[\phi(X)])\frac{\partial}{\partial\theta}\log f_{X|\Theta}(X|\theta)].\end{aligned}$$

By the Cauchy-Schwartz inequality,

$$\left(\frac{d}{d\theta}E_\theta[\phi(X)]\right)^2 \leq \mathrm{Var}_\theta(\phi(X))\mathcal{I}_X(\theta).$$

Now, divide by $\mathcal{I}_X(\theta)$.

For an unbiased estimator, $\frac{d}{d\theta}E_\theta[\phi(X)] = 1$.

Equality in the Cauchy-Schwartz occurs if and only if the estimator $\phi(X)$ and the score function $\partial \log f_{X|\Theta}(X|\theta)/\partial\theta$ are linearly related.

$$\frac{\partial}{\partial\theta}\log f_{X|\Theta}(x|\theta) = a(\theta)\phi(x) + d(\theta) \quad \text{a.s. } P_\theta.$$

or

$$f_{X|\Theta}(x|\theta) = c(\theta)h(x)\exp\left(\pi(x)\phi(x)\right),$$

the density of an exponential family with sufficient statistic ϕ.

Examples.

1. For X a $N(\theta, \sigma_0^2)$ random variable and $\phi(x) = x$,
$$\mathcal{I}_{\mathcal{X}}(\theta) = 1/\theta_0^2$$
and the Cramér-Rao bound is met.

2. For X an $Exp(\lambda)$ random variable, set $\theta = 1/\lambda$. Then
$$f_{X|\Theta}(x|\theta) = \frac{1}{\theta}e^{-x/\theta}I_{(0,\infty)}(x), \quad \frac{\partial}{\partial\theta}\log f_{X|\Theta}(x|\theta) = -\frac{1}{\theta} + \frac{x}{\theta^2}.$$
Thus ϕ must be a linear function to achieve the Cramér-Rao bound. The choice $\phi(x) = x$ is unbiased and the bound is achieved.

3. $t_d(\theta, 1)$-family of distributions has density
$$f_{X|\Theta}(x|\theta) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\sqrt{d\pi}}\left(1 + \frac{1}{d}(x-\theta)^2\right)^{-(d+1)/2}$$
with respect to Lebesgue measure. In order for the variance to exist, we must have that $d \geq 3$. Check that
$$\frac{\partial^2}{\partial\theta^2}\log f_{X|\Theta}(x|\theta) = -\frac{d+1}{d}\frac{1-(x-\theta)^2/d}{(1+(x-\theta)^2/d)^2}.$$
Then
$$\mathcal{I}_{\mathcal{X}}(\theta) = -E_\theta[\frac{\partial^2}{\partial\theta^2}\log f_{X|\Theta}(X|\theta)] = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\sqrt{d\pi}}\frac{d+1}{d}\int\frac{1-(x-\theta)^2/d}{(1+(x-\theta)^2/d)^{(d+5)/2}}\,dx.$$
Make the change of variables
$$\frac{z}{\sqrt{d+4}} = \frac{x-\theta}{\sqrt{d}}$$
to obtain
$$\mathcal{I}_{\mathcal{X}}(\theta) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\sqrt{d\pi}}\frac{d+1}{d}\sqrt{\frac{d}{d+4}}\int\frac{1-z^2/(d+4)}{(1+z^2/(d+4))^{(d+5)/2}}\,dz.$$
Multiplication of the integral by
$$\frac{\Gamma((d+5)/2)}{\Gamma((d+4)/2)\sqrt{(d+4)\pi}}$$
gives
$$E\left[1 - \frac{T_{d+4}^2}{d+4}\right] = 1 - \frac{1}{d+4}\frac{d+4}{d+2} = \frac{d+1}{d+2}.$$
Therefore,
$$\begin{aligned}
\mathcal{I}_{\mathcal{X}}(\theta) &= \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\sqrt{d\pi}}\frac{d+1}{d}\sqrt{\frac{d}{d+4}}\frac{\Gamma((d+4)/2)\sqrt{(d+4)\pi}}{\Gamma((d+5)/2)}\frac{d+1}{d+2} \\
&= \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\frac{\Gamma((d+4)/2)}{\Gamma((d+5)/2)}}\frac{d+1}{d}\frac{d+1}{d+2} \\
&= \frac{(d+2)/2\ d/2}{(d+3)/2\ (d+1)/2}\frac{d+1}{d}\frac{d+1}{d+2} = \frac{d+1}{d+3}
\end{aligned}$$

Theorem. (Chapman-Robbins lower bound). Let

$$m(\theta) = E_\theta[\phi(X)], \quad S(\theta) = \text{supp} f_{X|\Theta}(\cdot|\theta).$$

Assume for each $\theta \in \Omega$ there exists $\theta' \neq \theta$ such that $S(\theta') \subset S(\theta)$. Then

$$\text{Var}_\theta(\phi(X)) \geq \sup_{\{\theta': S(\theta') \subset S(\theta)\}} \left(\frac{(m(\theta) - m(\theta'))^2}{E_\theta[(f_{X|\Theta}(X|\theta')/f_{X|\Theta}(X|\theta) - 1)^2]} \right).$$

Proof. Define

$$U(X) = \frac{f_{X|\Theta}(X|\theta')}{f_{X|\Theta}(X|\theta)} - 1.$$

Then $E_\theta[U(X)] = 1 - 1 = 0$. Choose θ' so that $S(\theta') \subset S(\theta)$, then

$$\begin{aligned}
\sqrt{\text{Var}_\theta(\phi(X))}\sqrt{\text{Var}_\theta(U(X))} &\geq |\text{Cov}_\theta(U(X), \phi(X))| \\
&= |\int_{S(\theta)} (\phi(x) f_{X|\Theta}(x|\theta') - \phi(x) f_{X|\Theta}(x|\theta))\ \nu(dx)| = |m(\theta) - m(\theta')|.
\end{aligned}$$

Examples.

1. Let $X_1, \cdots, X_n$ be independent with density function

$$f_{X|\Theta}(x|\theta) = \exp(\theta - x) I_{(\theta,\infty)}$$

with respect to Lebesgue measure. Thus, $S(\theta') \subset S(\theta)$ whenever $\theta' \geq \theta$.

$$U(X) = \exp(n(\theta' - \theta)) I_{(\theta',\infty)}(\min_{1 \leq i \leq n} X_i) - 1,$$

and

$$E_\theta[U(X)^2] = (\exp(2n(\theta' - \theta)) - 2\exp(n(\theta' - \theta))) P_\theta\{\min_{1 \leq i \leq n} X_i \geq \theta'\} + 1.$$

Because

$$P_\theta\{\min_{1 \leq i \leq n} X_i \geq \theta'\} = P_\theta\{X_1 \geq \theta'\}^n = \exp(-n(\theta' - \theta)),$$

we have

$$E_\theta[U(X)^2] = \exp(n(\theta' - \theta)) - 1.$$

Thus, the Chapman-Robbins lower bound for an unbiased estimator is

$$\text{Var}_\theta(\phi(X)) \geq \sup_{\theta' \geq \theta} \left(\frac{(\theta - \theta')^2}{\exp(n(\theta' - \theta)) - 1} \right) = \frac{1}{n^2} \min_{t \geq 0} \left(\frac{t^2}{e^t - 1} \right).$$

2. For X a $U(0,\theta)$ random variable, $S(\theta') \subset S(\theta)$ whenever $\theta' \leq \theta$.

$$U(X) = \frac{\theta'^{-1} I_{(0,\theta')}(X)}{\theta^{-1} I_{(0,\theta)}(X)} - 1 = \frac{\theta}{\theta'} I_{(0,\theta')}(X) - 1,$$

and

$$E_\theta[U(X)^2] = ((\frac{\theta}{\theta'})^2 - 2\frac{\theta}{\theta'})P_\theta\{X \leq \theta'\} + 1 = \frac{\theta}{\theta'} - 1.$$

Thus, the Chapman-Robbins lower bound for an unbiased estimator is

$$\mathrm{Var}_\theta(\phi(X)) \geq \sup_{\theta' \leq \theta} \left(\frac{(\theta - \theta')^2}{(\theta/\theta') - 1} \right) == \sup_{\theta' \leq \theta} \theta'(\theta - \theta') = \frac{\theta^2}{4}$$

Lemma. Let $\phi(X)$ be an unbiased estimator of $g(\theta)$ and let $\psi(x,\theta),\ i = 1, \cdots, k$ be functions that are not linearly related. Set

$$\gamma_i = \mathrm{Cov}_\theta(\phi(X), \psi_i(X,\theta)), \quad C_{ij} = \mathrm{Cov}_\theta(\psi_i(X,\theta), \psi_j(X,\theta)).$$

Then $\mathrm{Var}_\theta \phi(X) \geq \gamma^T C^{-1} \gamma$.

Proof. The covariance matrix of $(\phi(X), \psi_1(X), \cdots, \psi_k(X))$ is

$$\begin{pmatrix} \mathrm{Var}_\theta \phi(X) & \gamma^T \\ \gamma & C \end{pmatrix}$$

Use the vector $(1, \alpha^T)$ to see that

$$\mathrm{Var}_\theta \phi(X) + \alpha^T \gamma + \gamma^T \alpha + \alpha^T C \alpha \geq 0.$$

Now the inequality follows by taking $\alpha = -C^{-1}\gamma$

Corollary (Bhattacharyya system of lower bounds). In addition to the hypotheses for the Cramér-Rao lower bound, assume that k partial derivatives can be performed under the integral sign. Define

$$\gamma_i(\theta) = \frac{d^i}{d\theta^i} E_\theta[\phi(X)], \quad J_{ij}(\theta) = \mathrm{Cov}_\theta(\psi_i(X,\theta), \psi_j(X,\theta)), \quad \psi_i(x,\theta) = \frac{1}{f_{X|\Theta}(x|\theta)} \frac{\partial^i}{\partial \theta^i} f_{X|\Theta}(x|\theta).$$

Assume that $J(\theta)$ is a nonsingular matrix, then

$$\mathrm{Var}_\theta(\phi(X)) \geq \gamma(\theta)^T J(\theta)^{-1} \gamma(\theta).$$

Proof. Note that by continuing to differentiate, we obtain

$$\frac{d^i}{d\theta^i} E_\theta[\phi(X)] = \mathrm{Cov}_\theta(\phi(X), \frac{\partial^i}{\partial \theta^i} f_{X|\Theta}(X|\theta)).$$

Now, apply the lemma.

Example. Let X be an $Exp(\lambda)$ random variable. Set $\theta = 1/\lambda$. Then, with respect to Lebesgue measure on $(0,\infty)$, X has density

$$f_{X|\Theta}(x|\theta) = \frac{1}{\theta}e^{-x/\theta}.$$

Thus,

$$\frac{\partial}{\partial\theta}f_{X|\Theta}(x|\theta) = \left(-\frac{1}{\theta} + \frac{x}{\theta^2}\right) f_{X|\Theta}(x|\theta), \quad \frac{\partial^2}{\partial\theta^2}f_{X|\Theta}(x|\theta) = \left(\frac{2}{\theta^2} - \frac{4x}{\theta^3} + \frac{x^2}{\theta^5}\right) f_{X|\Theta}(x|\theta).$$

With, $\phi(x) = x^2$,

$$E_\theta[\phi(X)] = 2\theta^2 \quad \text{and} \quad \mathrm{Var}_\theta(\phi(X)) = 20\theta^4.$$

Because

$$\mathcal{I}_{\mathcal{X}}(\theta) = \frac{1}{\theta^2} \quad \text{and} \quad \frac{d}{d\theta}E_\theta[\phi(X)] = 4\theta.$$

the Cramér-Rao lower bound on the variance of $\phi(X)$ is $16\theta^4$. However,

$$J(\theta) = \begin{pmatrix} \frac{1}{\theta^2} & 0 \\ 0 & \frac{4}{\theta^2} \end{pmatrix}, \qquad \gamma(\theta) = \begin{pmatrix} 4\theta \\ 4 \end{pmatrix}.$$

Thus,

$$\gamma(\theta)^T J(\theta)^{-1}\gamma(\theta) = 20\theta^4$$

and the Bhattacharayya lower bound is achieved.

Corollary (Multiparameter Cramér-Rao lower bound.) Assume the FI regularity conditions and let $\mathcal{I}_{\mathcal{X}}(\theta)$ be a positive definite Fisher information matrix. Suppose that $E_\theta|\phi(X)| < \infty$ and that $\int \phi(x)f_{X|\Theta}(x|\theta)\ \nu(dx)$ can be differentiated twice under the integral sign with respect to the coordinates of θ. Set

$$\gamma_i(\theta) = \frac{\partial}{\partial\theta_i}E_\theta[\phi(X)].$$

Then

$$\mathrm{Var}_\theta(\phi(X)) \geq \gamma(\theta)^T\mathcal{I}_{\mathcal{X}}(\theta)^{-1}\gamma(\theta).$$

Proof. Apply the lemma, noting that

$$\frac{\partial}{\partial\theta_i}E_\theta[\phi(X)] = \mathrm{Cov}_\theta(\phi(X), \frac{\partial}{\partial\theta_i}\log f_{X|\Theta}(X|\theta)).$$

Example. Let X be $N(\mu,\sigma^2)$. Then

$$\mathcal{I}_{\mathcal{X}} = \begin{pmatrix} \frac{2}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}.$$

If $\phi(X) = X^2$ then $E_\theta[\phi(X)] = \mu^2 + \sigma^2$. Then

$$\gamma(\mu, \sigma) = (2\mu, 2\sigma) \quad \gamma(\theta)^T \mathcal{I}_{\mathcal{X}}(\theta)^{-1}\gamma(\theta) = 4\mu^2\sigma^2 + 2\sigma^4.$$

This equals $\mathrm{Var}_\theta(\phi(X))$ and so the Cramér-Rao bound is met.

Definition. Let X be a sample from $P_\theta,\ \theta \in \Omega$ and assume that P_θ has density $f_{X|\Theta}(\cdot|\theta)$ with respect to a σ-finite measure ν.

1. If $X = x$ is observed, then the function

   $$L(\theta) = f_{X|\Theta}(x|\theta)$$

   is called the *likelihood function.*

2. $\hat{\theta}$ in the closure of Ω is called the *maximum likelihood estimate* of θ if

   $$L(\hat{\theta}) = \max_{\theta \in \bar{\Omega}} f_{X|\Theta}(x|\theta).$$

   Viewed as a function of x, $L(\hat{\theta})$ is called the *maximum likelihood estimator* of θ.

Theorem. Let $g : \Omega \to G$ be measurable. Suppose that there exists a space U and a function g^* so that

$$(g, g^*) : \Omega \to G \times U$$

is a one-to-one measurable function. If $\hat{\theta}$ is a maximum likelihood estimator of θ, then $g(\hat{\theta})$ is a maximum likelihood estimator of $g(\theta)$.

Proof. For ψ in the range of (g, g^*), set the likelihood

$$f_{X|\Psi}(x|\psi) = f_{X|\Theta}(x|(g, g^*)^{-1}(\psi)).$$

Fix x and let $f_{X|\Theta}(x|\theta)$ assume its maximum at $\hat{\theta}$. Define $\hat{\psi} = (g, g^*)(\hat{\theta})$. Then, the maximum occurs at $(g, g^*)^{-1}(\hat{\psi})$.

If the maximum of $f_{X|\Psi}(x|\psi)$ occurs at $\psi = \psi_0$, then

$$f_{X|\Theta}(x|(g, g^*)^{-1}(\psi_0)) = f_{X|\Psi}(x|\psi_0) \geq f_{X|\Psi}(x|\hat{\psi}) = f_{X|\Theta}(x|\hat{\theta}).$$

Because the last term is at least as large as the first, we have that $\hat{\psi}$ provides a maximum.

Consequently, $(g, g^*)(\hat{\theta})$ is a maximum likelihood estimate of ψ and so $g(\hat{\theta})$ is the MLE of $g(\theta)$, the first coordinate of ψ.

Example. Let $X_1 \cdots, X_n$ be independent $N(\mu, \sigma^2)$, then the maximum likelihood estimates of (μ, σ) are

$$\bar{X} \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

For $g(m) = m^2$, define $g^*(m, \sigma) = (\mathrm{sign}(\mu), \sigma)$, to see that $\bar{X}^2$ is a maximum likelihood estimator of m^2.

For the maximum, we have that $L(\hat{\theta}) > 0$. Thus, we can look to maximize $\log L(\theta)$. In an exponential family, using the natural parameter,

$$\log L(\theta) = \log c(\theta) + \langle x, \theta \rangle.$$

Thus, if $\partial L(\theta)/\partial \theta_i = 0$,

$$x_i = \frac{\partial}{\partial \theta_i} \log c(\theta) = E_\theta[X_i],$$

and $\hat{\theta}_i$ is chosen so that $x_i = E_\theta[X_i]$.

Example. Let $X_1 \cdots, X_n$ be independent $N(\mu, \sigma^2)$, then the natural parameter for this family is

$$(\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}).$$

The natural sufficient statistic is

$$(n\bar{X}, \sum_{i=1}^n X_i^2).$$

Also,

$$\log c(\theta) = \frac{n}{2} \log(-2\theta_2) + n\frac{\theta_1^2}{4\theta_2}.$$

Thus,

$$\frac{\partial}{\partial \theta_1} \log c(\theta) = \frac{\theta_1}{2\theta_2}, \qquad \frac{\partial}{\partial \theta_2} \log c(\theta) = \frac{n}{2\theta_2} - n\frac{\theta_1^2}{4\theta_2^2}.$$

Setting these equal to the negative of the coordinates of the sufficient statistic and solving for (θ_1, θ_2) gives

$$\hat{\theta}_1 = \frac{n\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\theta}_2 = -\frac{n}{2\sum_{i=1}^n (X_i - \bar{X})^2},$$

and

$$\hat{\mu} = -\frac{\hat{\theta}_1}{2\hat{\theta}_2} = \bar{X}, \quad \hat{\sigma}^2 = -\frac{1}{2\hat{\theta}_2} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

Given a loss function, the Bayesian method of estimation uses some fact about the posterior distribution. For example, if Ω is one dimensional, and we take a quadratic loss function, the the Bayes estimator is the posterior mean.

For the loss function $L(\theta, a) = |\theta - a|$ the rule is a special case of the following theorem.

Theorem. Suppose that Θ has finite posterior mean. For the loss function

$$L(\theta, a) = \begin{cases} c(a - \theta) & \text{if } a \geq \theta, \\ (1 - c)(\theta - a) & \text{if } a < \theta, \end{cases}$$

then a formal Bayes rule is any $1 - c$ quantile of the posterior distribution of Θ.

Proof. Let $\tilde{a}$ be a $1-c$ quantile of the posterior distribution of Θ. Then

$$P\{\Theta \le \tilde{a}|X=x\} \ge 1-c, \quad P\{\Theta \ge \tilde{a}|X=x\} \ge c.$$

If $a > \tilde{a}$, then

$$L(\theta,a) - L(\theta,\tilde{a}) = \begin{cases} c(a-\tilde{a}) & \text{if } \tilde{a} \ge \theta, \\ c(a-\tilde{a}) - (\theta - \tilde{a}) & \text{if } a \ge \theta > \tilde{a}, \\ (1-c)(\tilde{a}-a) & \text{if } \theta > a, \end{cases}$$

or

$$L(\theta,a) - L(\theta,\tilde{a}) = c(a-\tilde{a}) + \begin{cases} 0 & \text{if} \tilde{a} \ge \theta, \\ (\tilde{a}-\theta) & \text{if } a \ge \theta > \tilde{a}, \\ (\tilde{a}-a) & \text{if } \theta > a, \end{cases}$$

Thus, the difference in posterior risk is

$$\begin{aligned} r(a|x) - r(\tilde{|}x) &= c(a-\tilde{a}) + \int_{(\tilde{a},a]} (\tilde{a}-\theta) f_{\Theta|X}(\theta|x)\ \lambda(dx) + (\tilde{a}-a)P\{\Theta > a|X=x\} \\ &\ge c(a-\tilde{a}) + (\tilde{a}-a)P\{\Theta > a|X=x\} = (a-\tilde{a})(c - P\{\Theta > a|X=x\}) \ge 0. \end{aligned}$$

Similarly, if $a < \tilde{a}$, then

$$L(\theta,a) - L(\theta,\tilde{a}) = c(a-\tilde{a}) + \begin{cases} 0 & \text{if } \tilde{a} \ge \theta, \\ (\theta-a) & \text{if } \tilde{a} \ge \theta > a, \\ (\tilde{a}-a) & \text{if } \theta > a, \end{cases}$$

and

$$r(a|x) - r(\tilde{|}x) \ge (\tilde{a}-a)(P\{\Theta > a|X=x\} - c) \ge 0.$$

Consequently, $\tilde{a}$ provides the minimum posterior risk.

Note that if $c = 1/2$, then the Bayes is to take the median

7.2 Nonparametric Estimation

Let $\mathcal{P}_0$ be a collection of distributions on a Borel space $(\mathcal{X},\mathcal{B})$ and let

$$T : \mathcal{P}_0 \to R^k$$

be a functional. If we collect data $X_1,\cdots,X_n$, then we may estimate $P \in \mathcal{P}_0$ by its empirical distribution P_n. In this circumstance, the natural estimator of $T(P)$ is $T(P_n)$.

For example, if

$$T(P) = \int_{\mathcal{X}} \psi(x)\ P(dx),$$

then

$$T(P_n) = \int_{\mathcal{X}} \psi(x)\ P_n(dx) = \frac{1}{n}\sum_{i=1}^{n} \psi(X_n).$$

Methods of moments techniques are examples of this type of estimator with $\psi(x) = x^n$.

If $\mathcal{X} \subset R$, then we can look at the p-th quantile of P.

$$T(P) = \inf\{x : P[x, \infty) \geq p\}$$

Then $T(P_n)$ is the pth sample quantile.

Definition. Let $\mathcal{L}$ be a linear topological vector space and let

$$T : \mathcal{L} \to R^k$$

be a functional on $\mathcal{L}$.

1. T is *Gâteaux differentiable* at $Q \in \mathcal{L}$ if there is a linear functional $L(Q; \cdot)$ on $\mathcal{L}$ such that for $\Delta \in \mathcal{L}$,

$$\lim_{t \to 0} \frac{1}{t}(T(Q + t\Delta) - T(Q)) = L(Q; \Delta).$$

2. If $\mathcal{L}$ is a metric space with metric ρ, T is *Fréchet differentiable* at $P \in \mathcal{L}$ if there is a linear functional $L(P; \cdot)$ on $\mathcal{L}$ such that for $\{P_j; j \geq 0\}$ converging to P,

$$\lim_{j \to \infty} \frac{1}{\rho(P_j, P)}(T(P_j) - T(P) - L(P; P_j - P)) = 0.$$

3. If $\mathcal{L}$ is a Banach space with norm $|| \cdot ||$, T is *Hadamard differentiable* at $Q \in \mathcal{L}$ if there is a linear functional $L(Q; \cdot)$ on $\mathcal{L}$ such that for for any sequence of numbers $\{t_j; j \geq 0\}$ converging to 0, and $\{\Delta_j; j \geq 0\}$ converging in norm to Δ,

$$\lim_{j \to \infty} \frac{1}{t_j}(T(Q + t_j\Delta_j) - T(Q)) - L(Q; \Delta_j) = 0.$$

The functional $L(Q; \cdot)$ is call the *differential* of T at Q and is sometimes written $DT(Q; \cdot)$

One approach to robust estimation begins with the following definition.

Definition. Let $\mathcal{P}_0$ be a collection of distributions on a Borel space $(\mathcal{X}, \mathcal{B})$ and let

$$T : \mathcal{P}_0 \to R^k$$

be a functional. Let $\mathcal{L}$ be the linear span of the distributions in $\mathcal{P}_0$ The *influence function* of T at P is the Gâteaux derivative

$$IF(x; T, P) = DT(P; \delta_x - P)$$

for those x for which the limit exists.

Examples.

1. Let $T(P)$ be the mean functional, then

$$T(P + t(\delta_x - P)) = (1-t)T(P) + tx,$$

   and the influence function

$$IF(x; T, P) = x - T(P).$$

2. Let F be the cumulative distribution function for P, and let T be the median. Then,

$$T(P + t(\delta_x - P)) = \begin{cases} F^{-1}(\frac{1/2-t}{1-t}) & \text{if } x < F^{-1}(\frac{1/2-t}{1-t}) \\ x & \text{if } F^{-1}(\frac{1/2-t}{1-t}) \le x \le F^{-1}(\frac{1}{2(1-t)}) \\ F^{-1}(\frac{1}{2(1-t)}) & \text{if } x \le F^{-1}(\frac{1}{2(1-t)}). \end{cases}$$

   If F has a derivative f at the median, then

$$IF(x; T, P) = \frac{1}{2f(F^{-1}(1/2))} \text{sign}(x - F^{-1}(\frac{1}{2})).$$

Definition. The *gross error sensitivity* is

$$\gamma^*(T, P) = \sup_{x \in \mathcal{X}} |IF(x; T, P)|.$$

Example.

1. If T is the mean and the distributions have unbounded support, then

$$\gamma^*(T, P) = \infty.$$

2. If T is the median, then

$$\gamma^*(T, P) = \frac{1}{2f(F^{-1}(1/2))}.$$

   This is one way of arguing that the median is more robust with respect to gross errors than the mean.

We will now discuss some classes of statistical functionals based on independent and identically distributed observations $X_1, \cdots, X_n$. Denote its empirical distribution

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

If T is Gâteaux differentiable at P and P_n is an empirical distribution from an i.i.d. sum, then setting $t = n^{-1/2}$, and $\Delta = \sqrt{n}(P_n - P)$.

$$\begin{aligned} \sqrt{n}(T(P_n) - T(F)) &= DT(P; \sqrt{n}(P_n - P)) + r_n \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n DT(P; \delta_{X_i} - P) + r_n \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n IF(X_i; T, P) + r_n \end{aligned}$$

We can use the central limit theorem on the first part provided that

$$E[IF(X_1;T,P)] = 0 \quad \text{and} \quad \sigma_P^2 = \mathrm{Var}(IF(X_1;T,P)) < \infty.$$

Thus, by Slutsky theorem, as $n \to \infty$

$$\sqrt{n}(T(P_n) - T(P)) \to^{\mathcal{D}} X_\infty$$

where X_∞ is $N(0, \sigma_P^2)$ provided that

$$r_n \to^p 0.$$

Typically, Gâteaux differentiability is too weak to be useful in establishing the necessary convergence in probability.

Definition. Let P be a distribution on R with cumulative distribution function F and let J be a function on $[0,1]$. An *L-functional* is defined as

$$T(P) = \int xJ(F(x))\, dF(x).$$

$T(P_n)$ is called an L-estimator of $T(P)$.

Examples.

1. If $J \equiv 1$, then $T(P_n) = \bar{X}$.
2. If $J(t) = 4t - 2$, then $T(P_n)$ is the U-statistic called *Gini's mean difference*

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|.$$

3. If $J = (b-a)^{-1} I_{(a,b)}(t)$ for some constants $a < b$, then $T(P_n)$ is called the *trimmed sample mean.*

Theorem. Let T be an L-functional, and assume that

$$\mathrm{supp} J \subset [a,b], \quad 0 < a < b < 1.$$

and that the set

$$D = \{x : J \text{ is discontinuous at } F(x)\}$$

has Legesgue measure zero. Then T is Fréchet differentiable at P with influence function

$$IF(x;T,P) = -J(F(x)) + \int F(y)J(F(y))\, dy.$$

Definition. Let P be a probability measure on R^d with cumulative distribution function F and let

$$r : R^d \times R \to R$$

be a Borel function. An *M-functional* $T(P)$ is defined to be a solution of

$$\int r(x, T(P))\ P(dx) = \min_{t \in G} \int r(x,t)\ P(dx).$$

If P_n is the empirical distribution of i.i.d. observations $X_1, \cdots, X_n$, then $T(P_n)$ is called an *M-estimator* of $T(P)$.

Assume that $\psi_P(t) = \partial r(x,t)/\partial t$ exists a.e. and

$$\lambda_P(t) = \int \psi(x,t)\ P(dx) = \frac{\partial}{\partial t} \int r(x,t)\ P(dx).$$

Note that $\lambda_P(T(P)) = 0$.

Examples.

1. If $r(x,t) = (x-t)^p/p,\ 1 \leq p \leq 2$, then $T(P_n)$ is call the *minimum* L^p *distance estimator.* For $p = 2$, $T(P) = \int x\ P(dx)$ is the mean functional and $T(P_n) = \bar{X}$, the sample mean. For $p = 1$, $T(P_n)$ is the sample median.

2. If $\mathcal{P}_0$ is a paramteric family $\{P_\theta : \theta \in \Omega\}$, with densities $\{f_{X|\Theta} : \theta \in \Omega\}$. Set

$$r(x,t) = -\log f_{X|\Theta}(x|\theta).$$

   Then, $T(P_n)$ is a maximum likelihood estimator.

3. The choice

$$r(x,t) = \min\{\frac{1}{2}(x-t)^2, C\}$$

   and the corresponding $T(P_n)$ gives a trimmed sample mean.

4. The choice

$$r(x,t) = \left\{ \begin{array}{ll} \frac{1}{2}(x-t)^2 & |x-t| \leq C \\ C|x-t| - \frac{1}{2}c^2 & |x-t| > C \end{array} \right.$$

   yields an estimtor $T(P_n)$ that is a type of Winsorized sample mean.

Theorem. Let T be an M-functional and assume that ψ is bounded and continuous and that λ_P is differentiable at $T(P)$ with $\lambda_P'(T(P)) \neq 0$. Then T is Hadamard differentiable at P with influence function

$$IF(x; T, P) = \frac{\psi(x, T(P))}{\lambda_P'(T(P))}.$$

7.3 Set Estimation

Definition. Let

$$g : \Omega \to G$$

and let $\mathcal{G}$ be the collection of all subsets of G. The function

$$R : \mathcal{X} \to \mathcal{G}$$

is a *coefficient* γ confidence set for $g(\theta)$ if for every $\theta \in \Omega$

1. $\{x : g(\theta) \in R(x)\}$ is measurable, and
2. $P'_\theta\{g(\theta) \in R(X)\} \geq \gamma$.

The confidence set R is *exact* if $P'_\theta\{g(\theta) \in R(X)\} = \gamma$ for each $\theta \in \Omega$. If

$$\inf_{\theta \in \Omega} P'_\theta\{g(\theta) \in R(X)\} > \gamma,$$

the confidence set is called *conservative.*

Related confidence sets to nonrandomized tests gives us the following.

Proposition. Let $g : \Omega \to G$.

1. For each $y \in G$, let ϕ_y be a level α nonrandomized test of

   $$H : g(\Theta) = y \quad \text{versus} \quad A : g(\Theta) \neq y.$$

   Then R is a coefficient $1 - \alpha$ confidence set for $g(\theta)$. The confidence set R is exact if and only if ϕ_y is α similar for all y.

2. Let R be a coeffieicnet $1 - \alpha$ set for $g(\theta)$. For each $y \in G$, define

   $$\phi_y(x) = \begin{cases} 0 & \text{if } y \in R(x), \\ 1 & \text{otherwise.} \end{cases}$$

   Then, for each y, ϕ_y has level α for the hypothesis given above. The test ϕ_y is α-similar for all y if and only if R is exact.

Example. Let $X_1, \cdots, X_n$ be independent $N(\mu, \sigma^2)$. The usual UMP level α test of

$$H : M = \mu_0 \quad \text{versus} \quad A : M \neq \mu_0$$

is

$$\phi_{\mu_0}(x) = 1 \quad \text{if} \quad \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t^*$$

where $T_{n-1}^{-1}(1 - \alpha/2)$ and T_{n-1} is the cumulative distribution function of the $t_{n-1}(0, 1)$ distribution. This translates into the confidence interval

$$(\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}}).$$

This form of a confidence uses a *pivotal quantity.* A pivotal is a function

$$h : \mathcal{X} \times \Omega \to R$$

whose distribution does not depend on the parameter. The general method of forming a confidence set from a pivotal is to set

$$R(x) = \{\theta : h(x, \theta) \leq F_h^{-1}(\gamma)\}$$

where

$$F_h(c) = P_\theta\{h(X, \theta) \leq c\}$$

does not depend on θ.

We can similarly define *randomized confidence sets*

$$R^* : \mathcal{X} \times \mathcal{G} \to [0, 1]$$

and extend the relationship of similar tests and exact confidence sets to this setting.

Note that a nonrandomized confidence set R can be considered as th trivial randomized confidence set given by

$$R(x, y) = I_{R(x)}(y).$$

The concept of a uniformly most powerful test leads us to the following definition.

Definition. Let R be a coefficient γ confidence set for $g(\theta)$ and let $\mathcal{G}$ be the collection of all subsets of the range G of g. Let

$$B : G \to \mathcal{G}$$

be a function such that $y \notin B(y)$. Then R is *uniformly most accurate (UMA)* coefficient γ against B if for each $\theta \in \Omega$, each $y \in B(g(\theta))$ and each γ confidence set $\tilde{R}$ for $g(\theta)$,

$$P'_\theta\{y \in R(X)\} \leq P'_\theta\{y \in \tilde{R}(X)\}.$$

If R^* is a coefficient γ randomized confidence set for $g(\theta)$, then R^* is *UMA coefficient* γ *randomized against* B if for every coefficient γ randomized confidence set $\tilde{R}^*$, for each $\theta \in \Omega$, each $y \in B(g(\theta))$,

$$E_\theta[R^*(X, y)] \leq E_\theta[\tilde{R}^*(X, y)].$$

Theorem. For $g(\theta) = \theta$ and $B : G \to \mathcal{G}$. Suppose that

$$B^{-1}(\theta) = \{\theta' : \theta \in B(\theta')\} \neq \emptyset$$

for every θ. Let ϕ_θ be a test and define

$$R^*(x, \theta) = 1 - \phi_\theta(x).$$

Then ϕ_θ is UMP level α for the test

$$H : \Theta = \theta \quad \text{versus} \quad A : \Theta \in B^{-1}(\theta)$$

for all θ if and only if R^* is UMA coefficient $1-\alpha$ randomized against B.

Proof. For each $\theta \in \Omega$ let ϕ_θ be UMP for the test above and let $\tilde{R}^*$ be a coefficient $1-\alpha$ randomized confidence set. Define a test

$$\tilde{\phi}(x) = 1 - \tilde{R}^*(x, \theta').$$

Then by the proposition, $\tilde{\phi}$ has level α for $\{\theta'\} = \Omega_H$. Note that $\theta' \in B(\theta)$ implies $\theta \in B^{-1}(\theta')$ and therefore

$$1 - E_\theta[\tilde{R}^*(X, \theta')] = E_\theta[\tilde{\phi}(X)] = \beta_{\tilde{\phi}}(\theta) \leq \beta_{\phi_{\theta'})}(\theta) = E_\theta[\phi_{\theta'}(X)] = 1 - E_\theta[R^*(X, \theta')],$$

and the result follows.

Conversely, suppose that R^* is a UMA coefficient $1-\alpha$ randomized condifence set against B. For each $\theta \in \Omega$ and $\Omega_H = \{\theta\}$, let $\tilde{\phi}_\theta$ be a level α test. Define

$$\tilde{R}^*(X, \theta) = 1 - \tilde{\phi}_\theta(X).$$

Then, $\tilde{R}^*$ is a coefficient $1-\alpha$ randomized confidence set. Let

$$\tilde{\Omega} = \{(\tilde{\theta}, \theta) : \tilde{\theta} \in \Omega, \theta \in B(\tilde{\theta})\} = \{(\theta, \tilde{\theta}) : \theta \in \Omega, \tilde{\theta} \in B^{-1}(\theta)\}.$$

This uses $B^{-1}(\theta) \neq \emptyset$ for all θ. Thus, for $\tilde{\theta} \in B^{-1}(\theta)$,

$$\beta_{\phi_\theta}(\tilde{\theta}) = 1 - E_{\tilde{\theta}}[R^*(X, \theta)] \geq 1 - E_{\tilde{\theta}}[\tilde{R}^*(X, \theta)] \geq \beta_{\tilde{\phi}_\theta}(\tilde{\theta})$$

and therefore ϕ_θ is UMP level α for the test

Example. Let $X_1, \cdots, X_n$ be independent $N(\mu, 1)$ random variables. Let

$$R(X) = (-\infty, \bar{X} + \frac{1}{\sqrt{n}} z^*]$$

where $z^* = \Phi^{-1}(1-\alpha)$. Note that

$$P'_\mu\{\mu \in R(X)\} = 1 - \alpha,$$

and R is an exact $1-\alpha$ confidence set. Now, consider the test

$$\phi(x) = I_{(-\infty, \mu + z^*/\sqrt{n})}(\bar{x}).$$

Then

$$R(x) = \{\mu : \phi_\mu(x) = 0\}$$

and ϕ_μ is the UMP level α test of

$$H : M = \mu \quad \text{versus} \quad A : M < \mu.$$

Here,

$$B^{-1}(\mu) = (\infty, \mu) = \{\tilde{\mu} : \mu \in B(\tilde{\mu})\}, \quad B(\mu) = (\mu, \infty)$$

and R us UMA coefficient $1-\alpha$ against B, i.e. if $\tilde{\mu} < \mu$, then R has a smaller chance of covering $\tilde{\mu}$ than any other coefficient $1-\alpha$ confidence set.

Example. (Pratt). Let $X_1, \cdots, X_n$ be independent $U(\theta-1/2, \theta+1/2)$ random variables. The minimum sufficient statistic $T = (T_1, T_2) = (\min_i X_i, \max_i X_i)$ has density

$$f_{T_1,T_2|\Theta}(t_1, t_2|\theta) = n(n-1)(t_2 - t_2)^{n-2}, \quad \theta - \frac{1}{2} \leq t_1 \leq t_2 \leq \theta + \frac{1}{2}$$

with respect to Lebesgue measure.

Let's look to find the UMA coefficient $1-\alpha$ confidence set against $B(\theta) = (-\infty, \theta)$. Thus, for each θ, we look to find the UMP level α test for

$$H : \Theta \leq \theta \quad \text{versus} \quad A : \Theta > \theta$$

to construct the confidence set. Pick $\tilde{\theta} > \theta$ and considering the inequality

$$f_{T_1,T_2|\Theta}(t_1, t_2|\tilde{\theta}) > k f_{T_1,T_2|\Theta}(t_1, t_2|\theta)$$

from the Neyman-Pearson lemma.

1. For $k < 1$ and $\theta < \tilde{\theta} < \theta + 1$, this inequality holds if

$$t_1 > \tilde{\theta} - \frac{1}{2} \quad \text{or if} \quad t_2 > \theta + \frac{1}{2}.$$

2. For k =1, then the densities are equal on the intersections of their supports.

3. If $\tilde{\theta} \geq \theta + 1$, then the inequality holds if for

$$t_1 > \tilde{\theta} + \frac{1}{2}.$$

For a size α test, we take

$$\phi(t_1, t_2) = \begin{cases} 1 & \text{if } t_2 > \theta + \frac{1}{2} \text{ or } t_1 > \theta + \frac{1}{2} - \alpha^{1/n} \\ 0 & \text{if } t_2 \leq \theta + \frac{1}{2} \text{ and } t_1 > \theta + \frac{1}{2} - \alpha^{1/n}. \end{cases}$$

To check that it is most powerful for each $\tilde{\theta} > \theta$,

1. take $k = 1$ if $\tilde{\theta} - 1/2 < \theta + 1/2 - \alpha^{1/n}$, and
2. take $k = 0$ if $\tilde{\theta} - 1/2 < \theta + 1/2 - \alpha^{1/n}$.

Set

$$T^* = \max\{T_1 - 1/2 + \alpha^{1/n}, T_2 - 1/2\},$$

then the UMA coefficient $1-\alpha$ confidence set against B is $[T^*, \infty)$. Note that

1. $\Theta \geq T_2 - 1/2$, and
2. $T^* \leq T_2 - 1/2$ whenever $T_1 - 1/2 + \alpha^{1/n}$ or $T_2 - T_1 \geq \alpha^{1/n}$ Thus, we are 100% confident that $\Theta \geq T^*$ rather than $100(1-\alpha)$%.

For two-sided or multiparameter confidence sets, we need to extend the concept of unbiasedness.

Definition. Let R be a coefficient γ confidence set for $g(\theta)$. Let $\mathcal{G}$ be the power set of G and let $B : G \to \mathcal{G}$ be a function so that $y \notin B(y)$. Then R is *unbiased* against B if, for each $\theta \in \Omega$,

$$P'_\theta\{y \in R(X)\} \leq \gamma \quad \text{for all } y \in B(g(\theta)).$$

R is a *uniformly most accurate unbiased (UMAU)* coefficient γ confidence set for $g(\theta)$ against B if its UMA against B among unbiased coefficient γ confidence sets.

Proposition. For each $\theta \in \Omega$, let $B(\theta)$ be a subset of Ω such that $\theta \notin B(\theta)$, and let ϕ_θ be a nonrandomized level α test of

$$H : \Theta = \theta \quad \text{versus} \quad A : \Theta \in B^{-1}(\theta).$$

Set

$$R(x) = \{\theta : \phi_\theta(x) = 0\}.$$

Then R is a UMAU coefficient $1-\alpha$ confidence set against B if ϕ_θ is UMPU level α for the hypothesis above.

In the Bayesian framework, we want to choose a set C such that

$$Pr\{V \in C | X = x\} = \gamma.$$

Some of the approaches in choosing C are:

1. If V has posterior density $f_{V|X}(\cdot|x)$, choose d so that

   $$C = \{v : f_{V|X}(v|x) \geq d\}.$$

   This choice, called the *highest posterior density (HPD)* region, is sensitive to the choice of reference measure. Indeed, C may be disconnected if $f_{V|X}(\cdot|x)$ is multi-modal.

2. If V is real valued, choose c_- and c_+ so that

   $$Pr\{V \leq c_- | X = x\} = \frac{1-\gamma}{2} \quad \text{and} \quad Pr\{V \geq c_+ | X = x\} = \frac{1+\gamma}{2}$$

3. For a given loss function, choose C with the smallest posterior expected loss.

To exhibit the use of the loss function in choosing the confidence set, consider a one-dimenstional parameter set Ω. To obtain a bounded connected confidence interval, choose the action space

$$A = \{(a_-, a_+), a_- < a_+\}$$

and the loss function

$$L(\theta, a_-, a_+) = a_+ - a_- + \begin{cases} c_-(a_- - \theta) & \text{if } \theta < a_-, \\ 0 & \text{if } a_- \leq \theta \leq a_+, \\ c_+(\theta - a_+) & \text{if } a_+ < \theta. \end{cases}$$

Theorem. Suppose that the posterior mean of Θ is finite and the loss is as above with c_- and c_+ at least 1. Then the formal Bayes rule is the interval between the $1/c_-$ and $1 - 1/c_+$ quantiles of the posterior distribution of Θ.

Proof. Write the loss function above by $L_- + L_+$ where

$$L_-(\theta, a_-) = \begin{cases} (c_- - 1)(a_- - \theta) & \text{if } a_{>}\theta, \\ (\theta - a_-) & \text{if } a_{\leq}\theta. \end{cases}$$

and

$$L_+(\theta, a_+) = \begin{cases} (a_+ - \theta) & \text{if } a_+ \geq \theta, \\ (c_+ - 1)(\theta - a_+) & \text{if } a_+ < \theta. \end{cases}$$

Because each of these loss functions depends only on one action, the posterior means can be minimized separately. Recall, in this case that the posterior mean of $L_-(\Theta, a_-)/c_-$ is minimized at a_- equal to the $1/c_-$ quantile of the posterior. Similarly, the posterior mean of $L_+(\Theta, a_+)/c_+$ is minimized at a_+ equal to the $(c_+ - 1)1/c_+$ quantile of the posterior.

7.4 The Bootstrap

The strategy of the bootstap is to say that one can use a calculation performed by using a cumulative distribution function $\hat{F}_n$ obtained from an observed sample as an estimate of the calculation one would like to perform using F.

Let $X = (X_1, \cdots, X_n)$ be an i.i.d. sample.

1. If the empirical distriubtion function

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

is used, then the method is the *nonparametric bootstrap.*

2. If $\hat{\Theta}_n$ is an estimate of Θ and

$$\hat{F}_n(x) = F_{X_1|\Theta}(x|\hat{\Theta}_n)$$

is used, then the method is the *parametric bootstrap.*

Let $\mathcal{F}$ be an appropriate space of cumulative distribution functions and let

$$R : \mathcal{X} \times \mathcal{F} \to R$$

be some function of interest, e.g., the difference between the sample median of X and the median of F. Then the bootstrap replaces

$$R(X,F) \qquad \text{by} \qquad R(X^*,\hat{F}_n)$$

where X^* is an i.i.d. sample of size n from $\hat{F}_n$.

The bootstrap was originally designed as a tool for estimating bias and standard error of a statistic.

Examples.

1. Assume that the sample is real values having CDF F satisfying $\int x^2\, dF(x) < \infty$. Let

$$R(X,F) = \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 - \left(\int x\ dF(x)\right)^2,$$

then

$$R(X^*,\hat{F}_n) = \left(\frac{1}{n}\sum_{i=1}^n X_i^*\right)^2 - (\bar{x}_n)\,,$$

where $\bar{x}_n$ is the observed sample average. Use

$$s_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$$

as an estimate of the variance. Now

$$E[R(X,F)] = \frac{1}{n}\sigma^2 \quad \text{and} \quad E[R(X^*,\hat{F}_n)|X=x] = \frac{1}{n}s_n^2.$$

2. (Bickel and Freedman) Suppose that $X_1,\cdots,X_n$ are independent $U(0,\theta)$ random variables. Take

$$R(X,F) = \frac{n}{F^{-1}(1)}(F^{-1}(1) - \max_i X_i).$$

The distribution of $\max_i X_i/F^{-1}(1)$ is $Beta(n,1)$. This has cumulative distribution function t^n. Thus,

$$P_\theta\{R(X,F) \le t\} = 1 - (1 - \frac{t}{n})^n \approx 1 - e^{-t}.$$

For the nonparametric bootstrap,

$$R(X^*,\hat{F}_n) = \frac{n}{\max_i X_i}(\max_i X_i - \max_i X_i^*)$$

and

$$P_\theta\{R(X^*,\hat{F}_n) = 0|\hat{F}_n\} = 1 - (1 - \frac{1}{n})^n \approx 1 - e^{-1} \approx 0.6321.$$

For any parametric bootstrap, we compute

$$P_\theta\{R(X^*,\hat{F}_n) \le t|\hat{F}_n\} = P_\theta\{F^{-1}_{X_1|\Theta}(1|\hat{\Theta}) \le \max_i X_i^*|F_{X_1|\Theta}(\cdot|\hat{\Theta})\}.$$

To construct a bootstrap confidence interval, we proceed as follows. Let

$$h : \mathcal{F} \to R.$$

Then the confidence interval for $h(F_\theta)$ may take the form

$$(-\infty, h(\hat{F}_n) + Y] \quad \text{or} \quad [h(\hat{F}_n) - Y_-, h(\hat{F}_n) + Y_+]$$

where, for a coefficient γ confidence interval

$$P_\theta\{h(\hat{F}_n) + Y \geq h(F_\theta)\} = \gamma \quad \text{or} \quad P_\theta\{h(\hat{F}_n) - Y_- \leq h(F) \leq h(\hat{F}_n) + Y_+\} = \gamma.$$

The goal is to find Y, Y_- and Y_+.

In the case that there is an available formula for the variance of F, $\sigma^2(F)$, then we can write, for example,

$$Y = \sigma(\hat{F}_n)\tilde{Y}.$$

The acknowledges that $\tilde{Y}$ may depend less on the underlying distribution than Y. Thus, we want $\tilde{Y}$ to satisfy

$$P_\theta\{\frac{h(F_\theta) - h(\hat{F}_n)}{\sigma(\hat{F}_n)} \leq \tilde{Y}\} = \gamma.$$

This lead to the *percentile-t bootstrap confidence interval* for $h(F)$,

$$(-\infty, h(\hat{F}_n) + \sigma(\hat{F}_n)\hat{Y}].$$

To determine $\hat{Y}$, note that

$$R(X, F) = \frac{h(F) - h(\hat{F}_n)}{\sigma(\hat{F}_n)} \quad \text{and} \quad R(X^*, \hat{F}_n) = \frac{h(\hat{F}_n) - h(\hat{F}_n^*)}{\sigma(\hat{F}_n^*)}.$$

Let $\hat{F}_{R^*}$ be the empirical cumulative distribution function of $R(X^*, \hat{F}_n)$, then

$$(-\infty, \sigma(\hat{F}_n)\hat{F}_{R^*}^{-1}(\gamma)]$$

will serve to give the bootstrap confidence interval.

One can use a similar procedure to detect bias. If

$$E_\theta[\phi(X)] = h(F_\theta)$$

choose

$$R(X, F) = \phi(X) - h(F)$$

and find $\hat{F}_{R^*}$.

8 Large Sample Theory

Large sample theory relies on the limit theorems of probability theory. We begin the study of large with empirical distribution functions.

8.1 Empirical Distribution Functions

Definition. Suppose that $X_1, \cdots, X_n$ are independent and identically distributed and let

$$X_{(1)}, \cdots, X_{(n)}$$

be the order statistics.

Define the *empirical cumulative distribution function* F_n by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i).$$

For any distribution define the *p-th quantile* to be

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}.$$

If $X_1, \cdots, X_n$ and independent real valued random variables hypothesized to follow some continuous probability distribution function, F, then we might plot

$$X_{(k)} \quad \text{versus} \quad F^{-1}(\frac{k}{n+1})$$

or equivalently

$$F(X_{(k)}) \quad \text{versus} \quad \frac{k}{n+1}.$$

By the *probability integral transform*, $Y_i = F(X_i)$ has a uniform distribution on $[0,1]$. In addition,

$$Y_{(k)} = F(X_{(k)}).$$

This transform allows us to reduce our study to the uniform distribution. The fact that $Y_{(k)}$ and $k/(n+1)$ are close is the subject of the following law of large numbers stheorem.

Theorem. (Glivenko-Cantelli)

$$\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0 \quad \text{a.s.}$$

Proof. (F continuous.) Fix x.

$$E[F_n(x)] = \frac{1}{n}\sum_{i=1}^{n} E[I_{(\infty,x]}(X_i)] = P\{X_1 \leq x\} = F(x).$$

Thus by the strong law of large numbers

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \text{a.s.}$$

If a sequence of nondecreasing functions converges pointwise to a bounded continuous function, then the convergence is uniform.

We can use the delta method to obtain results for quantiles and order statistics. The Glivenko-Cantelli theorem holds for general cumulative distribution functions F. One must also look at the limits of $F_n(x-)$.

We know look for the central limit behavior that accompanies this. If we look at the limiting distribution at a finite number of values x, we expect to find a mutivariate normal random variable.

Definition. Let $\mathcal{I}$ be an index set. An R^d-valued stochastic process $\{Z(t) : t \in \mathcal{I}\}$ is called a *Gaussian process* if each of its finite dimensional distributions

$$Z(t_1), \cdots, Z(t_n)$$

is a multivariate normal random variable. Consequently, the distribution of Z is determined by its

$$\text{mean function } \mu_Z(t) = E[Z(t)], \text{ and its covariance function } \Gamma_Z(s,t) = \text{Cov}(Z(s).Z(t)).$$

Note that the covariance of a Gaussian process is a positive definite function. Conversely, any positive semidefinite function is the covariance function of a Gaussian process.

Definition.

1. A real valued Gaussion process $\{W(t) : t \geq 0\}$ is called a *Brownian motion* if $t \mapsto W(t)$ is continuous a.s.,
$$\mu_W(t) = 0 \quad \text{and} \quad \Gamma_W(s,t) = \min\{s,t\}.$$

2. A real valued Gaussion process $\{B(t) : 0 \leq t \leq 1\}$ is called a *Brownian bridge* if $t \mapsto B(t)$ is continuous a.s.,
$$\mu_B(t) = 0 \quad \text{and} \quad \Gamma_B(s,t) = \min\{s,t\} - st.$$

Given the existence of Brownian motion, we can deduce the properties of the Brownian bridge by setting

$$\tilde{B}(t) = W(t) - tW(1).$$

Because $\tilde{B}$ is the sum of two Gaussian processes, it is also a Gaussian process.

$$\mu_{\tilde{B}}(t) = E[W(t)] - tE[W(1)] = 0,$$

and for $s \leq t$

$$\begin{aligned}\Gamma_{\tilde{B}}(s,t) &= E[(W(s) - sW(1))(W(t) - tW(1))] \\ &= E[W(s)W(t)] - tE[W(s)W(1)] - sE[W(1)W(t)] + stE[W(1)^2] \\ &= s - ts - st + st = s(1-t).\end{aligned}$$

Theorem. Let $Y_1, Y_2, \cdots$ be independent $U(0,1)$ and define

$$B_n(t) = \sqrt{n}(F_n(t) - t),$$

then for $t_1 < \cdots < t_k$,

$$(B_n(t_1), \cdots, B_n(t_k)) \to^{\mathcal{D}} (B(t_1), \cdots, B(t_k))$$

as $n \to \infty$ where B is the Brownian bridge.

Proof. The number of observations among $Y_1, \cdots, Y_n$ below t is

$$U_n(t) = nF_n(t).$$

Set $t_0 = 0$ and assume that $t_k = 1$. Note that the random variables

$$D_n(i) = U_n(t_i) - U_n(t_{i-1})$$

are $Multi(n, t_1 - t_0, \cdots, t_k - t_{k-1})$. By the central limit theorem

$$\tilde{D}(i) \equiv \frac{1}{\sqrt{n}}(U_n(t_i) - U_n(t_{i-1}) - n(t_i - t_{i-1})), \quad i = 1, \cdots, k$$

converges to a multivariate normal random variable $\tilde{D}$ with mean zero and covariance matrix

$$\Gamma_{\tilde{D}}(i,j) = (t_i - t_{i-1})(\delta_{ij} - (t_j - t_{j-1})).$$

Now,

$$\sum_{i=1}^{j} \tilde{D}_n(j) = \sum_{i=1}^{j} \frac{1}{\sqrt{n}}(U_n(t_i) - U_n(t_{i-1}) - n(t_i - t_{i-1})) = \frac{1}{\sqrt{n}}(U(t_j) - nt_j) = B_n(t_j).$$

Thus,

$$B_n(t_j) = \tilde{D}_n A$$

where the matrix A has entries $A(i,j) = I_{\{i \leq j\}}$. Consequently set,

$$B(t_j) = \tilde{D}A.$$

Because $B(t_j)$ has mean zero, we need only check that these random variables have the covariance structure of the Brownian bridge.

To check this, let $i \leq j$

$$\begin{aligned}
\Gamma_{\tilde{B}}(t_i, t_j) &= (A^T \Gamma_D A)(i, j) \\
&= \sum_{m=1}^{j} \sum_{\ell=1}^{j} A(m, i) \Gamma_D(m, \ell) A(\ell, j) \\
&= \sum_{m=1}^{i} \sum_{\ell=1}^{j} \Gamma_D(m, \ell) \\
&= \sum_{m=1}^{i} \sum_{\ell=1}^{j} (t_k - t_{k-1})(\delta_{k\ell} - (t_\ell - t_{\ell-1})) \\
&= t_i - t_i t_j = t_j(1 - t_i)
\end{aligned}$$

We would like to extend this result to say that the the entire path converges in distribution. First, we take the empirical distribution F_n and create a continuous version $\tilde{F}_n$ of this by linear interpolation. Then the difference between F_n and $\tilde{F}_n$ is at most $1/n$. This will allows us to discuss convergence on the separable Banach space $C([0, 1], R)$ under the supremum norm. Here is the desired result.

Theorem. Let $Y_1, Y_2, \cdots$ be independent $U(0, 1)$ and define

$$B_n(t) = \sqrt{n}(\tilde{F}_n(t) - t),$$

then

$$B_n \to^{\mathcal{D}} B$$

as $n \to \infty$ where B is the Brownian bridge.

The plan is to show that the distribution of the processes $\{B_n; n \geq 1\}$ forms a relatively compact set in the space of probability measures on $C([0, 1], R)$. If this holds, then we know that the distributios of $\{B_n; n \geq 1\}$ have limit points. We have shown that all of these limit points have the same finite dimensional distributions. Because this characterizes the process, we have only one limit point, the Brownian bridge.

The strategy is due to Prohorov and begins with the following definition:

Definition. A set $\mathcal{M}$ of probability measures on a metric space S is said to be *tight* if for every $\epsilon > 0$, there exists a compact set K so that

$$\inf_{P \in \mathcal{M}} P(K) \geq 1 - \epsilon.$$

Theorem. (Prohorov)

1. If $\mathcal{M}$ is tight, then it is relatively compact.
2. If S is complete and separable, and if $\mathcal{M}$ is relatively compact, then it is tight.

Thus, we need a characterization of compact sets in $C([0,1],R)$. This is provided by the following:

Definition. Let $x \in C([0,1],R)$. Then the *modulus of continuity* of x is defined by

$$\omega_x(\delta) = \sup_{|s-t|<\delta} |x(s) - x(t)|$$

.

Note that

$$|\omega_x(\delta) - \omega_y(\delta)| \leq 2||x-y||,$$

and therefore, for fixed δ, $\omega_x(\delta)$ is continuous in x. Because x is uniformly continuous,

$$\lim_{\delta \to 0} \omega_x(\delta) = 0.$$

Theorem. (Arzela-Ascoli) A subset $A \in C([0,1],R)$ has compact closure if and only if

$$\sup_{x \in A} |x(0)| < \infty, \quad \text{and} \quad \lim_{\delta \to 0} \sup_{x \in A} \omega_x(\delta) = 0.$$

In brief terms, any collection of uniformly bounded and equicontinuous functions has compact closure. This leads to the following theorem.

Theorem. The sequence $\{P_n : n \geq 1\}$ of probability measures on $C([0,1],R)$ is tight if and only if:

1. For each positive η, there exist M so that
$$\limsup_{n \to \infty} P_n\{x : |x(0)| > M\} \leq \eta.$$
2. For each postive ϵ and η, there exists δ such that
$$\limsup_{n \to \infty} P_n\{x : \omega_x(\delta) \geq \epsilon\} \leq \eta.$$

We can apply this criterion with P_n being the distribution of B_n to obtain the convergence in distribution to the Brownian bridge.

Because $B_n(0) = 0$, the first property is easily satisfied.

For the second criterion, we estimate

$$\begin{aligned}
P\{\sup_{s \leq t \leq t+\delta} |B_n(s) - B_n(t)| \geq \epsilon\} &= P\{\sup_{t \leq \delta} |B_n(t)| \geq \epsilon\} \\
&= P\{\sup_{t \leq \delta} \sqrt{n}|F_n(t) - t| \geq \epsilon\} \\
&= P\{\sup_{t \leq \delta} \sqrt{n}|\frac{1}{n}\sum_{i=1}^{n} I_{[0,t]}(Y_i) - t| \geq \epsilon\}.
\end{aligned}$$

and use this to show that given $\eta > 0$ and $\epsilon > 0$, the exists $\delta > 0$ so that

$$\limsup_{n\to\infty} P\{\omega_{B_n}(\delta) \geq \epsilon\} \leq \limsup_{n\to\infty} \frac{1}{\delta} P\{\sup_{t\leq\delta} \sqrt{n}|\frac{1}{n}\sum_{i=1}^{n} I_{[0,t]}(Y_i) - t| \geq \epsilon\} \leq \eta.$$

Write the Brownian bridge $B(t) = W(t) - tW(1)$, where W is Brownian motion. Now let $\{P_\epsilon : \epsilon > 0\}$ be a family of probability measures defined by

$$P_\epsilon(A) = P\{W \in A|W(1) \in [0,\epsilon]\}.$$

Proposition. $P_\epsilon \to^{\mathcal{D}} P_0$ where P_0 is the distribution of the Brownian bridge.

Proof. Let F be a closed subset of $C([0,1],R)$. We show that

$$\limsup_{\epsilon\to 0} P\{W \in F|W(1) \in [0,\epsilon]\} \leq P\{B \in F\}.$$

Fix $t_1 < \cdots < t_k$ note that the correlation of $W(1)$ with each component of $(B(t_1),\cdots,B(t_k))$ is zero and thus $W(1)$ independent of B. Therefore

$$P\{B \in A|W(1) \in [0,\epsilon]\} = P\{B \in A\}.$$

Note that

$$|B(t) - W(t)| = t|W(1)|$$

and therefore

$$||B - W|| = |W(1)|$$

Choose $\eta < \epsilon$, then

$$P\{W \in F|W(1) \in [0,\epsilon]\} \leq P\{B \in F^\eta|W(1) \in [0,\epsilon]\} = P\{B \in F^\eta\}.$$

Because F is closed,

$$\lim_{\eta\to 0} P\{B \in F^\eta\} = P\{B \in F\}$$

We will use this to compute the Kolmogorov-Smirnov test.

Theorem. For $c > 0$,

$$\lim_{n\to\infty} P\{\sup_{0\leq t\leq 1} |B_n(t)| \leq c\} = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2c^2}.$$

To begin, let $X_1, X_2, \cdots$ be an i.i.d. sequence of mean zero, finite variance, and let $S_0, S_1, S_2, \cdots$ be the sequence of partial sums. Set

$$m_n = \min_{0 \leq i \leq n} S_i \qquad M_n = \max_{0 \leq i \leq n} S_i$$

$$m = \min_{0 \leq t \leq 1} W(t) \quad M = \max_{0 \leq t \leq 1} W(t)$$

Because the mapping

$$x \mapsto (\min_{0 \leq t \leq 1} x(t), \max_{0 \leq t \leq 1} x(t), x(1))$$

from $C([0,1], R)$ to R^3 is continuous, we have

Theorem.

$$\frac{1}{\sigma \sqrt{n}}(m_n, M_n, S_n) \to^{\mathcal{D}} (m, M, W(1)).$$

For a simple random walk, we find an explicit formula for

$$p_n(a,b,v) = P\{a < m_n \leq M_n < b, S_n = v\}.$$

Claim. If $q_n(j) = P\{S_n = j\}$, then for integer $a \leq 0 \leq b$, $a \leq v \leq b$, $a < b$,

$$p_n(a,b,v) = \sum_{k=-\infty}^{\infty} q_n(v + 2k(b-a)) - \sum_{k=-\infty}^{\infty} q_n(2b - v + 2k(b-a)).$$

Note that the sum above is finite if $a < b$. To prove by induction, check that $p_0(a,b,0) = 1$, and $p_0(a,b,v) \neq 0$ for $v \neq 0$.

Now assume that the formula above holds for p_{n-1}.

Case 1. $a = 0$

Because $S_0 = 0$, $p_n(0,b,v) = 0$. Because $q_n(j) = q_n(-j)$, the two sums in the formula are equal.

Case 2. $b = 0$ is similar.

Case 3. $a < 0 < b$, $a \leq v \leq b$.

Because $a + 1 \leq 0$ and $b - 1 \geq 0$, we have the formula for

$$p_{n-1}(a-1, b-1, v-1) \quad \text{and} \quad p_{n-1}(a+1, b+1, v+1)$$

Use

$$q_n(j) = \frac{1}{2}(q_{n-1}(j-1) + q_{n-1}(j+1))$$

and

$$p_n(a,b,v) = \frac{1}{2}(p_{n-1}(a-1, b-1, v-1) + p_{n-1}(a+1, b+1, v+1))$$

to obtain

$$\sum_{k=-\infty}^{\infty}\frac{1}{2}(q_{n-1}(v-1+2k(b-a))+q_{n-1}(v+1+2k(b-a)))$$

$$-\sum_{k=-\infty}^{\infty}\frac{1}{2}(q_{n-1}(2(b-1)-(v-1)+2k(b-a))+q_{n-1}(2(b+1)-(v+1)+2k(b-a)))$$

$$=\frac{1}{2}(p_{n-1}(a-1,b-1,v-1)+p_{n-1}(a+1,b+1,v+1)).$$

Therefore,

$$P\{a<m_n\le M_n<b,u<S_n<v\}$$

$$=\sum_{k=-\infty}^{\infty}P\{u+2k(b-a)<S_n<v+2k(b-a)\}-\sum_{k=-\infty}^{\infty}P\{2b-v+2k(b-a)<S_n<2b-u+2k(b-a)\}.$$

By the continuity of the normal distribution, a termwise passage to the limit as $n\to\infty$ yields

$$P\{a<m\le M<b,u<W(1)<v\}$$

$$=\sum_{k=-\infty}^{\infty}P\{u+2k(b-a)<W(1)<v+2k(b-a)\}-\sum_{k=-\infty}^{\infty}P\{2b-v+2k(b-a)<W(1)<2b-u+2k(b-a)\}.$$

Let $-a=b=c$, $u=0$, and $v=\epsilon$, then

$$P\{\sup_{0<t<1}|W(t)|<c,0<W(1)<\epsilon\}$$

$$=\sum_{k=-\infty}^{\infty}P\{4kc<W(1)<\epsilon+4kc\}-\sum_{k=-\infty}^{\infty}P\{2c-\epsilon+4kc<W(1)<2c+4kc\}.$$

Use

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}P\{x<W(1)<x+\epsilon\}=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

to obtain

$$P\{|\sup_{0<t<1}|B(t)|<c\}=\sum_{k=-\infty}^{\infty}e^{-2(kc)^2}-\sum_{k=-\infty}^{\infty}e^{-2(c+kc)^2}=1+2\sum_{k=-\infty}^{\infty}e^{-2k^2c^2}$$

We can convert this limit theorem on empirical cumulative distribution function to sample quantiles using the following lemma.

Lemma. Let $Y_1,Y_2,\cdots$ be independent $U(0,1)$. Fix $t\in[0,1]$ then for each $z\in R$ there exists a sequence A_n such that, for every $\epsilon>0$,

$$\lim_{n\to\infty}P\{|A_n|>\sqrt{n}\epsilon\}=0.$$

and

$$\sqrt{n}(\tilde{F}_n^{-1}(t) - t) \leq z, \quad \text{if and only if} \quad \sqrt{n}(t - \tilde{F}(t)) \leq z + \sqrt{n}A_n.$$

Proof. Set

$$A_n = \tilde{F}_n(t + \frac{z}{\sqrt{n}}) - \tilde{F}_n(t) - \frac{z}{\sqrt{n}} = \frac{1}{n}(U_n(t + \frac{z}{\sqrt{n}}) - U_n(t)) - \frac{z}{\sqrt{n}} + \delta_n,$$

where $\delta_n \leq 2/n$ and $U_n(t)$ is the number of observations below t. Check, using, for example characteristic functions, that A_n has the desired property. To complete the lemma, consider the following equivalent inequalities.

$$\begin{array}{rcl} \sqrt{n}(\tilde{F}_n^{-1}(t) - t) & \leq & z \\ \tilde{F}_n^{-1}(t) & \leq & t + \frac{z}{\sqrt{n}} \\ t = \tilde{F}_n(\tilde{F}_n^{-1}(t)) & \leq & \tilde{F}_n(t + \frac{z}{\sqrt{n}}) \\ t & \leq & A_n + \tilde{F}_n(t) + \frac{z}{\sqrt{n}} \\ \sqrt{n}(t - \tilde{F}_n(t)) & \leq & z + \sqrt{n}A_n \end{array}$$

Corollary. Let $Y_1, Y_2, \cdots$ be independent $U(0,1)$ and define

$$\hat{B}_n(t) = \sqrt{n}(\tilde{F}_n^{-1}(t) - t),$$

then

$$\hat{B}_n \to^{\mathcal{D}} B$$

as $n \to \infty$, where B is the Brownian bridge.

Use the delta method to obtain the following.

Corollary. Let $0 < t_1 < \cdots < t_k < 1$ and let $X_1, X_2, \cdots$ be independent with cumulative distribution function F. Let $x_t = F^{-1}(t)$ and assume that F has derivative f in a neighborhood of each x_{t_i}. $i = 1, \cdots, k$, $0 < f(x_{t_i}) < \infty$. Then

$$\sqrt{n}(\tilde{F}_n^{-1}(t_1) - x_{t_1}, \cdots, \tilde{F}_n^{-1}(t_k) - x_{t_k}) \to^{\mathcal{D}} W,$$

where W is a mean zero normal random vector with covariance matrix

$$\Gamma_W(i,j) = \frac{\min\{t_i, t_j\} - t_i t_j}{f(t_i)f(t_j)}.$$

Examples. For the first and third quartiles, Q_1 and Q_3, and the median we have the following covariance matrix.

$$\Gamma_W = \left(\begin{array}{ccc} \frac{3}{16}\frac{1}{f(x_{1/4})^2} & \frac{1}{8}\frac{1}{f(x_{1/4})f(x_{1/2})} & \frac{1}{16}\frac{1}{f(x_{1/4})f(x_{3/4})} \\ \frac{1}{8}\frac{1}{f(x_{1/4})f(x_{1/2})} & \frac{1}{4}\frac{1}{f(x_{1/2})^2} & \frac{1}{8}\frac{1}{f(x_{1/2})f(x_{3/4})} \\ \frac{1}{16}\frac{1}{f(x_{1/4})f(x_{3/4})} & \frac{1}{8}\frac{1}{f(x_{1/2})f(x_{3/4})} & \frac{3}{16}\frac{1}{f(x_{3/4})^2} \end{array} \right)$$

1. For the Cauchy distribution, the density is

$$f(x) = \frac{1}{\sigma\pi}\frac{1}{1 + (x-\mu)^2/\sigma^2}.$$

   Therefore,

$$\begin{array}{ccc} x_{1/4} = \mu - \sigma, & x_{1/2} = \mu, & x_{3/4} = \mu + \sigma, \\ f(x_{1/4}) = \frac{1}{4\sigma\pi} & f(x_{1/2}) = \frac{1}{\sigma\pi}, & f(x_{3/4}) = \frac{1}{4\sigma\pi} \end{array}$$

2. For the normal distribution, the density is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

   Therefore,

$$\begin{array}{ccc} x_{1/4} = \mu - 0.6745\sigma, & x_{1/2} = \mu, & x_{3/4} = \mu + 0.6745\sigma, \\ f(x_{1/4}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{0.6745^2}{2}\right) & f(x_{1/2}) = \frac{1}{\sigma\sqrt{2\pi}}, & f(x_{3/4}) = \frac{1}{4\sigma\pi} \exp\left(-\frac{0.6745^2}{2}\right) \end{array}$$

For the extreme statistics, we have:

Theorem. For $x_-, x_+ \in R$, $\alpha_-, \alpha_+ > 0$, and assume

$$\lim_{x \to x_- +}(x - x_-)F(x) = c_- > 0, \quad \text{and} \quad \lim_{x \to x_+ -}(x - x_+)F(x) = c_+ > 0.$$

Let $X_1, X_2, \cdots$ be i.i.d. observations with cumulative distribution function F,

$$m_n = \min\{X_1, \cdots, X_n\} \quad \text{and} \quad M_n = \max\{X_1, \cdots, X_n\}.$$

then

$$\lim_{n\to\infty} P\{n^{1/\alpha_-}(m_n - x_-) < t_-, n^{1/\alpha_+}(x_+ - M_n) < t_+\} = (1 - \exp(-c_- t_-^{\alpha_-}))(1 - \exp(-c_+ t_+^{\alpha_+})).$$

8.2 Estimation

The sense that an estimator ought to arrive eventually at the parameter as the data increases is captured by the following definition.

Definition. Let $\{P_\theta : \theta \in \Omega\}$ be a parametric family of distributions on a sequence space $\mathcal{X}^\infty$. Let G be a metric space with the Borel σ-field. Let

$$g : \Omega \to G$$

and

$$Y_n : \mathcal{X}^\infty \to G$$

be a measurable function that depends on the first n coordinates of $\mathcal{X}^\infty$. We say that Y_n is *consistent for* $g(\theta)$ if

$$Y_n \to^P g(\theta), \quad P_\theta$$

.

Suppose that Ω is k-dimensional and suppose that the FI regularity conditions hold and that $\mathcal{I}_{X_1}(\theta)$ is the Fisher information matrix for a single observation. Assume, in addition,

$$\sqrt{n}(\hat{\Theta}_n - \theta) \to^{\mathcal{D}} Z$$

where Z is $N(0, V_\theta)$.

If g is differentiable, then, by the delta method

$$\sqrt{n}(g(\hat{\Theta}_n) - g(\theta)) \to^{\mathcal{D}} \langle \nabla g(\theta), Z\rangle.$$

In particluar, $g(\hat{\Theta})$ is a consistent estimator of $g(\theta)$.

The variance of $\langle \nabla g(\theta), Z\rangle$ is

$$\nabla g(\theta)^T V_\theta \nabla g(\theta).$$

We know that the smallest possible variance for an unbiased estimator of $g(\theta)$ is

$$\nabla g(\theta)^T \mathcal{I}_{X_1}(\theta)^{-1} \nabla g(\theta).$$

The ratio of these two variance is a measure of the quality of a consistent estimator.

Definition. For each n, let G_n be and estimator of $g(\theta)$ satisfying

$$\sqrt{n}(G_n) - g(\theta)) \to^{\mathcal{D}} W$$

where W is $N(0, v_\theta)$. Then the ratio

$$\frac{\nabla g(\theta)^T \mathcal{I}_{X_1}(\theta)^{-1} \nabla g(\theta)}{v_\theta}$$

is called the *asymptotic efficiency* of $\{G_n : n \geq 1\}$. If this ratio is one then the sequence of estimators is *asymptotically efficient.*

To compare estimating sequences, we have

Definition. Let $\{G_n : n \geq 1\}$ and $\{G'_n : n \geq 1\}$ and let C_ϵ be a criterion for an estimator. Fix ϵ and let

$$n(\epsilon) \quad \text{and} \quad n'(\epsilon)$$

be the first value for which the respective estimator satisfies C_ϵ. Assume

$$\lim_{\epsilon \to 0} n(\epsilon) = \infty \quad \text{and} \quad \lim_{\epsilon \to 0} n'(\epsilon) = \infty.$$

Then

$$r = \lim_{\epsilon \to 0} \frac{n'(\epsilon)}{n(\epsilon)}$$

is called the *asymptotically relative efficiency (ARE)* of $\{G_n : n \geq 1\}$ and $\{G'_n : n \geq 1\}$.

Examples.

1. Let $X_1, X_2, \cdots$ be independent $N(\mu, \sigma^2)$ random variables. Let $g(\mu, \sigma) = \mu$. Let G_n be the sample mean and G'_n be the sample median. Then,

$$\sqrt{n}(G_n - \mu) \to^{\mathcal{D}} \sigma Z \quad \text{and} \quad \sqrt{n}(G_n - \mu) \to^{\mathcal{D}} \sqrt{\frac{\pi}{2}}\sigma Z,$$

where Z is a standard normal. Assume that C_ϵ is that the estimator have variance below ϵ. Then the asymptotic relative efficiency is $\sqrt{2/\pi} \approx 0.79788$.

2. Let $X_1, X_2, \cdots$ be independent $U(0, \theta)$ random variables. The maximum likelihood estimator of θ is

$$\hat{\Theta} = \max_{1 \leq i \leq n} X_i.$$

A second estimator is

$$2\bar{X}_n.$$

Set C_ϵ to be having the stimator have variance below $\theta^2\epsilon$. We have

$$\text{Var}(\hat{\Theta}_n) = \frac{\theta^2 n}{(n+1)^2(n+2)} \quad \text{and} \quad \text{Var}(2\hat{X}_n) = \frac{\theta^2}{3n}.$$

Therefore,

$$n'(\epsilon) = \frac{(n(\epsilon)+1)^2(n(\epsilon)+2)}{n(\epsilon)},$$

and the ARE of $\hat{\Theta}_n$ to $2\bar{X}_n$ is ∞.

3. Let $X_1, X_2, \cdots$ be independent $N(\mu, 1)$ random variables. Then $\mathcal{I}_{X_1}(\theta) = 1$ and

$$\sqrt{n}(\bar{X}_n - \theta) \to^{\mathcal{D}} Z,$$

a standard normal. Fix θ_0 and for $0 < a < 1$ define a new estimator of Θ

$$G_n = \begin{cases} \bar{X}_n & \text{if } |\bar{X}_n - \theta_0| \geq n^{-1/4}, \\ \theta_0 + a(\bar{X} - \theta_0) & \text{if } |\bar{X}_n - \theta_0| < n^{-1/4}. \end{cases}$$

This is like using a posterior mean of Θ when $\bar{X}_n$ is close to θ_0. To calculate the effieiency of G_n, consider:

Case 1. $\theta \neq \theta_0$.

$$\sqrt{n}|\bar{X}_n - G_n| = \sqrt{n}(1-a)|\bar{X}_n - G_n|I_{[0,n^{-1/4}]}(|\bar{X}_n - G_n|).$$

Hence, for $\epsilon > 0$,

$$\begin{aligned} P'_\theta\{\sqrt{n}|\bar{X}_n - \theta_0| > \epsilon\} &\leq P'_\theta\{|\bar{X}_n - \theta_0| > n^{-1/4}\} \\ &= P'_\theta\{(\theta_0 - \theta)\sqrt{n} - n^{-1/4} \leq \sqrt{n}(\bar{X}_n - \theta) \leq (\theta_0 - \theta)\sqrt{n} + n^{-1/4}\}. \end{aligned}$$

Because $\sqrt{n}(\bar{X}_n - \theta)$ is a standard normal and the endpoints both tend to either $+\infty$ or $-\infty$, this last quantity has zero limit as $n \to \infty$. Therefore

$$\lim_{n\to\infty} P'_\theta\{\sqrt{n}|G_n - \bar{X}_n| > \epsilon\} = 0.$$

Case 2. $\theta = \theta_0$.

$$\sqrt{n}|(\bar{X}_n - \theta_0) + (\theta_0 - G_n)| = \sqrt{n}(1-a)|\bar{X}_n - \theta_0|I_{[n^{-1/4},\infty)}(|\bar{X}_n - \theta_0|).$$

Hence, for $\epsilon > 0$,

$$P'_{\theta_0}\{\sqrt{n}|(\bar{X}_n - \theta_0) + (\theta_0 - G_n)| > \epsilon\} \leq P'_{\theta_0}\{|(\bar{X}_n - \theta_0)| > n^{-1/4}\} = P'_{\theta_0}\{\sqrt{n}|(\bar{X}_n - \theta_0)| > n^{1/4}\}$$

Again, this last quantity has zero limit as $n \to \infty$. Therefore

$$\lim_{n\to\infty} P'_\theta\{\sqrt{n}|(G_n - \theta_0) + a(\bar{X}_n - \theta_0)| > \epsilon\} = 0.$$

Therefore,

$$\sqrt{n}(G_n - \theta) \to^{\mathcal{D}} W$$

where W is $N(0, v_\theta)$ where $v_\theta = 1$ except at θ_0 where $v_{\theta_0} = a^2$. Thus, the effieiency at θ_0 is $1/a^2 > 1$.

This phenomenon is called *superefficiency.* LeCam proved that under conditions slightly stronger that the FI regularity conditions, superefficiency can occur only on sets of Lebsegue measure 0.

8.3 Maximum Likelihood Estimators

Theorem. Assume that $X_1, X_2, \cdots$ are independent with density

$$f_{X_1|\Theta}(x|\theta)$$

with respect to some σ-finite measure ν. Then for each θ_0 and for each $\theta \neq \theta_0$,

$$\lim_{n\to\infty} P'_{\theta_0}\{\prod_{i=1}^n f_{X_1|\Theta}(X_i|\theta_0) > \prod_{i=1}^n f_{X_1|\Theta}(X_i|\theta)\} = 1.$$

Proof. The event above is equivalent to

$$R(x) = \frac{1}{n}\sum_{i=1}^n \log\frac{f_{X_1|\Theta}(x_i|\theta)}{f_{X_1|\Theta}(x_i|\theta_0)} < 0.$$

By the law of large numbers

$$R(X) \to E_{\theta_0}[\log\frac{f_{X|\Theta}(X_i|\theta)}{f_{X_1|\Theta}(X_i|\theta_0)}] = -\mathcal{I}_{X_1}(\theta_0;\theta), \quad \text{a.s. } P_{\theta_0},$$

the Kullback-Leibler information, which is negative whenever $\theta \neq \theta_0$.

Consistency follows from the following.

Theorem. (Wald) Assume that $X_1, X_2, \cdots$ are independent with density

$$f_{X_1|\Theta}(x|\theta)$$

with respect to some σ-finite measure ν. Fix $\theta_0 \in \Omega$, and define, for each $M \subset \Omega$ and $x \in \mathcal{X}$.

$$Z(M,x) = \inf_{\psi\in M} \log\frac{f_{X_1|\Theta}(x|\theta_0)}{f_{X_1|\Theta}(x|\psi)}.$$

Assume,

1. for each $\theta \neq \theta_0$, the is an open neighborhood N_θ of θ such that $E_{\theta_0}[Z(N_\theta, X_1)] > 0$. and
2. for Ω not compact, there exists a compact set K containing θ_0 such that $E_{\theta_0}[Z(K^c, X_1)] = c > 0$.

Then, then maximum likelihood estimator

$$\hat{\Theta}_n \to \theta_0 \quad \text{a.s. } P_{\theta_0}.$$

Proof. For Ω compact, take $K = \Omega$. Let $\epsilon > 0$ and let G_0 be the open ϵ ball about θ_0. Because $K\backslash G_0$ is compact, we can find a finite open cover $G_1, \cdots, G_{m-1} \subset \{N_\theta : \theta \neq \theta_0\}$. Thus, writing $G_m = K^c$,

$$\Omega = G_0 \cup G_1 \cup \cdots \cup G_m, \quad \text{and} \quad E_{\theta_0}[Z(G_j, X_1)] = c_j > 0.$$

Let $B_j \subset \mathcal{X}^\infty$ satisfy

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n Z(G_j, x_i) = c_j.$$

Similarly define the set B_0 for K^c. Then, by the strong law of large numbers

$$P_{\theta_0}(B_j) = 1 \quad j = 0, \cdots, m.$$

Then,

$$\begin{aligned}
\{x : \limsup_{n\to\infty} ||\hat{\Theta}_n(x_1, \cdots, x_n) - \theta_0|| > \epsilon\} &\subset \cup_{j=1}^m \{x : \hat{\Theta}_n(x_1, \cdots, x_n) \in G_j \text{ i.o.}\} \\
&\subset \cup_{j=1}^m \{x : \inf_{\psi \in G_j} \frac{1}{n}\sum_{i=1}^n \log \frac{f_{x_1|\Theta}(x_i|\theta_0)}{f_{x_1|\Theta}(x_i|\psi)} \le 0 \text{ i.o.}\} \\
&\subset \cup_{j=1}^m \{x : \frac{1}{n}\sum_{i=1}^n Z(G_j, x_i) \le 0 \text{ i.o.}\} \subset \cup_{j=1}^m B_j^c.
\end{aligned}$$

This last event has probability zero and thus we have the theorem.

Examples.

1. Let $X_1, X_2, \cdots$ be independent $U(0,\theta)$ random variables. Then

$$\log \frac{f_{X_1|\Theta}(x|\theta_0)}{f_{X_1|\Theta}(x|\psi)} = \begin{cases} \log \frac{\psi}{\theta_0} & \text{if } x \le \min\{\theta_0, \psi\} \\ \infty & \text{if } \psi < x \le \theta_0 \\ -\infty & \text{if } \theta_0 < x \le \psi \\ \text{undefined} & \text{otherwise.} \end{cases}$$

We need not consider the final two case which have P_{θ_0} probability 0. Choose

$$\begin{array}{lll} N_\theta = (\frac{\theta+\theta_0}{2}, \infty) & \text{for } \theta > \theta_0, & Z(N_\theta, x) = \log(\frac{\theta+\theta_0}{2\theta_0}) > 0 \\ N_\theta = (\frac{\theta}{2}, \frac{\theta+\theta_0}{2}) & \text{for } \theta < \theta_0, & Z(N_\theta, x) = \infty \end{array}$$

For the compact set K consider $[\theta_0/a, a\theta_0]$, $a > 1$. Then,

$$\inf_{\theta \in \Omega\setminus K} \log \frac{f_{X_1|\Theta}(x|\theta_0)}{f_{X_1|\Theta}(x|\theta)} = \begin{cases} \log \frac{x}{\theta_0} & \text{if } x < \frac{\theta_0}{a} \\ \log a & \text{if } \theta_0 \ge x \ge \frac{\theta_0}{a} \end{cases}$$

This has P_{θ_0} mean

$$\frac{1}{\theta_0}\left(\int_0^{\theta_0/a} \log \frac{x}{\theta_0}\, dx + \int_{\theta_0/a}^{\theta_0} \log a \, dx\right).$$

As $a \to \infty$, the first integral has limit 0, the second has limit ∞. Choose a such that the mean is positive.

The conditions on the theorem above can be weakened if $f_{X_1|\Theta}(x|\cdot)$ is upper semicontinuous. Note that the sum of two upper semicontinuous functions is upper semicontinuous and that an upper semicontinuous function takes its maximum on a compact set.

Theorem. Replace the first condition of the previous theorem with

1. $E_{\theta_0}[Z(N_\theta, X_i)] > -\infty$.

Further, assume that $f_{X_1|\Theta}(x|\cdot)$ is upper semicontinuous in θ for every x a.s. P_{θ_0}. Then

$$\lim_{n\to\infty} \hat{\Theta} = \theta_0, \quad \text{a.s. } P_{\theta_0}.$$

Proof. For Ω compact, take $K = \Omega$. For each $\theta \neq \theta_0$, let $N_{\theta,k}$ be a closed ball centered at θ, having radius at most $1/k$ and satisfying

$$N_{\theta,k+1} \subset N_{\theta,k} \subset N_\theta.$$

Then by the finite intersection property,

$$\cap_{k=1}^{\infty} N_{\theta,k} = \{\theta\}.$$

Note that $Z(N_{\theta,k}, x)$ increases with k and that $\log\left(f_{X_1|\Theta}(x|\theta_0)/f_{X_1|\Theta}(x|\psi)\right)$ is is upper semicontinuous in θ for every x a.s. P_{θ_0}. Consequently, for each k, there exists $\theta_k(x) \in N_{\theta,k}$ such that

$$Z(N_{\theta,k}, x) = \log\left(\frac{f_{X_1|\Theta}(x|\theta_0)}{f_{X_1|\Theta}(x|\theta_k)}\right)$$

and therefore

$$Z(N_\theta, x) \geq \lim_{k\to\infty} Z(N_{\theta,k}, x) \geq \log\left(\frac{f_{X_1|\Theta}(x|\theta_0)}{f_{X_1|\Theta}(x|\theta)}\right),$$

If $Z(N_\theta, x) = \infty$, then $Z(N_{\theta,k}, x) = \infty$ for all k. If $Z(N_\theta, x) < \infty$, then an application of Fatou's lemma to $\{Z(N_{\theta,k}, x) - Z(N_\theta, x)\}$ implies

$$\liminf_{k\to\infty} E_{\theta_0}[Z(N_{\theta,k}, X_i)] \geq E_{\theta_0}[\liminf_{k\to\infty} Z(N_{\theta,k}, X_i)] \geq \mathcal{I}_{X_1}(\theta_0; \theta) > 0.$$

Now choose k^* so that $E_{\theta_0}[Z(N_{\theta,k^*}, X_i)] > 0$ and apply the previous theorem.

Theorem. Suppose that $X_1, X_2, \cdots$ and independent random variables from a nondegenerate exponential family of distributions whose density with respect to a measure ν is

$$f_{X_1|\Theta}(X|\theta) = c(\theta)\exp\langle\theta, x\rangle.$$

Suppose that the natural parameter space Ω is an open subset of R^k and let $\hat{\Theta}_n$ be the maximum likelihood estimate of θ based on $X_1, \cdots, X_n$ if it exists. Then

1. $\lim_{n\to\infty} P_\theta\{\hat{\Theta}_n \text{exists}\} = 1$,
2. and under P_θ
$$\sqrt{n}(\hat{\Theta}_n - \theta) \to^{\mathcal{D}} Z,$$
where Z is $N(0, \mathcal{I}_{X_1}(\theta)^{-1})$ and $\mathcal{I}_{X_1}(\theta)$ is the Fisher information matrix.

Proof. Fix θ.

$$\nabla \log f_{X|\Theta}(x|\theta) = n\bar{x}_n + n\nabla \log c(\theta).$$

Thus if the MLE exists, it must be a solution to

$$-\nabla \log c(\theta) = \bar{x}.$$

The Hessian matrix is the Fisher information matrix,

$$\mathcal{I}_X(\theta)_{i,j} = -\frac{\partial^2}{\partial\theta_i\theta_j} \log c(\theta)$$

Because the family is nondegenerate, this matrix is is positive definite and therefore $-\log c(\theta)$ is strictly convex. By the implicit function theorem, v has a continuously differentiable inverse h in a neighborhood of θ. If $\bar{X}_n$ is in the domain of h, then the MLE is $h(\bar{X}_n)$.

By the weak law of large numbers,

$$\bar{X}_n \to^P E_\theta[X] = -\nabla \log c(\theta) \quad P_\theta.$$

Therefore, $\bar{X}_n$ will be be in the domain of h with probability approaching 1. This proves 1.

By the central limit theorem,

$$\sqrt{n}(\bar{X}_n + \nabla c(\theta)) \to^{\mathcal{D}} Z$$

as $n \to \infty$ where Z is $N(0, \mathcal{I}_X(\theta))$. Thus, by the delta method,

$$\sqrt{n}(\hat{\Theta}_n - \theta) \to^{\mathcal{D}} AZ$$

where the matrix A has (i,j) entry $\partial h(t)/\partial t_j$ evaluated at $t = v(\theta)$, i.e., $A = \mathcal{I}_X(\theta)^{-1}$. Therefore AZ has covariance matrix $\mathcal{I}_X(\theta)^{-1}$.

Corollary. Under the conditions of the theorem above, the maximum likelihood estimate of θ is consistent.

Corollary. Under the conditions of the theorem above, and suppose that $g \in C^1(\Omega, R)$, then $g(\hat{\Theta}_n)$ is an asymptotically efficient estimator of $g(\theta)$.

Proof. Using the delta method

$$\sqrt{n}(g(\hat{\Theta}_n) - g(\theta)) \to^{\mathcal{D}} \nabla g(\theta) W$$

where W is $N(0, \mathcal{I}_X(\theta)^{-1})$. Therefore, the estimator is asymtotically efficient.

We can obtain inconsistent maximum likelihood estimators. This is easy in cases in which the mapping

$$\theta \to P_\theta$$

does not have good continuity properties.

Example. Let $(X_1, Y_1), (X_2, Y_2), \cdots$ be independent random variales, X_i, Y_i are $N(\mu_i, \sigma^2)$ The log likelihood function

$$\begin{aligned}\log L(\theta) &= -n\log(2\pi) - 2n\log\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}((x_i-\mu_i)^2 + (y_i-\mu_i)^2) \\ &= -n\log(2\pi) - 2n\log\sigma + \frac{1}{2\sigma^2}\left(2\sum_{i=1}^{n}\left(\frac{x_i+y_i}{2} - \mu_i\right)\frac{1}{2}\sum_{i=1}^{n}(x_i-y_i)^2\right)\end{aligned}$$

Thus the maximum likelihood estimators are

$$\hat{M}_{i,n} = \frac{X_i+Y_i}{2}, \quad \hat{\Sigma}^2_{i,n} = \frac{1}{4n}\sum_{i=1}^{n}(X_i-Y_i)^2.$$

Because $X_i - Y_i$ is $N(0, 2\sigma^2)$ and thus

$$\hat{\Sigma}_{i,n} \to^P \frac{\sigma^2}{2} \quad P_\theta.$$

Thus, the estimator is not consistent.

To obtain general sufficient conditions for the asymptotically normality of maximum likelihood estimators, we have.

Theorem. For the parameter space $\Omega \in R^p$, let $X_1, X_2, \cdots$ have density $f_{X|\Theta}$ be C^2 in θ and that this differentiation can be passed under the integral sign. Assume

1. the Fisher information matrix $\mathcal{I}_{X_1}(\theta)$ is finite and non-singular,
2. for each θ, the MLE $\hat{\Theta}_n \to^P \theta$, P_θ, as $n \to \infty$, and
3. there exists H_r such that for each $\theta_0 \in \text{int}(\Omega)$,

$$\sup_{||\theta-\theta_0||\le r}|\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f_{X_1|\Theta}(x|\theta_0) - \frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f_{X_1|\Theta}(x|\theta)| \le H_r(x, \theta_0),$$

with $\lim_{r\to 0} E_{\theta_0}[H_r(X_1, \theta_0)] = 0$. Then, under P_{θ_0},

$$\sqrt{n}(\hat{\Theta}_n - \theta_0) \to^{\mathcal{D}} W$$

as $n \to \infty$ where W is $N(0, \mathcal{I}^{-1}_{X_1}(\theta_0))$.

Proof. For $\theta_0 \in \text{int}(\Omega)$, $\hat{\Theta}_n \to^P \theta_0$, P_{θ_0}. Thus, with P_{θ_0} probability 1, there exists N such that $I_{\text{int}(\Omega)^c}(\hat{\Theta}_n) = 0$ for all $n \ge N$. Thus, for every sequence of random variables $\{Z_n : n \ge 1\}$ and $\epsilon > 0$,

$$\lim_{n\to\infty} P_{\theta_0}\{Z_n I_{\text{int}(\Omega)^c}(\hat{\Theta}_n) > \epsilon\sqrt{n}\} = 0.$$

Set

$$\ell(\theta|x) = \frac{1}{n}\sum_{i=1}^{n}\log f_{X_1|\Theta}(x_i|\theta).$$

For $\hat{\Theta}_n \in \text{int}(\Omega)$,

$$\nabla_\theta \ell(\hat{\Theta}_n|X) = 0.$$

Thus,

$$\nabla_\theta \ell(\hat{\Theta}_n|X) = \nabla_\theta \ell(\hat{\Theta}_n|X) I_{\text{int}(\Omega)^c}(\hat{\Theta}_n),$$

and

$$\lim_{n\to\infty} P_{\theta_0}\{\nabla_\theta \ell(\hat{\Theta}_n|X) > \epsilon\sqrt{n}\} = 0.$$

By Taylor's theorem,

$$\frac{\partial}{\partial\theta_j}\ell(\hat{\Theta}_n|X) = \frac{\partial}{\partial\theta_j}\ell(\theta_0|X) + \sum_{k=1}^{p} \frac{\partial}{\partial\theta_k}\frac{\partial}{\partial\theta_j}\ell(\theta^*_{n,k}|X)(\hat{\Theta}_{n,k} - \theta_{0,k}),$$

where $\theta^*_{n,k}$ is between $\theta_{0,k}$ and $\hat{\Theta}_{n,k}$.

Because $\hat{\Theta}_n \to^P \theta_0$, we have that $\theta^*_n \to^P \theta_0$. Set B_n to be the matrix above. Then,

$$\lim_{n\to\infty} P_{\theta_0}\{|\nabla_\theta \ell(\theta_0|X) + B_n(\hat{\Theta}_n - \theta_0)| > \epsilon\sqrt{n}\} = 0.$$

Passing the derivative with respect to θ under the integral sign yields

$$E_{\theta_0}[\nabla_\theta \ell(\theta_0|X)] = 0.$$

Pass a second derivative with respect to θ under the integral sign to obtain the covariance matrix $\mathcal{I}_{X_1}(\theta_0)$ for $\nabla_\theta \ell(\theta_0|X)$. By the multivariate central limit theorem,

$$\sqrt{n}\nabla_\theta(\theta_0|X) \to^{\mathcal{D}} W$$

as $n \to \infty$ where W is $N(0, \mathcal{I}_{X_1}(\theta_0))$.

Consequently, for $\epsilon > 0$, there exists c so that

$$\limsup_{n\to\infty} P_{\theta_0}\{\sqrt{n}B_n(\hat{\Theta}_n - \theta_0) > c\} < \epsilon.$$

Write

$$B_n(k,j) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta_k}\frac{\partial}{\partial\theta_j}\log f_{X_1|\Theta}(X_i|\theta) + \Delta_n.$$

Then, by hypothesis,

$$|\Delta_n| \leq \frac{1}{n}\sum_{i=1}^{n} H_r(X_i, \theta_0)$$

whenever $||\theta_0 - \theta^*_n|| \leq r$. The law of large numbers gives

$$\frac{1}{n}\sum_{i=1}^{n} H_r(X_i, \theta_0) \to^P E_{\theta_0}[H_r(X_1, \theta_0)] \quad P_{\theta_0}.$$

Let $\epsilon > 0$ and choose r so that

$$E_{\theta_0}[H_r(X_1, \theta_0)] < \frac{\epsilon}{2}.$$

$$\begin{aligned} P'_{\theta_0}\{|\Delta_n| > \epsilon\} &\leq P'_{\theta_0}\{\frac{1}{n}\sum_{i=1}^n H_r(X_i, \theta_0) > \epsilon\} + P'_{\theta_0}\{||\theta_0 - \theta^*_n|| \leq r\} \\ &\leq P'_{\theta_0}\{|\frac{1}{n}\sum_{i=1}^n H_r(X_i, \theta_0) - E_{\theta_0}[H_r(X_1, \theta_0)]| > \frac{\epsilon}{2}\} + P'_{\theta_0}\{||\theta_0 - \theta^*_n|| \leq r\} \end{aligned}$$

Therefore,

$$\Delta_n \to^P 0, \quad P_{\theta_0}, \quad \text{and} \quad B_n \to^P -\mathcal{I}_{X_1}(\theta_0), \quad P_{\theta_0}.$$

Write $B_n = -\mathcal{I}_{X_1}(\theta_0) + C_n$, then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P_{\theta_0}\{|\sqrt{n}C_n(\hat{\Theta}_n - \theta_0)| > \epsilon\} = 0.$$

Consequently,

$$\sqrt{n}(\nabla_\theta \ell(\theta_0|X) + \mathcal{I}_{X_1}(\theta_0)(\hat{\Theta}_n - \theta_0)) \to^P 0, \quad P_{\theta_0}.$$

By Slutsky's theorem,

$$-\mathcal{I}_{X_1}(\theta_0)\sqrt{n}(\hat{\Theta}_n - \theta_0)) \to^{\mathcal{D}} Z, \quad P_{\theta_0}.$$

By the continuity of matrix multiplication,

$$\sqrt{n}(\hat{\Theta}_n - \theta_0)) \to^{\mathcal{D}} -\mathcal{I}_{X_1}(\theta_0)^{-1}Z, \quad P_{\theta_0},$$

which is the desired distribution.

Example. Suppose that

$$f_{X_1|\Theta}(x|\theta) = \frac{1}{\pi(1 + (x - \theta)^2)}.$$

Then,

$$\frac{\partial^2}{\partial\theta^2} \log f_{X_1|\Theta}(x|\theta) = -2\frac{1 - (x - \theta)^2}{(1 + (x - \theta)^2)^2}.$$

This is a differentiable function with finite mean. Thus, H_r exists in the theorem above.

Because θ_0 is not known, a candidate for $\mathcal{I}_{X_1}(\theta_0)$ must be chosen. The choice

$$\mathcal{I}_{X_1}(\hat{\Theta}_n)$$

is called the*expected Fisher information.*

A second choice is the matrix with (i, j) entry

$$-\frac{1}{n}\frac{\partial^2}{\partial\theta_i\theta_j} \log f_{X_1|\Theta}(X|\hat{\Theta}_n)$$

is called the *observed Fisher information.*

The reason given by Efron and Hinkley for this choice is that the inverse of the observed information is closer to the conditional variance of the maximum likelihood estimator given an ancillary.

8.4 Bayesian Approaches

Theorem. Let $(S, \mathcal{A}, \mu)$ be a probability space, $(\mathcal{X}, \mathcal{B})$ a Borel space, and Ω be a finite dimensional parameter space endowed with a Borel σ-field. Let

$$\Theta : S \to \Omega \quad \text{and} \quad X_n : S \to \mathcal{X} \; n = 1, 2, \cdots,$$

be measurable. Suppose that there exists a sequence of functions

$$h_n : \mathcal{X}^n \to \Omega$$

such that

$$h_n(X_1, \cdots, X_n) \to^P \Theta.$$

Given $(X_1, \cdots, X_n) = (x_1, \cdots, x_n)$, let

$$\mu_{\Theta|X_1,\cdots,X_n}(\cdot|x_1, \cdots, x_n)$$

denote the posterior probability distribution on Ω. Then for each Borel set $B \in \Omega$,

$$\lim_{n\to\infty} \mu_{\Theta|X_1,\cdots,X_n}(B|x_1, \cdots, x_n) = I_B(\Theta) \text{ a.s. } \mu.$$

Proof. By hypothesis Θ is measurable with respect to the completion of $\sigma\{X_n : n \geq 1\}$. Therefore, with μ probability 1, by Doob's theorem on uniformly integrable martingales,

$$I_B(\Theta) = \lim_{n\to\infty} E[I_B(\Theta)|X_1, \cdots, X_n] = \lim_{n\to\infty} \mu_{\Theta|X_1,\cdots,X_n}(B|X_1, \cdots, X_n).$$

Theorem. Assume the conditions on Wald's consistency theorem for maximum likelihood estimators. For $\epsilon > 0$, assume that the prior distribution μ_Θ satisfies

$$\mu_\Theta(C_\epsilon) > 0,$$

where $C_\epsilon = \{\theta : \mathcal{I}_{X_1}(\theta_0; \theta) < \epsilon\}$. Then, for any $\epsilon > 0$ and open set G_0 containing C_ϵ, the posterior satisfies

$$\lim_{n\to\infty} \mu_{\Theta|X_1,\cdots,X_n}(G_0|X_1, \cdots, X_n) = 1 \quad \text{a.s. } P_{\theta_0}.$$

Proof. For each sequence $x \in \mathcal{X}^\infty$, define

$$D_n(\theta, x) = \frac{1}{n} \sum_{i=1}^n \log \frac{f_{X_1|\Theta}(x_i|\theta_0)}{f_{X_1|\Theta}(x_i|\theta)}.$$

Write the posterior odds of G_0 as

$$\frac{\mu_{\Theta|X_1,\cdots,X_n}(G_0|x_1, \cdots, x_n)}{\mu_{\Theta|X_1,\cdots,X_n}(G_0^c|x_1, \cdots, x_n)} = \frac{\int_{G_0} \prod_{i=1}^n f_{X_1|\Theta}(x_i|\theta) \; \mu_\Theta(d\theta)}{\int_{G_0^c} \prod_{i=1}^n f_{X_1|\Theta}(x_i|\theta) \; \mu_\Theta(d\theta)} = \frac{\int_{G_0} \exp(-nD_n(\theta, x)) \; \mu_\Theta(d\theta)}{\int_{G_0^c} \exp(-nD_n(\theta, x)) \; \mu_\Theta(d\theta)}.$$

To show that the posterior odds go to ∞, we find a lower bound for the numerator and an upper bound for the denominator.

As in the proof of Wald's theorem, we construct sets $G_1, \cdots, G_m$ so that

$$\Omega = G_0 \cup G_1 \cup \cdots \cup G_m, \quad \text{and} \quad E_{\theta_0}[Z(G_j, X_1)] = c_j > 0.$$

For $M \in \Omega$,

$$\inf_{\theta \in M} D_n(\theta, x) \geq \frac{1}{n} \sum_{i=1}^{n} Z(M, x_i).$$

Thus the denominator is at most

$$\sum_{j=1}^{m} \int_{G_j} e^{-nD_n(\theta,x)} \ \mu_\Theta(d\theta) \leq \sum_{j=1}^{m} \sup_{\theta \in G_j} e^{-nD_n(\theta,x)} \ \mu_\Theta(G_j) \leq \sum_{j=1}^{m} \exp\left(-\sum_{i=1}^{n} Z(G_j, x_i)\right) \mu_\Theta(G_j).$$

Set $c = \min\{c_1, \cdots, c_m\}$, then for all x in a set of P_{θ_0} probability 1, we have, by the strong law of large numbers, a number $N(x)$ so that for $n \geq N(x)$,

$$\sum_{i=1}^{n} Z(G_j, x_i) > \frac{nc}{2}$$

and, thus, we have a bound on the denominator of $\exp(-nc/2)$.

For the numerator, let $0 < \delta < \min\{\epsilon, c/2\}/4$. For each $x \in \mathcal{X}^\infty$ or $\theta \in \Omega$, define

$$W_n(x) = \{\theta : D_\ell(\theta, x) \leq \mathcal{I}_{X_1}(\theta_0; \theta) + \delta, \text{ for all } \ell \geq n\},$$

and

$$V_n(\theta) = \{x : D_\ell(\theta, x) \leq \mathcal{I}_{X_1}(\theta_0; \theta) + \delta, \text{ for all } \ell \geq n\}.$$

Clearly

$$x \in V_n(\theta) \text{ if and only if } \theta \in W_n(x).$$

Note that the strong law says that

$$D_n(\theta, x) \to \mathcal{I}_{X_1}(\theta_0; \theta) \text{ a.s. } P_{\theta_0}.$$

Thus, $V_n(\theta)$ is a nested sequence of events whose P_{θ_0} probability converges to 1.

$$\begin{aligned}
\mu_\Theta(C_\delta) &= \textstyle\lim_{n\to\infty} \int_{C_\delta} P_{\theta_0}(V_n(\theta)) \ \mu_\Theta(d\theta) &&= \lim_{n\to\infty} \int_{C_\delta} \int_{\mathcal{X}^\infty} I_{V_n(\theta)}(x) \ P_{\theta_0}(dx)\mu_\Theta(d\theta) \\
&= \textstyle\lim_{n\to\infty} \int_{\mathcal{X}^\infty} \int_{C_\delta} I_{W_n(x)}(\theta) \ \mu_\Theta(d\theta) P_{\theta_0}(dx) &&= \lim_{n\to\infty} \int_{\mathcal{X}^\infty} \mu_\Theta(C_\delta \cap W_n(x)) P_{\theta_0}(dx)
\end{aligned}$$

Therefore,

$$\lim_{n\to\infty} \mu_\Theta(C_\delta \cap W_n(x)) = \mu_\Theta(C_\delta) \text{ a.s. } P_{\theta_0}.$$

For x in the set in which this limit exists, there exists $\tilde{N}(x)$ such that

$$\mu_\Theta(C_\delta \cap W_n(x)) > \frac{1}{2}\mu_\Theta(C_\delta)$$

whenever $n \geq \tilde{N}(x)$. Use the fact that $\mathcal{I}_{X_1}(\theta_0;\theta) < \delta$ for $\theta \in C_\delta$ to see that the numerator is at least

$$\int_{C_\delta \cap W_n(x)} \exp(-n(\mathcal{I}_{X_1}(\theta_0;\theta)+\delta))\ \mu_\Theta(d\theta) \geq \frac{1}{2}\exp(-2n\delta)\mu_\Theta(C_\delta) \geq \frac{1}{2}\exp(-\frac{nc}{4})\mu_\Theta(C_\delta).$$

Therefore, for x in a set having P_{θ_0} probability 1 and $n \geq \max\{N(x), \tilde{N}(x)\}$, the posterior odds are at least

$$\frac{1}{2}\mu_\theta(C_\delta)\exp(\frac{nc}{4})$$

which goes to infinity with n.

Example. For $X_1, X_2, \cdots$ independent $U(0,\theta)$ random variables, the Kullback-leibler information is

$$\mathcal{I}_{X_1}(\theta_0;\theta) = \begin{cases} \log\frac{\theta}{\theta_0} & \text{if } \theta \geq \theta_0 \\ \infty & \text{if } \theta < \theta_0 \end{cases}$$

The set

$$C_\epsilon = [\theta, e^\epsilon\theta_0).$$

Thus, for some $\delta > 0$,

$$(\theta_0 - \delta, e^\epsilon\theta_0) \subset G_0.$$

Consequently is the prior distribution assigns positive mass to every open interval, then the posterior probability of any open interval containing θ_0 will tend to 1 a.s. P_{θ_0} as $n \to \infty$.

To determine the asymptotic normality for posterior distributions, we adopt the following general notation and regularity conditions.

1. $X_n : S \to \mathcal{X}_n$, for $n = 1, 2, \cdots$
2. The parameter space $\Omega \subset R^k$ for some k.
3. $\theta_0 \in \text{int}(\Omega)$.
4. The conditional distribution has density $f_{x_n|\Theta}(X_n|\theta)$ with respect to some σ-finite measure ν.
5. $\ell_n(\theta) = \log f_{X_n|\Theta}(X_n|\theta)$.
6. $\mathcal{H}\ell_n(\theta)$ is the Hessian of $\ell_n(\theta)$.
7. $\hat{\Theta}_n$ is the maximum likelihood estimator of Θ if it exists.
8. $$\Sigma_n = \begin{cases} -\mathcal{H}\ell_n(\hat{\Theta}_n)^{-1} & \text{if the inverse and } \hat{\Theta}_n \text{ exist,} \\ I_k & \text{otherwise.} \end{cases}$$

9. The prior distribution of Θ has a density with respect to Lebesgue measure that is positive and continuous at θ_0.
10. The largest eigenvalue of Σ_n goes to zero as $n \to \infty$.
11. Let λ_n be the smallest eigenvalue of Σ_n. If the open ball $B(\theta_0, \delta) \subset \Omega$, then there exists $K(\delta)$ such that
$$\lim_{n\to\infty} P'_{\theta_0}\{ \sup_{\theta \notin B(\theta_0,\delta)} \lambda_n(\ell_n(\theta) - \ell_n(\theta_0)) < -K(\delta)\} = 1.$$
12. For each $\epsilon > 0$, there exists $\delta > 0$ such that
$$\lim_{n\to\infty} P'_{\theta_0}\{ \sup_{\theta \in B(\theta_0,\delta), ||\gamma||=1} |1 + \gamma^T \Sigma_n^{1/2} \mathcal{H}\ell_n(\theta) \Sigma_n^{1/2} \gamma| < \epsilon\} = 1.$$

Theorem. Under the regularity conditions given above, set
$$\Psi_n = \Sigma_n^{-1/2}(\Theta - \hat{\Theta}_n).$$
Then, for each compact set $K \subset R^k$ and each $\epsilon > 0$,
$$\lim_{n\to\infty} P'_{\theta_0}\{ \sup_{\psi \in K} |f_{\Psi_n|X_n}(\psi|X_n) - \phi(\psi)| > \epsilon\} = 0.$$
where ϕ is the $N_k(0, I_k)$ density.

Proof. Note the regularity conditions guarantee that $\hat{\Theta}_n$ is consistent. By Taylor's theorem,
$$\begin{aligned} f_{X_n|\Theta}(X_n|\theta) &= f_{X_n|\Theta}(X_n|\hat{\Theta}_n)\exp(\ell(\theta) - \ell(\hat{\Theta}_n)) \\ &= f_{X_n|\Theta}(X_n|\hat{\Theta}_n)\exp\left(-\frac{1}{2}(\theta - \hat{\Theta}_n)^T \Sigma_n^{-1/2}(I_k - R_n(\theta, X_n))\Sigma_n^{-1/2}(\theta - \hat{\Theta}_n) + \Delta_n\right), \end{aligned}$$
where
$$\Delta_n = (\theta - \hat{\Theta}_n)^T \nabla\ell(\hat{\Theta}_n) I_{\text{int}(\Omega)}(\hat{\Theta}_n), \text{ and } R_n(\theta, X_n) = I_k + \Sigma_n^{1/2}\mathcal{H}\ell_n(\theta_n^*)\Sigma_n^{1/2},$$
with θ_n^* between θ and $\hat{\Theta}_n$. Note that $\theta_0 \in \text{int}(\Omega)$ and the consistency of $\hat{\Theta}_n$ imply that
$$\lim_{n\to\infty} P'_{\theta_0}\{\Delta_n = 0, \text{ for all } \theta\} = 1.$$

By Bayes' theorem, we can write the posterior density of Θ as
$$f_{\Theta|X_n}(\theta|X_n) = f_\Theta(\theta)\frac{f_{X_n|\Theta}(X_n|\theta)}{f_{X_n}(X_n)}.$$

The posterior density
$$\begin{aligned} f_{\Psi_n|X_n}(\psi|X_n) &= \frac{\det(\Sigma_n)^{1/2} f_\Theta(\Sigma_n^{1/2}\psi + \hat{\Theta}_n) f_{X_n|\Theta}(X_n|\Sigma_n^{1/2}\psi + \hat{\Theta}_n))}{f_{X_n}(X_n)} \\ &= \frac{\det(\Sigma_n)^{1/2} f_{X_n|\Theta}(X_n|\hat{\Theta}_n) f_\Theta(\Sigma_n^{1/2}\psi + \hat{\Theta}_n)}{f_{X_n}(X_n)} \frac{f_{X_n|\Theta}(X_n|\Sigma_n^{1/2}\psi + \hat{\Theta}_n)}{f_{X_n|\Theta}(X_n|\hat{\Theta}_n)}. \end{aligned}$$

To consider the first factor, choose $0 < \epsilon < 1$. Let η satisfy

$$1 - \epsilon \le \frac{1-\eta}{(1+\eta)^{k/2}}, \quad 1 + \epsilon \ge \frac{1+\eta}{(1-\eta)^{k/2}}.$$

Because the prior is continuous and nonnegative at θ_0, and by the regularity conditions there exists $\delta > 0$ such that

$$||\theta - \theta_0|| < \delta \quad \text{implies } |f_\Theta(\theta) - f_\Theta(\theta_0)| < \eta f_\Theta(\theta_0)$$

and

$$\lim_{n\to\infty} P'_{\theta_0}\{ \sup_{\theta \in B(\theta_0,\delta), ||\gamma||=1} |1 + \gamma^T \Sigma_n^{1/2} \mathcal{H}\ell_n(\theta) \Sigma_n^{1/2}\gamma| < \eta\} = 1.$$

Clearly

$$f_{X_n}(X_n) = J_1 + J_2 = \int_{B(\theta_0,\delta)} f_\Theta(\theta) f_{X_n|\Theta}(X_n|\theta)\, d\theta + \int_{B(\theta_0,\delta)^c} f_\Theta(\theta) f_{X_n|\Theta}(X_n|\theta)\, d\theta.$$

Claim I.

$$\frac{J_1}{\det(\Sigma)^{1/2} f_{X_n|\Theta}(X_n|\hat\Theta_n)} \to^P (2\pi)^{k/2} f_\Theta(\theta_0).$$

$$J_1 = f_{X_n|\Theta}(X_n|\hat\Theta_n) \int_{B(\theta_0,\delta)} f_\Theta(\theta) \exp\left(-\frac{1}{2}(\theta - \hat\Theta_n)^T \Sigma_n^{-1/2}(I_k - R_n(\theta, X_n))\Sigma_n^{-1/2}(\theta - \hat\Theta_n) + \Delta_n \right)\, d\theta$$

and therefore

$$(1-\eta) J_3 < \frac{J_1}{f_\Theta(\theta_0) f_{X_n|\Theta}(X_n|\hat\Theta_n)} < (1+\eta) J_3$$

where

$$J_3 = \int_{B(\theta_0,\delta)} \exp\left(-\frac{1}{2}(\theta - \hat\Theta_n)^T \Sigma_n^{-1/2}(I_k - R_n(\theta, X_n))\Sigma_n^{-1/2}(\theta - \hat\Theta_n) + \Delta_n \right)\, d\theta.$$

Consider the events with limiting probability 1,

$$\{\Delta_n = 0\} \cap \{\int_{B(\theta_0,\delta)} \exp\left(-\frac{1+\eta}{2}(\theta - \hat\Theta_n)^T \Sigma_n^{-1}(\theta - \hat\Theta_n) \right)\, d\theta$$
$$\le J_3 \le \int_{B(\theta_0,\delta)} \exp\left(-\frac{1-\eta}{2}(\theta - \hat\Theta_n)^T \Sigma_n^{-1}(\theta - \hat\Theta_n) \right)\, d\theta\}.$$

These two integrals equal

$$(2\pi)^{k/2}(1 \pm \eta)^{-k/2} \det(\Sigma_n)^{1/2} \Phi(C_n^{\pm}).$$

where $\Phi(C_n^{\pm})$ is the probability that a $N_k(0, I_k)$ random variable takes values in

$$C_n = \{z : \hat\Theta_n + (1 \pm \eta)^{-k/2} \Sigma_n^{1/2} z \in B(\theta_0, \delta)\}.$$

By the condition on the largest eignevalue, we have for all z,

$$\Sigma_n^{1/2} z \to^P 0 \quad \text{and hence } \Phi(C_n^{\pm}) \to^P 1$$

as $n \to \infty$. Consequently,

$$\lim_{n\to\infty} P'_{\theta_0}\{(2\pi)^{k/2}\frac{\det(\Sigma)^{1/2}}{(1+\eta)^{k/2}} < J_3 < (2\pi)^{k/2}\frac{\det(\Sigma)^{1/2}}{(1-\eta)^{k/2}}\} = 1,$$

or

$$\lim_{n\to\infty} P'_{\theta_0}\{(2\pi)^{k/2}\det(\Sigma)^{1/2}(1-\epsilon) < \frac{J_1}{f_\Theta(\theta_0) f_{X_n|\Theta}(X_n|\hat{\Theta}_n)} < (2\pi)^{k/2}\det(\Sigma)^{1/2}(1+\epsilon)\} = 1.$$

Claim II.

$$\frac{J_2}{\det(\Sigma)^{1/2} f_{X_n|\Theta}(X_n|\hat{\Theta}_n)} \to^P 0.$$

Write

$$J_2 = f_{X_n|\Theta}(X_n|\hat{\Theta}_n)\exp(\ell_n(\theta_0) - \ell_n(\hat{\Theta}_n)) \int_{B(\theta_0,\delta)^c} f_\Theta(\theta)\exp(\ell_n(\theta) - \ell_n(\theta_0))\, d\theta.$$

Because $\lambda_n \leq \det(\Sigma_n)^{1/k}$, we have by the regularity conditions

$$\ell_n(\theta) - \ell_n(\theta_0) < -\det(\Sigma_n)^{-1/k} K(\delta).$$

Use this to bound the integral above by

$$\exp(-\det(\Sigma_n)^{-1/k}K(\delta)) \int_{B(\theta_0,\delta)} f_\Theta(\theta)\, d\theta \leq \exp(-\det(\Sigma_n)^{-1/k}K(\delta)),$$

with probability tending to 1.

Because $\hat{\Theta}_n$ is a maximum likelihood estimator,

$$\exp(\ell_n(\theta_0) - \ell_n(\hat{\Theta}_n)) \leq 1.$$

The condition on the largest eigenvalue guarantees us

$$\frac{\exp(-\det(\Sigma_n)^{-1/k}K(\delta))}{\det(\Sigma_n)^{1/2}} \to^P 0$$

giving the claim.

Combining the claims gives

$$\frac{f_{X_n}(X_n)}{\det(\Sigma)^{1/2} f_{X_n|\Theta}(X_n|\hat{\Theta}_n)} \to^P (2\pi)^{k/2} f_\Theta(\theta_0).$$

Because $\hat{\Theta}_n$ is consistent and the prior is continuous at θ_0, we have that

$$f_\Theta(\Sigma^{1/2}\psi + \hat{\Theta}_n) \to^P f_\Theta(\theta)$$

uniformly for ψ in a compact set.

Combine this with the results of the claims to obtain

$$\frac{\det(\Sigma)^{1/2} f_{X_n|\Theta}(X_n|\hat{\Theta}_n) f_\Theta(\Sigma^{1/2}\psi + \hat{\Theta}_n)}{f_{X_n}(X_n)} \to^P (2\pi)^{-k/2}$$

uniformly on compact sets.

To complete the proof, we need to show that the second fraction in the posterior density converges in probability to $\exp(-||\psi||^2/2)$ uniformly on compact sets. Referring to the Taylor expansion,

$$\frac{f_{X_n|\Theta}(X_n|\Sigma_n^{1/2}\psi + \hat{\Theta}_n)}{f_{X_n|\Theta}(X_n|\hat{\Theta}_n)} = \exp\left(-\frac{1}{2}(\psi^T(I_k - R_n(\Sigma_n^{1/2}\psi + \hat{\Theta}_n, X_n))\psi + \Delta_n)\right).$$

Let $\eta, \epsilon > 0$ and let $K \subset B(0,k)$ be compact. Then by the regularity conditions. Choose δ and M so that $n \geq M$ implies

$$P'_{\theta_0}\{\sup_{\theta \in B(\theta_0,\delta), ||\gamma||=1} |1 + \gamma^T \Sigma_n^{1/2} \mathcal{H}\ell_n(\theta)\Sigma_n^{1/2}\gamma| < \frac{\eta}{k}\} > 1 - \frac{\epsilon}{2}.$$

Now choose $N \geq M$ so that $n \geq N$ implies

$$P'_{\theta_0}\{\Sigma_n^{1/2}\psi + \hat{\Theta}_n \in B(\theta_0,\delta), \text{ for all } \psi \in K\} > 1 - \frac{\epsilon}{2}.$$

Consequently, if $n \geq N$,

$$P'_{\theta_0}\{|\psi^T(I_k - R_n(\Sigma_n^{1/2}\psi + \hat{\Theta}_n, X_n))\psi - ||\psi||^2| < \eta \text{ for all } \psi \in K\} > 1 - \epsilon.$$

Because

$$\lim_{n\to\infty} P'_{\theta_0}\{\Delta_n = 0, \text{ for all } \psi\} = 1$$

the second fraction in the representation of the posterior distribution for Ψ_n is between

$$\exp(-\eta)\exp(-||\psi||^2/2) \quad \text{and} \quad \exp(\eta)\exp(-||\psi||^2/2)$$

with probability tending to 1, uniformly on compact sets.

Examples.

1. Let $Y_1, Y_2, \cdots$ be conditionally IID given Θ and set $X_n = (Y_1, \cdots, Y_n)$. In addition, suppose that the Fisher information $\mathcal{I}_{X_1}(\theta_0)$.
   - Because $n\Sigma_n \to^P \mathcal{I}_{X_1}(\theta_0)^{-1}$, the largest eigenvalue of Σ_n goes to 0 in probability as $n \to \infty$.
   - Using the notation from Wald's theorem on consistency, we have

$$\begin{aligned}\sup_{\theta \in B(\theta_0,\delta)^c} (\ell_n(\theta) - \ell_n(\theta_0)) &= -\inf_{\theta \in B(\theta_0,\delta)^c} (\ell_n(\theta_0) - \ell_n(\theta)) \\ &\leq -\min_{j=1,\cdots,m}\{-\inf_{\theta \in G_j}(\ell_n(\theta_0) - \ell_n(\theta))\} \\ &\leq -\min_{j=1,\cdots,m}\{\sum_{i=1}^n Z(G_j, Y_i)\}.\end{aligned}$$

Note that

$$\frac{1}{n}\sum_{i=1}^{n} Z(G_j, Y_i) \to^P E_{\theta_0}[Z(G_j, Y_i)].$$

If these means are all postive, and if λ is the smallest eigenvalue of $\mathcal{I}_{Y_1}(\theta_0)$, then take

$$K(\delta) \leq \frac{1}{2\lambda} \min_{j=1,\cdots,m}\{E_{\theta_0}[Z(G_j, Y_i)]\}$$

- Recalling the conditions on the asymptotically normality of maximum likelihood estimators, let $\epsilon > 0$ and choose δ fo that
$$E_{\theta_0}[H_\delta(Y_i, \theta)] < \frac{\epsilon}{\mu + \epsilon}$$
where μ is the largest eigenvalue of $\mathcal{I}_{Y_1}(\theta_0)$. Let μ_n be the largest eigenvalue of Σ_n and $\theta \in B(\theta_0, \delta)$.
$$\sup_{||\gamma=1||} |1 + \gamma^T \Sigma_n^{1/2} \mathcal{H}\ell_n(\theta) \Sigma_n^{1/2}\gamma|$$
$$= \sup_{||\gamma||=1} |\gamma^T \Sigma_n^{1/2}(\Sigma_n^{-1} + \mathcal{H}\ell_n(\theta)\Sigma_n^{1/2})\gamma| \leq \mu_n \sup_{||\gamma||=1} |\gamma^T(\Sigma_n^{-1} + \mathcal{H}\ell_n(\theta))\gamma|$$
$$\leq \mu \left(\sup_{||\gamma||=1} |\gamma^T(\Sigma_n^{-1} + \mathcal{H}\ell_n(\theta_0))\gamma| + \sup_{||\gamma||=1} |\gamma^T(\mathcal{H}\ell_n(\theta_0) - \mathcal{H}\ell_n(\theta_0))\gamma| \right)$$
If $\hat{\Theta}_n \in B(\theta_0, \delta)$ and $|\mu - n\mu_n| < \epsilon$, then the expression above is bounded above by
$$(\mu + \epsilon)\frac{2}{n}\sum_{i=1}^{n} H_\delta(Y_i, \theta_0)$$
which converges in probability to a value no greater than ϵ.

2. Let $\Omega = (-1, 1)$ and let $Z_1, Z_2, \cdots$ be independent standard normal random variables. Define the *first order autoregressive process*
$$Y_n = \theta Y_{n-1} + Z_n, \quad n = 1, 2, \cdots.$$
Let $X_n = (Y_1, \cdots, Y_n)$. Then
   - $$\ell_n(\theta) = k - \frac{1}{2}(Y_n^2 + (1+\theta^2)\sum_{i=1}^{n-1} Y_i^2 - 2\theta\sum_{i=1}^{n} Y_i Y_{i-1}).$$
   - $$\hat{\Theta}_n = \frac{\sum_{i=1}^{n} Y_i Y_{i-1}}{\sum_{i=1}^{n-1} Y_i^2}$$
   - Several of the regularity conditions can be satisfied by taking a prior having a positve continuous density.

- $$\ell_n''(\theta) = -\sum_{i=1}^n Y_{i-1}^2$$

  does not depend on θ.

- $$\mathrm{Cov}_\theta(Y_i^2, Y_{i-k}^2) = \theta^{2k}\mathrm{Var}_\theta(Y_{i-k}^2) \leq \frac{2\theta^{2k}}{(1-\theta^2)^2}$$

  and therefore

  $$\lim_{n\to\infty} \mathrm{Var}_\theta(\frac{1}{n}\sum_{i=1}^n Y_{i-1}^2) = 0.$$

  Consequently,

  $$n\Sigma_n \text{ converges in probability} \quad \text{and} \quad \Sigma_n \to^P 0.$$

- Under P_{θ_0},

  $$\frac{1}{n}\sum_{i=1}^n Y_iY_{i-1} \to^P \theta_0.$$

- $$\ell_n(\theta) - \ell_n(\theta_0) = -\frac{\theta-\theta_0}{2}\left((\theta+\theta_0)\sum i = 1^n Y_i^2 - 2\sum_{i=1}^n Y_iY_{i-1}\right).$$

3. Let $Y_1, Y_2, \cdots$ be conditionally IID given $\Theta = \theta$ with Y_i a $N(\theta, 1/i)$ random variable. Set $X_n = (Y_1, \cdots, Y_n)$. Then, for some constant K,

- $$\ell_n(\theta) = K - \frac{1}{2}\left(\sum_{i=1}^n \log i + \frac{1}{i}(Y_i-\theta)^2\right).$$

- $$\hat{\Theta}_n = \sum_{i=1}^n \frac{Y_i}{i} / \sum_{i=1}^n \frac{1}{i}.$$

- $$\ell_n''(\theta) = -\sum_{i=1}^n \frac{1}{i}$$

  which does not depend on θ.

- $$\Sigma_n = 1/\sum_{i=1}^n \frac{1}{i} \;\sim\; \frac{1}{\log n}.$$

- $$\lambda_n(\ell_n(\theta) - \ell_n(\theta_0)) = \frac{\theta - \theta_0}{\sum_{i=1}^n 1/i} \sum_{i=1}^n \frac{1}{i}(Y_i - \frac{\theta + \theta_0}{2}).$$

  and
  $$\sum_{i+1}^n \frac{Y_i}{i} / \sum_{i=1}^n \frac{1}{i} \quad \text{is} \quad N(\theta_0, 1/\sum_{i=1}^n \frac{1}{i}).$$

  Thus,
  $$\lambda_n(\ell_n(\theta) - \ell_n(\theta_0)) \to^P -\frac{1}{2}(\theta - \theta_0)^2.$$

- If the prior is continuous, the the remaining conditions are satisfied.

Note that $\hat{\Theta}_n$ is not $\sqrt{n}$ consistent, but the posterior distribution is asymptotically normal.

8.5 Classical Chi-square Tests

Definition. Let $\Omega \subset R^p$ and let

$$\Omega_H = \{\theta = (\theta_1, \cdots, \theta_p) : \theta_i = c_i, 1 \leq i \leq k\}.$$

Then the *likelihood ratio criterion*

$$L_n = \frac{\sup_{\theta \in \Omega_H} f_{X|\Theta}(x|\theta)}{\sup_{\theta \in \Omega} f_{X|\Theta}(x|\theta)} = \frac{f_{X|\Theta}(x|\hat{\Theta}_{n,H})}{f_{X|\Theta}(x|\hat{\Theta}_n)}$$

where

$\hat{\Theta}_{n,H}$ is the MLE of θ on Ω_H and
$\hat{\Theta}_n$ is the (unrestricted) MLE.

Theorem. Assume the conditions in the theorem for the asymptotic normality of maximum likelihood estimators and let L_n be the likelihood ratio criterion for

$$H : \Theta_i = c_i \text{ for all } i = 1, \cdots, k \quad \text{versus} \quad A : \Theta_i \neq c_i \text{ for all } i = 1, \cdots, k.$$

Then, under H
$$-2 \log L_n \to^{\mathcal{D}} \chi^2_k$$
as $n \to \infty$.

For the case $p = k = 1$,

$$-2 \log L_n = -2\ell_n(c) + 2\ell_n(\hat{\Theta}_n) = 2(c - \hat{\Theta}_n)\ell_n(\hat{\Theta}_n) + (c - \hat{\Theta}_n)^2 \ell''_m(\theta^*_n)$$

for some θ_n^* between c and $\hat{\Theta}_n$. Note that $\ell_n'(\hat{\Theta}_n) = 0$ and under P_c,

$$\frac{1}{n}\ell_n''(\theta_n^*) \to^P \mathcal{I}_{X_1}(c), \quad \sqrt{n}(c - \hat{\Theta}_n) \to^{\mathcal{D}} Z/\mathcal{I}_{X_1}(c),$$

where Z is a standard normal random variable. Therefore,

$$-2\log L_n \to^{\mathcal{D}} Z^2$$

which is χ_1^2.

Proof. Let ψ_0 be a $p-k$-dimensional vector and set

$$\left(\begin{array}{c} c \\ \psi_0 \end{array}\right).$$

If the conditions of the theorem hold for Ω, then they also hold for Ω_H. Write

$$\hat{\Theta}_{n,H} = \left(\begin{array}{c} c \\ \hat{\Psi}_{n,H} \end{array}\right) \quad \hat{\Theta}_{n,H} = \left(\begin{array}{c} \hat{c} \\ \hat{\Psi}_n \end{array}\right).$$

Then for some θ^* between $\hat{\Theta}_n$ and $\hat{\Theta}_{n,H}$,

$$\ell_n(\hat{\Theta}_{n,H}) = \ell_n(\hat{\Theta}_n) + (\hat{\Theta}_{n,H} - \hat{\Theta}_n)^T\nabla_\theta\ell_n(\hat{\Theta}_n) + \frac{1}{2}(\hat{\Theta}_{n,H} - \hat{\Theta}_n)^T\mathcal{H}\ell_n(\hat{\Theta}_n)(\hat{\Theta}_{n,H} - \hat{\Theta}_n).$$

In addition, for some $\tilde{\theta}_n$ between θ_0 and $\hat{\Theta}_n$ and some $\tilde{\theta}_{n,H}$ between θ_0 and $\hat{\Theta}_{n,H}$,

$$0 = \nabla_\theta\ell_n(\hat{\Theta}_n) + \mathcal{H}_\theta\ell_n(\tilde{\theta}_n)(\hat{\Theta}_n - \theta_0)$$

and

$$0 = \nabla_\psi\ell_n(\hat{\Theta}_{n,H}) + \mathcal{H}_\psi\ell_n(\tilde{\theta}_{n,H})(\hat{\Psi}_n - \theta_0).$$

Write

$$\left(\begin{array}{cc} \hat{A}_n & \hat{B}_n \\ \hat{B}_n^T & \hat{D}_n \end{array}\right) = \frac{1}{n}\mathcal{H}_\theta\ell_n(\tilde{\theta}_n) \to^P -\mathcal{I}_{X_1}(\theta_0) = \left(\begin{array}{cc} A_0 & B_0 \\ B_0^T & D_0 \end{array}\right)$$

and

$$\hat{D}_{n,H} = \frac{1}{n}\mathcal{H}_\theta\ell_n(\tilde{\theta}_{n,H}) \to^P D_0.$$

Taking the last $p-k$ coordinates form the first expansion and equating it to the second yields

$$\hat{D}_{n,H} = (\Psi_{n,H} - \psi_0) = \hat{B}_n^T(\hat{c} - c) + \hat{D}_n(\hat{\Psi}_n - \psi_0).$$

From the asymptotic normality of the MLE we have that

$$\sqrt{n}(D_0(\Psi_{n,H} - \psi_0) - (B_0^T(\hat{c} - c) + D_0(\Psi_n - \psi_0))).$$

or equivalently that

$$\sqrt{n}((\hat{\Psi}_{n,H} - \hat{\Psi}_n) - D_0^{-1}B_0^T(\hat{c} - c))$$

is bounded in probability. Therefore,

$$\ell_n(\hat{\Theta}_{n,H}) - \ell(\hat{\Theta}) - \frac{n}{2}\begin{pmatrix} c-\hat{c} \\ D_0^{-1}B_0^T(\hat{c}-c) \end{pmatrix}^T \begin{pmatrix} A_0 & B_0 \\ B_0^T & D_0 \end{pmatrix} \begin{pmatrix} c-\hat{c} \\ D_0^{-1}B_0^T(\hat{c}-c) \end{pmatrix} \to^P 0.$$

This last term simplifies to

$$\frac{n}{2}(c-\hat{c})(A_0 - B_0 D_0^{-1} B_0^T)(c-\hat{c}).$$

The matrix $A_0 - B_0 D_0^{-1} B_0^T$ is the upper left $k \times k$ corner of $-\mathcal{I}_{X_1}(\theta_0)^{-1}$, the asymptotic covariance of c. Therefore independent of the choice of ψ_0,

$$-2\log L_n \to^{\mathcal{D}} C^2,$$

a χ^2_k random variable

Theorem. Let $\Gamma \subset R^k$ be a parameter space with parameter Ψ and let $R_1, \cdots, R_p$ be a partition of $\mathcal{X}$. For $i = 1, \cdots, p$, let

$$Y_i = \sum_{j=1}^{n} I_{R_i}(X_k)$$

be the number of observations in R_i (called the *reduced data*). Define

$$q_i(\psi) = P_\psi(R_i)$$

and

$$q(\psi) = (q_1(\psi), \cdots, q_k(\psi)).$$

Assume that $q \in C^2(\Gamma)$ is one-to-one. Let $\hat{\Psi}_n$ be the maximum likelihood estimate based on the reduced data and let $\mathcal{I}_{X_1}(\psi)$ be the Fisher information matrix. Assume that

$$\sqrt{n}(\hat{\Psi}_n - \psi) \to^{\mathcal{D}} W$$

where W is a $N(0, \mathcal{I}_{X_1}(\psi)^{-1})$ random vector. Define

$$\hat{q}_{i,n} = q_i(\hat{\Psi}_n)$$

and

$$C_n = \sum_{i=1}^{p} \frac{(Y_i - n\hat{q}_{i,n})^2}{\hat{q}_{i,n}}.$$

Then,

$$C_n \to^{\mathcal{D}} C$$

where C is a χ^2_{p-k-1} random variable.

9 Hierarchical Models

Definition. A sequence $X_1, X_2, \cdots$ of random variables is called *exchangeable* if the distribution of the sequence is invariant under finite permutations.

DiFinetti's theorem states that the sequence $X_1, X_2, \cdots$ can be represented as the mixture of IID sequences. Thus, the Bayes viewpoint is to decide using the data, what choice form the mixture is being observed.

Hierarchical models requires a sense of partial exchangeability.

Definition. A sequence $X_1, X_2, \cdots$ of random variables is called *marginally partially exchangeable* if it can be partitioned deterministically, into subsequences

$$X_1^{(k)}, X_2^{(k)}, \cdots \quad \text{for } k = 1, 2, \cdots.$$

Example. (One way analysis of variance) Let $\{j_n; n \geq 1\}$ be a sequence with each $j_n \in \{0, 1\}$. Then

$$f_{X_1 \cdots, X_n|M}(x_1, \cdots, x_n|\mu) = \frac{1}{(2\pi\sigma)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i - \mu_{1-j_i})^2\right).$$

We introduce the *general hierarchical model* by considering a protocal in a series of clinical trials in which several treatments are considered. The observations inside each treatment group are typically modeled as exchangeable. If we view the treatment groups symmetrically prior to observing the data, then we may take the set of parameters corresponding to different groups as a sample from another population. Thus, the parameters are exchangeable. This second level parameters necessary to model the joint distribution of the parameters are called the *hyperparameters*.

Denote the data by X, the parameters by Θ and the hyperparameters by Ψ, then the conditional density of the parameters given the hyperparamters is

$$f_{\Theta|X,\Psi}(\theta|x,\psi) = \frac{f_{X|\Theta,\Psi}(x|\theta,\psi)f_{\Theta|\Psi}(\theta|\psi)}{f_{X|\Psi}(x|\psi)}$$

where the density of the data given the hyperparameters alone is

$$f_{X|\Psi}(x|\psi) = \int f_{X|\Theta,\Psi}(x|\theta,\psi)f_{\Theta|\Psi}(\theta|\psi)\ \nu_\Theta(d\theta).$$

The marginal posterior distribution of the parameters can be found from

$$f_{\Theta|X}(\theta|x) = \int f_{\Theta|X,\Psi}(\theta|x,\psi)f_{\Psi|X}(\psi|x)\ \nu_\Psi(d\psi)$$

where the posterior density of Ψ given $X = x$ is

$$f_{\Psi|X}(\psi|x) = \frac{f_{X|\Psi}(x|\psi)f_\Psi(\psi)}{f_X(x)}$$

and the marginal density of X is

$$f_X(x) = \int f_{X|\Psi}(x|\psi) f_\Psi(\psi)\ \nu_\Psi(d\psi).$$

Example.

1. Let $X_{i,j}$ denote the observed response of the j-th subject in treatment group $i = 1, \cdots, k$. We have parameters
$$M = (M_1, \cdots, M_k).$$
If $(M_1, \cdots, M_k) = (\mu_1, \cdots, \mu_k)$, we can model $X_{i,j}$ as independent $N(\mu_i, 1)$ random variables. M itself can be modeled as exhangeable $N(\Theta, 1)$ random variables. Thus, Θ is the hyperparameter.

2. Consider $X_{i,j}$ denote the observed answer of the j-th person to a "yes-no" question in city $i = 1, \cdots, k$. The observations in a single city can be modeled as exhangeable Bernoulli random variables with parameters
$$P = (P_1, \cdots, P_k).$$
These parameters can be modeled as $Beta(A, B)$. Thus P is the parameter and (A, B) are the hyperparameters.

9.1 Normal Linear Models

For one-way analysis of variance (ANOVA), consider independent real valued observations

$$X_{i,j},\ j = 1, \cdots, n_i,\ i = 1, \cdots, k$$

that are $N(\mu_i, \sigma^2)$ given $M = (\mu_1, \cdots, \mu_n)$ and

$$M_1, \cdots, M_k$$

are independent $N(\psi, \tau^2)$ given $\Psi = \psi$ and $T = \tau$. We model Ψ as $N(\psi_0, \tau^2/\zeta_0)$. The distribution of T and the joint distribution of (Σ, T) remained unspecified for now.

Thus

Stage	Density
Data	$(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^k (n_i(\bar{x}_i - \mu_i)^2 + (n_1 - 1)s_i^2)\right)$
Parameter	$(2\pi\tau^2)^{-k/2} \exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^k (\mu_i - \psi)^2\right)$
Hyperparameter	$(2\pi\tau^2/\zeta_0)^{-1/2} \exp\left(-\frac{\zeta_0}{2\tau^2}(\psi - \psi_0)^2\right)$
Variance	$f_{\Sigma,T}(\sigma, \tau)$

From the usual updating of normal distributions, we see that, conditioned on $(\Sigma, T, \Psi) = (\sigma, \tau, \psi)$, the posterior of $M_1, \cdots, M_k$ are independent

$$N(\mu_i(\sigma, \tau, \psi), \frac{\tau^2\sigma^2}{n_i\tau^2 + \sigma^2}), \quad \mu_i(\sigma, \tau, \psi) = \frac{n_i\bar{x}_i\tau^2 + \psi\sigma^2}{n_i\tau^2 + \sigma^2}.$$

Consequently, given $(\Sigma, T, \Psi) = (\sigma, \tau, \psi)$, the data $\bar{X}_i$ and S_i are independent with

$$\bar{X}_i, \text{ a } N(\psi, \frac{\sigma^2}{n_i} + \tau^2) \text{ random variable, and } (n_i - 1)\frac{S_i^2}{\sigma^2}, \text{ a } \chi^2_{n_i-1} \text{ random variable.}$$

This gives a posterior on Ψ conditioned on $\Sigma = \sigma$ and $T = \tau$ that is

$$N\left(\psi_1(\sigma,\tau), \left(\frac{\zeta_0}{\tau^2} + \sum_{i=1}^k \frac{n_i}{\sigma^2 + \tau^2 n_i}\right)^{-1}\right), \quad \text{where} \quad \psi_1(\sigma,\tau) = \frac{\frac{\zeta_0\psi_0}{\tau^2} + \sum_{i=1}^k \frac{n_i \bar{x}_i}{\sigma^2+\tau^2 n_i}}{\frac{\zeta_0}{\tau^2} + \sum_{i=1}^k \frac{n_i}{\sigma^2+\tau^2 n_i}}.$$

To find the posterior distribution of (Σ, T), let $\bar{X} = (\bar{X}_1, \cdots, \bar{X}_k)$. Then, given $(\Sigma, T) = (\sigma, \tau)$

$$\bar{X} \text{ is } N(\psi_0 1, W(\sigma,\tau)) \quad \frac{n_i - 1}{\sigma} S_i^2 \text{ is } \chi^2_{n_i-1},$$

where $W(\sigma,\tau)$ has diagonal elements

$$\frac{\sigma^2}{n_i} + \tau^2(1 + \frac{1}{\zeta_0}), \ i = 1, \cdots, k.$$

and off diagonal elements

$$\frac{\tau^2}{\zeta_0}.$$

Consequently,

$$f_{\Sigma,T|\bar{X},s_1^2,\cdots,S_k^2}(\sigma,\tau|\bar{x}, s_1^2, \cdots, s_k^2)$$

is proportional to

$$f_{\Sigma,T}(\sigma,\tau)\sigma^{-(n_1+\cdots n_k - k)}\det(W(\sigma,\tau))^{-1/2}\exp\left(-\sum_{i=1}^k \left(\frac{s_i^2(n_i-1)}{2\sigma^2}\right) - \frac{1}{2}(\bar{x} - \psi_0 1)^T W^{-1}(\sigma,\tau)(\bar{x} - \psi_0 1)\right).$$

This situation simplifies in the case

$$T = \frac{\Sigma}{\sqrt{\lambda}}.$$

In this case, write $\gamma_0 = \lambda\zeta_0$, $\lambda_i = \lambda + n_i$, and $\gamma_i = n_i\lambda/\lambda_i$.
so that

$$\mu_i(\psi,\sigma,\tau) = \frac{n_i\bar{x}_i + \psi\lambda}{\lambda_i} = \mu_i(\psi), \ \frac{\tau^2\sigma^2}{n_i\tau^2 + \sigma^2} = \frac{\sigma^2}{\lambda_i}, \ \psi_1(\sigma,\tau) = \frac{\gamma_0\psi_0 \sum_{i=1}^k \gamma_i \bar{x}_i}{\gamma_0 + \sum_{i=1}^k \gamma_i} = \psi_1,$$

$$\frac{\zeta_0}{\tau^2} + \sum_{i=1}^k \frac{n_i}{\sigma^2 + \tau^2 n_i} = \gamma_0 + \sum_{i=1}^k \gamma_i.$$

Therefore

$$W(\sigma,\tau)_{ij} = \frac{1}{\gamma_0} + \left(\frac{1}{\lambda} + \frac{1}{n_i}\right)\delta_{ij}$$

and after some computation

$$\det(W(\sigma,\tau)) = \sigma^{2k}\prod_{i=1}^{k}\frac{1}{\gamma_i}(1+\frac{1}{\gamma_0}\sum_{j=1}^{l}\gamma_j).$$

With the prior $\Gamma^{-1}(a_0/2, b_0/2)$ for Σ^2, we have the posterior $\Gamma^{-1}(a_1/2, b_1/2)$ with

$$a_1 = a_0 + \sum_{i=1}^{k} n_i,\ |\gamma| = \sum_{i=1}^{k}\gamma_i,\ b_1 = b_0 + \sum_{i=1}^{k}((n_i-1)s_i^3 + \gamma_i(\bar{x}-u)^2) + \frac{|\gamma|\gamma_0}{\gamma_0+|\gamma|}(u-\psi_0)^2$$

where

$$|\gamma| = \sum_{i=1}^{k}\gamma_i, \quad u = \sum_{i=1}^{k}\frac{\gamma_i\bar{x}_i}{|\gamma|}.$$

Posterior distributions for linear functions are now t distributions. For example

$$\Psi \text{ is } t_{a_1}(\psi_1, \frac{b_1}{a_1}\frac{1}{\gamma+|\gamma|}), \quad M_i \text{ is } t_{a_1}(\mu_i(\psi_1), \frac{b_1}{a_1}\left(\frac{1}{\lambda_i}\left(\frac{\lambda}{\lambda_i}^2\frac{1}{\gamma+\gamma^*}\right)\right)),$$

9.2 Bernoulli Process Data

The data from group $i = 1, \cdots, k$ is

- n_i, the number of subjects, and
- X_i, the number of succeses.

The successes can be modeled by the success parameters

$$(P_1, \cdots, P_k).$$

The hyperparameters are (Θ, R) and given $(\Theta, R) = (\theta, r)$, the parameters are independent

$$Beta(\theta r, (1-\theta)r).$$

This random variable has

$$\text{mean } \theta, \text{ and variance } \frac{\theta(1-\theta)}{r+1}.$$

Thus, θ is the mean value of the P_i. The larger R is, the more similar are the P_i. Given $(\Theta, R, X) = (\theta, r, x)$. the P_i are independent

$$Beta(\theta r + x_i, (1-\theta)r + n_i - x_i)$$

random variables.

The posterior distribution of (Θ, R) is proportional to

$$f_{\Theta,R}(\theta, r)\frac{\Gamma(r)^k}{\Gamma(\theta r)^k\Gamma((1-\theta)r)^k}\prod_{i=1}^{k}\frac{\Gamma(\theta r + x_i)\Gamma((1-\theta)r + n_i - x_i)}{\Gamma(r+n_i)}$$

The next step uses numerical techniques or approximations. One possible approximation for large n_i is to note that

$$\frac{d}{dp}\arcsin\sqrt{p} = \frac{1}{2\sqrt{p(1-p)}},$$

and

$$\mathrm{Var}g(Z) \approx g'(\mu_Z)^2\mathrm{Var}(Z)$$

to obtain

$$Y_i = 2\arcsin\sqrt{\frac{X_i}{n_i}} \text{ is approximately } N(2\arcsin\sqrt{p_i}, \frac{1}{n_i})$$

random variable.

$$M_i = 2\arcsin\sqrt{P_i} \text{ is approximately } N(\mu, \frac{1}{\tau})$$

random variable given $(M, T) = (\mu, \tau) = (2\arcsin\sqrt{\theta}, r+1)$ Finally,

$$M \text{ is } N(\mu_0, \frac{1}{\lambda\tau}),$$

and T is unspecified.

9.3 Empirical Bayes Analysis

The naïve approach to empirical Bayes analysis is to estimate the hyperparameters at some level of the hierarchical model, treat them as known and use resulting posterior distributions for lower levels of the hierarchy.

Examples

1. The simpliest empirical Bayes method is to estimate prior parameters by viewing the data $x = (x_1, \cdots, x_n)$ as a sample from the marginal distribution of X given the hyperparameter Ψ on H.

   Let $X_1, \cdots, X_n$ in independent $N(\mu, \sigma_0^2)$ random variables. Assume that σ_0^2 is known and that μ has a prior distribution that is $N(\mu_0, \tau^2)$. To obtain moment estimates of $\psi = (\mu_0, \tau^2)$, we need to calculate

   $$\int_{R^n} x_1 f_X(x)\ dx = \int_{R_n}\int_H x_1 f_{X|\Psi}(x|\psi)f_\Psi(\psi)\ d\psi dx = \int_H \mu f_\Psi(\psi)\ d\psi = \mu_0,$$

   and

   $$\int_{R^n} x_1^2 f_X(x)\ dx = \int_{R_n}\int_H x_1^2 f_{X|\Psi}(x|\psi)f_\Psi(\psi)\ d\psi dx = \sigma_0^2 + \int_H \mu^2 f_\Psi(\psi)\ d\psi = \sigma_0^2 + \mu_0^2 + \tau^2.$$

This gives moment estimates

$$\hat{\mu}_0 = \bar{x} \text{ and } \hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \sigma_0^2$$

Under the usual Bayesian approach, the conditional distribution of μ given $X = x$ is $N(\mu_1(x), \tau_1(x)^2)$ where

$$\mu_1(x) = \frac{\sigma_0\mu_0 + n\tau^2\bar{x}}{n\tau^2 + \sigma_0^2}, \text{ and } \tau_1(x)^2 = \frac{\tau^2\sigma_0^2}{n\tau^2 + \sigma_0^2}.$$

The empirical Bayes approach replaces μ_0 with $\hat{\mu}_0$ and τ_0 with $\hat{\tau}_0$. Note that $\hat{\tau}^2$ can be negative.

2. In the case of one way analysis of variance, we can say that

$$\bar{X} \text{ is } N_k(\psi_0 1, W(\sigma, \tau)), \text{ and } (n_i - 1)\frac{S_i^2}{\sigma^2} \text{ is } \chi^2_{n_i - 1}$$

given $(\Psi, T, \Sigma) = (\psi_0, \tau, \sigma)$. Set

$$\Lambda = \Sigma^2/T^2 \text{ and } n = \sum_{i=1}^{k} n_i,$$

then the likelihood of $(\Psi, \Sigma^2, \Lambda)$ is

$$\prod_{i=1}^{k}\left(\frac{1}{n_i} + \frac{1}{\lambda}\right)^{-1/2}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\left(\frac{(\bar{x}_i - \psi)^2}{1/n_i + 1/\lambda} + (n_i - 1)s_i^2\right)\right).$$

Fix σ^2 and λ, then the expression above is maximized over ψ be taking

$$\hat{\Psi}(\lambda) = \sum_{i=1}^{k}\frac{\bar{x}_i}{1/n_i + 1/\lambda} \Big/ \sum_{i=1}^{k}\frac{1}{1/n_i + 1/\lambda}.$$

Using this value for ψ and fixing λ, we see that the expression is maximized over σ^2 by taking

$$\hat{\Sigma}^2(\lambda) = \frac{1}{n}\sum_{i=1}^{k}\left(\frac{(\bar{x}_i - \Psi(\lambda))^2}{1/n_i + 1/\lambda} + (n_i - 1)s_i^2\right).$$

Using this value for σ^2 we obtain the function

$$\prod_{i=1}^{k}\left(\frac{1}{n_i} + \frac{1}{\lambda}\right)^{-1/2}\hat{\Sigma}^2(\lambda)^{-n/2}$$

to obtain the MLE estimate $\hat{\Lambda}$. Then set

$$\hat{\Sigma}^2 = \hat{\Sigma}^2(\hat{\Lambda}) \text{ and } \hat{\Psi} = \hat{\Psi}(\hat{\Lambda}).$$

For the special case $n_i = m$ for all i,

$$\hat{\Psi}(\lambda) = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i,$$

which does not depend on λ. Set

$$\gamma = \frac{1}{m} + \frac{1}{\lambda},$$

then

$$\hat{\Sigma}^2(\gamma) = \frac{1}{n\gamma}\sum_{i=1}^{k}(\bar{x}_i - \hat{\Psi})^2 + \frac{m-1}{n}\sum_{i=1}^{k} s_i^2.$$

Substitute this in for the likelihood above, and take the derivative of the logarithm to obtain

$$-\frac{k}{2\gamma} + \frac{n}{2\hat{\Sigma}^2(\gamma)}\frac{\sum_{i=1}^{k}(\bar{x}_i - \hat{\Psi})^2}{n\gamma^2}.$$

Setting this equal to zero gives

$$\gamma = \frac{1}{k\Sigma^2(\gamma)}\sum_{i=1}^{k}(\bar{x}_i - \psi)^2$$

or

$$\hat{\Gamma} = \frac{(n-k)\sum_{i=1}^{k}(\bar{x}_i - \psi)^2}{k(m-1)\sum_{i=1}^{l} s_i^2} = \frac{k-1}{km}F$$

where F is an F-distribution.

Note that $\gamma \geq 1/m$. If $F < k/(k-1)$, then the derivative above is negative at $\gamma = 1/m$ and so the maximum occurs at $1/m$. Consequently, the maximum likelihood estimator

$$\hat{\Lambda} = \begin{cases} \frac{mk}{(k-1)F-k} & \text{if } F > \frac{k}{k-1} \\ \infty & \text{otherwise.} \end{cases}$$

Consequently, $\hat{T}^2 = 0$ if $F \leq k/(k-1)$.

This naïve approach, because it estimates the hyperparameters and then takes them as known, underestimates the variances of the parameters. In the case on one way ANOVA, to reflect the fact that Ψ is not known, the posterior variance of M_i should be increased by

$$\left(\frac{\Sigma^4}{n_i T^2 + \Sigma^2}\right)^2 \mathrm{Var}(\Psi) = \left(\frac{\Lambda^2}{n_i T^2 + \Lambda}\right)^2 \mathrm{Var}(\Psi).$$

We can estimate Λ from the previous analysis and we can estimate $\mathrm{Var}(\Psi)$ by

$$1/\sum_{i=1}^{k}\frac{n_i}{\hat{\Sigma}^2 + n_i\hat{T}^2}.$$