**Sufficiency (and Ancillarity and Completeness)**

Roughly, $T(X)$ is sufficient for $\{P_\theta\}_{\theta \in \Theta}$ (or for $\theta$) provided the conditional distribution of $X$ given $T(X)$ is the same for each $\theta \in \Theta$.

**Example** Let $X = (X_1, X_2, \ldots, X_n)$, where the $X_i$ are iid Bernoulli $(p)$ and $T(X) = \sum_{i=1}^n X_i$. You may show that $P_p[X = x|T = t]$ is the same for each $p$.

In some sense, a sufficient statistic carries all the information in $X$. (A variable with the same distribution as $X$ can be constructed from $T(X)$ and some randomization whose form does not depend upon $\theta$.) The $64$ question is how to recognize one. The primary tool is the "Factorization Theorem" in its various incarnations. This says roughly that $T$ is sufficient iff

$$f_\theta(x) = g_\theta(T(x))h(x).$$

**Example** (Page 20 Lehmann TSH) Suppose that $X$ is discrete. Then $T(X)$ is sufficient for $\theta$ iff $\exists$ nonnegative functions $g_\theta(\cdot)$ and $h(\cdot)$ such that $P_\theta[X = x] = g_\theta(T(x))h(x)$.

(You can and should prove this yourself.)

The Stat 643 version of the above result is considerably more general and relies on measure-theoretic machinery. Basically, one can "easily" push the factorization result as far as "families of distributions dominated by a $\sigma$-finite measure." The following, standard **notation** will be used throughout these notes:

$(\mathcal{X}, \mathcal{B}, P_\theta)$ a probability space (for fixed $\theta$)

$$\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$$

$$E_\theta(g(X)) = \int g dP_\theta$$

$$T : (\mathcal{X}, \mathcal{B}) \rightarrow (T, \mathcal{F})$$ a statistic

$\mathcal{B}_0 = \mathcal{B}(T) = \{T^{-1}(F)\}_{F \in \mathcal{F}}$ the $\sigma$-algebra generated by $T$

**Definition 1** $T$ (or $\mathcal{B}_0$) is sufficient for $\mathcal{P}$ if for every $B \in \mathcal{B}$ $\exists$ a $\mathcal{B}_0$-measurable function $Y$ such that for all $\theta \in \Theta$

$$Y = E_\theta(I_B|\mathcal{B}_0) \text{ a.s. } P_\theta.$$
(Note that $E_\theta (I_B |\mathcal{B}_0 ) = P_\theta (B |\mathcal{B}_0 )$. This definition says that there is a single random variable that will serve as a conditional probability of $B$ given $\mathcal{B}_0$ for all $\theta$ in the parameter space.)

**Definition 2** \( \mathcal{P} \) is dominated by a $\sigma$-finite measure $\mu$ if $P_\theta \ll \mu \forall \theta \in \Theta$.

Then assuming that $\mathcal{P}$ is dominated by $\mu$, standard **notation** is to let

$$f_\theta = \frac{dP_\theta}{d\mu}$$

and it is these objects (R-N derivatives) that factor nicely iff $T$ is sufficient. To prove this involves some nontrivial technicalities.

**Definition 3** $\mathcal{P}$ has a countable equivalent subset if $\exists$ a sequence of positive constants \( \{c_i\}_{i=1}^\infty \) with $\sum c_i = 1$ and a corresponding sequence $\{\theta_i\}_{i=1}^\infty$ such that $\lambda = \sum c_i P_{\theta_i}$ is equivalent to $\mathcal{P}$ in the sense that $\lambda(B) = 0$ iff $P_\theta (B) = 0 \forall \theta$.

**Notation:** Assuming that $\mathcal{P}$ has a countable equivalent subset, let $\lambda$ be a choice of $\sum c_i P_{\theta_i}$ as in Definition 3.

**Theorem 1** (Page 575 Lehmann TSH) $\mathcal{P}$ is dominated by a $\sigma$-finite measure $\mu$ iff it has a countable equivalent subset.

**Theorem 2** (Page 54 Lehmann TSH) Suppose that $\mathcal{P}$ is dominated by a $\sigma$-finite measure and $T : (\mathcal{X}, B) \to (T, \mathcal{F})$. Then $T$ is sufficient for $\mathcal{P}$ iff $\exists$ nonnegative $\mathcal{F}$-measurable functions $g_\theta$ such that $\forall \theta$

$$\frac{dP_\theta}{d\lambda} (x) = g_\theta (T(x)) \quad \text{a.s. } \lambda .$$

(This last equality is equivalent to $P_\theta (B) = \int_B g_\theta (T(x)) d\lambda(x) \forall B \in \mathcal{B}$.)

**Theorem 3** (Page 55 Lehmann TSH) ("The" Factorization Theorem) Suppose that $\mathcal{P} \ll \mu$ where $\mu$ is $\sigma$-finite. Then $T$ is sufficient for $\mathcal{P}$ iff $\exists$ a nonnegative $\mathcal{B}$-measurable function $h$ and nonnegative $\mathcal{F}$-measurable functions $g_\theta$ such that $\forall \theta$

$$\frac{dP_\theta}{d\mu} (x) = f_\theta (x) = g_\theta (T(x)) h(x) \quad \text{a.s. } \mu .$$

Standard/elementary applications of Theorem 3 are to cases where:

1) $\mu$ is Lebesgue measure on $\mathbb{R}^d$, $\mathcal{P}$ is some parametric family of $d$-dimensional distributions, and the $f_\theta$ are $d$-dimensional probability densities, or
2) \( \mu \) is counting measure on some countable set (like, e.g., \((\mathbb{Z}^+)^d\)) \( \mathcal{P} \) is some parametric family of discrete distributions and the \( f_\theta \) are probability mass functions.

But it can also be applied much more widely (e.g. to cases where \( \mu \) is the sum of \( d \)-dimensional Lebesgue measure and counting measure on some countable subset of \( \mathcal{R}^d \)).

**Example** Theorem 3 can be used to show that for \( \mathcal{X} = \mathcal{R}^d \), \( X = (X_1, X_2, ..., X_d), B = B_d \) the \( d \)-dimensional Borel \( \sigma \)-algebra, \( \pi = (\pi_1, \pi_2, ..., \pi_d) \) a generic permutation of the first \( d \) positive integers, \( \pi(X) = (X_{\pi_1}, X_{\pi_2}, ..., X_{\pi_d}) \), \( \mu \) the \( d \)-dimensional Lebesgue measure and \( \mathcal{P} = \{ P | P \ll \mu \text{ and } f = \frac{dP}{d\mu} \text{ satisfies } f(\pi(x)) = f(x) \text{ a.s. } \mu \text{ for all permutations } \pi \} \), the order statistic 

\[ T(X) = (X_{(1)}, X_{(2)}, ..., X_{(d)}) \]

is sufficient.

Actually, it is an important fact (that does not follow from the factorization theorem) that one can establish the sufficiency of the order statistic even more broadly. With \( \mathcal{P} = \{ P \text{ on } \mathcal{R}^d \text{ such that } P(B) = P(\pi(B)) \text{ for all permutations } \pi \} \), direct argument shows that the order statistic is sufficient. (You should try to prove this. For \( B \in B \), try \( Y(x) \) equal to the fraction of all permutations that place \( x \) in \( B \).)

**(Obvious) Result** If \( T(X) \) is sufficient for \( \mathcal{P} \) and \( \mathcal{P}_0 \) is a subset of \( \mathcal{P} \), \( T(X) \) is sufficient for \( \mathcal{P}_0 \).

A natural question is "How far can one go in the direction of 'reducing' \( X \) without throwing something away?" The notion of minimal sufficiency is aimed in the direction of answering this question.

**Definition 4** A sufficient statistic \( T : (\mathcal{X}, \mathcal{B}) \rightarrow (T, \mathcal{F}) \) is minimal sufficient (for \( \theta \) or for \( \mathcal{P} \)) provided for every sufficient statistic \( S : (\mathcal{X}, \mathcal{B}) \rightarrow (S, \mathcal{G}) \exists \) a (?measurable?) function \( U : S \rightarrow T \) such that

\[ T = U \circ S \text{ a.e. } \mathcal{P} \].

(It is probably equivalent or close to equivalent that \( B(T) \subset B(S) \).)

There are a couple of ways of identifying a sufficient statistic.

**Theorem 4** (Like page 92 of Schervish ... other development on pages 41-43 of Lehmann TPE) Suppose that \( \mathcal{P} \) is dominated by a \( \sigma \)-finite measure \( \mu \) and \( T : (\mathcal{X}, \mathcal{B}) \rightarrow (T, \mathcal{F}) \) is sufficient for \( \mathcal{P} \). Suppose that there exist versions of the R-N derivatives \( \frac{dP}{d\mu}(x) = f_\theta(x) \) such that

\[ \text{(the existence of } k(x, y) > 0 \text{ such that } f_\theta(y) = f_\theta(x)k(x, y) \forall \theta \text{ implies } T(x) = T(y) \text{)} \]

then \( T \) is minimal sufficient.
(A trivial "generalization" of Theorem 4 is to suppose that the condition only holds for a set \( \mathcal{X}^* \subset \mathcal{X} \) with \( P_\theta \) probability 1 for all \( \theta \). The condition in Theorem 4 is essentially that \( T(x) = T(y) \) iff \( \frac{f_\theta(y)}{f_\theta(x)} \) is free of \( \theta \).

Note that sufficiency of \( T(X) \) means that \( T(x) = T(y) \) implies (the existence of \( k(x, y) > 0 \) such that \( f_\theta(y) = f_\theta(x)k(x, y) \forall \theta \)). So the condition in Theorem 4 could be phrased as an iff condition.

**Example** Let \( X = (X_1, X_2, \ldots, X_n) \), where the \( X_i \) are iid Bernoulli \( (p) \) and \( T(X) = \sum_{i=1}^{n} X_i \). You may use Theorem 4 to show that \( T \) is minimal sufficient.

Lehmann emphasizes a second ("more technical and less illuminating?") method of establishing minimality based on the following two theorems.

**Theorem 5** (Page 41 of Lehmann TPE) Suppose that \( \mathcal{P} = \{P_i\}_{i=0}^{k} \) is finite and let \( \{f_i\} \) be the R-N derivatives with respect to some dominating \( \sigma \)-finite measure (like, e.g., \( \sum_{i=0}^{k} P_i \)). Suppose that all \( k + 1 \) derivatives \( f_i \) are positive on all of \( \mathcal{X} \). Then

\[
T(X) = \left( \frac{f_1(x)}{f_0(x)}, \frac{f_2(x)}{f_0(x)}, \ldots, \frac{f_k(x)}{f_0(x)} \right)
\]

is minimal sufficient.

(Again it is a trivial "generalization" to observe that the minimality also holds if there is a set \( \mathcal{X}^* \subset \mathcal{X} \) with \( P_\theta \) probability 1 for all \( \theta \) where the \( k + 1 \) derivatives \( f_i \) are positive on all of \( \mathcal{X}^* \).)

**Theorem 6** (Page 42 of Lehmann TPE) Suppose that \( \mathcal{P} \) is a family of mutually absolutely continuous distributions and that \( \mathcal{P}_0 \subset \mathcal{P} \). If \( T \) is sufficient for \( \mathcal{P} \) and minimal sufficient for \( \mathcal{P}_0 \), then it is minimal sufficient for \( \mathcal{P} \).

The way that Theorems 5 and 6 are used together is to look at a few \( \theta \)'s, invoke Theorem 5 to establish minimal sufficiency for the finite family of a vector made up of a few likelihood ratios and then use Theorem 6 to infer minimality for the whole family.

**Example** Let \( X = (X_1, X_2, \ldots, X_n) \), where the \( X_i \) are iid Beta \((\alpha, \beta)\). Using \( P_0 \) the Beta \((1, 1)\) distribution, \( P_1 \) the Beta \((2, 1)\) distribution and \( P_2 \) the Beta \((1, 2)\) distribution, Theorems 5 and 6 can be used to show that \( \left( \prod_{i=1}^{n} X_i, \prod_{i=1}^{n} (1 - X_i) \right) \) is minimal sufficient for the Beta family.
**Definition 5** A statistic $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{F})$ is said to be ancillary for $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ if the distribution of $T$ (i.e., the measure $P_\theta^T$ on $(\mathcal{T}, \mathcal{F})$ defined by $P_\theta^T(F) = P_\theta(T^{-1}(F)))$ does not depend upon $\theta$ (i.e. is the same for all $\theta$).

**Definition 6** A statistic $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{R}^1, \mathcal{B}_1)$ is said to be first order ancillary for $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ if the mean of $T$ (i.e., $E_\theta(T(X)) = \int T dP_\theta$) does not depend upon $\theta$ (i.e. is the same for all $\theta$).

**Example** Let $X = (X_1, X_2, \ldots, X_n)$, where the $X_i$ are iid $\text{N}(\theta, 1)$. The sample variance of the $X_i$'s is ancillary.

Lehmann argues that roughly speaking models with nice (low dimensional) minimal sufficient statistics $T$ are ones where no nontrivial function of $T$ is ancillary (or first order ancillary). This line of thinking brings up the notion of completeness.

**Notation:** Continuing to let $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{F})$ and $P_\theta^T$ be the measure on $(\mathcal{T}, \mathcal{F})$ defined by $P_\theta^T(F) = P_\theta(T^{-1}(F))$, let $\mathcal{P}^T = \{P_\theta^T\}_{\theta \in \Theta}$.

**Definition 7a** $\mathcal{P}^T$ or $T$ is complete (boundedly complete) for $\mathcal{P}$ or $\theta$ if
  i) $U : (\mathcal{X}, \mathcal{B}_0) \to (\mathcal{R}^1, \mathcal{B}_1)$ $\mathcal{B}_0$-measurable (and bounded), and
  ii) $E_\theta U(X) = 0 \ \forall \theta$

imply that $U = 0$ a.s. $P_\theta \ \forall \theta$.

By virtue of Lehmann's Theorem (see Cressie's Probability Summary) this is equivalent to the next definition.

**Definition 7b** $\mathcal{P}^T$ or $T$ is complete (boundedly complete) for $\mathcal{P}$ or $\theta$ if
  i) $h : (\mathcal{T}, \mathcal{F}) \to (\mathcal{R}^1, \mathcal{B}_1)$ $\mathcal{F}$-measurable (and bounded), and
  ii) $E_\theta h(T(X)) = 0 \ \forall \theta$

imply that $h \circ T = 0$ a.s. $P_\theta \ \forall \theta$.

These definitions say that there is no nontrivial function of $T$ that is an unbiased estimator of 0. This is equivalent to saying that there is no nontrivial function of $T$ that is first order ancillary. Notice also that completeness implies bounded completeness, so that bounded completeness is a slightly weaker condition than completeness.

**Example** Let $X = (X_1, X_2, \ldots, X_n)$, where the $X_i$ are iid Bernoulli $(p)$ and $T(X) = \sum_{i=1}^n X_i$. You may use the fact that polynomials can be identically 0 on an interval only if their coefficients are all 0, to show that $T(X)$ is complete for $p \in [0, 1]$.

**(Obvious) Result 7** If $T$ is complete for $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ and $\Theta' \subset \Theta$, $T$ need not be complete for $\mathcal{P}' = \{P_\theta\}_{\theta \in \Theta'}$. 


**Theorem 8** If $\Theta \subset \Theta'$ and $\Theta$ dominates $\Theta'$ in the sense that $P_\theta(B) = 0 \forall \theta \in \Theta$ implies that $P_\theta(B) = 0 \forall \theta \in \Theta'$, then $T$ (boundedly) complete for $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ implies that $T$ is (boundedly) complete for $\mathcal{P}' = \{P_\theta\}_{\theta \in \Theta'}$.

Completeness of a statistic says that there is no part of it that is somehow irrelevant in and of itself to inference about $\theta$. That is reminiscent of minimal sufficiency.

**Theorem 9** (Bahadur's Theorem) Suppose that $T : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathcal{T}, \mathcal{F})$ is sufficient and boundedly complete.

i) (Pages 94-95 of Schervish) If $(T, \mathcal{F}) = (\mathcal{R}^d, \mathcal{B}_d)$ then $T$ is minimal sufficient.

ii) If there is any minimal sufficient statistic, $T$ is minimal sufficient.

**Example** Suppose that $\mathcal{P} \ll \mu$ (where $\mu$ is $\sigma$-finite) and that for $\theta \in \Theta \subset \mathcal{R}^k$

$$
\frac{dP_\theta}{d\mu}(x) = c(\theta) \exp \sum_{i=1}^{k} \theta_i T_i(x) .
$$

Provided, for example, that $\Theta$ contains an open rectangle in $\mathcal{R}^k$, Theorem 1 page 142 Lehmann TSH implies that $T(X) = (T_1(X), T_2(X), ..., T_k(X))$ is complete. It is clearly sufficient by the factorization criterion. So by part i) of Theorem 9, $T$ is minimal sufficient.

**Some Facts About Common Statistical Models**

**Bayes Models**

The "Bayes" approach to statistical problems is to add additional model assumptions to the basic structure introduced thus far. That is, to the standard assumptions concerning $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ "Bayesians" add a distribution on the parameter space $\Theta$. That is, suppose that

$$G$$

is a distribution on $(\Theta, \mathcal{C})$

where it may be convenient to suppose that $G$ is dominated by some $\sigma$-finite measure $\nu$ and that

$$\frac{dG}{d\nu}(\theta) = g(\theta)$$

If considered as a function of both $x$ and $\theta$, $f_\theta(x)$ is $\mathcal{B} \times \mathcal{C}$ measurable, then $\exists$ a probability (a joint distribution for $(X, \theta)$ on $(\mathcal{X} \times \Theta, \mathcal{B} \times \mathcal{C})$ defined by

$$
\pi^{X,\theta}(A) = \int_A f_\theta(x) d(\mu \times G)(x, \theta) = \int_A f_\theta(x) g(\theta) d(\mu \times \nu)(x, \theta)
$$
This distribution has marginals

\[ \pi^X(B) = \int_{B \times \Theta} f_\theta(x) d(\mu \times G)(x, \theta) = \int_B \int_{\Theta} f_\theta(x) dG(\theta) d\mu(x) \]

(so that \( \frac{d\pi^X}{d\mu} = \int_{\Theta} f_\theta(x) dG(\theta) = \int_{\Theta} f_\theta(x) g(\theta) d\nu(\theta) \)) and

\[ \pi^\theta(C) = \int_{X \times C} f_\theta(x) d(\mu \times G)(x, \theta) = \int_C \int_X f_\theta(x) d\mu(x) dG(\theta) = G(C) \]

Further, the conditionals (regular conditional probabilities) are computable as

\[ \pi^{X|\theta}(B|\theta) = \int_B f_\theta(x) d\mu(x) = P_\theta(B) \]

and

\[ \pi^{\theta|X}(C|x) = \int_C \frac{f_\theta(x)}{\int_{\Theta} f_\theta(x) dG(\theta)} dG(\theta) \]

Note then that

\[ \frac{d\pi^{\theta|X}}{dG}(\theta|x) = \frac{f_\theta(x)}{\int_{\Theta} f_\theta(x) dG(\theta)} \]

and thus that

\[ \frac{d\pi^{\theta|X}}{d\nu}(\theta|x) = \frac{f_\theta(x) g(\theta)}{\int_{\Theta} f_\theta(x) g(\theta) d\nu(\theta)} \]

is the usual "posterior density" for \( \theta \) given \( X = x \).

Bayesians use the posterior distribution or density (THAT DEPENDS ON ONE’S CHOICE OF \( G \)) as their basis of inference for \( \theta \). Pages 84-88 of Schervish concern a Bayesian formulation of sufficiency and argue (basically) that the Bayesian form is equivalent to the "classical" form used above.

**Exponential Families**

Many common families of distributions can be treated with a single set of analyses by recognizing them to be of a common "exponential family" form. The following is a list of facts about such families.

**Definition 8** Suppose that \( \mathcal{P} = \{ P_\theta \} \) is dominated by a \( \sigma \)-finite measure \( \mu \). \( \mathcal{P} \) is called an exponential family if for some \( h(x) \geq 0 \),
\[ f_\theta(x) \triangleq \frac{dP_\theta}{d\mu}(x) = \exp\left\{ a(\theta) + \sum_{i=1}^{k} \eta_i(\theta)T_i(x) \right\} h(x) \ \forall \theta. \]

**Definition 9** A family of probability measures \( \mathcal{P} = \{P_\theta\} \) is said to be identifiable if \( \theta_1 \neq \theta_2 \) implies that \( P_{\theta_1} \neq P_{\theta_2} \).

**Fact(s) 10** Let \( \eta = (\eta_1, \eta_2, \ldots, \eta_k), \) \( \Gamma = \{\eta \in \mathbb{R}^k | \int h(x) \exp\left\{ \sum_{i=1}^{k} \eta_i T_i(x) \right\} d\mu(x) < \infty\} \), and consider also the family of distributions (on \( \mathcal{X} \)) with R-N derivatives wrt \( \mu \) of the form

\[ f_\eta(x) = K(\eta) \exp\left\{ \sum_{i=1}^{k} \eta_i T_i(x) \right\} h(x), \ \eta \in \Gamma. \]

Call this family \( \mathcal{P}^* \). This set of distributions on \( \mathcal{X} \) is at least as large as \( \mathcal{P} \) (and this second parameterization is mathematically nicer than the first). \( \Gamma \) is called the natural parameter space for \( \mathcal{P}^* \). It is a convex subset of \( \mathbb{R}^k \). (TSH, page 57.) If \( \Gamma \) lies in a subspace of dimension less than \( k \), then \( f_\eta(x) \) (and therefore \( f_\theta(x) \)) can be written in a form involving fewer than \( k \) statistics \( T_i \). We will henceforth assume \( \Gamma \) to be fully \( k \)-dimensional. Note that depending upon the nature of the functions \( \eta_i(\theta) \) and the parameter space \( \Theta \), \( \mathcal{P} \) may be a proper subset of \( \mathcal{P}^* \). That is, defining \( \Gamma_\Theta = \{(\eta_1(\theta), \eta_2(\theta), \ldots, \eta_k(\theta)) \in \mathbb{R}^k | \theta \in \Theta\} \), \( \Gamma_\Theta \) can be a proper subset of \( \Gamma \).

**Fact 11** The "support" of \( P_\theta \), defined as \( \{x | f_\theta(x) > 0\} \), is clearly \( \{x | h(x) > 0\} \), which is independent of \( \theta \). The distributions in \( \mathcal{P} \) are thus mutually absolutely continuous.

**Fact 12** From the Factorization Theorem, the statistic \( T = (T_1, T_2, \ldots, T_k) \) is sufficient for \( \mathcal{P} \).

**Fact 13** \( T \) has induced measures \( \{P^T_\theta | \theta \in \Theta\} \) which also form an exponential family.

**Fact 14** If \( \Gamma_\Theta \) contains an open rectangle in \( \mathbb{R}^k \), then \( T \) is complete for \( \mathcal{P} \). (See pages 142-143 of TSH.)

**Fact 15** If \( \Gamma_\Theta \) contains an open rectangle in \( \mathbb{R}^k \) (and actually under the much weaker assumptions given on page 44 of TPE) \( T \) is minimal sufficient for \( \mathcal{P} \).

**Fact 16** If \( g \) is any measurable real valued function such that \( E_\eta |g(X)| < \infty \), then

\[ E_\eta g(X) = \int g(x) f_\eta(x) d\mu(x) \]

is continuous on \( \Gamma \) and has continuous partial derivatives of all orders on the interior of \( \Gamma \). These can be calculated as
\[
\frac{\partial^{\alpha_1+\alpha_2+\cdots+\alpha_k}}{\partial \eta_1^{\alpha_1} \partial \eta_2^{\alpha_2} \cdots \partial \eta_k^{\alpha_k}} E_{\eta} g(X) = \int \frac{\partial^{\alpha_1+\alpha_2+\cdots+\alpha_k}}{\partial \eta_1^{\alpha_1} \partial \eta_2^{\alpha_2} \cdots \partial \eta_k^{\alpha_k}} f_\eta(x) d\mu(x) .
\]

(See page 59 of TSH.)

**Fact 17** If for \( \mathbf{u} = (u_1, u_2, \ldots, u_k) \), both \( \eta_0 \) and \( \eta_0 + \mathbf{u} \) belong to \( \Gamma \)

\[
E_{\eta_0} \exp \left\{ u_1 T_1(X) + u_2 T_2(X) + \cdots + u_k T_k(X) \right\} = \frac{K(\eta_0)}{K(\eta_0 + \mathbf{u})} .
\]

Further, if \( \eta_0 \) is in the interior of \( \Gamma \), then

\[
E_{\eta_0}(T_1^{\alpha_1}(X) T_2^{\alpha_2}(X) \cdots T_k^{\alpha_k}(X)) = K(\eta_0)\frac{\partial^{\alpha_1+\alpha_2+\cdots+\alpha_k}}{\partial \eta_1^{\alpha_1} \partial \eta_2^{\alpha_2} \cdots \partial \eta_k^{\alpha_k}} \left( \frac{1}{K(\eta)} \right) \bigg|_{\eta = \eta_0} .
\]

In particular, \( E_{\eta_0} T_j(X) = \frac{\partial}{\partial \eta_j} (-\log K(\eta)) \bigg|_{\eta = \eta_0} \), \( \text{Var}_{\eta_0} T_j(X) = \frac{\partial^2}{\partial \eta_j^2} (-\log K(\eta)) \bigg|_{\eta = \eta_0} \), and \( \text{Cov}_{\eta_0}(T_j(X), T_i(X)) = \frac{\partial^2}{\partial \eta_j \partial \eta_i} (-\log K(\eta)) \bigg|_{\eta = \eta_0} \).

**Fact 18** If \( \mathbf{X} = (X_1, X_2, \ldots, X_n) \) has iid components, with each \( X_i \sim P_\theta \), then \( \mathbf{X} \) generates a \( k \) dimensional exponential family on \( (\mathcal{X}^n, \mathcal{B}^n) \) wrt \( \mu^n \). The \( k \) dimensional statistic \( \sum_{i=1}^n \mathbf{T}(X_i) \) is sufficient for this family. Under the condition that \( \Gamma_\Theta \) contains an open rectangle, this statistic is also complete and minimal sufficient.

"**Fisher Information**"

**Definition 10** We will say that the model \( \mathcal{P} \) with dominating \( \sigma \)-finite measure \( \mu \) is FI regular at \( \theta_0 \in \Theta \subset \mathcal{R}^1 \), provided there exists an open neighborhood of \( \theta_0 \), say \( O \), such that

i) \( f_\theta(x) > 0 \ \forall x \) and \( \forall \theta \in O \),

ii) \( \forall x, f_\theta(x) \) is differentiable at \( \theta_0 \) and

iii) \( 1 = \int f_\theta(x) d\mu(x) \) can be differentiated wrt \( \theta \) under the integral at \( \theta_0 \), i.e.

\[
0 = \int f_\theta(x) \frac{d}{d\theta} f_\theta(x) \bigg|_{\theta = \theta_0} d\mu(x) .
\]

(A definition like that on page 111 of Schervish, that requires i) and ii) only on a set of \( P_\theta \) probability 1 \( \forall \theta \) in \( O \), is really no more general than this. WOLOG, just throw away the exceptional set.)

**Jargon:** The (random) function of \( \theta, \frac{d}{d\theta} \log f_\theta(X) \) is called the score function. iii) of Definition 10 says that the mean of the score function is 0 at \( \theta_0 \).
Definition 11 If the model $\mathcal{P}$ is FI regular at $\theta_0$, then

$$I(\theta_0) = E_{\theta_0} \left( \frac{d}{d\theta} \log f_{\theta}(X) \bigg|_{\theta=\theta_0} \right)^2$$

is called the Fisher Information contained in $X$ at $\theta_0$.

(If a model is FI regular at $\theta_0$, $I(\theta_0)$ is the variance of the score function at $\theta_0$.)

Theorem 19 Suppose that the model $\mathcal{P}$ is FI regular at $\theta_0$. If $\forall x$, $f_{\theta}(x)$ is in fact twice differentiable at $\theta_0$ and $1 = \int f_{\theta}(x) d\mu(x)$ can be differentiated twice wrt $\theta$ under the integral at $\theta_0$, i.e. both $0 = \int \frac{d}{d\theta} f_{\theta}(x) \bigg|_{\theta=\theta_0} d\mu(x)$ and $0 = \int \frac{d^2}{d\theta^2} f_{\theta}(x) \bigg|_{\theta=\theta_0} d\mu(x)$ then

$$I(\theta_0) = -E_{\theta_0} \left( \frac{d^2}{d\theta^2} \log f_{\theta}(X) \bigg|_{\theta=\theta_0} \right).$$

$k$-dimensional versions of Definitions 10 and 11 are next.

Definition 10b We will say that the model $\mathcal{P}$ with dominating $\sigma$-finite measure $\mu$ is FI regular at $\theta_0 \in \Theta \subset \mathcal{R}^k$, provided there exists an open neighborhood of $\theta_0$, say $O$, such that

i) $f_{\theta}(x) > 0 \ \forall x$ and $\forall \theta \in O$,

ii) $\forall x$, $f_{\theta}(x)$ has first order partial derivatives at $\theta_0$ and

iii) $1 = \int f_{\theta}(x) d\mu(x)$ can be differentiated wrt to each $\theta_i$ under the integral at $\theta_0$,

i.e. $0 = \int \frac{\partial}{\partial \theta_i} f_{\theta}(x) \bigg|_{\theta=\theta_0} d\mu(x)$.

Definition 11b If the model $\mathcal{P}$ is FI regular at $\theta_0$ and $E_{\theta_0} \left( \frac{\partial}{\partial \theta_i} \log f_{\theta}(X) \bigg|_{\theta=\theta_0} \right)^2 < \infty \ \forall i$, then the $k \times k$ matrix

$$I(\theta_0) = \left( E_{\theta_0} \frac{\partial}{\partial \theta_i} \log f_{\theta}(X) \bigg|_{\theta=\theta_0} , \frac{\partial}{\partial \theta_j} \log f_{\theta}(X) \bigg|_{\theta=\theta_0} \right)$$

is called the Fisher Information contained in $X$ at $\theta_0$.

There is a multiparameter version of Theorem 19.

Theorem 19b Suppose that the model $\mathcal{P}$ is FI regular at $\theta_0$. If $\forall x$, $f_{\theta}(x)$ in fact has continuous second order partials in the neighborhood $O$, and $1 = \int f_{\theta}(x) d\mu(x)$ can be differentiated twice under the integral at $\theta_0$, i.e. both $0 = \int \frac{\partial}{\partial \theta_i} f_{\theta}(x) \bigg|_{\theta=\theta_0} d\mu(x)$ and $0 = \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_{\theta}(x) \bigg|_{\theta=\theta_0} d\mu(x) \ \forall i$ and $j$, then
\[ I(\theta_0) = -E_{\theta_0}\left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_\theta(X) \Bigg|_{\theta = \theta_0} \right). \]

**Fact 20** If \( X_1, X_2, \ldots, X_n \) are independent, \( X_i \sim P_{i\theta} \), then \( X = (X_1, X_2, \ldots, X_n) \) (that has distribution \( P_{1\theta} \times P_{2\theta} \times \cdots \times P_{n\theta} \)) carries Fisher Information \( I(\theta) = \sum_{i=1}^{n} I_i(\theta) \), where in the obvious way, \( I_i(\theta) \) is the Fisher information in \( X_i \).

**Fact 21** \( I(\theta) \) doesn't depend upon the dominating measure.

**Observation 22** The Fisher Information *does* depend upon the parameterization one uses. For example, suppose that the model \( \mathcal{P} \) is FI regular at \( \theta_0 \in \mathcal{R}^1 \) and that
\[ \eta = h(\theta) \text{ for some "nice" } h \text{ (say, 1-1 with derivative } h') \]

Then for \( P_{\theta} \) with R-N derivative \( f_{\theta} \) with respect to \( \mu \), the distributions \( Q_{\eta} \) for \( \eta \) in the range of \( \Theta \) are defined by
\[ Q_{\eta} = P_{h^{-1}(\eta)} \]
and have R-N derivatives (again with respect to \( \mu \))
\[ g_{\eta} = \frac{dQ_{\eta}}{d\mu} \]

Then continuing to let "prime" denote differentiation with respect to \( \theta \) or \( \mu \),
\[ g'_{\eta} = f'_{h^{-1}(\eta)} \left( \frac{1}{h'(h^{-1}(\eta))} \right) \]
and
\[ I(\eta) = \left( \left( \frac{1}{h'(h^{-1}(\eta))} \right) \right)^2 I(h^{-1}(\eta)) \]
(where the first "\( I \)" here is information about \( \eta \) and the second is information about \( \theta \)).

**Observation 23** If \( T(x) \) is a 1-1 function on \( \mathcal{X} \), then the Fisher Information in \( T(X) \) is the same as that in \( X \). This follows since the distributions of \( T(X), P_{\theta}^T \), are dominated by the measure \( \mu^T \) defined by \( \mu^T(F) = \mu(T^{-1}(F)) \) and have R-N derivatives
\[ \frac{dP_{\theta}^T}{d\mu^T} = f_\theta \circ T^{-1} \]
which will satisfy the FI regularity conditions and have the right derivatives and integrals.
**Theorem 24** Suppose the model \( P \) is FI regular at \( \theta_0 \). (And suppose that the family \( P^T = \{ P^T_\theta \}_{\theta \in \Theta} \) is also FI regular at \( \theta_0 \).) Then
\[
I_X(\theta_0) \geq I_{T(X)}(\theta_0)
\]
and if \( T(X) \) is sufficient for \( \theta \) there is equality.

**Theorem 24b** (Theorem 2.86 of Schervish) Suppose the model \( P \) is FI regular at \( \theta_0 \in \Theta \subset R^k \). (And suppose that the family \( P^T = \{ P^T_\theta \}_{\theta \in \Theta} \) is also FI regular at \( \theta_0 \).) Then the matrix
\[
I_X(\theta_0) - I_{T(X)}(\theta_0)
\]
is nonnegative definite
and if \( T(X) \) is sufficient for \( \theta \) the matrix contains all 0 elements.

**Fact 25** In exponential families, the Fisher Information is very simple. For a family of distributions as in **Fact(s) 10** above with densities \( f_\eta(x) \) with respect to \( \mu \),
\[
I(\eta_0) = \text{Var}_{\eta_0} T(X) = -\left( \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log K(\eta) \Bigg|_{\eta=\eta_0} \right)
\]

"Kullback-Leibler Information"

There is another notion of statistical information, intended to measure how far apart 2 distributions are in the sense of likelihood, or in some sense how easy it will be to discriminate between them on the basis of an observable, \( X \). (This measure is also useful in the asymptotics of maximum likelihood estimation.)

**Definition 12** If \( P \) and \( Q \) are probability measures on \( X \) with R-N derivatives with respect to a common dominating measure \( \mu \), \( p \) and \( q \) respectively, then the Kullback-Leibler Information \( I(P;Q) \) is the \( P \) expected log likelihood ratio
\[
I(P;Q) = \mathbb{E}_P \log \frac{p(X)}{q(X)} = \int \log \left( \frac{p(x)}{q(x)} \right) p(x) d\mu(x)
\]

**Fact 26** \( I(P;Q) \) is not in general the same as \( I(Q;P) \)

**Fact 27** \( I(P;Q) \geq 0 \) and there is equality only when \( P = Q \)

This fact follows from the next (slightly more general) lemma, upon noting that \( C(\cdot) = -\log(\cdot) \) is strictly convex.

**Lemma 28** (Silvey page 75) Suppose that \( P \) and \( Q \) are two different distributions on a space \( X \) with R-N derivatives \( p \) and \( q \) respectively with respect to a common measure \( \mu \). Suppose further that a function \( C : (0,\infty) \to R^1 \) is convex. Then
\[ E_P C \left( \frac{q(X)}{p(X)} \right) \geq C(1) \]

with strict inequality if \( C \) is strictly convex.

**Theorem 29**  If \( X = (X_1, X_2) \) and under both \( P \) and \( Q \) the variables \( X_1 \) and \( X_2 \) are independent, then
\[
I_X(P, Q) = I_{X_1}(P^{X_1}; Q^{X_1}) + I_{X_2}(P^{X_2}; Q^{X_2})
\]

So the K-L Information has a kind of additive property similar to that possessed by the Fisher Information. It also has the kind of non-increasing property associated with reduction to a statistic \( T(X) \) we saw with the Fisher Information.

**Theorem 30**  For any statistic \( T(X) \)
\[
I_X(P, Q) \geq I_{T(X)}(P^T; Q^T)
\]
and there is equality iff \( T(X) \) is sufficient for \( \{P, Q\} \).

Finally, there is also an important connection between the notions of Fisher and K-L Information.

**Theorem 31**  Under the hypotheses of Theorem 19 and supposing that one can reverse the order of limits in
\[
\frac{d^2}{d\theta^2} E_{\theta_0} \log f_\theta(X) \bigg|_{\theta = \theta_0}
\]
to produce
\[
E_{\theta_0} \frac{d^2}{d\theta^2} \log f_\theta(X) \bigg|_{\theta = \theta_0}
\]
it follows that
\[
\frac{d^2}{d\theta^2} I_X(P_{\theta_0}, P_{\theta}) \bigg|_{\theta = \theta_0} = I_X(\theta_0)
\]
(the first "\( I \)" is the K-L Information and the second is the Fisher Information).

The multiparameter version of this is, of course, similar.

**Theorem 31b**  Under the hypotheses of Theorem 19b (and assuming that it is permissible to interchange the order of second order partial differentiation and integration \( d\mu(x) \))
(the first "I" is the K-L Information and the second is the Fisher Information).

**Types of Statistical Problems**

In the framework introduced above (an identifiable family of distributions $\mathcal{P} = \{P_\theta\}$ dominated by a $\sigma$-finite measure $\mu$ with R-N derivatives $f_\theta = \frac{dP_\theta}{d\mu}$) there are several formulations of the basic notion of "using $X$ to tell us about $\theta$" that is statistical theory. Here we briefly describe those.

**Point Estimation** is the business of using a statistic $\delta(X)$ (where $\delta : \mathcal{X} \rightarrow \Theta$) to guess at the value of $\theta$ (or some function of $\theta$). In those (common) cases where $\Theta \subset \mathcal{R}^k$ one can adopt criteria like "'small' $\theta$ expected difference between $\delta(X)$ and $\theta$ (or a function of $\theta$)" to compare choices of "estimators" $\delta(X)$. Common methods of inventing estimators are taught in 543 and include such methods as maximum likelihood and the method of moments. Bayesians typically use some feature of their posterior distribution (like its mean or mode) to estimate $\theta$ (or a function thereof).

**Set Estimation** is the business of using a set-valued statistic $S(X)$ (where for $\mathcal{C}$ a $\sigma$-algebra on $\Theta$, $S : \mathcal{X} \rightarrow \mathcal{C}$) to guess at the true value of $\theta$ (or some function of $\theta$). One (obvious) criterion for judging choices of "confidence procedures" $S$ is that one wants "'large' $\theta$ coverage probability" for all $\theta$. In those (common) cases where $\Theta \subset \mathcal{R}^k$ one can adopt criteria like "'small' $\theta$ expected set 'size'" to add to the notion of large coverage probability. Common methods of inventing set estimators include looking for "pivots," inverting families of tests, looking for "standard errors" for good point estimators and using them to provide limits around those point estimators, and bootstrap methods. Bayesian, having a posterior probability distribution for $\theta$ typically use that to choose what they call "credible sets" with large posterior probability of including $\theta$ (or a function thereof).

**Prediction** is the business of positing a model for $X = (X_1, X_2)$, supposing only $X_1$ to be observable and using it to predict $X_2$ (or some function thereof). There are both point and set versions of this problem. In the common case where $X_2$ takes values in $\mathcal{R}^k$, one can adopt criteria like "'small' $\theta$ expected difference between the prediction $\delta(X_1)$ and $X_2$ (or a function thereof)" to compare choices of "predictors" $\delta(X_1)$. Bayesians have a conditional distribution for $X_2$ (called the predictive distribution) and can use some feature of it to predict $X_2$. "Set prediction" methods are usually judged on the basis of their $\theta$ coverage probabilities and $\theta$ expected set sizes. Bayesian predictions are really formally no different from Bayesian estimation, being based on conditional distributions of the quantity of interest given the observable $X_1$.  

\[
\left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} I_X(P_{\theta_0}, P_\theta) \right)_{\theta \rightarrow \theta_0} = I_X(\theta_0)
\]
Hypothesis Testing is the business of partitioning the parameter space $\Theta$ into two pieces, $\Theta_0$ and $\Theta_1$, and using a statistic $\phi(X)$ (where $\phi : \mathcal{X} \rightarrow \{0, 1\}$) to decide which of the two pieces is more plausible. Criteria like small "size" (maximum $\theta \in \Theta_0$ probability that $\phi(X) = 1$) and large "power" ($\theta \in \Theta_1$ probability that $\phi(X) = 1$) are usually used to compare "tests" $\phi$. Classical methods of inventing tests (taught in 543) include appeal to the Neyman-Pearson Lemma and its extensions to convenient families (like monotone likelihood ratio families) and use of the "likelihood ratio criterion." Bayesians, armed as they are with a posterior distribution over $\Theta$, have posterior probabilities for $\Theta_0$ and $\Theta_1$ to guide their choice between $\Theta_0$ and $\Theta_1$ (and therefore their construction of $\phi$'s).

There are optimality theories that can be brought to bear on these statistical problems. Those are part of the somewhat broader world of "statistical decision theory." We will do a fair amount with this subject before the term is over. However, because the situations in which the theory provides clean results do not begin to span the space of important statistical applications, we will first continue with some more general theoretical material of wider applicability, based primarily on asymptotic considerations.

**Maximum Likelihood, Its Implications in Likelihood Ratio Testing and Other Contexts (Primarily the Simple Asymptotics Thereof)**

"Maximum Likelihood"

Maximum likelihood is a method of point estimation dating at least to Fisher and probably beyond. As usual, for an identifiable family of distributions $\mathcal{P} = \{P_\theta\}$ dominated by a $\sigma$-finite measure $\mu$, we let $f_\theta = \frac{dP_\theta}{d\mu}$.

**Definition 13** A statistic $T : \mathcal{X} \rightarrow \Theta$ is called a maximum likelihood estimator (MLE) of $\theta$ if

$$f_{T(x)}(x) = \sup_{\theta \in \Theta} f_\theta(x) \quad \forall x \in \mathcal{X}.$$  

This definition is not without practical problems in terms of existence and computation of MLEs. $f_\theta(x)$ can fail to be bounded in $\theta$, and even when it is, can fail to have a maximum value as one searches over $\Theta$. One way to get rid of the problem of potential lack of boundedness (favored and extensively practiced by Prof. Meeker for example) is to admit that one can measure only "to the nearest something," so that "real" $\mathcal{X}$ are discrete (finite or countable) so that $\mu$ can be counting measure and

$$f_\theta(x) = P_\theta(\{x\}) < 1$$

(so that one is just maximizing the probability of observing the data in hand). Of course, the possibility that the sup is not achieved still remains even when this device is used (like, for example, when $X \sim \text{Binomial}(n, p)$ for $p \in (0, 1)$ when $X = 0$ or $n$).
Note that in the common situation where $\Theta \subset \mathcal{R}^k$ and the $f_\theta$ are nicely differentiable in $\theta$, a maximum likelihood estimator must satisfy the likelihood equations

$$\frac{\partial}{\partial \theta_i} \log f_\theta(x) \bigg|_{\theta = \hat{\theta}(x)} = 0 \quad \text{for } i = 1, 2, \ldots, k.$$ 

Much of the theory connected with maximum likelihood estimation has to do with roots of the likelihood equations (as opposed to honest maximizers of $f_\theta(x)$). Further, much of the attractiveness of the method derives from its asymptotic properties. A simple version of the theory (primarily for the $k = 1$ case) follows.

Adopt the following notation for the discussion of the asymptotics of maximum likelihood.

- $X^n$ represents all information on $\theta$ available through the $n$th stage of observation.
- $\mathcal{X}_n$ the observation space for $X^n$ (through the $n$th stage).
- $B_n$ a $\sigma$-algebra on $\mathcal{X}_n$.
- $P_{\theta}^{X^n}$ the distribution of $X^n$ (dominated by a $\sigma$-finite measure $\mu_n$).
- $\Theta$ a (fixed) parameter space.
- $f_{\theta}^{n}$ the R-N derivative $\frac{dP_{\theta}^{X^n}}{d\mu_n}$.

For this material one really does need the basic probability space $(\Omega, \mathcal{A}, P_\theta)$ in the background, as it is necessary to have a fixed space upon which to prove convergence results. Often (i.e. in the iid case) we may take

$$\mathcal{X} = \mathcal{X}^\infty$$

$X^n = (X_1, X_2, \ldots, X_n)$ for $X_i(\omega) =$ the $i$th coordinate of $\omega$.

- $\mathcal{A} = \mathcal{B}^\infty$.
- $P_\theta = (P_{\theta}^{X_i})^\infty$.
- $P_{\theta}^{X^n} = (P_{\theta}^{X_i})^n$.
- $\mu_n = \mu^n$.

and

$$f_{\hat{\theta}}^{n}(x^n) = \prod_{i=1}^{n} f_{\hat{\theta}}(x_i).$$

For such a set-up, the usual drill is for appropriate sequences of estimators $\{\hat{\theta}_n(x^n)\}$ to try to prove consistency and asymptotic normality results.

**Definition 14** Suppose that $\{T_n\}$ is a sequence of statistics $T_n : \mathcal{X}_n \rightarrow \Theta$, where $\Theta \subset \mathcal{R}^k$.

i) $\{T_n\}$ is (weakly) consistent at $\theta_0 \in \Theta$ if for every $\epsilon > 0$,

$$P_{\theta_0}(\|T_n - \theta_0\| > \epsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and

ii) $\{T_n\}$ is strongly consistent at $\theta_0 \in \Theta$ if $T_n \rightarrow \theta_0$ a.s. $P_{\theta_0}$. 

16
There are two standard paths for developing asymptotic results for likelihood-based estimators in iid contexts. These are

i) the path blazed by Wald based on very strong regularity assumptions (that can be extremely hard to verify for a specific model) but that deal with "real" maximum likelihood estimators, and

ii) the path blazed by Cramér based on much weaker assumptions and far easier proofs that deal with solutions to the likelihood equations.

We'll wave our hands very briefly at a heuristic version of the Wald argument and defer to pages 415-424 of Schervish for technical details. Then we'll do a version of the Cramér results (probably assuming more than is absolutely necessary to get them).

All that follows regarding the asymptotics of maximum likelihood refers to the iid case, where $x^n = (x_1, x_2, \ldots, x_n)$, $\Theta \subset \mathcal{R}^k$ and $f^n_\theta(x^n) = \prod_{i=1}^n f_\theta(x_i)$.

Now suppose that the $f_\theta$ are positive everywhere on $\mathcal{X}$ and define the function of $\theta$

$$z_\theta(\theta) = \mathbb{E}_{\theta_0} \log f_\theta(X) = \int \log f_\theta(x) f_{\theta_0}(x) \, d\mu(x) = - I(\theta_0; \theta) + \int \log f_{\theta_0}(x) f_{\theta_0}(x) \, d\mu(x) .$$

**Corollary 32** If the $f_\theta$ are positive everywhere on $\mathcal{X}$, then for $\theta \neq \theta_0$

$$z_{\theta_0}(\theta_0) > z_{\theta_0}(\theta) .$$

Then, as a matter of notation, let

$$l_n(x^n, \theta) = \log f^n_\theta(x^n) = \sum_{i=1}^n \log f_\theta(x_i) .$$

Then clearly $\mathbb{E}_{\theta_0} \frac{1}{n} l_n(X^n, \theta) = z_{\theta_0}(\theta)$ and in fact, the law of large numbers says that for every $\theta$,

$$\frac{1}{n} l_n(X^n, \theta) = \frac{1}{n} \sum_{i=1}^n \log f_\theta(X_i) \to z_{\theta_0}(\theta) \text{ a.s. } P_{\theta_0} .$$

So assuming enough regularity conditions to make this almost sure convergence *uniform in $\theta$*, when $\theta_0$ governs the distribution of the observations there is perhaps hope that a single maximizer of $\frac{1}{n} l_n(X^n, \theta)$ will eventually exist and be close to the maximizer of $z_{\theta_0}(\theta)$, namely $\theta_0$. To make this precise is not so easy. See Theorems 7.49 and 7.54 of Schervish.

Now turn to Cramér type results.
Theorem 33 Suppose that $X_1, X_2, ..., X_n$ are iid $P_\theta$, $\Theta \subset \mathcal{R}^1$ and there exists an open neighborhood of $\theta_0$, say $O$, such that

\begin{itemize}
  \item[i)] $f_\theta(x) > 0 \ \forall x$ and $\forall \theta \in O$,
  \item[ii)] $\forall x$, $f_\theta(x)$ is differentiable at every point $\theta \in O$, and
  \item[iii)] $E_{\theta_0}\log f_\theta(X)$ exists for all $\theta \in O$ and $E_{\theta_0}\log f_\theta(X)$ is finite.
\end{itemize}

Then, if $\epsilon > 0$ and $\delta > 0$, $\exists n$ such that $\forall m > n$, the $P_{\theta_0}$ probability that the equation

$$\frac{d}{d\theta} \sum_{i=1}^{m} \log f_\theta(X_i) = 0$$

has a root within $\epsilon$ of $\theta_0$ is at least $1 - \delta$.

(Essentially this theorem ducks the uniformity issue raised earlier by considering $l_n$ at only $\theta_0$, $\theta_0 - \epsilon$ and $\theta_0 + \epsilon$.)

In general, Theorem 33 doesn't immediately do what we'd like. The root guaranteed by the Theorem isn't necessarily an MLE. And in fact, when the likelihood equation can have multiple roots, we haven't even yet defined a sequence of random variables, let alone a sequence of estimators converging to $\theta_0$. But the theorem can form the basis for simple results that are of statistical importance.

Corollary 34 Under the hypotheses of Theorem 33, suppose that (in addition) $\Theta$ is open and $\forall x$, $f_\theta(x)$ is positive and differentiable at every point $\theta \in \Theta$ so that the likelihood equation

$$\frac{d}{d\theta} \sum_{i=1}^{n} \log f_\theta(X_i) = 0$$

makes sense at all points $\theta \in \Theta$. Then define $\rho_n$ to be the root of the likelihood equation when there is exactly one (and otherwise adopt any arbitrary definition for $\rho_n$). If with $P_{\theta_0}$ probability approaching 1 the likelihood equation has a single root

$$\rho_n \rightarrow \theta_0 \ \text{in} \ P_{\theta_0} \text{probability} .$$

Where there is no guarantee that the likelihood equation will eventually have only a single root, one could use Theorem 33 to argue that defining

$$r_n(\theta_0) = \begin{cases} 
0 & \text{if the likelihood equation has no root} \\
\text{the root closest to } \theta_0 & \text{otherwise} 
\end{cases},$$

$r_n(\theta_0) \rightarrow \theta_0$ in $P_{\theta_0}$ probability. (The continuity of the likelihood implies that limits of roots are roots, so for $\theta_0$ in an open $\Theta$, the definition of $r_n(\theta_0)$ above eventually makes sense.) But, of course, this is of no statistical help. What is similar in nature but of more help is the next corollary.
**Corollary 35** Under the hypotheses of Theorem 33, if \( \{T_n\} \) is a sequence of estimators consistent at \( \theta_0 \) and

\[
\hat{\theta}_n = \begin{cases} 
T_n & \text{if the likelihood equation has no roots} \\
\text{the root closest to } T_n & \text{otherwise}
\end{cases},
\]

then

\[
\hat{\theta}_n \xrightarrow{P} \theta_0 \text{ in } P_{\theta_0} \text{ probability}.
\]

In practical terms, finding the root of the likelihood equation nearest to \( T_n \) may be a numerical problem. (Actually, if it is, it's not so clear how you know you have solved the problem. But we'll gloss over that.) Finding a root beginning from \( T_n \) might be approached using something like Newton's method. That is, abbreviating

\[
L_n(\theta) \equiv l_n(X^n, \theta) = \sum_{i=1}^{n} \log f_{\theta}(X_i),
\]

the likelihood equation is

\[
L_n'(\theta) = 0.
\]

One might try looking for a root of this equation by iterating using linearizations of \( L_n'(\theta) \) (presumably beginning at \( T_n \)). Or, assuming that \( L_n''(T_n) \neq 0 \) taking a single step in this iteration process, a 1-step likelihood-based modification/improvement on \( T_n \),

\[
\tilde{\theta}_n = T_n - \frac{L_n'(T_n)}{L_n''(T_n)},
\]

is suggested. It is plausible that an estimator like \( \tilde{\theta}_n \) might have asymptotic behavior similar to that of an MLE. (See page 442 of TPE or Theorem 7.75 of Schervish, that treats a more general multiparameter M-estimation version of this idea.)

The next matter up is the asymptotic distribution of likelihood-based estimators. The following is representative of what people have proved. (I suspect that the hypotheses here are stronger than are really needed to get the conclusion, but they make for a simple proof.)

**Theorem 36** Suppose that \( X_1, X_2, \ldots, X_n \) are iid \( P_{\theta}, \Theta \subset \mathbb{R}^1 \) and there exists an open neighborhood of \( \theta_0 \), say \( O \), such that

i) \( f_{\theta}(x) > 0 \ \forall x \text{ and } \forall \theta \in O \),

ii) \( \forall x, f_{\theta}(x) \) is three times differentiable at every point \( \theta \in O \),

iii) \( \exists M(x) \geq 0 \text{ with } \mathbb{E}_{\theta_0} M(X) < \infty \) and

\[
\left| \frac{d^3}{d\theta^3} \log f_{\theta}(x) \right| \leq M(x) \ \forall x \text{ and } \forall \theta \in O,
\]
iv) \( 1 = \int f_\theta(x) d\mu(x) \) can be differentiated twice wrt \( \theta \) under the integral at \( \theta_0 \), i.e. both \( 0 = \int \frac{d}{d\theta} f_\theta(x) \bigg|_{\theta=\theta_0} d\mu(x) \) and \( 0 = \int \frac{d^2}{d\theta^2} f_\theta(x) \bigg|_{\theta=\theta_0} d\mu(x) \), and

v) \( I_1(\theta) = \mathbb{E}_\theta \left( \frac{d}{d\theta} \log f_\theta(X) \right)^2 \in (0, \infty) \forall \theta \in \mathcal{O} \).

If with \( P_{\theta_0} \) probability approaching 1, \( \{\hat{\theta}_n\} \) is a root of the likelihood equation, and \( \hat{\theta}_n \rightarrow \theta_0 \) in \( P_{\theta_0} \) probability, then under \( P_{\theta_0} \)

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\rightarrow} N(0, \frac{1}{I_1(\theta_0)}) .
\]

There are multiparameter versions of this theorem (indeed the Schervish theorem is a multiparameter one). And there are similar (multiparameter) versions of this for the "Newton 1-step approximate MLE's" based on consistent estimators of \( \theta \) (i.e. \( \hat{\theta}_n \)'s).

The practical implication of a result like this theorem is that one can make approximate confidence intervals for \( \theta \) of the form

\[
\hat{\theta}_n \pm z \frac{1}{\sqrt{nI_1(\theta_0)}} .
\]

Well, it is almost a practical implication of the theorem. One needs to do something about the fact that \( I_1(\theta_0) \) is unknown when making a confidence interval so that it can't be used in the formula. There are two common routes to fixing this problem. The first is to use \( I_1(\hat{\theta}_n) \) in its place. This method is often referred to as using the "expected Fisher information" in the making of the interval. (Note that \( I_1(\theta_0) \) is a measure of curvature of \( z_{\theta_0}(\theta) \) at its maximum (located at \( \theta_0 \)) while \( I_1(\hat{\theta}_n) \) is a measure of curvature of \( z_{\hat{\theta}_n}(\theta) \) at its maximum (located at \( \hat{\theta}_n \)).) Dealing with the theory of the first method is really very simple. It is easy to prove the following.

**Corollary 37** Under the hypotheses of Theorem 36, if \( I_1(\theta) \) is continuous at \( \theta_0 \), then under \( P_{\theta_0} \)

\[
\sqrt{nI_1(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) \overset{d}{\rightarrow} N(0, 1) .
\]

(Actually, the continuity of \( I_1(\theta) \) at \( \theta_0 \) may well follow from the hypotheses of Theorem 36 and therefore be redundant in the hypotheses of this corollary. I've not tried to think this issue through.)

The second method of approximating \( I_1(\theta_0) \) comes about by realizing that in the proof of Theorem 36, \( I_1(\theta_0) \) arises as the limit of \( -\frac{1}{n} L''_n(\theta_0) \). One might then hope to approximate it by
\[- \frac{1}{n} L_n''(\hat{\theta}_n) = - \frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f_\theta(X_i) \bigg|_{\theta=\hat{\theta}_n}, \]

which is sometimes called the "**observed Fisher information**." (Note that \(- \frac{1}{n} L_n''(\hat{\theta}_n)\) is a measure of curvature of the average log likelihood function \(\frac{1}{n} l_n(X^n, \theta)\) at \(\hat{\theta}_n\), which one is secretly thinking of as essentially a maximizer of the average log likelihood.) Schervish says that Efron and others claim that this second method is generally better than the first (in terms of producing intervals with real coverage properties close to the nominal properties). Proving an analogue of Corollary 37 where \(- \frac{1}{n} L_n''(\hat{\theta}_n)\) replaces \(I_1(\hat{\theta}_n)\) is somewhat harder, but not by any means impossible. That is, one can prove the following.

**Corollary 38** Under the hypotheses of Theorem 36 and \(P_{\theta_0}\),

\[ \sqrt{- L_n''(\hat{\theta}_n)} (\hat{\theta}_n - \theta_0) \overset{d}{\to} N(0, 1). \]

Under the hypotheses of Theorem 36, convergence in distribution of \(\hat{\theta}_n\) implies the convergence of the second moment of \(\hat{\theta}_n\), so that

\[ \text{Var}_{\theta_0} \sqrt{n} (\hat{\theta}_n - \theta_0) \to \frac{1}{I_1(\theta_0)}, \]

i.e.

\[ \text{Var}_{\theta_0} (\sqrt{n} (\hat{\theta}_n)) \to \frac{1}{I_1(\theta_0)}. \]

Then the fact that the C-R inequality (recall Stat 543) declares that under suitable conditions no unbiased estimator of \(\theta\) has variance less than \(1/(n I_1(\theta))\), makes one suspect that for a consistent sequence of estimators \(\{\theta_0^n\}\), convergence under \(\theta_0\)

\[ \sqrt{n} (\theta_0^n - \theta_0) \overset{d}{\to} N(0, V(\theta_0)) \]

might imply that \(V(\theta_0) \geq \frac{1}{I_1(\theta_0)}\), but in fact this inequality does **not** have to hold up.

**Likelihood Ratio Tests, Related Tests and Confidence Procedures**

One set of general purpose tools for making tests and confidence procedures consists of the "likelihood ratio test," its relatives and related confidence procedures. The following is an introduction to these methods and a little about their asymptotics.
For $\mathcal{P} = \{P_\theta\}$, $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ one might plausibly consider a test criterion like

$$LR(x) = \sup_{\theta \in \Theta_1} \frac{f_\theta(x)}{\sup_{\theta \in \Theta_0} f_\theta(x)}.$$ 

(Notice, by the way, that Bayes tests look at ratios of $G$ averages rather than sups.)

Or equivalently, one might consider

$$\lambda(x) = \max(LR(x), 1) = \sup_{\theta \in \Theta} \frac{f_\theta(x)}{\sup_{\theta \in \Theta_0} f_\theta(x)},$$

where we will want to reject $H_0$ for large $\lambda(x)$. This is an intuitively reasonable criterion, and many common tests (including optimal ones) turn out to be likelihood ratio tests. On the other hand, it is easy enough to invent problems where LRTs need not be in any sense optimal.

It is a useful fact that one can find some general results about how to set critical values for LRTs (at least for iid models and large $n$). This is because if there is an MLE, say $\hat{\theta}$,

$$\sup_{\theta \in \Theta} f_\theta(x) = f_{\hat{\theta}}(x)$$

and we can make use of MLE-like estimation results. For example, in the iid context of Theorem 36, there is the following.

**Theorem 39** Suppose that $X_1, X_2, \ldots, X_n$ are iid $P_\theta$, $\Theta \subset \mathcal{R}^1$ and there exists an open neighborhood of $\theta_0$, say $O$, such that

i) $f_\theta(x) > 0 \forall x$ and $\forall \theta \in O$,

ii) $\forall x$, $f_\theta(x)$ is three times differentiable at every point $\theta \in O$,

iii) $\exists M(x) \geq 0$ with $E_{\theta_0}M(X) < \infty$ and

$$\left| \frac{d^3}{d\theta^3} \log f_\theta(x) \right| \leq M(x) \ \forall x \text{ and } \forall \theta \in O,$$

iv) $1 = \int f_\theta(x) d\mu(x)$ can be differentiated twice wrt $\theta$ under the integral at $\theta_0$,

i.e. both $0 = \int \frac{d}{d\theta} f_\theta(x) \bigg|_{\theta = \theta_0} d\mu(x)$ and $0 = \int \frac{d^2}{d\theta^2} f_\theta(x) \bigg|_{\theta = \theta_0} d\mu(x)$, and

v) $I_1(\theta) \equiv E_{\theta} \left( \frac{d}{d\theta} \log f_\theta(X) \right)^2 \in (0, \infty) \ \forall \theta \in O$.

If with $P_{\theta_0}$ probability approaching 1, $\{\theta_n\}$ is a root of the likelihood equation, and $\theta_n \to \theta_0$ in $P_{\theta_0}$ probability, then under $P_{\theta_0}$
\[2 \log \lambda_n^* = 2 \log \left( \frac{f_{\hat{\theta}_n}^n(X^n)}{f_{\theta_0}^n(X^n)} \right) \xrightarrow{\mathcal{L}} \chi_1^2.\]

(The hypotheses here are exactly the hypotheses of Theorem 36.)

The application that people often make of this theorem is that they reject \(H_0 : \theta = \theta_0\) in favor of \(H_1 : \theta \neq \theta_0\) when \(2 \log \lambda_n^*\) exceeds the \((1 - \alpha)\) quantile of the \(\chi_1^2\) distribution. Notice also that upon inverting these tests, the theorem gives a way of making large sample confidence sets for \(\theta\). When \(\hat{\theta}_n\) is a maximizer of the log likelihood, the set of \(\theta_0\) for which \(H_0 : \theta = \theta_0\) will be accepted using the LRT with an (approximate) \(\chi_1^2\) cut-off value (say \(\chi_1^2\)) is

\[\{\theta | L_n(\theta) \geq L_n(\hat{\theta}_n) - \frac{1}{2} \chi_1^2\} .\]

There are versions of the \(\chi^2\) limit that cover (composite) null hypotheses about some part of a multi-dimensional parameter vector. For example, Schervish (page 459) proves a result like the following.

**Result 40** Suppose that \(\Theta \subset \mathcal{R}^p\) and appropriate regularity conditions hold for iid observations \(X_1, X_2, ..., X_n\). If \(k \leq p\) and

\[\lambda_n = \sup_{\theta} \frac{\sup_{\theta \text{ s.t. } \theta_i = \theta_0^i, i = 1, 2, ..., k} f_{\theta}^n(X^n)}{\sup_{\theta} f_{\theta}^n(X^n)},\]

then under any \(P_{\theta}\) with \(\theta_i = \theta_0^i, i = 1, 2, ..., k\)

\[2 \log \lambda_n \xrightarrow{\mathcal{L}} \chi_k^2 .\]

This result (of course) allows one to set approximate critical points for testing \(H_0 : \theta_i = \theta_0^i, i = 1, 2, ..., k\) vs \(H_1 : \text{not } H_0\), by rejecting \(H_0\) when

\[2 \log \lambda_n > \text{(1 - \(\alpha\)) } \chi_k^2 \text{ quantile } .\]

And, one can make large \(n\) approximate confidence sets for \((\theta_1, \theta_2, ..., \theta_k)\) of the form

\[\{(\theta_1, ..., \theta_k) | \sup_{\theta_{k+1}, ..., \theta_p} L_n((\theta_1, ..., \theta_k, \theta_{k+1}, ..., \theta_p)) \geq \sup_{\theta} L_n(\theta) - \frac{1}{2} \chi_k^2\} \]

for \(\chi_k^2\) the \((1 - \alpha)\) quantile of the \(\chi_k^2\) distribution.

There is some standard jargon associated with the function of \((\theta_1, ..., \theta_k)\) appearing in the above expressions.
Definition 15 The function of \((\theta_1, \ldots, \theta_k)\)
\[
L^*_n((\theta_1, \ldots, \theta_k)) \doteq \sup_{\theta_{k+1}, \ldots, \theta_p} L_n((\theta_1, \ldots, \theta_k, \theta_{k+1}, \ldots, \theta_p))
\]
is called a profile likelihood function for the parameters \(\theta_1, \ldots, \theta_k\).

With this jargon/notation the form of the large \(n\) confidence set becomes
\[
\{(\theta_1, \ldots, \theta_k) \mid L^*_n((\theta_1, \ldots, \theta_k)) \geq \sup_{(\theta_1, \ldots, \theta_k)} L^*_n((\theta_1, \ldots, \theta_k)) - \frac{1}{2} \chi^2_k \},
\]
which looks more like the form from the case of a one-dimensional parameter.

There are alternatives to the LRT of \(H_0 : \theta_i = \theta^0_i, i = 1, 2, \ldots, k\) vs \(H_1 : \text{not } H_0\). One such test is the so called "Wald Test." See, for example, pages 115-120 of Silvey. The following is borrowed pretty much directly from Silvey.

Consider \(\Theta \subset \mathcal{R}^p\) and \(k\) restrictions on \(\theta\)
\[
h_1(\theta) = h_2(\theta) = \cdots = h_k(\theta) = 0 ,
\]
(like, for instance, \(\theta_1 - \theta^0_1 = \theta_2 - \theta^0_2 = \cdots = \theta_k - \theta^0_k = 0\). Suppose that \(\hat{\theta}_n\) is an MLE of \(\theta\) and define
\[
h(\theta) \doteq (h_1(\theta), h_2(\theta), \ldots, h_k(\theta)) .
\]
Then if \(H_0 : h_1(\theta) = h_2(\theta) = \cdots = h_k(\theta) = 0\) is true, \(h(\hat{\theta}_n)\) ought to be near 0. The question is how to measure "nearness" and to set a critical value in order to have a test with size approximately \(\alpha\). An approach to doing this is as follows.

We expect (under suitable conditions) that under \(P_\theta\) the \((p\text{-dimensional})\) estimator \(\hat{\theta}_n\) has
\[
\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N_p(0, I_1^{-1}(\theta)) .
\]
Then if
\[
H_{k \times p}(\theta) \doteq \left( \frac{\partial h_i}{\partial \theta_j \mid \theta} \right) ,
\]
the \(\Delta\) method suggests that
\[
\sqrt{n}(h(\hat{\theta}_n) - h(\theta)) \xrightarrow{L} N_k(0, H(\theta)I_1^{-1}(\theta)H'(\theta)) .
\]
Abbreviate \(H(\theta)I_1^{-1}(\theta)H'(\theta)\) as \(B(\theta)\). Since \(h(\theta) = 0\) if \(H_0\) is true, the above suggests that
\[ nh'(\hat{\theta}_n)B^{-1}(\theta)h(\hat{\theta}_n) \xrightarrow{d} \chi_k^2 \] under \( H_0 \).

Now \( nh'(\hat{\theta}_n)B^{-1}(\theta)h(\hat{\theta}_n) \) cannot serve as a test statistic, since it involves \( \theta \), which is not completely specified by \( H_0 \). But it is plausible to consider the statistic

\[ W_n = nh'(\hat{\theta}_n)B^{-1}(\hat{\theta}_n)h(\hat{\theta}_n) \]

and hope that under suitable conditions

\[ W_n \xrightarrow{d} \chi_k^2 \] under \( H_0 \).

If this can be shown to hold up, one may reject for \( W_n > (1 - \alpha) \) quantile of the \( \chi_k^2 \) distribution.

The Wald tests for \( h_i(\theta) = \theta_i - \theta_i^0 \) \( i = 1, 2, \ldots, k \) can be inverted to get approximate confidence regions for \( k \leq p \) of the parameters \( \theta_1, \theta_2, \ldots, \theta_p \). The resulting confidence regions will, by construction, be ellipsoidal.

There is a final alternative to the LRT and Wald tests that deserves mention. That is the "test" or "score test." See again Silvey. The notion here is that on occasion it can be easier to maximize \( L_n(\theta) \) subject to \( h(\theta) = 0 \) (again taking this to be a statement of \( k \) constraints on a \( p \)-dimensional parameter vector \( \theta \)) than to simply maximize \( L_n(\theta) \). Let \( \hat{\theta}_n \) be a "restricted" MLE (i.e. a maximizer of \( L_n(\theta) \) subject to \( h(\theta) = 0 \)). One might expect that if \( H_0 : h_i(\theta) = 0 \) is true, then \( \hat{\theta}_n \) ought to be nearly an unrestricted maximizer of \( L_n(\theta) \) and the partial derivatives of \( L_n(\theta) \) should be nearly 0 at \( \hat{\theta}_n \). On the other hand, if \( H_0 \) is not true, there is little reason to expect the partials to be nearly 0. So (again/still with \( L_n(\theta) = \sum_{i=1}^{n} \log f_{\theta}(X_i) \)) one might consider the statistic

\[ X^2 \doteq \frac{1}{n} \left( \frac{\partial}{\partial \theta_i} L_n(\theta) \right)_{\theta=\hat{\theta}_n} \left( I_1^{-1}(\hat{\theta}_n) \right) \left( \frac{\partial}{\partial \theta_i} L_n(\theta) \right)_{\theta=\hat{\theta}_n} \].

Then, for

\[ \lambda_n = \sup_{\theta} \frac{\sup \ f^n_{\theta}(X^n)}{\sup_{\theta \ s.t. \ h(\theta) = 0 \ f^n_{\theta}(X^n)}} \]

it is typically the case that under \( H_0 \), \( X^2 \) differs from \( 2\log \lambda_n \) by a quantity tending to 0 in \( P_\theta \) probability. That is, this statistic can be calibrated using \( \chi^2 \) critical points and can form the basis of a test of \( H_0 \) asymptotically equivalent to a "Result 40 type" likelihood ratio type test.

Once more (as for the Wald tests) for \( h_i(\theta) = \theta_i - \theta_i^0 \) \( i = 1, 2, \ldots, k \), one can invert these tests to get confidence regions for \( k \leq p \) of the parameters \( \theta_1, \theta_2, \ldots, \theta_p \).
As a final note on this topic, observe that

i) an LRT of the style in Result 40 requires computation of an MLE, \( \hat{\theta}_n \), and a restricted MLE, \( \tilde{\theta}_n \),

ii) a Wald test requires only computation of the MLE, \( \hat{\theta}_n \), and

iii) a "\( \chi^2 \) test" requires only computation of the restricted MLE, \( \tilde{\theta}_n \).

M-Estimation

From one point of view, this method of estimation is simply a generalization of maximum likelihood. ("Robust" estimation proponents have other motivations for the method.) Here is a very brief introduction to the subject.

Consider a statistical model with parameter space \( \Theta \subset \mathbb{R}^k \) and observations \( X_1, X_2, \ldots, X_n \) iid \( P_\theta \). Likelihood-based estimation uses the loglikelihood

\[
L_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i)
\]

and typically chooses \( \theta \) to maximize \( L_n(\theta) \) or to solve the \( k \)-dimensional system of likelihood equations

\[
\nabla L_n(\theta) = 0
\]

M-estimation uses some suitable function \( \rho(x, \theta) \) and

\[
r_n(\theta) = \sum_{i=1}^n \rho(X_i, \theta)
\]

and chooses \( \theta \) to maximize \( r_n(\theta) \) or to solve the \( k \)-dimensional system of likelihood equations

\[
\nabla r_n(\theta) = 0
\]

(With \( \psi_j = \frac{\partial L}{\partial \theta_j} \), this is \( \sum_{i=1}^n \psi_j(X_i, \theta) = 0 \) \( \forall j = 1, 2, \ldots, k. \)

\( \rho \) could be \( (x - \theta)^2 \), \( \log f_\theta \) or even \( \log g_\theta \) for \( g_\theta \) the density for some other family of models.

Provided that \( \rho \) (or the \( \psi_j \)) interacts properly with the \( f_\theta \), one gets properties for M-estimation like those we've deduced for maximum likelihood. For example,

\[
E_{\theta_0, \rho}(X, \theta) > E_{\theta_0, \rho}(X, \theta) \forall \theta_0, \forall \theta
\]
plus other conditions implies that an M-estimator is consistent. (Note that with \( \rho \) the function \( \log f_\theta \), this inequality comes from Fact 27 about K-L information.)

Further, under appropriate conditions, one can prove asymptotic normality of i) one-step Newton improvements on a consistent estimator of \( \theta \), or ii) consistent roots of \( \nabla r_n(\theta) = 0 \). In such results, the counterpart of \( I_1^{-1}(\theta_0) \) in the limiting \( k \)-variate normal distribution is

\[
J^{-1}(\theta_0) K(\theta_0) J^{-1}(\theta_0)'
\]

for

\[
J(\theta_0) = -E_{\theta_0} \left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} \rho(X, \theta) \bigg|_{\theta = \theta_0} \right)
\]

and

\[
K(\theta_0) = \left( \text{Cov}_{\theta_0} \left( \frac{\partial}{\partial \theta_i} \rho(X, \theta) \bigg|_{\theta = \theta_0}, \frac{\partial}{\partial \theta_j} \rho(X, \theta) \bigg|_{\theta = \theta_0} \right) \right)
\]

(In the case of \( \rho(x, \theta) = \log f_\theta(x) \), \( J = K = I_1 \).)

**Simple Implications of the MLE Results for the Shape of the Loglikelihood and Posteriors**

In rough terms, the statistical folklore says that "likelihoods pile up around true values and are locally quadratic near their peaks, while provided a prior is sufficiently diffuse, posteriors will pile up near true values and look Gaussian." Doing a really good job of even saying this precisely is nontrivial (involving uniformities of convergence of random functions, etc.). What follows are some very simpleminded (and non-uniform type) consequences of the material on maximum likelihood. For more precise and harder theorems, see Schervish §7.4.

**Lemma 41** For \( X_1, X_2, \ldots, X_n \) iid \( P_\theta \)

\[
L_n(\theta_0) - L_n(\theta) \xrightarrow{P} \infty \text{ for } \theta \neq \theta_0
\]

**Lemma 42** For \( X_1, X_2, \ldots, X_n \) iid \( P_\theta \) and \( G \) a prior on \( \Theta \) with density \( g(\theta) \), if

\[
g(\theta|x^n) = \frac{f_\theta^n(x^n)g(\theta)}{\int_{\Theta} f_\theta^n(x^n)g(\theta)d\nu(\theta)}
\]

is the posterior density of \( \theta|X^n \) (with respect to the measure \( \nu \)) and \( g(\theta) > 0 \) then
\[
\frac{g(\theta|X^n)}{g(\theta_0|X^n)} \xrightarrow{P_{\theta_0}} 0 \text{ for } \theta \neq \theta_0
\]

**Corollary 43** Under the hypotheses of Lemma 42, if \( \Theta \) is finite, \( \nu \) is counting measure and \( g(\theta) > 0 \ \forall \theta \), the (random) posterior distribution of \( \theta|X^n \) is consistent in the sense that for \( A \) a subset of \( \Theta \)

\[
\pi^{\theta|X^n}(A) \xrightarrow{P_{\theta_0}} I[\theta_0 \in A]
\]

Next, here's a simple observation about the local shape of the loglikelihood near its maximum.

**Lemma 44** Under the hypotheses of Theorem 36, suppose that \( \hat{\theta}_n \) is an MLE for \( \theta \) that is consistent at \( \theta_0 \). Then

\[
Q_n(\Delta) = L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n + \frac{\Delta}{\sqrt{n}}) \xrightarrow{P_{\theta_0}} \frac{1}{2} \Delta^2 I_1(\theta_0)
\]

(This lemma can be interpreted as saying that if one rescales properly so that there is a limiting shape, the loglikelihood is for large \( n \) essentially quadratic near its maximum. This is not surprising, since it must have 0 derivative at the maximum and there is enough differentiability assumed to allow a 2nd order Taylor expansion.)

**Corollary 45** Under the hypotheses of Lemma 44, suppose that a prior \( G \) has a density \( g(\theta) \) with respect to Lebesgue measure on \( \Theta \subset \mathcal{R}^1 \) such that \( g(\theta_0) > 0 \) and \( g \) is continuous at \( \theta_0 \). The posterior density of \( \theta|X^n \) has the property that

\[
R_n(\Delta) = \log \left( \frac{g(\hat{\theta}_n|X^n)}{g(\hat{\theta}_n + \frac{\Delta}{\sqrt{n}}|X^n)} \right) \xrightarrow{P_{\theta_0}} \frac{1}{2} \Delta^2 I_1(\theta_0)
\]

Now in some very rough sense, this says that the near the MLE, the (random) log posterior density tends to look approximately quadratic in the distance from the MLE. But now note that if \( Y \) is a continuous random variable with pdf \( f \) and (exactly)

\[
\log \frac{f(0)}{f(y)} = \frac{c}{2} y^2
\]

it is the case that \( Y \) must be normal with mean 0 and variance \( \frac{1}{c} \). This then suggests that under \( P_{\theta_0} \) one can expect posterior densities to start to look as if \( \theta \) were normal with mean \( \hat{\theta}_n \) and variance \( \frac{1}{nI_1(\theta_0)} \). A Bayesian who wants a convenient approximate form for a posterior density (perhaps because the exact one is analytically intractable) might then hope to get approximate posterior probabilities from the
distributions. (Schervish proves that the posterior density of \( \sqrt{-L''(\hat{\theta}_n)} \left( \theta - \hat{\theta}_n \right) \), a random function of \( \theta \), converges in some appropriate sense to the standard normal density. It is another, harder, matter to argue that approximate posterior probabilities for \( \theta \) can be obtained using normal distributions.)

**Some Simple Theory of Markov Chain Monte Carlo**

It is sometimes the case that one has a "known" but unwieldy (joint) distribution for a large dimensional random vector \( Y \), and would like to know something about the distribution (like, for example, the marginal distribution of \( Y_1 \)). While in theory it might be straightforward to compute the quantity of interest, in practice the necessary computation can be impossible. With the advent of cheap computing, given appropriate theory the possibility exists of simulating a sample of \( n \) realizations of \( Y \) and computing an empirical version of the quantity of interest to approximate the (intractable) theoretical one.

This possibility is especially important to Bayesians, where \( X|\theta \sim P_\theta \) and \( \theta \sim G \) produce a joint distribution for \( (X, \theta) \) and then (in theory) a posterior distribution \( G(\theta|X) \), the object of primary Bayesian interest. Now (in the usual kind of notation)

\[
g(\theta|X = x) \propto f_\theta(x)g(\theta)
\]

so that in this sense the posterior of the possibly high dimensional \( \theta \) is "known." But it may well be analytically intractable, with (for example) no analytical way to compute \( E\theta_1 \). Simulation is a way of getting (approximate) properties of an intractable/high dimensional posterior distribution.

"Straightforward" simulation from an arbitrary multivariate distribution is, however, typically not (straightforward). Markov Chain Monte Carlo methods have recently become very popular as a solution to this problem. The following is some simple theory covering the case where the (joint) distribution of \( Y \) is discrete.

**Definition 16** A (discrete time/discrete state space) Markov Chain is a sequence of random quantities \( \{X_k\} \), each taking values in a (finite or) countable set \( \mathcal{X} \), with the property that

\[
P[X_n = x_n|X_1 = x_1, ..., X_{n-1} = x_{n-1}] = P[X_n = x_n|X_{n-1} = x_{n-1}]
\]

**Definition 17** A Markov Chain is stationary provided \( P[X_n = x|X_{n-1} = x'] \) is independent of \( n \).
WOLOG we will henceforth name the elements of \( \mathcal{X} \) with the integers 1, 2, 3, ... and call them "states."

**Definition 18** With \( p_{ij} = P[X_n = j|X_{n-1} = i] \), the square matrix \( P = (p_{ij}) \) is called the transition matrix for a stationary Markov Chain and the \( p_{ij} \) are called transition probabilities.

Note that a transition matrix has nonnegative entries and row sums of 1. Such matrices are often called "stochastic" matrices. As a matter of further notation for a stationary Markov Chain, let

\[ p_{ij}^k = P[X_{n+k} = j|X_n = i] \]

and

\[ f_{ij}^k = P[X_{n+k} = j, X_{n+k-1} \neq j, \ldots, X_{n+1} \neq j|X_n = i] \, . \]

(These are respectively the probabilities of moving from \( i \) to \( j \) in \( k \) steps and first moving from \( i \) to \( j \) in \( k \) steps.)

**Definition 19** We say that a MC is irreducible if for each \( i \) and \( j \) \( \exists \ k \) (possibly depending upon \( i \) and \( j \)) such that \( p_{ij}^k > 0 \).

(A chain is irreducible if it is possible to eventually get from any state \( i \) to any other state \( j \).)

**Definition 20** We say that the \( i \)th state of a MC is transient if \( \sum_{k=1}^{\infty} f_{ii}^k < 1 \) and say that the state is persistent if \( \sum_{k=1}^{\infty} f_{ii}^k = 1 \). A chain is called persistent if all of its states are persistent.

**Definition 21** We say that state \( i \) of a MC has period \( t \) if \( p_{ii}^k = 0 \) unless \( k = vt \) (\( k \) is a multiple of \( t \)) and \( t \) is the largest integer with this property. The state is aperiodic if no such \( t > 1 \) exists. And a MC is called aperiodic if all of its states are aperiodic.

Many sources (including Chapter 15 of the 3rd Edition of Feller Volume 1) present a number of useful simple results about MC's. Among them are the following.

**Theorem 46** All states of an irreducible MC are of the same type (with regard to persistence and periodicity).

**Theorem 47** A finite state space irreducible MC is persistent.
Theorem 48  Suppose that a MC is irreducible, aperiodic and persistent. Suppose further that for each state $i$ the mean recurrence time is finite, i.e.

$$\sum_{k=1}^{\infty} k f_{ii}^k < \infty .$$

Then an invariant/stationary distribution for the MC exists, i.e. $\exists \{u_j\}$ with $u_j > 0$ and $\sum u_j = 1$ such that

$$u_j = \sum_i u_i p_{ij} .$$

(If the chain is started with distribution $\{u_j\}$, after one transition it is in states 1, 2, 3, ... with probabilities $\{u_j\}$.) Further, this distribution $\{u_j\}$ satisfies

$$u_j = \lim_{k \to \infty} p_{ij}^k \forall i ,$$

and

$$u_j = \frac{1}{\sum_{k=1}^{\infty} k f_{jj}^k} .$$

There is a converse of this theorem.

Theorem 49  An irreducible, aperiodic MC for which $\exists \{u_j\}$ with $u_j > 0$ and $\sum u_j = 1$ such that $u_j = \sum_i u_i p_{ij}$ must be persistent with $u_j = \frac{1}{\sum_{k=1}^{\infty} k f_{jj}^k}$.

And there is an important "ergodic" result that guarantees that "time averages" have the right limits.

Theorem 50  Under the hypotheses of Theorem 48, if $g$ is a real-valued function such that

$$\sum_j |g(j)| u_j < \infty$$

then for any $j$, if $X_0 = j$

$$\frac{1}{n} \sum_{k=1}^{n} g(X_k) \xrightarrow{a.s.} \sum_j g(j) u_j$$

(Note that the choice of $g$ as an indicator provides approximations for stationary probabilities.)
With this background, the basic idea of MCMC is the following. If we wish to simulate from a distribution \( \{u_j\} \) or approximate properties of the distribution that can be expressed as moments of some function \( g \), we find a convenient MC \( \{X_k\} \) whose invariant distribution is \( \{u_j\} \). From a starting state \( X_0 = i \), one uses \( P \) to simulate \( X_1 \). Using the realization \( X_1 = x_1 \) and \( P \), one simulates \( X_2 \), etc. One applies Theorem 50 to approximate the quantity of interest. Actually, it is common practice to use a "burn in" of \( m \) periods before starting the kind of time average indicated in Theorem 50. Two important questions about this plan (only one of which we'll address) are:

1) How does one set up an appropriate/convenient \( P \)?
2) How big should \( m \) be? (We'll not touch this matter.)

In answer to question 1), there are presumably a multitude of chains that would do the job. But there is the following useful sufficient condition (that has application in the original motivating problem of simulating from high dimensional distributions) for a chain to have \( \{u_j\} \) for an invariant distribution.

**Lemma 51** If \( \{X_k\} \) is a MC with transition probabilities satisfying
\[
u_i P_{ij} = u_j P_{ji},
\]
then it has invariant distribution \( \{u_j\} \).

Note then that if a candidate \( P \) satisfies ★ and is irreducible and aperiodic, Theorem 49 shows that it is persistent and Theorem 48 then shows that any arbitrary starting value can be used and yields approximate realizations from \( \{u_j\} \) and Theorem 50 implies that "time averages" can be used to approximate properties of \( \{u_j\} \).

Several useful MCMC schemes can be shown to have the "right" invariant distributions by noting that they satisfy ★.

For example, Lemma 51 can be applied to the "**Metropolis-Hastings Algorithm.**" That is, let \( T = (t_{ij}) \) be any stochastic matrix corresponding to an irreducible aperiodic MC. (Presumably, some choices will be better that others in terms of providing quick convergence to a target invariant distribution. But that issue is beyond the scope of the present exposition.) (Note that in a finite case, one can take \( t_{ij} = 1/(\text{the number of states}) \.) The Metropolis-Hastings algorithm simulates as follows:

- Supposing that \( X_{n-1} = i \), generate \( J \) (at random) according to the distribution over the state space specified by row \( i \) of \( T \), that is according to \( \{t_{ij}\} \).
- Then generate \( X_n \) based on \( i \) and (the randomly generated) \( J \) according to
\[
X_n = \begin{cases} 
J & \text{with probability } \min\left(1, \frac{u_{ij} t_{ij}}{u_{i,j}} \right) \\
i & \text{with probability } \max\left(0, 1 - \frac{u_{ij} t_{ij}}{u_{i,j}} \right)
\end{cases}.
\]

★ ★
Note that for \( \{X_k\} \) so generated, for \( j \neq i \)

\[
p_{ij} = P[X_n = j|X_{n-1} = i] = \min\left(1, \frac{u_j t_{ji}}{u_i t_{ij}}\right) t_{ij}
\]

and

\[
p_{ii} = P[X_n = i|X_{n-1} = i] = t_{ii} + \sum_{j \neq i} \max\left(0, 1 - \frac{u_j t_{ji}}{u_i t_{ij}}\right) t_{ij}.
\]

So, for \( i \neq j \)

\[
u_i p_{ij} = \min(u_i t_{ij}, u_j t_{ji}) = u_j p_{ji}.
\]

That is, \( \star \) holds and the MC \( \{X_k\} \) has stationary distribution \( \{u_j\} \). (Further, the assumption that \( T \) corresponds to an irreducible aperiodic chain implies that \( \{X_k\} \) is irreducible and aperiodic.)

Notice, by the way, that in order to use the Metropolis-Hastings Algorithm one only has to have the \( u_j \)'s up to a multiplicative constant. That is, if we have \( u_j \propto u_i \), we may compute the ratios \( u_j/u_i \) as \( v_j/v_i \) and are in business as far as simulation goes. (This fact is especially attractive in Bayesian contexts where it can be convenient to know a posterior only up to a multiplicative constant.) Notice also that if \( T \) is symmetric, (i.e. \( t_{ij} = t_{ji} \)), step \( \star \star \) of the Metropolis algorithm becomes

\[
X_n = \begin{cases} 
J & \text{with probability } \min\left(1, \frac{u_j}{u_i}\right) \\
i & \text{with probability } \max\left(0, 1 - \frac{u_j}{u_i}\right)
\end{cases}.
\]

An algorithm similar to the Metropolis-Hastings algorithm is the "Barker Algorithm."

The Barker algorithm modifies the above by replacing \( \min\left(1, \frac{u_j t_{ji}}{u_i t_{i,j}}\right) \) with \( \frac{u_j t_{ji}}{u_i t_{i,j} + u_j t_{ji}} \) in \( \star \star \). Note that for this modification of the Metropolis algorithm, for \( j \neq i \)

\[
p_{ij} = \left(\frac{u_j t_{ji}}{u_i t_{ij} + u_j t_{ji}}\right) t_{ij},
\]

so

\[
u_i p_{ij} = \frac{(u_i t_{ij})(u_j t_{ji})}{u_i t_{ij} + u_j t_{ji}} = u_j p_{ji}.
\]

That is, \( \star \) holds and thus Lemma 51 guarantees that \( \{X_k\} \) has invariant distribution \( \{u_j\} \). (And \( T \) irreducible and aperiodic continues to imply that \( \{X_k\} \) is also.)
Note also that since
\[
\frac{u_j t_{ji}}{u_i t_{ij} + u_j t_{ji}} = \frac{u_j t_{ji}}{t_{ij} + \frac{u_j}{u_i} t_{ji}}
\]

once again it suffices to know the \( u_j \) up to a multiplicative constant in order to be able to implement Barker's algorithm. (This is again of comfort to Bayesians.) Finally, note that if \( T \) is symmetric, step ★★ of the Barker algorithm becomes
\[
X_n = \begin{cases} 
J & \text{with probability } \frac{u_j}{u_i + u_j} \\
i & \text{with probability } \frac{u_i}{u_i + u_j}. 
\end{cases}
\]

Finally, consider now the "Gibbs Sampler," or as Schervish correctly argues that it should be termed, "Successive Substitution Sampling," for generating an observation from a high dimensional distribution. For sake of concreteness, consider the situation where the distribution of a discrete 3-dimensional random vector \((Y, V, W)\) with probability mass function \( f(y, v, w) \) is at issue. In the general MC setup, a possible outcome \((y, v, w)\) represents a "state" and the distribution \( f(y, v, w) \) is the "\( \{u_j\} \)" from which one wants to simulate. One defines a MC \( \{X_k\} \) as follows. For an arbitrary starting state \((y_0, v_0, w_0)\):

• Generate \( X_1 = (Y_1, v_0, w_0) \) by generating \( Y_1 \) from the conditional distribution of \( Y|V = v_0 \) and \( W = w_0 \), i.e. from the (conditional) distribution with probability function \( f_{Y|V,W}(y|v_0, w_0) = \frac{f(y,v_0,w_0)}{\sum_i f(i,v_0,w_0)} \). Let \( y_1 \) be the realized value of \( Y_1 \).

• Generate \( X_2 = (y_1, V_1, w_0) \) by generating \( V_1 \) from the conditional distribution of \( V|Y = y_1 \) and \( W = w_0 \), i.e. from the (conditional) distribution with probability function \( f_{V|Y,W}(v|y_1, w_0) = \frac{f(y_1,v,w_0)}{\sum_i f(y_1,i,w_0)} \). Let \( v_1 \) be the realized value of \( V_1 \).

• Generate \( X_3 = (y_1, v_1, W_1) \) by generating \( W_1 \) from the conditional distribution of \( W|Y = y_1 \) and \( V = v_1 \), i.e. from the (conditional) distribution with probability function \( f_{W|Y,V}(w|y_1, v_1) = \frac{f(y_1,v_1,w)}{\sum_i f(y_1,v_1,i)} \). Let \( w_1 \) be the realized value of \( W_1 \).

• Generate \( X_4 = (Y_2, v_1, w_1) \) by generating \( Y_2 \) from the conditional distribution of \( Y|V = v_1 \) and \( W = w_1 \), i.e. from the (conditional) distribution with probability function \( f_{Y|V,W}(y|v_1, w_1) = \frac{f(y,v_1,w_1)}{\sum_i f(i,v_1,w_1)} \). Let \( y_2 \) be the realized value of \( Y_2 \).

And so on, for some appropriate number of cycles.

Note that with this algorithm, a typical transition probability (for a step where a \( Y \) realization is going to be generated) is
\[ P[X_n = (y', v, w)|X_{n-1} = (y, v, w)] = \frac{f(y', v, w)}{\sum_l f(l, v, w)} \]

so if \( X_{n-1} \) has distribution \( f \), the probability that \( X_n = (y, v, w) \) is

\[ \sum_{l'} f(l', v, w) \frac{f(y, v, w)}{\sum_l f(l, v, w)} = f(y, v, w) , \]

that is, \( X_n \) also has distribution \( f \). And analogous results hold for transitions of all 3 types (\( Y', V \) and \( U \) realizations).

But it is not absolutely obvious how to apply the results 48 and 49 here since the chain is not stationary. The transition mechanism is not the same for a "Y" transition as it is for a "V" transition. The key to extricating oneself from this problem is to note that if \( P_Y, P_V \) and \( P_W \) are respectively the transition matrices for \( Y, V \) and \( W \) exchanges, then

\[ P = P_Y P_V P_W \]

describes an entire cycle of the SSS algorithm. Then \( \{X'_k\} \) defined by \( X'_n = X_{3n} \) is a stationary Markov Chain with transition matrix \( P \). The fact that \( P_Y, P_V \) and \( P_W \) all leave the distribution \( f \) invariant then implies that \( P \) also leaves \( f \) invariant. So one is in the position to apply Theorems 49 and 48. If \( P \) is irreducible and aperiodic, Theorem 49 says that the chain \( \{X'_k\} \) is persistent and then Theorems 48 and 50 say that \( f \) can be simulated using an arbitrary starting state.

The non discrete \( Y \) version of all this is somewhat more subtle, but in many ways parallel to the exposition given above. There is a nice exposition in Tierney's December 1994 Annals of Statistics paper (available at ISU through JSTOR at http://www.jstor.org/cgi-bin/jstor/listjournal). The following is borrowed from that source.

Consider now a (possibly non discrete) state space \( \mathcal{X} \) and a target distribution \( \pi \) on \((\mathcal{X}, \mathcal{B})\) (for which one wishes to approximate some property via simulation). Suppose that \( P(x, A) : \mathcal{X} \times \mathcal{B} \rightarrow [0, 1] \) is a "transition kernel" (a regular conditional probability, a function such that for each \( x \), \( P(x, \cdot) \) is a probability measure, and for each \( A \), \( P(\cdot, A) \) is \( \mathcal{B} \) measurable). Consider a stochastic process \( \{X_k\} \) with the property that given \( X_1 = x_1, \ldots, X_{n-1} = x_{n-1} \), the variable \( X_n \) has distribution \( P(x_{n-1}, \cdot) \). \( \{X_k\} \) is a general stationary discrete time MC with transition kernel \( P(x, A) \), and provided one can simulate from \( P(x, A) \), one can simulate realizations of \( \{X_k\} \). If \( P(x, A) \) is properly chosen, empirical properties of a realization of \( \{X_k\} \) can approximate theoretical properties of \( \pi \).
Notice that the $n$-step transition kernels can be computed recursively beginning with $P(x, A)$ using

$$ P^n(x, A) = \int P^{n-1}(y, A)dP(x, y) $$

**Definition 22** $P$ has invariant distribution $\pi$ means that $\forall A \in \mathcal{B}$

$$ \pi(A) = \int P(x, A)d\pi(x) $$

**Definition 23** $P$ is called $\pi$-irreducible if for each $x$ and each $A$ with $\pi(A) > 0 \exists n(x, A) \geq 1$ such that $P^{n(x,A)}(x, A) > 0$.

**Definition 24** A $\pi$-irreducible transition kernel $P$ is called periodic if $\exists$ an integer $d \geq 2$ and a sequence of sets $A_0, A_1, ..., A_{d-1}, A_d = A_0$ in $\mathcal{B}$ such that for $i = 0, 1, ..., d-1$ and all $x \in A_i$, $P(x, A_{i+1}) = 1$. If a $\pi$-irreducible transition kernel $P$ is not periodic it is called aperiodic.

**Definition 25** A $\pi$-irreducible transition kernel $P$ with invariant distribution $\pi$ is called Harris recurrent if the corresponding MC, $\{X_k\}$, has the property that $\forall A \in \mathcal{B}$ with $\pi(A) > 0$ and all $x \in \mathcal{X}$

$$ P_{X_0=x}[X_n \in A \text{ infinitely often}] = 1 $$

**Theorem 52** Suppose that transition kernel $P$ has invariant distribution $\pi$, is $\pi$-irreducible, aperiodic and Harris recurrent. If $g$ is such that

$$ \int |g(x)|d\pi(x) < \infty $$

then for $X_0 = x$ any element of $\mathcal{X}$

$$ \frac{1}{n} \sum_{k=1}^{n} g(X_k) \overset{\text{a.s.}}{\rightarrow} \int g(x)d\pi(x) $$

Theorem 52 says that integrals of $\pi$ (like, for example, probabilities and moments) can be approximated by long run sample averages of a sample path of the MC $\{X_k\}$ beginning from any starting value, provided a suitable $P$ can be identified. "Suitable" means, of course, that it has the properties hypothesized in Theorem 52 and one can simulate from it knowing no more about $\pi$ than is known in practice. The game of MCMC is then to identify useable $P$'s. As Tierney's article lays out, the general versions of SSS, the Barker algorithm and the Metropolis-Hastings algorithm (and others) can often be shown to satisfy the hypotheses of Theorem 52.
The meaning of "SSS" in the general case is obvious (one generates in turn from each conditional of a coordinate of a multidimensional $Y$ given all other current coordinates and considers what one has at the end of each full cycle of these simulations). To illustrate what another version of MCMC means in the general case, consider the Metropolis-Hastings algorithm.

Suppose that $Q(x, A)$ is some convenient transition kernel with the property that for some $\sigma$-finite measure $\mu$ on $(\mathcal{X}, \mathcal{B})$ (dominating $\pi$) and a nonnegative function $q(x, y)$ on $\mathcal{X} \times \mathcal{X}$ (that is $\mathcal{B} \times \mathcal{B}$ measurable) with $\int_{\mathcal{X}} q(x, y) d\mu(y) = 1$ for all $x$,

$$Q(x, A) = \int_A q(x, y) d\mu(y)$$

Let

$$\frac{d\pi}{d\mu} = p$$

and suppose that $\pi$ is not concentrated on a single point. Then with

$$\alpha(x, y) = \min \left( \begin{array}{ll} \frac{p(y)q(y,x)}{p(x)q(x,y)}, 1 \end{array} \right)$$

if $p(x)q(x, y) > 0$

if $p(x)q(x, y) = 0$

and

$$p(x, y) = \begin{cases} q(x, y)\alpha(x, y) & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

a Metropolis-Hastings transition kernel is

$$P(x, A) = \int_A p(x, y) d\mu(y) + \left( 1 - \int p(x, y) d\mu(y) \right) \Delta_x(A)$$

where $\Delta_x$ is a probability measure degenerate at $x$. The "simulation interpretation" of this is that if $X_{n-1} = x$, one generates $y$ from $Q(x, \cdot)$ and computes $\alpha(x, y)$. Then with probability $\alpha(x, y)$, one sets $X_n = y$ and otherwise takes $X_n = x$. Note that as in the discrete version, the algorithm depends on $p$ only through the ratios $p(y)/p(x)$ so it suffices to know $p$ only up to a multiplicative constant, a fact of great comfort to Bayesians wishing to study analytically intractable posteriors.

It is straightforward to verify that the kernel defined by $\Delta_x$ is $\pi$-invariant. It turns out that sufficient conditions for the kernel $P$ to be $\pi$-irreducible are that $Q$ is $\pi$-irreducible and either that i) $q(x, y) > 0 \forall x$ and $y$ or that ii) $q(x, y) = q(y, x) \forall x$ and $y$. In turn, $\pi$-irreducibility of $P$ is sufficient to imply Harris recurrence for $P$. And, if $P$ is $\pi$-irreducible, a sufficient condition for aperiodicity of $P$ is that
so there are some fairly concrete conditions that can be checked to verify the hypotheses of Theorem 52.

A (Very) Little Bit About Bootstrapping

Suppose that the entries of \( X = (X_1, X_2, \ldots, X_n) \) are iid according to some unknown (possibly multivariate) distribution \( F \), and for a random function

\[
R(X, F)
\]

there is some aspect of the distribution of \( R(X, F) \) that is of statistical interest. For example, in a univariate context, we might be interested in the \( (F) \) probability that

\[
|X - \mu_F| < S \sqrt{n}
\]

has absolute value less than 2. Often, even if one knew \( F \), the problem of finding properties of the distribution of \( R(X, F) \) would be analytically intractable. (Of course, knowing \( F \), simulation might provide at least some approximation for the property of interest ...)

As a means of possibly approximating the distribution of \( R(X, F) \), Efron suggested the following. Suppose that \( \widehat{F} \) is an estimator of \( F \). In parametric contexts, where \( \hat{\theta}(X) \) is an estimator of \( \theta \), this could be \( \widehat{F} = F_{\hat{\theta}} \). In nonparametric contexts, if \( \hat{F}_n \) is the empirical distribution of the sample \( X \), we might take \( \widehat{F} = \hat{F}_n \). Suppose then that \( X^* = (X^*_1, \ldots, X^*_n) \) has entries that are an iid sample from \( \widehat{F} \). One might hope that (at least in some circumstances)

\[
\text{the } F \text{ distribution of } R(X, F) \approx \text{the } \widehat{F} \text{ distribution of } R(X^*, \widehat{F})
\]

(This will not always be a reasonable approximation.) Now even the right side of this expression will typically be analytic intractable. What, however, can be done is to now simulate from this distribution. (Where one can not simulate from the distribution on the left because \( F \) is not known, simulation from the distribution on the right is possible.) That is, suppose that \( X^{*1}, X^{*2}, \ldots, X^{*B} \) are \( B \) (bootstrap) iid samples of size \( n \) from \( \widehat{F} \). Using these, one may compute

\[
R^*_i = R(X^{*i}, \widehat{F})
\]

and use the empirical distribution of \( \{R^*_i|i = 1, \ldots, B\} \) to approximate the \( \widehat{F} \) distribution of \( R(X^*, \widehat{F}) \).
General Decision Theory

This is a formal framework for posing problems of finding good statistical procedures as optimization problems.

**Basic Elements of a (Statistical) Decision Problem:**

- $X$ data: formally a mapping $(\Omega, \mathcal{H}) \rightarrow (\mathcal{X}, \mathcal{B})$
- $\Theta$ parameter space: often one needs to do integrations over $\Theta$, so let $\mathcal{C}$ be a $\sigma$-algebra on $\Theta$ and assume that for $B \in \mathcal{B}$, the mapping taking $\theta \mapsto P_\theta(B)$ from $(\Theta, \mathcal{C}) \rightarrow (\mathbb{R}^1, \mathcal{B}_1)$ is measurable
- $\mathcal{A}$ decision/action space: the class of possible decisions or actions we will also need to average over (random) actions, so let $\mathcal{E}$ be $\sigma$-algebra on $\mathcal{A}$
- $L(\theta, \alpha)$ loss function: $L : \Theta \times \mathcal{A} \rightarrow [0, \infty)$ and needs to be suitably measurable
- $\delta(x)$ decision rule: $\delta(x) : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathcal{A}, \mathcal{E})$ and for data $X$, $\delta(X)$ is the decision taken on the basis of observing $X$

The fact that $L(\theta, \delta(X))$ is random prevents any sensible comparison of decision rules until one has averaged out over the distribution of $X$, producing something called a risk function.

**Definition 26** $R(\theta, \delta) \doteq E_\theta L(\theta, \delta(X)) = \int L(\theta, \delta(x)) dP_\theta(x)$ maps $\Theta \rightarrow [0, \infty)$ and is called the risk function (of the decision rule $\delta$).

**Definition 27** $\delta$ is at least as good as $\delta'$ if $R(\delta, \theta) \leq R(\delta', \theta) \ \forall \theta$.

**Definition 28** $\delta$ is better than $\delta'$ if $R(\delta, \theta) \leq R(\delta', \theta) \ \forall \theta$ with strict inequality for at least one $\theta \in \Theta$.

**Definition 29** $\delta$ and $\delta'$ are risk equivalent if $R(\theta, \delta) = R(\theta, \delta') \ \forall \theta$.

(Note that risk equivalency need not imply equality of decision rules.)

**Definition 30** $\delta$ is best in a class of decision rules $\Delta$ if

i) $\delta \in \Delta$

and ii) $\delta$ is at least as good as any other $\delta' \in \Delta$.

Typically, only smallish classes of decision rules will have best elements. (For example, in some point estimation problems there are best unbiased estimators, where no estimator is best overall. Here the class $\Delta$ is the class of unbiased estimators.) An alternative to looking for best elements in small classes of rules is to somehow reduce the function of $\theta, R(\theta, \delta)$, to a single number and look for rules minimizing the number. (Reduction by
averaging amounts to looking for Bayes rules. Reduction by taking a maximum value amounts to looking for minimax rules.)

**Definition 31** $\delta$ is inadmissible in $\Delta$ if $\exists \delta' \in \Delta$ which is better than $\delta$ (which "dominates" $\delta$).

**Definition 32** $\delta$ is admissible in $\Delta$ if it is not inadmissible in $\Delta$ (if there is no $\delta' \in \Delta$ which is better).

On the face of it, it would seem that one would never want to use an inadmissible decision rule. But, as it turns out, there can be problems where all rules are inadmissible. (More in this direction later.)

Largely for technical reasons, it is convenient/important to somewhat broaden one's understanding of what constitutes a decision rule. To this end, as a matter of notation, let

$$\mathcal{D} = \{\delta(x)\} = \text{the class of (nonrandomized) decision rules}.$$ 

Then one can define two fairly natural notions of randomized decision rules.

**Definition 33** Suppose that for each $x \in \mathcal{X}$, $\phi_x$ is a distribution on $(\mathcal{A}, \mathcal{E})$. Then $\phi_x$ is called a behavioral decision rule.

The idea is that having observed $X = x$, one then randomizes over the action space according to the distribution $\phi_x$ in order to choose an action. (This should remind you of what is done in some theoretical arguments in hypothesis testing.)

**Notation:** Let $\mathcal{D}^* = \{\phi_x\} = \text{the class of behavioral decision rules}$

It is possible to think of $\mathcal{D}$ as a subset of $\mathcal{D}^*$. For $\delta \in \mathcal{D}$, let $\phi_{\delta}^\delta$ be a point mass distribution on $\mathcal{A}$ concentrated at $\delta(x)$. Then $\phi_{\delta}^\delta \in \mathcal{D}^*$ and is "clearly" equivalent to $\delta$.

The risk of a behavioral decision rule is (abusing notation and reusing the symbols $R(\cdot, \cdot)$ in this context as well as that of Definition 8) clearly

$$R(\theta, \phi) = \int_\mathcal{X} \int_\mathcal{A} L(\theta, a) d\phi_x(a) dP_\theta(x).$$

A somewhat less intuitively appealing notion of randomization for decision rules is next. Let $\mathcal{F}$ be a $\sigma$-algebra on $\mathcal{D}$ that contains the singleton sets.

**Definition 34** A randomized decision function (or decision rule) $\psi$ is a probability measure on $(\mathcal{D}, \mathcal{F})$. ($\delta$ with distribution $\psi$ becomes a random object.)

The idea here is that prior to observation, one chooses an element of $\mathcal{D}$ by use of $\psi$, say $\delta$, and then having observed $X = x$ decides $\delta(x)$.
**Notation:** Let $\mathcal{D}_r = \{ \psi \} = \text{the class of randomized decision rules.}$

It is possible to think of $\mathcal{D}$ as a subset of $\mathcal{D}_r$ by identifying $\delta \in \mathcal{D}$ with $\psi_\delta$ placing mass 1 on $\delta$. $\psi_\delta$ is "clearly" equivalent to $\delta$.

The risk of a randomized decision rule is (once again abusing notation and reusing the symbols $R(\cdot, \cdot)$) clearly

$$R(\theta, \psi) = \int_{\mathcal{D}} R(\theta, \delta) d\psi(\delta) = \int_{\mathcal{D}} \int_{\mathcal{X}} L(\theta, \delta(x)) dP_\theta(x) d\psi(\delta) ,$$

assuming that $R(\theta, \psi)$ is properly measurable.

The behavioral decision rules are probably more appealing/natural than the randomized decision rules. On the other hand, the randomized decision rules are often easier to deal with in proofs. So a reasonably important question is "When are $\mathcal{D}$ and $\mathcal{D}_r$ equivalent in the sense of generating the same set of risk functions?"

**"Result"53** (Some properly qualified version of this is true ... See Ferguson pages 26-27) If $\mathcal{A}$ is a complete separable metric space with $\mathcal{E}$ the Borel $\sigma$-algebra, appropriate conditions hold on the $P_\theta$ and $\phi_x$ hold, then $\mathcal{D}^*$ and $\mathcal{D}_r$ are equivalent in terms of generating the same set of risk functions.

It is also of interest to know when the extra complication of considering randomized rules is really not necessary. One simple result in that direction concerns cases where $L$ is convex in its second argument.

**Lemma 54** (See page 151 of Schervish, page 40 of Berger and page 78 of Ferguson) Suppose that $\mathcal{A}$ is a convex subset of $\mathcal{R}^d$ and $\phi_x$ is a behavioral decision rule. Define a nonrandomized decision rule $\delta$ by

$$\delta(x) = \int_{\mathcal{A}} a d\phi_x(a) .$$

(In the case that $d > 1$, interpret $\delta(x)$ as vector-valued, the integral as a vector of integrals over the $d$ coordinates of $a \in \mathcal{A}$.) Then

i) If $L(\theta, \cdot) : \mathcal{A} \to [0, \infty)$ is convex, then

$$R(\theta, \delta) \leq R(\theta, \phi) .$$

ii) If $L(\theta, \cdot) : \mathcal{A} \to [0, \infty)$ is strictly convex, $R(\theta, \phi) < \infty$ and

$$P_\theta\{ x | \phi_x \text{ is nondegenerate} \} > 0 ,$$

then

$$R(\theta, \delta) < R(\theta, \phi) .$$
(Strict convexity of a function $g$ means that for $\alpha \in (0, 1)$
\[
g(\alpha x + (1 - \alpha)y) < \alpha g(x) + (1 - \alpha)g(y)
\]
for $x$ and $y$ in the convex domain of $g$. See Berger pages 38 through 40. There is an
extension of Jensen’s inequality that says that if $g$ is strictly convex and the distribution of
$X$ is not concentrated at a point, then $Eg(X) > g(E(X))$.

**Corollary 55** Suppose that $A$ is a convex subset of $\mathbb{R}^d$ and $\phi_x$ is a behavioral decision
rule. Define a nonrandomized decision rule $\delta$ by
\[
\delta(x) = \int_A ad\phi_x(a).
\]

i) If $L(\theta, \cdot) : A \rightarrow [0, \infty)$ is convex $\forall \theta$, then $\delta$ is at least as good as $\phi$.

ii) If $L(\theta, \cdot) : A \rightarrow [0, \infty)$ is convex $\forall \theta$ and for some $\theta_0$ the function
$L(\theta_0, \cdot) : A \rightarrow [0, \infty)$ is strictly convex, $R(\theta_0, \phi) < \infty$ and
\[
P_{\theta_0}\{x | \phi_x \text{ is nondegenerate} \} > 0
\]
then $\delta$ is better than $\phi$.

Corollary 55 says roughly that in the case of a convex loss function, one need not worry
about considering randomized decision rules.

**Finite Dimensional** Geometry of Decision Theory

A very helpful device for understanding some of the basics of decision theory is to
consider the geometry associated with cases where $\Theta$ is finite. So until further notice,
assume
\[
\Theta = \{ \theta_1, \theta_2, ..., \theta_k \}
\]
and
\[
R(\theta, \psi) < \infty \forall \theta \in \Theta \text{ and } \psi \in D_x.
\]
As a matter of notation, let
\[
S = \{ y = (y_1, y_2, ..., y_k) \in \mathbb{R}^k \mid y_i = R(\theta_i, \psi) \forall i \text{ and } \psi \in D_x \}.
\]
$S$ is the set of all randomized risk vectors.
Theorem 56  $S$ is convex.

In fact, it turns out to be the case, and is more or less obvious from Theorem 56, that if

$$S^0 = \{ y | y_i = R(\theta_i, \psi) \ \forall i \text{ and } \psi \in \mathcal{D} \} ,$$

then $S$ is the convex hull of $S^0$, i.e. the smallest convex set containing $S^0$.

Definition 35  For $x \in \mathbb{R}^k$, the lower quadrant of $x$ is

$$Q_x = \{ z | z_i \leq x_i \ \forall i \} .$$

Theorem 57  $y \in S$ (or the decision rule giving rise to $y$) is admissible iff

$$Q_y \cap S = \{ y \} .$$

Definition 36  The lower boundary of $S$ is

$$\lambda(S) = \{ y | Q_y \cap \overline{S} = \{ y \} \}$$

for $\overline{S}$ the closure of $S$.

Definition 37  $S$ is closed from below if $\lambda(S) \subset S$.

Notation:  Let $A(S)$ stand for the class of admissible rules (or risk vectors corresponding to them) in $S$.

Theorem 58  If $S$ is closed from below, then $A(S) = \lambda(S)$.

Theorem 58'  If $S$ is closed, then $A(S) = \lambda(S)$.

Complete Classes of Decision Rules

Now drop the finiteness assumption on $\Theta$.

Definition 38  A class of decision rules $C \subset \mathcal{D}^*$ is a complete class if for any $\phi \notin C$, $\exists$ a $\phi' \in C$ such that $\phi'$ is better than $\phi$.

Definition 39  A class of decision rules $C \subset \mathcal{D}^*$ is an essentially complete class if for any $\phi \notin C$, $\exists$ a $\phi' \in C$ such that $\phi'$ is at least as good as $\phi$.

Definition 40  A class of decision rules $C \subset \mathcal{D}^*$ is a minimal complete class if it is complete and is a subset of any other complete class.

Notation:  Let $A(\mathcal{D}^*)$ be the set of admissible rules in $\mathcal{D}^*$.

Theorem 59  If a minimal complete class $C$ exists, then $C = A(\mathcal{D}^*)$. 
**Theorem 60** If $A(D^*)$ is complete, then it is minimal complete.

Theorems 59 and 60 are the properly qualified versions of the rough (and not quite true) statement "$A(D^*)$ equals the minimal complete class".

**Sufficiency and Decision Theory**

We worked very hard on the notion of sufficiency, with the goal in mind of establishing that one "doesn't need" $X$ if one has $T(X)$. That hard work ought to have some implications for decision theory. Presumably, in most cases one can get by with basing a decision rule on a sufficient statistic.

**Result 61** (See page 120 of Ferguson, page 36 of Berger and page 151 of Schervish)

Suppose a statistic $T$ is sufficient for $P = \{P_0\}$. Then if $\phi$ is a behavioral decision rule there exists another behavioral decision rule $\phi'$ that is a function of $T$ and has the same risk function as $\phi$.

Note that $\phi'$ a function of $T$ means that $T(x) = T(y)$ implies that $\phi'_x$ and $\phi'_y$ are the same measures on $A$. Note also that Result 61 immediately implies that the class of decision rules in $D^*$ that are functions of $T$ is essentially complete. In terms of risk, one need only consider rules that are functions of sufficient statistics.

The construction used in the proof of Result 61 doesn't aim to improve on the original (behavioral) decision rule by moving to a function of the sufficient statistic. But in convex loss cases, there is a construction that produces rules depending on a sufficient statistic and potentially improving on an initial (nonrandomized) decision rule. This is the message of Theorem 64. But to prove Theorem 64, a lemma of some independent interest is needed. That lemma and an immediate corollary are stated before Theorem 64.

**Lemma 62** Suppose that $A \subset \mathcal{R}^k$ is convex and $\delta_1(x)$ and $\delta_2(x)$ are two nonrandomized decision rules. Then $\delta(x) = \frac{1}{2}(\delta_1(x) + \delta_2(x))$ is also a decision rule. Further,

1) if $L(\theta, a)$ is convex in $a$ and $R(\theta, \delta_1) = R(\theta, \delta_2)$, then $R(\theta, \delta) \leq R(\theta, \delta_1)$, and

2) if $L(\theta, a)$ is strictly convex in $a$, $R(\theta, \delta_1) = R(\theta, \delta_2) < \infty$ and $P_\theta(\delta_1(X) \neq \delta_2(X)) > 0$, then $R(\theta, \delta) < R(\theta, \delta_1)$.

**Corollary 63** Suppose $A \subset \mathcal{R}^k$ is convex and $\delta_1(x)$ and $\delta_2(x)$ are two nonrandomized decision rules with identical risk functions. If $L(\theta, a)$ is convex in $a \forall \theta$, the existence of a $\theta \in \Theta$ such that $L(\theta, a)$ is strictly convex in $a$, $R(\theta, \delta_1) = R(\theta, \delta_2) < \infty$ and $P_\theta(\delta_1(X) \neq \delta_2(X)) > 0$, implies that $\delta_1$ and $\delta_2$ are inadmissible.
**Theorem 64** (See Berger page 41, Ferguson page 121, TPE page 50, Schervish page 152) (Rao-Blackwell Theorem) Suppose that $A \subset \mathcal{R}^k$ is convex and $\delta$ is a nonrandomized decision function with $E_\theta \| \delta(X) \| < \infty \forall \theta$. Suppose further that $T$ is sufficient for $\theta$ and with $B_0 = B(T)$, let

$$
\delta_0(x) = E[\delta|B_0](x).
$$

Then $\delta_0$ is a nonrandomized decision rule. Further,

i) if $L(\theta, a)$ is convex in $a$, then $R(\theta, \delta_0) \leq R(\theta, \delta)$, and

ii) for any $\theta$ for which $L(\theta, a)$ is strictly convex in $a, R(\theta, \delta) < \infty$ and $P_\theta(\delta(X) \neq \delta_0(X)) > 0$, one has that $R(\theta, \delta_0) < R(\theta, \delta)$.

**Bayes Rules**

The Bayes approach to decision theory is one means of reducing risk functions to numbers so that they can be compared in a straightforward fashion.

**Notation:** Let $G$ be a distribution on $(\Theta, \mathcal{C})$. $G$ is usually (for philosophical reasons) called the "prior" distribution on $\Theta$.

**Definition 41** The Bayes risk of $\phi \in \mathcal{D}^*$ with respect to the prior $G$ is (abusing notation again and once more using the symbols $R(\cdot, \cdot)$)

$$
R(G, \phi) = \int_\Theta R(\theta, \phi)dG(\theta).
$$

**Notation:** Again abusing notation somewhat, let

$$
R(G) = \inf_{\phi' \in \mathcal{D}^*} R(G, \phi').
$$

**Definition 42** $\phi$ is said to be Bayes with respect to $G$ (or to be a Bayes rule with respect to $G$) provided

$$
R(G, \phi) = R(G).
$$

**Definition 43** $\phi$ is said to be $\epsilon$-Bayes with respect to $G$ provided

$$
R(G, \phi) \leq R(G) + \epsilon.
$$

Bayes rules are often admissible, at least if the prior involved "spreads its mass around sufficiently."

**Theorem 65** If $\Theta$ is finite (or countable), $G$ is a prior distribution with $G(\theta) > 0 \forall \theta$ and $\phi$ is Bayes with respect to $G$, then $\phi$ is admissible.
Theorem 66 Suppose $\Theta \subset \mathcal{R}^k$ is such that every neighborhood of any point $\theta \in \Theta$ has nonempty intersection with the interior of $\Theta$. Suppose further that $R(\theta, \phi)$ is continuous in $\theta$ for all $\phi \in \mathcal{D}^*$. Let $G$ be a prior distribution that has support $\Theta$ in the sense that every open ball that is a subset of $\Theta$ has positive $G$ probability. Then if $R(G) < \infty$ and $\phi$ is a Bayes rule with respect to $G$, $\phi$ is admissible.

Theorem 67 If every Bayes rule with respect to $G$ has the same risk function, they are all admissible.

Corollary 68 In a problem where Bayes rules exist and are unique, they are admissible.

A converse of Theorem 65 (for finite $\Theta$) is given in Theorem 70. Its proof requires the use of a piece of $k$-dimensional analytic geometry (called the separating hyperplane theorem) that is stated next.

Theorem 69 Let $S_1$ and $S_2$ be two disjoint convex subsets of $\mathcal{R}^k$. Then $\exists$ a $p \in \mathcal{R}^k$ such that $p \neq 0$ and $\sum p_i x_i \leq \sum p_i y_i \; \forall y \in S_1$ and $x \in S_2$.

Theorem 70 If $\Theta$ is finite and $\phi$ is admissible, then $\phi$ is Bayes with respect to some prior.

For nonfinite $\Theta$, admissible rules are "usually" Bayes or "limits" of Bayes rules. See Section 2.10 of Ferguson or page 546 of Berger.

As the next result shows, Bayesians don't need randomized decision rules, at least for purposes of achieving minimum Bayes risk, $R(G)$.

Result 71 (See also page 147 of Schervish) Suppose that $\psi \in \mathcal{D}_*$ is Bayes versus $G$ and $R(G) < \infty$. Then there exists a nonrandomized rule $\delta \in \mathcal{D}$ that is also Bayes versus $G$.

There remain the questions of 1) when Bayes rules exist and 2) what they look like when they do exist. These issues are (to some extent) addressed next.

Theorem 72 If $\Theta$ is finite, $\mathcal{S}$ is closed from below and $G$ assigns positive probability to each $\theta \in \Theta$, there is a rule Bayes versus $G$.

Definition 44 A formal nonrandomized Bayes rule versus a prior $G$ is a rule $\delta(x)$ such that $\forall x \in \mathcal{X}$,

$$\delta(x) \text{ is an } a \in \mathcal{A} \text{ minimizing } \int_{\Theta} L(\theta, a) \left( \frac{f_{\theta}(x)}{\int_{\Theta} f_{\theta}(x) dG(\theta)} \right) dG(\theta).$$
**Definition 45** If $G$ is a $\sigma$-finite measure, a formal nonrandomized generalized Bayes rule versus $G$ is a rule $\delta(x)$ such that $\forall x \in X$, 

$$
\delta(x) \text{ is an } a \in A \text{ minimizing } \int_{\Theta} L(\theta, a) f_{\theta}(x) dG(\theta) .
$$

**Minimax Decision Rules**

An alternative to the Bayesian reduction of $R(\theta, \phi)$ to a number by averaging according to the distribution $G$, is to instead reduce $R(\theta, \phi)$ to a number by maximization over $\theta$.

**Definition 46** A decision rule $\phi \in D^*$ is said to be minimax if

$$
\sup_{\theta} R(\theta, \phi) = \inf_{\phi'} \sup_{\theta} R(\theta, \phi') .
$$

**Definition 47** If a decision rule $\phi$ has constant (in $\theta$) risk, it is called an equalizer rule.

It should make some sense that if one is trying to find a minimax rule by so to speak pushing down the highest peak in the profile $R(\theta, \phi)$, doing so will tend to drive one towards rules with "flat" $R(\theta, \phi)$ profiles.

**Theorem 73** If $\phi$ is an equalizer rule and is admissible, then it is minimax.

**Theorem 74** Suppose that $\{\phi_i\}$ is a sequence of decision rules, each $\phi_i$ Bayes against a prior $G_i$. If $R(G_i, \phi_i) \rightarrow C$ and $\phi$ is a decision rule with $R(\theta, \phi) \leq C \ \forall \theta$, then $\phi$ is minimax.

**Corollary 75** If $\phi$ is Bayes versus $G$ and $R(\theta, \phi) \leq R(G) \ \forall \theta$, then $\phi$ is minimax.

**Corollary 75'** If $\phi$ is an equalizer rule and is Bayes with respect to a prior $G$, then it is minimax.

A reasonable question to ask, in the light of Corollaries 75 and 75', is "Which $G$ should I be looking for in order to produce a Bayes rule that might be minimax?" The answer to this question is "Look for a least favorable prior."

**Definition 48** A prior distribution $G$ is said to be least favorable if

$$
R(G) = \sup_{G'} R(G') .
$$

Intuitively, one might think that if an intelligent adversary were choosing $\theta$, he/she might well generate it according to a least favorable $G$, therefore maximizing the minimum possible risk. In response a decision maker ought to use $\phi$ Bayes versus $G$. Perhaps then, this Bayes rule might tend to minimize the worst that could happen over choice of $\theta$?

**Theorem 76** If $\phi$ is Bayes versus $G$ and $R(\theta, \phi) \leq R(G) \ \forall \theta$, then $G$ is least favorable.
(Corollary 75 already shows $\phi$ in Theorem 76 to be minimax.)

**Optimal Point Estimation (Unbiasedness, Information Inequalities and Invariance and Equivariance)**

**Unbiasedness**

One possible limited class of "sensible" decision rules in estimation problems is the class of "unbiased" estimators. (Here, we are tacitly assuming that $\mathcal{A}$ is $\mathcal{R}^1$ or some subset of it.)

**Definition 49** The bias of an estimator $\delta(x)$ for $\gamma(\theta) \in \mathcal{R}^1$ is

$$E_{\theta}(\delta(X) - \gamma(\theta)) = \int_{\mathcal{X}} (\delta(x) - \gamma(\theta)) dP_{\theta}(x) .$$

**Definition 50** An estimator $\delta(x)$ is unbiased for $\gamma(\theta) \in \mathcal{R}^1$ provided

$$E_{\theta}(\delta(X) - \gamma(\theta)) = 0 \forall \theta .$$

Before jumping into the business of unbiased estimation with too much enthusiasm, an early caution is in order. We supposedly believe that "Bayes admissible" and that admissibility is sort of a minimal "goodness" requirement for a decision rule ... but there are the following theorem and corollary.

**Theorem 77** (See Ferguson pages 48 and 52) If $L(\theta, a) = w(\theta)(\gamma(\theta) - a)^2$ for $w(\theta) > 0$, and $\delta$ is both Bayes wrt $G$ with finite risk function and unbiased for $\gamma(\theta)$, then (according to the joint distribution of $(X, \theta)$)

$$\delta(X) = \gamma(\theta) \text{ a.s.} .$$

**Corollary 78** Suppose $\Theta \subset \mathcal{R}^1$ has at least two elements and the family $\mathcal{P} = \{P_{\theta}\}$ is mutually absolutely continuous. If $L(\theta, a) = (\theta - a)^2$ no Bayes estimator of $\theta$ is unbiased.

**Definition 51** $\delta$ is best unbiased for $\gamma(\theta) \in \mathcal{R}^1$ if it is unbiased and at least as good as any other unbiased estimator.

There need not be any unbiased estimators in a given estimation problem, let alone a best one. But for many common estimation loss functions, if there is a best unbiased estimator it is unique.
**Theorem 79** Suppose that $A \subset \mathcal{R}^1$ is convex, $L(\theta, a)$ is convex in $a \forall \theta$ and $\delta_1$ and $\delta_2$ are two best unbiased estimators of $\gamma(\theta)$. Then for any $\theta$ for which $L(\theta, a)$ is strictly convex in $a$ and $R(\theta, \delta_1) < \infty$

$$P_{\theta}[\delta_1(X) = \delta_2(X)] = 1.$$  

The special case of $L(\theta, a) = (\gamma(\theta) - a)^2$ is, of course, one of most interest. For this case of unweighted squared error loss, for estimators with second moments $R(\theta, \delta) = \text{Var}_\theta(\delta(X))$.

**Definition 52** $\delta$ is called a UMVUE (uniformly minimum variance unbiased estimator) of $\gamma(\theta) \in \mathcal{R}^1$ if it is unbiased and

$$\text{Var}_\theta(\delta(X)) \leq \text{Var}_\theta(\delta'(X)) \forall \theta$$

for any other unbiased estimator of $\gamma(\theta)$, $\delta'$.

A much weaker notion is next.

**Definition 53** $\delta$ is called locally minimum variance unbiased at $\theta_0 \in \Theta$ if it is unbiased and

$$\text{Var}_{\theta_0}(\delta(X)) \leq \text{Var}_{\theta_0}(\delta'(X))$$

for any other unbiased estimator of $\gamma(\theta)$, $\delta'$.

**Notation:** Let $\mathcal{U}_0 = \{\delta | \text{E}_\theta(\delta(X)) = 0$ and $\text{E}_\theta(\delta^2(X)) < \infty \forall \theta\}$. ($\mathcal{U}_0$ is the set of unbiased estimators of 0 with finite variance.)

**Theorem 80** Suppose that $\delta$ is an unbiased estimator of $\gamma(\theta) \in \mathcal{R}^1$ and $\text{E}_\theta(\delta^2(X)) < \infty \forall \theta$. Then

$$\delta \text{ is a UMVUE of } \gamma(\theta) \iff \text{Cov}_\theta(\delta(X), u(X)) = 0 \forall \theta \forall u \in \mathcal{U}_0.$$  

**Theorem 81** (Lehmann/Scheffe) Suppose that $\mathcal{A}$ is a convex subset of $\mathcal{R}^1$, $T$ is a sufficient statistic for $\mathcal{P} = \{P_\theta\}$ and $\delta$ is an unbiased estimator of $\gamma(\theta)$. Let

$$\delta_0 = \text{E}[\delta|T] = \phi \circ T.$$  

Then $\delta_0$ is an unbiased estimator of $\gamma(\theta)$. If in addition, $T$ is complete for $\mathcal{P}$,

i) $\phi' \circ T$ unbiased for $\gamma(\theta)$ implies that $\phi' \circ T = \phi \circ T$ a.s. $P_\theta \forall \theta$, and

ii) $L(\theta, a)$ convex in $a \forall \theta$ implies that

a) $\delta_0$ is best unbiased, and

b) if $\delta'$ is any other best unbiased estimator of $\gamma(\theta)$

$$\delta' = \delta_0 \text{ a.s. } P_\theta$$

for any $\theta$ at which $R(\theta, \delta_0) < \infty$ and $L(\theta, a)$ is strictly convex in $a$.  

49
(Note that iia) says in particular that $\delta_0$ is UMVUE, and that iib) says that $\delta_0$ is the unique UMVUE.)

Basu's Theorem (recall Stat 543) is on occasion helpful in doing the conditioning required to get $\delta_0$ of Theorem 81. The most famous example of this is in the calculation of the UMVUE of $\Phi\left(\frac{\bar{x} - \mu}{\sigma}\right)$ based on $n$ iid $N(\mu, \sigma^2)$ observations.

**Information Inequalities**

These are inequalities about variances of estimators, some of which involve information measures. (See Schervish Sections 2.3.1, 5.1.2 and TPE Sections 2.6, 2.7.)

The fundamental tool here is the Cauchy-Schwarz inequality.

**Lemma 82** (Cauchy-Schwarz) If $U$ and $V$ are random variables with $EU^2 < \infty$ and $EV^2 < \infty$, then

$$EU^2 \cdot EV^2 \geq (EUV)^2 .$$

Applying this to $U = X - EX$ and $V = Y - EY$, we get the following.

**Corollary 83** If $X$ and $Y$ have finite second moments, $(\text{Cov}(X, Y))^2 \leq \text{Var}X \cdot \text{Var}Y$.

**Corollary 84** Suppose that $\delta(Y)$ takes values in $\mathcal{R}^1$ and $E(\delta(Y))^2 < \infty$. If $g$ is a measurable function such that $Eg(Y) = 0$ and $0 < E(g(Y))^2 < \infty$, defining $C = E\delta(Y)g(Y)$

$$\text{Var}\delta(Y) \geq \frac{C^2}{E(g(Y))^2} .$$

A restatement of Corollary 84 in terms that look more "statistical" is next.

**Corollary 84'** Suppose that $\delta(X)$ takes values in $\mathcal{R}^1$ and $E(\delta(X))^2 < \infty$. If $g_\theta$ is a measurable function such that $E_\theta\delta(X) = 0$ and $0 < E_\theta(g_\theta(X))^2 < \infty$, defining $C(\theta) = E_\theta\delta(X)g_\theta(X)$

$$\text{Var}_{\theta}\delta(X) \geq \frac{C^2(\theta)}{E_\theta(g_\theta(X))^2} .$$

**Theorem 85** (Cramer-Rao) Suppose that the model $\mathcal{P}$ is FI regular at $\theta_0$ and that $0 < I(\theta_0) < \infty$. If $E_\theta\delta(X) = \int \delta(x)f_\theta(x)d\mu(x)$ can be differentiated under the integral at $\theta_0$, i.e. if

$$\left.\frac{d}{d\theta}E_\theta\delta(X)\right|_{\theta = \theta_0} = \int \delta(x)\left.\frac{d}{d\theta}f_\theta(x)\right|_{\theta = \theta_0}d\mu(x) ,$$

then
For the special case of $\delta(X)$ viewed as an estimator of $\theta$, we sometimes write $E_{\theta}\delta(X) = \theta + b(\theta)$, and thus Theorem 85 says $\text{Var}_{\theta_0}\delta(X) \geq \frac{(1 + b'(\theta_0))^2}{I(\theta_0)}$.

On the face of it, showing that the C-R lower bound is achieved looks like a general method of showing that an estimator is UMVUE. But the fact is that the circumstances in which the C-R lower bound is achieved are very well defined. The C-R inequality is essentially the Cauchy-Schwarz inequality, and one gets equality in Lemma 82 iff $V = aU$ a.s.. This translates in the proof of Theorem 85 to equality iff

$$
\frac{d}{d\theta}\log f_{\theta}(X) \bigg|_{\theta = \theta_0} = a(\theta_0) \delta(X) + b(\theta_0) \text{ a.s. } P_{\theta_0}.
$$

And, as it turns out, this can happen for all $\theta_0$ in a parameter space $\Theta$ iff one is dealing with a one parameter exponential family of distributions and the statistic $\delta(X)$ is the natural sufficient statistic. So the method works only for UMVUEs of the means of the natural sufficient statistics in one parameter exponential families.

**Result 86** Suppose that for $\eta$ real, $\{P_{\eta}\}$ is dominated by a $\sigma$-finite measure $\mu$, and with $h(x) > 0$,

$$f_{\eta}(x) = \exp(a(\eta) + \eta T(x)) \text{ a.s. } \mu \forall \eta.
$$

Then for any $\eta_0$ in the interior of $\Gamma$ where $\text{Var}_{\eta}(T(X) < \infty, T(X)$ achieves the C-R lower bound for its variance.

(A converse of Result 86 is true, but is also substantially harder to prove.)

The C-R inequality is Corollary 84 with $g_{\theta}(x) = \frac{d}{d\theta}\log f_{\theta}(x)$. There are other interesting choices of $g_{\theta}$. For example (assuming it is permissible to move enough derivatives through integrals), $g_{\theta}(x) = \left(\frac{d}{d\theta}\log f_{\theta}(x)\right) / f_{\theta}(x)$ produces an interesting inequality.

And the choice $g_{\theta}(x) = (f_{\theta}(x) - f_{\theta}(x)) / f_{\theta}(x)$ leads to the next result.

**Theorem 87** If $P_{\theta'} \ll P_{\theta}$,

$$\text{Var}_{\theta}\delta(X) \geq \frac{(E_{\theta}\delta(X) - E_{\theta'}\delta(X))^2}{E_{\theta}\left(\frac{f_{\theta}(X) - f_{\theta'}(X)}{f_{\theta}(X)}\right)^2}.
$$

So, (the Chapman-Robbins Inequality)
There are multiparameter versions of the information inequalities (and for that matter, multiple $\delta$ versions involving the covariance matrix of the $\delta$'s). The fundamental result needed to produce multiparameter versions of the information inequalities is the simple generalization of Corollary 84.

**Lemma 88** Suppose that $\delta(Y)$ takes values in $\mathcal{R}^1$ and $\text{E}(\delta(Y))^2 < \infty$. Suppose further that $g_1, g_2, ..., g_k$ are measurable functions such that $\text{E}g_i(Y) = 0$ and $0 < \text{E}(g_i(Y))^2 < \infty$ $\forall i$ and the $g_i$ are linearly independent as square integrable functions on $\mathcal{X}$ according to the probability measure $P$. Let $C_i = \text{E}\delta(Y)g_i(Y)$ and define a $k \times k$ matrix

$$D = (\text{E}g_i(Y)g_j(Y)) .$$

Then for $C = (C_1, C_2, ..., C_k)$

$$\text{Var}\delta(Y) \geq C'D^{-1}C .$$

The "statistical-looking" version of Lemma 88 is next.

**Lemma 88'** Suppose that $\delta(X)$ takes values in $\mathcal{R}^1$ and $\text{E}_\theta(\delta(X))^2 < \infty$. Suppose further that $g_{\theta_1}, g_{\theta_2}, ..., g_{\theta_k}$ are measurable functions such that $\text{E}_{\theta_0}g_{\theta_i}(X) = 0$ and $0 < \text{E}_{\theta_0}(g_{\theta_i}(X))^2 < \infty$ $\forall i$ and the $g_{\theta_i}$ are linearly independent as square integrable functions on $\mathcal{X}$ according to the probability measure $P_{\theta_0}$. Let $C_i(\theta) = \text{E}_{\theta_0}\delta(X)g_{\theta_i}(X)$ and define a $k \times k$ matrix

$$D(\theta) = (\text{E}_{\theta_0}g_{\theta_i}(X)g_{\theta_j}(X)) .$$

Then for $C(\theta) = (C_1(\theta), C_2(\theta), ..., C_k(\theta))$

$$\text{Var}_\theta\delta(X) \geq C'(\theta)'D(\theta)^{-1}C(\theta) .$$

And there is the multiparameter version of the C-R inequality.

**Theorem 89** (Cramer-Rao) Suppose that the model $\mathcal{P}$ is FI regular at $\theta_0 \in \Theta \subset \mathcal{R}^k$ and that $I(\theta_0)$ exists and is nonsingular. If $\text{E}_\theta\delta(X) = \int \delta(x)f_\theta(x)d\mu(x)$ can be differentiated wrt to each $\theta_i$ under the integral at $\theta_0$, i.e. if

$$\left. \frac{\partial}{\partial \theta_i} \text{E}_\theta \delta(X) \right|_{\theta = \theta_0} = \int \delta(x) \left. \frac{\partial}{\partial \theta_i} f_\theta(x) \right|_{\theta = \theta_0} d\mu(x) ,$$

then with

$$C(\theta_0) = \left( \left. \frac{\partial}{\partial \theta_1} \text{E}_\theta \delta(X) \right|_{\theta = \theta_0} , \left. \frac{\partial}{\partial \theta_2} \text{E}_\theta \delta(X) \right|_{\theta = \theta_0} , ..., \left. \frac{\partial}{\partial \theta_k} \text{E}_\theta \delta(X) \right|_{\theta = \theta_0} \right)'$$
\[ \text{Var}_{\theta_0} \delta (X) \geq C(\theta_0)' I(\theta_0)^{-1} C(\theta_0). \]

There are multiple \( \delta \) versions of this information inequality stuff. See Rao, pages 326+.

**Invariance and Equivariance**

Unbiasedness (of estimators) is one means of restricting a class of decision rules down to a subclass small enough that there is hope of finding a "best" rule in the subclass. Invariance/equivariance considerations is another. These notions have to do with symmetries in a decision problem. "Invariance" means that (some) things don't change. "Equivariance" means that (some) things change similarly or equally.

We'll do the case of Location Parameter Estimation first, and then turn to generalities about invariance/equivariance that apply more broadly (to both other estimation problems and to other types of decision problems entirely).

**Definition 54** When \( \mathcal{X} = \mathcal{R}^k \) and \( \Theta \subset \mathcal{R}^k \), to say that \( \theta \) is a location parameter means that \( X - \theta \) has the same distribution \( \forall \theta \).

**Definition 55** When \( \mathcal{X} = \mathcal{R}^k \) and \( \Theta \subset \mathcal{R}^1 \), to say that \( \theta \) is a 1-dimensional location parameter means that \( X - \theta \mathbf{1} \) (for \( \mathbf{1} \) a \( k \)-vector of 1's) has the same distribution \( \forall \theta \).

Note that for \( \theta \) a 1-dimensional location parameter, \( X \sim P_\theta \) means that \( (X + c \mathbf{1}) \sim P_{c+\theta} \). It therefore seems that a "reasonable" restriction on estimators \( \delta \) is that they should satisfy \( \delta(x + c \mathbf{1}) = \delta(x) + c \). How to formalize this?

Until further notice suppose that \( \Theta = \mathcal{A} = \mathcal{R}^1 \).

**Definition 56** \( L(\theta, \alpha) \) is called location invariant if \( L(\theta, \alpha) = \rho(\theta - \alpha) \). If \( \rho(z) \) is increasing in \( z \) for \( z > 0 \) and decreasing in \( z \) for \( z < 0 \), the problem is called a location estimation problem.

**Definition 57** A decision rule \( \delta \) is called location equivariant if \( \delta(x + c \mathbf{1}) = \delta(x) + c \).

**Theorem 90** If \( \theta \) is a 1-dimensional location parameter, \( L(\theta, \alpha) \) is location invariant and \( \delta \) is location equivariant, then \( R(\theta, \delta) \) is constant in \( \theta \). (That is, \( \delta \) is an equalizer rule.)

This theorem says that if I'm working on a location estimation problem with a location invariant loss function and restrict attention to location equivariant estimators, I can compare estimators by comparing numbers. In fact, as it turns out, there is a fairly explicit representation for the best location equivariant estimator in such cases.

**Definition 58** A function \( g \) is called location invariant if \( g(x + c \mathbf{1}) = g(x) \).
Lemma 91 Suppose that $\delta_0$ is location equivariant. Then $\delta_1$ is location equivariant iff $\exists$ a location invariant function $u$ such that $\delta_1 = \delta_0 + u$.

Lemma 92 A function $u$ is location invariant iff $u$ depends on $x$ only through $y(x) = (x_1 - x_k, x_2 - x_k, \ldots, x_{k-1} - x_k)$.

Theorem 93 Suppose that $L(\theta, a) = \rho(\theta - a)$ is location invariant and $\exists$ a location equivariant estimator $\delta_0$ for $\theta$ with finite risk. Let $\rho'(z) = \rho(-z)$. Suppose that for each $y$ $\exists$ a number $v^*(y)$ that minimizes

$$E_{\theta=0}[\rho'(\delta_0 - v)|Y = y]$$

over choices of $v$. Then

$$\delta^*(x) = \delta_0(x) - v^*(y(x))$$

is a best location equivariant estimator of $\theta$.

In the case of $L(\theta, a) = (\theta - a)^2$ the estimator of Theorem 93 is Pittman's estimator, which can for some situations be written out in a more explicit form.

Theorem 94 Suppose that $L(\theta, a) = (\theta - a)^2$ and the distribution of $X = (X_1, X_2, \ldots, X_k)$ has R-N derivative wrt to Lebesgue measure on $\mathcal{R}^k$.

$$f_\theta(x) = f(x_1 - \theta, x_2 - \theta, \ldots, x_k - \theta) = f(x - \theta1)$$

where $f$ has second moments. Then the best location equivariant estimator of $\theta$ is

$$\delta^*(x) = \frac{\int_{-\infty}^{\infty} \theta f(x - \theta1)d\theta}{\int_{-\infty}^{\infty} f(x - \theta1)d\theta}$$

if this is well-defined and has finite risk. (Notice that this is the formal generalized Bayes estimator vs Lebesgue measure.)

Lemma 95 For the case of $L(\theta, a) = (\theta - a)^2$ a best location equivariant estimator of $\theta$ must be unbiased for $\theta$.

Now cancel the assumption that $\Theta = A = \mathcal{R}^1$. The following is a more general story about invariance/equivariance.

Suppose that $\mathcal{P} = \{P_\theta\}$ is identifiable and (to avoid degeneracies) that

$$L(\theta, a) = L(\theta, a') \forall \theta \Rightarrow a = a' .$$

Invariance/equivariance theory is phrased in terms of groups of transformations

i) $\mathcal{X} \rightarrow \mathcal{X}'$ ,

ii) $\Theta \rightarrow \Theta$ , and

iii) $A \rightarrow A$ ,

where $\mathcal{X}$, $\Theta$, $A$ are subsets of $\mathcal{R}$, $\mathcal{R}$, and $\mathcal{R}$, respectively.
and how $L(\theta, a)$ and decision rules behave under these groups of transformations. The
natural starting place is with transformations on $\mathcal{X}$.

**Definition 59** For 1-1 transformations of $\mathcal{X}$ onto $\mathcal{X}$, $g_1$ and $g_2$, the composition of $g_1$ and
$g_2$ is another 1-1 onto transformation of $\mathcal{X}$ defined by $g_2 \circ g_1(x) = g_2(g_1(x)) \ \forall x \in \mathcal{X}$.

**Definition 60** For a 1-1 transformation of $\mathcal{X}$ onto $\mathcal{X}$, if $h$ is another 1-1 onto
transformation such that $h \circ g(x) = x \ \forall x \in \mathcal{X}$, $h$ is called the inverse of $g$ and denoted
as $h = g^{-1}$.

It is then an easy fact that not only is $g^{-1} \circ g(x) = x \ \forall x$, but $g \circ g^{-1}(x) = x \ \forall x$ as well.

**Definition 61** The identity transformation on $\mathcal{X}$ is defined by $e(x) = x \ \forall x$.

Another easy fact is then that $e \circ g = g \circ e = g$ for any 1-1 transformation of $\mathcal{X}$ onto $\mathcal{X}$, $g$.

**(Obvious) Result 96** If $\mathcal{G}$ is a class of 1-1 transformations of $\mathcal{X}$ onto $\mathcal{X}$ that is closed
under composition ($g_1, g_2 \in \mathcal{G} \Rightarrow g_1 \circ g_2 \in \mathcal{G}$) and inversion ($g \in \mathcal{G} \Rightarrow g^{-1} \in \mathcal{G}$), then
with the operation "$\circ$" (composition) $\mathcal{G}$ is a group in the usual mathematical sense.

A group such as that described in Result 96 will be called a transformation group. Such a
group may or may not be Abelian/commutative (one is if $g_1 \circ g_2 = g_2 \circ g_1 \ \forall g_1, g_2 \in \mathcal{G}$).

**Jargon/notation:** Any class of transformations $\mathcal{C}$ can be thought of as generating a
group. That is, $\mathcal{G}(\mathcal{C})$ consisting of all compositions of finite numbers of elements of $\mathcal{C}$
and their inverses is a group under the operation "$\circ$".

To get anywhere with this theory, we then need transformations on $\Theta$ that fit together
nicely with those on $\mathcal{X}$. At a minimum, one will need that the transformations on $\mathcal{X}$ don't
get one outside of $\mathcal{P}$.

**Definition 62** If $g$ is a 1-1 transformation of $\mathcal{X}$ onto $\mathcal{X}$ such that $\{P_{g(\theta)}(X)\}_{\theta \in \Theta} = \mathcal{P}$, then
the transformation $g$ is said to leave the model invariant. If each element of a group of
transformations $\mathcal{G}$ leaves $\mathcal{P}$ invariant, we'll say that the group leaves the model invariant.

One circumstance in which $\mathcal{G}$ leaves $\mathcal{P}$ invariant is where the model $\mathcal{P}$ can be thought of
as "generated by $\mathcal{G}$" in the following sense. Suppose that $U$ has some fixed probability
distribution $P_0$ and consider

$$\mathcal{P} = \{P_{g(U)} \mid g \in \mathcal{G}\}.$$ 

If a group of transformations on $\mathcal{X}$ (say $\mathcal{G}$) leaves $\mathcal{P}$ invariant, there is a natural
corresponding group of transformations on $\Theta$. That is, for $g \in \mathcal{G}$
\[ X \sim P_\theta \Rightarrow g(X) \sim P_{\theta'} \text{ for some } \theta' \in \Theta. \]

In fact, because of the identifiability assumption, there is only one such \( \theta' \). So, as a matter of notation, adopt the following.

**Definition 63** If \( G \) leaves the model \( \mathcal{P} \) invariant, for each \( g \in G \), define \( \tilde{g} : \Theta \rightarrow \Theta \) by
\[
\tilde{g}(\theta) = \text{ the element } \theta' \text{ of } \Theta \text{ such that } X \sim P_\theta \Rightarrow g(X) \sim P_{\theta'}.
\]

(According to the notation established in Definition 51, for \( G \) leaving \( \mathcal{P} \) invariant
\[
X \sim P_\theta \Rightarrow g(X) \sim P_{\tilde{g}(\theta)}.
\]

**Obvious** Result 97 For \( G \) a group of 1-1 transformations of \( \mathcal{X}' \) onto \( \mathcal{X}' \) leaving \( \mathcal{P} \) invariant, \( \tilde{G} = \{ \tilde{g} \mid g \in G \} \) with the operation "\( \circ \)" (composition) is a group of 1-1 transformations of \( \Theta \).

(I believe that the group \( \tilde{G} \) is the homomorphic image of \( G \) under the transformation "\( - \)" so that, for example, \( \tilde{g}^{-1} = \tilde{g}^{-1} \) and \( \tilde{g}_2 \circ \tilde{g}_1 = \tilde{g}_2 \circ \tilde{g}_1 \).)

We next need transformations on \( \mathcal{A} \) that fit nicely together with those on \( \mathcal{X} \) and \( \Theta \). If that's going to be possible, it would seem that the loss function will have to fit nicely with the transformations already considered.

**Definition 64** A loss function \( L(\theta, a) \) is said to be invariant under the group of transformations \( G \) provided that for every \( g \in G \) and \( a \in \mathcal{A} \), \( \exists \) a unique \( a' \in \mathcal{A} \) such that
\[
L(\theta, a) = L(\tilde{g}(\theta), a') \ \forall \theta.
\]

Now then, as another matter of jargon/notation adopt the following.

**Definition 65** If \( G \) leaves the model invariant and the loss \( L(\theta, a) \) is invariant, define \( \tilde{g} : \mathcal{A} \rightarrow \mathcal{A} \) by
\[
\tilde{g}(a) = \text{ the element } a' \text{ of } \mathcal{A} \text{ such that } L(\theta, a) = L(\tilde{g}(\theta), a') \ \forall \theta.
\]

(According to the notation established in Definition 53, for \( G \) leaving \( \mathcal{P} \) invariant and invariant loss \( L(\theta, a) \),
\[
L(\theta, a) = L(\tilde{g}(\theta), \tilde{g}(a)).
\]

**Obvious** Result 98 For \( G \) a group of 1-1 transformations of \( \mathcal{X}' \) onto \( \mathcal{X}' \) leaving \( \mathcal{P} \) invariant and invariant loss \( L(\theta, a) \), \( \tilde{G} = \{ \tilde{g} \mid g \in G \} \) with the operation "\( \circ \)" (composition) is a group of 1-1 transformations of \( \mathcal{A} \).

(I believe that the group \( \tilde{G} \) is the homomorphic image of \( G \) under the transformation "\( - \)" so that, for example, \( \tilde{g}^{-1} = \tilde{g}^{-1} \) and \( \tilde{g}_2 \circ \tilde{g}_1 = \tilde{g}_2 \circ \tilde{g}_1 \).)
So finally we come to the notion of equivariance for decision functions. The whole business is slightly subtle for behavioral rules (see pages 150-151 of Ferguson) so for this elementary introduction, we’ll confine attention to nonrandomized rules.

**Definition 66** In an invariant decision problem (one where $\mathcal{G}$ leaves $\mathcal{P}$ invariant and $L(\theta, a)$ is invariant) a nonrandomized decision rule $\delta: \mathcal{X} \rightarrow \mathcal{A}$ is said to be equivariant provided

$$\delta(g(x)) = \tilde{g}(\delta(x)) \quad \forall x \in \mathcal{X} \text{ and } g \in \mathcal{G}.$$  

(One could make a definition here involving a single $g$ or perhaps a 3 element group $\mathcal{G} = \{g, g^{-1}, e\}$ and talk about equivariance under $g$ instead of under a group. I'll not bother.)

The basic elementary theorem about invariance/equivariance is the generalization of Theorem 90 (that cuts down the size of the comparison problem that one faces when comparing risk functions).

**Theorem 99** If $\mathcal{G}$ leaves $\mathcal{P}$ invariant, $L(\theta, a)$ is invariant and $\delta$ is an equivariant nonrandomized decision rule, then  

$$R(\theta, \delta) = R(\tilde{g}(\theta), \tilde{g}(\delta)) \quad \forall \theta \in \Theta \text{ and } g \in \mathcal{G}.$$  

**Definition 67** In an invariant decision problem, for $\theta \in \Theta$, $O_\theta = \{\theta' \in \Theta | \theta' = \tilde{g}(\theta) \text{ for some } g \in \mathcal{G}\}$ is called the orbit in $\Theta$ containing $\theta$.

The set of orbits $\{O_\theta\}_{\theta \in \Theta}$ partitions $\Theta$. The distinct orbits are equivalence classes.

Theorem 99 says that an equivariant nonrandomized decision rule in an invariant decision problem has a risk function that is constant on orbits. When looking for good equivariant rules, the problem of comparing risk functions over $\Theta$ is reduced somewhat to comparing risk functions on orbits. Of course, that comparison is especially easy when there is only one orbit.

**Definition 68** A group of 1-1 transformations of a space onto itself is said to be transitive if for any two points in the space there is a transformation in the group taking the first point into the second.

**Corollary 100** If $\mathcal{G}$ leaves $\mathcal{P}$ invariant, $L(\theta, a)$ is invariant and $\mathcal{G}$ is transitive over $\Theta$, then every equivariant nonrandomized decision rule has constant risk.

**Definition 69** In an invariant decision problem, a decision rule $\delta$ is said to be a best (nonrandomized) equivariant (or MRE ... minimum risk equivariant) decision rule if it is equivariant and for any other nonrandomized equivariant rule $\delta'$
The import of Theorem 99 is that one needs only check for this risk inequality for one \( \theta \) from each orbit.

**Optimal Hypothesis Testing (Simple vs Simple, Composite Hypotheses, UMP Tests, Optimal Confidence Regions and UMPU Tests)**

We now consider the classical hypothesis testing problem. Here we treat

\[
\Theta = \Theta_0 \cup \Theta_1 \text{ where } \Theta_0 \cap \Theta_1 = \emptyset
\]

and the hypotheses \( H_i; \theta \in \Theta_i \ i = 0, 1 \). As is standard, we will call \( H_0 \) the "null hypothesis" and \( H_1 \) the "alternative hypothesis." \( \Theta_i \) with one element is called "simple" and with more than one element is called "composite."

Deciding between \( H_0 \) and \( H_1 \) on the basis of data can be thought of in terms of a two-action decision problem. That is, with \( A = \{0, 1\} \) one is using \( X \) to decide in favor of one of the two hypotheses. Typically one wants a loss structure where correct decisions are preferable to incorrect ones. An obvious (and popular) possibility is 0-1 loss,

\[
L(\theta, a) = \mathbb{I}[\theta \in \Theta_0, a = 1] + \mathbb{I}[\theta \in \Theta_1, a = 0]
\]

While it is perfectly sensible to simply treat hypothesis testing as a two-action decision problem, being older than formal statistical decision theory, it has its own peculiar/specialized set of jargon and emphases that we will have to recognize. These include the following.

**Definition 70** A nonrandomized decision function \( \delta : \mathcal{X} \to \{0, 1\} \) is called a hypothesis test.

**Definition 71** The rejection region for a nonrandomized test \( \delta \) is \( \{x|\delta(x) = 1\} \).

**Definition 72** A type I error in a testing problem is taking action 1 when \( \theta \in \Theta_0 \). A type II error in a testing problem is taking action 0 when \( \theta \in \Theta_1 \).

Note that in a two-action decision problem, a behavioral decision rule \( \phi_x \) is for each \( x \) a distribution over \( \{0, 1\} \), which is clearly characterized by \( \phi_x(\{1\}) \in [0, 1] \). This notion can be expressed in other terms.

**Definition 73** A randomized hypothesis test is a function \( \phi : \mathcal{X} \to [0, 1] \) with the interpretation that if \( \phi(x) = \pi \), one rejects \( H_0 \) with probability \( \pi \).
Note that for 0-1 loss,

\[ R(\theta, \phi) = E_\theta I[a = 1] = E_\theta \phi(X) \]

for \( \theta \in \Theta_0 \),

\[ R(\theta, \phi) = E_\theta I[a = 0] = E_\theta (1 - \phi(X)) = 1 - E_\theta \phi(X) \]

for \( \theta \in \Theta_1 \),

the type I error probability, while

the type II error probability. Most of testing theory is phrased in terms of \( E_\theta \phi(X) \) that appears in both of these expressions.

**Definition 74** The power function of the (possibly randomized) test \( \phi \) is

\[ \beta_{\phi}(\theta) = E_\theta \phi(X) = \int \phi(x) dP_\theta(x) . \]

**Definition 75** The size or level of a test \( \phi \) is \( \sup_{\theta \in \Theta_0} \beta_{\phi}(\theta) \).

This is the maximum probability of a type I error. Standard testing theory proceeds (somewhat asymmetrically) by setting a maximum allowable size and then (subject to that restriction) trying to maximize the power (uniformly) over \( \Theta_1 \). (That is, subject to a size limitation, standard theory seek to minimize type II error probabilities.)

Recall that Result 61 says that if \( T \) is sufficient for \( \theta \), the set of behavioral decision rules that are functions of \( T \) is essentially complete. Arguing that fact in the present case is very easy/straightforward. For any randomized test \( \phi \), consider

\[ \phi^*(x) = E[\phi|T](x) , \]

which is clearly a \( B(T) \) measurable function taking values in \([0, 1]\), i.e. is a hypothesis test based on \( T \). Note that

\[ \beta_{\phi^*}(\theta) = E_\theta \phi^*(X) = E_\theta E[\phi|T] = E_\theta \phi(X) = \beta_{\phi}(\theta) , \]

and \( \phi^* \) and \( \phi \) have the same power functions and are thus equivalent.

**Simple vs Simple Testing Problems**

Consider first testing problems where \( \Theta_0 = \{\theta_0\} \) and \( \Theta_1 = \{\theta_1\} \) (with 0-1 loss). In this finite \( \Theta \) decision problem, we can for each test \( \phi \) define the risk vector

\[ (R(\theta_0, \phi), R(\theta_1, \phi)) = (\beta_{\phi}(\theta_0), 1 - \beta_{\phi}(\theta_1)) \]

and make the standard two-dimensional plot of risk vectors. A completely equivalent device is to plot instead the points

\[ (\beta_{\phi}(\theta_0), \beta_{\phi}(\theta_1)) . \]
Lemma 101 For a simple versus simple testing problem, the risk set $S$ has the following properties:

i) $S$ is convex,

ii) $S$ is symmetric wrt the point $(\frac{1}{2}, \frac{1}{2})$,

iii) $(0, 1) \in S$ and $(1, 0) \in S$, and

iv) $S$ is closed.

Lemma 101' (see page 77 of TSH) For a simple versus simple testing problem, the set $V = \{ (\beta_\phi(\theta_0), \beta_\phi(\theta_1)) | \phi \text{ is a test} \}$ has the following properties:

i) $V$ is convex,

ii) $V$ is symmetric wrt the point $(\frac{1}{2}, \frac{1}{2})$,

iii) $(0, 0) \in V$ and $(1, 1) \in V$, and

iv) $V$ is closed.

As indicated earlier, the usual game in testing is to fix $\alpha \geq 0$ and look for a test with size $\alpha$ and maximum power on $\Theta_1$.

Definition 76 For testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, a test is called most powerful of size $\alpha$ if $\beta_\phi(\theta_0) = \alpha$ and for any other test $\phi^*$ with $\beta_{\phi^*}(\theta_0) \leq \alpha$, $\beta_{\phi^*}(\theta_1) \leq \beta_\phi(\theta_1)$.

The fundamental result of all of testing theory is the Neyman-Pearson Lemma.

Theorem 102 (The Neyman-Pearson Lemma) Suppose that $\mathcal{P} \ll \mu$, a $\sigma$-finite measure and let $f_i = \frac{dP_i}{d\mu}$.

i) (Sufficiency) Any test of the form

$$\phi(x) = \begin{cases} 1 & \text{if } f_1(x) > kf_0(x) \\ \nu(x) & \text{if } f_1(x) = kf_0(x) \\ 0 & \text{if } f_1(x) < kf_0(x) \end{cases} \quad (\star)$$

for some $k \in [0, \infty)$ and $\nu : X \to [0, 1]$ is most powerful of its size. Further, corresponding to the $k = \infty$ case, the test

$$\phi(x) = \begin{cases} 1 & \text{if } f_0(x) = 0 \\ 0 & \text{otherwise} \end{cases} \quad (\star\star)$$

is most powerful of size $\alpha = 0$.

ii) (Existence) For every $\alpha \in [0, 1]$, there exists a test of form $(\star)$ with $\nu(x) = \nu$ (a constant in $[0, 1]$) or a test of form $(\star\star)$ with $\beta_\phi(\theta_0) = \alpha$.

iii) (Uniqueness) If $\phi$ is a most powerful test of size $\alpha$, then it is of form $(\star)$ or form $(\star\star)$ except possibly for $x \in N$ with $P_{\theta_0}(N) = P_{\theta_1}(N) = 0$.

Note that for any two distributions there is always a dominating $\sigma$-finite measure ($\mu = P_{\theta_0} + P_{\theta_1}$ will do the job). Further, it's easy to verify that the tests prescribed by this theorem do not depend upon the dominating measure.
Note also that one could take a more explicitly decision theoretic slant on this whole business. Thinking of testing as a two-decision problem with 0-1 loss and prior
distribution \( G \), abbreviating \( G(\{\theta_0\}) = g_0 \) and \( G(\{\theta_1\}) = g_1 \), it's easy to see that a Bayesian is a Neyman-Pearson tester with \( k = g_0/g_1 \) and that a Neyman-Pearson tester is a closet Bayesian with \( g_0 = k/(k + 1) \) and \( g_1 = 1/(1 + k) \).

There is a kind of "finite \( \Theta_0 \) vs simple \( \Theta_1 \)" version of the N-P Lemma. See TSH page 96.

A final small fact about simple versus simple testing (that turns out to be useful in some composite versus composite situations) is next.

**Lemma 103** Suppose that \( \alpha \in (0, 1) \) and \( \phi \) is most powerful of size \( \alpha \). Then
\[
\beta_\phi(\theta_1) \geq \alpha \quad \text{and the inequality is strict unless} \quad f_0 = f_1 \quad \text{a.e.} \mu.
\]

**UMP Tests for Composite vs Composite Problems**

Now consider cases where \( \Theta_0 \) and \( \Theta_1 \) need not be singletons.

**Definition 77** \( \phi \) is UMP (uniformly most powerful) of size \( \alpha \) for testing \( H_0 : \theta \in \Theta_0 \) vs \( H_1 : \theta \in \Theta_1 \) provided \( \phi \) is of size \( \alpha \) (i.e. \( \sup \beta_\phi(\theta) = \alpha \)) and for any other \( \phi' \) of size
no more than \( \alpha \), \( \beta_\phi'(\theta) \leq \beta_\phi(\theta) \quad \forall \theta \in \Theta_1 \).

Problems for which one can find UMP tests are not all that many. Known results cover
situations where \( \Theta \subset \mathcal{R}^1 \) involving one-sided hypotheses in monotone likelihood ratio
families, a few special cases that yield to a kind of "least favorable distribution" argument
and an initially somewhat odd-looking two-sided problem in one parameter exponential
families.

**Definition 78** Suppose \( \Theta \subset \mathcal{R}^1 \) and \( \mathcal{P} \ll \mu \), a \( \sigma \)-finite measure on \((\mathcal{X}, \mathcal{B})\). Suppose that
\( T : \mathcal{X} \rightarrow \mathcal{R}^1 \). The family of R-N derivatives \( \{f_\theta\} \) is said to have the monotone likelihood ratio (MLR) property in \( T \) if for every \( \theta_1 < \theta_2 \) belonging to \( \Theta \), the ratio \( f_{\theta_2}(x)/f_{\theta_1}(x) \) is
nondecreasing in \( T(x) \) on \( \{x|f_{\theta_1}(x) + f_{\theta_2}(x) > 0\} \). (\( \mathcal{P} \) is also said to have the MLR
property.)

**Theorem 104** Suppose that \( \{f_\theta\} \) has MLR in \( T \). Then for testing \( H_0 : \theta \leq \theta_0 \) vs
\( H_1 : \theta > \theta_0 \), for each \( \alpha \in (0, 1] \exists \) a test of the form
\[
\phi_\alpha(x) = \begin{cases} 
1 & \text{if } T(x) > c \\
\nu & \text{if } T(x) = c \\
0 & \text{if } T(x) < c
\end{cases}
\]
for \( c \in [\ -\infty, \infty) \) chosen to satisfy \( \beta_\phi(\theta_0) = \alpha \). Such a test
i) has strictly increasing power function,
ii) is UMP size \( \alpha \) for these hypotheses, and
iii) uniformly minimizes \( \beta_\phi(\theta) \) for \( \theta < \theta_0 \) for tests with \( \beta_\phi(\theta_0) = \alpha \).
Theorem 105 Suppose that \( \{f_\theta\} \) has MLR in \( T \) and for state space \( \Theta \) and action space \( \mathcal{A} = \{0, 1\} \) the loss function \( L(\theta, a) \) satisfies
\[
L(\theta, 1) - L(\theta, 0) < 0 \text{ for } \theta > \theta_0, \text{ and}
\]
\[
L(\theta, 1) - L(\theta, 0) > 0 \text{ for } \theta < \theta_0.
\]
Then the class of size \( \alpha \) tests (\( \alpha \in (0, 1) \)) identified in Theorem 86 is essentially complete in the class of tests with size in \( (0, 1) \).

One two-sided problem a single real parameter for which it possible to find UMP tests is covered by the next theorem (that is proved in Section 4.3.4 of Schervish using the generalization of the N-P lemma alluded to earlier).

Theorem 106 Suppose \( \Theta \subset \mathcal{R}^1 \) and \( \mathcal{P} \ll \mu \), a \( \sigma \)-finite measure on \((\mathcal{X}, \mathcal{B})\) and
\[
f_\theta(x) = \exp(\alpha(\theta) + \theta T(x)) \text{ a.e. } \mu \forall \theta.
\]
Suppose further that \( \theta_1 < \theta_2 \) belong to \( \Theta \). A test of the form
\[
\phi(x) = \begin{cases} 
1 & \text{if } c_1 < T(x) < c_2 \\
\nu_i & \text{if } T(x) = c_i \text{ } i = 1, 2 \\
0 & \text{if } T(x) < c_1 \text{ or } T(x) > c_2
\end{cases}
\]
for \( c_1 \leq c_2 \) minimizes \( \beta_{\phi}(\theta) \forall \theta < \theta_1 \text{ and } \forall \theta > \theta_2 \) over choices of tests \( \phi' \) with \( \beta_{\phi'}(\theta_1) = \beta_{\phi}(\theta_1) \) and \( \beta_{\phi'}(\theta_2) = \beta_{\phi}(\theta_2) \). Further, if \( \beta_{\phi}(\theta_1) = \beta_{\phi}(\theta_2) = \alpha \), then \( \phi \) is UMP level \( \alpha \) for testing \( H_0 : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2 \) vs \( H_1 : \theta_1 < \theta < \theta_2 \).

As far as finding UMP tests for parameter spaces other than subsets of \( \mathcal{R}^1 \) is concerned, about the only "general" technique known is one discussed in Section 3.8 of TSH. The method is not widely applicable, but does yield a couple of important results. It is based on an "equalizer rule"/"least favorable distribution" type argument.

The idea is that for a few composite vs composite problems, we can for each \( \theta_1 \in \Theta_1 \) guess at a distribution \( \Lambda \) over \( \Theta_0 \) so that with
\[
f_\Lambda(x) = \int_{\Theta_0} f_\theta(x) d\Lambda(\theta)
\]
simple versus simple testing of \( H_0 : f_\Lambda \text{ vs } H_1 : f_{\theta_1} \) produces a size \( \alpha \) N-P test whose power over \( \Theta_0 \) is bounded by \( \alpha \). (Notice by the way, that this is a Bayes test in the composite vs composite problem against a prior that is a mixture of \( \Lambda \) and a point mass at \( \theta_1 \).) Anyhow, when that test is independent of \( \theta_1 \) we get UMP size \( \alpha \) for the original problem.
**Definition 79** For testing $H_0 : X \sim f_{\Lambda}$ vs $H_1 : X \sim f_{\theta_1}$, $\Lambda$ is called least favorable at level $\alpha$ if for $\phi_{\Lambda}$ the N-P size $\alpha$ test associated with $\Lambda$, 

$$
\beta_{\phi_{\Lambda}}(\theta_1) < \beta_{\phi_{\theta_1}}(\theta_1)
$$

for any other distribution over $\Theta_0, \Lambda$.

**Theorem 107** If $\phi$ is of size $\alpha$ for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta = \theta_1$, then it is of size no more than $\alpha$ for testing $H_0 : X \sim f_{\Lambda}$ vs $H_1 : X \sim f_{\theta_1}$. Further, if $\phi$ is most powerful of size $\alpha$ for testing $H_0 : X \sim f_{\Lambda}$ vs $H_1 : X \sim f_{\theta_1}$ and is also size $\alpha$ for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta = \theta_1$, then it is a MP size $\alpha$ test for this set of hypotheses and $\Lambda$ is least favorable at level $\alpha$.

Clearly, if in the context of Theorem 107, the same $\phi$ is most powerful size $\alpha$ for each $\theta_1$, that test is then UMP size $\alpha$ for the hypotheses $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$.

**Confidence Sets Derived From Tests (and Optimal Ones Derived From UMP Tests)**

**Definition 80** Suppose that for each $x \in \mathcal{X}$, $S(x)$ is a (measurable) subset of $\Theta$. The family $\{S(x) | x \in \mathcal{X}\}$ is called a $\gamma$ level family of confidence sets for $\theta$ provided 

$$
P_{\theta}[\theta \in S(x)] \geq \gamma \ \forall \theta \in \Theta .
$$

It is possible (and quite standard) to use tests to produce confidence procedures. That goes as follows. For each $\theta_0 \in \Theta$, let $\phi_{\theta_0}$ be a size $\alpha$ test of $H_0 : \theta = \theta_0$ vs $H_1 : \theta \in \Theta_1$ (where $\theta_0 \in \Theta - \Theta_1$) and define 

$$
S_{\phi}(x) = \{\theta | \phi_{\theta}(x) < 1\}
$$

($S_{\phi}(x)$ is the set of $\theta$'s for which there is positive conditional probability of accepting $H_0$ having observed $X = x$.) $\{S_{\phi}(x) | x \in \mathcal{X}\}$ is a level $1 - \alpha$ family of confidence sets for $\theta$. This rigmarole works perfectly generally and inversion of tests of point null hypotheses is a time-honored means of constructing confidence procedures. In the context of UMP testing, the construction produces confidence sets with an optimality property.

**Theorem 108** Suppose that for each $\theta_0$ , $K(\theta_0) \subset \Theta$ with $\theta_0 \notin K(\theta_0)$ and $\phi_{\theta_0}$ is a nonrandomized UMP size $\alpha$ test of $H_0 : \theta = \theta_0$ vs $H_1 : \theta \in K(\theta_0)$. Then if $S_{\phi}(x)$ is the confidence procedure derived from the tests $\phi_{\theta_0}$ and $S(x)$ is any other level $(1 - \alpha)$ confidence procedure 

$$
P_{\theta}[\theta_0 \in S(X)] \geq P_{\theta}[\theta_0 \in S_{\phi}(X)] \ \forall \theta \in K(\theta_0).
$$

(That is, the $S_{\phi}$ construction produces confidence procedures that minimize the probabilities of covering incorrect values.)
**UMPU Tests for Composite versus Composite Problems**

Cases in which one can find UMP tests are not all that many. This should not be terribly surprising given our experience to date in Stat 643. We are used to not being able to find uniformly best statistical procedures. In estimation problems we found that by restricting the class of estimators under consideration (e.g. to the unbiased or equivariant estimators) we could sometimes find uniformly best rules in the restricted class. Something similar is sometimes possible in composite versus composite testing problems.

**Definition 81** A size $\alpha$ test of $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ is unbiased provided

$$\beta_\phi(\theta) \geq \alpha \quad \forall \theta \in \Theta_1 .$$

**Definition 82** An unbiased size $\alpha$ test of $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ is UMPU (uniformly most powerful unbiased) level $\alpha$ if for any other size $\alpha$ unbiased test $\phi'$

$$\beta_\phi(\theta) \geq \beta_{\phi'}(\theta) \quad \forall \theta \in \Theta_1 .$$

The conditions under which people can identify UMPU tests require that $\Theta$ be a topological space. (One needs notions of interior, boundary, continuity of functions, etc.) So in this section make that assumption.

**Definition 83** The boundary set is $\Theta_B = \Theta_0 \cap \Theta_1$.

**Lemma 109** If $\beta_\phi(\theta)$ is continuous and $\phi$ is unbiased of size $\alpha$, then $\beta_\phi(\theta) = \alpha \quad \forall \theta \in \Theta_B$.

**Definition 84** A test is said to be $\alpha$-similar on the boundary if

$$\beta_\phi(\theta) = \alpha \quad \forall \theta \in \Theta_B .$$

When looking for UMPU tests it often suffices to restrict attention to tests $\alpha$-similar on the boundary.

**Lemma 110** Suppose that $\beta_\phi(\theta)$ is continuous for every $\phi$. If $\phi$ is UMP among tests $\alpha$-similar on the boundary and has level $\alpha$, then it is UMPU level $\alpha$.

This lemma and Theorem 106 (or the generalized Neyman-Pearson lemma that stands behind it) can be used to prove a kind of "reciprocal" of Theorem 106 for one parameter exponential families.
**Theorem 111** Suppose $\Theta \subset \mathcal{R}^1$ and $\mathcal{P} \ll \mu$, a $\sigma$-finite measure on $(\mathcal{X}, \mathcal{B})$, and
\[
f_\theta(x) = \exp(a(\theta) + \theta T(x)) \quad \text{a.e. } \mu \forall \theta.
\]
Suppose further that $\theta_1 < \theta_2$ belong to $\Theta$. A test of the form
\[
\phi(x) = \begin{cases} 
  1 & \text{if } T(x) < c_1 \text{ or } T(x) > c_2 \\
  \nu_i & \text{if } T(x) = c_i \quad i = 1, 2 \\
  0 & \text{if } c_1 < T(x) < c_2
\end{cases}
\]
for $c_1 \leq c_2$ maximizes $\beta_\phi(\theta) \forall \theta < \theta_1$ and $\forall \theta > \theta_2$ over choices of tests $\phi'$ with $\beta_{\phi'}(\theta_1) = \beta_\phi(\theta_1)$ and $\beta_{\phi'}(\theta_2) = \beta_\phi(\theta_2)$. Further, if $\beta_\phi(\theta_1) = \beta_\phi(\theta_2) = \alpha$, then $\phi$ is UMPU level $\alpha$ for testing $H_0 : \theta \in [\theta_1, \theta_2]$ vs $H_1 : \theta \notin [\theta_1, \theta_2]$.

Theorem 111 deals with an interval null hypothesis. Dealing with a point null hypothesis turns out to require slightly different arguments.

**Theorem 112** Suppose that for $\Theta$ an open interval in $\mathcal{R}^1$, $\mathcal{P} \ll \mu$, a $\sigma$-finite measure on $(\mathcal{X}, \mathcal{B})$ and
\[
f_\theta(x) = \exp(a(\theta) + \theta T(x)) \quad \text{a.e. } \mu \forall \theta.
\]
Suppose that $\phi$ is a test of the form given in Theorem 93 and that
\[
\beta_\phi(\theta_0) = \alpha
\]
and
\[
\beta_\phi'(\theta_0) = 0.
\]
Then $\phi$ is UMPU size $\alpha$ for testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

Theorems 111 and 112 are about cases where $\Theta \subset \mathcal{R}^1$ (and, of course, limited to one exponential families). It is possible to find some UMPU tests for hypotheses with "$(k - 1)$-dimensional" boundaries in "$k$-dimensional" $\Theta$ problems using a methodology exploiting "Neyman structure" in a statistic $T$. I may write up some notes on this methodology (and again I may not). It is discussed in Section 4.5 of Schervish. A general kind of result that comes from the arguments is next.
**Theorem 113** Suppose that $X = (X_1, X_2, ..., X_k)$ takes values in $\mathcal{R}^k$ and the distributions $P_\eta$ have densities with respect to a $\sigma$-finite measure $\mu$ on $\mathcal{R}^k$ of the form

$$f_\eta(x) = K(\eta)\exp\sum_{i=1}^{k} \eta_i x_i.$$ 

Suppose further that the natural parameter space $\Gamma$ contains an open rectangle in $\mathcal{R}^k$ and let $U = (X_2, ..., X_k)$. Suppose $\alpha \in (0, 1)$.

i) For $H_0 : \eta_1 \leq \eta_1^*$ vs $H_1 : \eta_1 > \eta_1^*$

or for $H_0 : \eta_1 \leq \eta_1^*$ or $\eta_1 \geq \eta_1^{**}$ vs $H_1 : \eta_1 \in (\eta_1^*, \eta_1^{**})$,

for each $u \not\in \rho_a$ a conditional on $U = u$ UMP level $\alpha$ test $\phi'_a(x_1)$, and the test

$$\psi(x) = \phi'_a(x_1)$$

is UMPU level $\alpha$.

ii) For $H_0 : \eta_1 \in [\eta_1^*, \eta_1^{**}]$ vs $H_1 : \eta_1 \not\in [\eta_1^*, \eta_1^{**}]$

or for $H_0 : \eta_1 = \eta_1^*$ vs $H_1 : \eta_1 \neq \eta_1^*$

for each $u \not\in \rho_a$ a conditional on $U = u$ UMPU level $\alpha$ test $\phi'_a(x_1)$, and the test

$$\phi(x) = \phi'_a(x_1)$$

is UMPU level $\alpha$.

This theorem says that if one thinks of $\eta = (\eta_1, \eta_2, ..., \eta_k)$ as containing the parameter of interest $\eta_1$ and "nuisance parameters" $\eta_2, ..., \eta_k$, at least in some exponential families it is possible to find "optimal" tests about the parameter of interest.