Math 37600 PR - (19364) - Lectures 02

Ethan Akin
Email: eakin@ccny.cuny.edu

Fall, 2024

Sec. 8.1: Most Powerful Simple Tests

In this section we consider a simple test for a family $f(x;\theta)$ of pdf's or pmf's. That is, the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$.

The samples $\mathbf{X} = X_1, \ldots, X_n$ lie in $\mathcal{S}^n$ where $\mathcal{S}$ is the support of the family of rv's. We are assuming that the support does not depend on θ. A critical region is a subset of $\mathcal{S}^n$ so that we reject H_0 when $\mathbf{X} \in C$ and accept it when $\mathbf{X} \in C^c$.

For a critical region C, the size or significance $\alpha = P_{\theta_0}(C)$, the probability of a Type I error, and the power of the test is $\gamma_C(\theta_1) = P_{\theta_1}(C)$, the probability of correctly rejecting the null hypothesis. That is, that the power is 1 minus the probability of a Type II error.

HMC Definition 8.1.1: A critical region C of size α is a *best critical region of size* α if, whenever A is a critical region of size α,

$$P_{\theta_1}(C) \geq P_{\theta_1}(A).$$

That is, the power of the region C is the maximum power possible for a critical region of size α.

We use the likelihood function $L(\theta)$

$$L(\theta; \mathbf{x}) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

Since θ_0 and θ_1 are known values, we can compute

$$\Lambda(\theta_0, \theta_1; \mathbf{x}) = \frac{L(\theta_0; \mathbf{x})}{L(\theta_1; \mathbf{x})}$$

and use it to define a critical region.

Neyman-Pearson Theorem

HMC Theorem 8.1.1: (**Neyman-Pearson Theorem**) If a subset C of the sample space satisfies:

- $P_{\theta_0}(C) = \alpha$;
- $\Lambda(\theta_0, \theta_1; \mathbf{x}) \leq k$ for all $\mathbf{x} \in C$;
- $\Lambda(\theta_0, \theta_1; \mathbf{x}) \geq k$ for all $\mathbf{x} \in C^c$;

for some positive constant k, then C is a best critical region of size α.

Proof: We will temporarily write $L(\theta; A)$ for $\int_A L(\theta; \mathbf{x}) d\mathbf{x}$. This is just alternate notation for $P_\theta(A)$. So, for example,

$$L(\theta; C) = L(\theta; A^c \cap C) + L(\theta; A \cap C).$$
$$L(\theta; A) = L(\theta; A \cap C^c) + L(\theta; A \cap C),$$

Subtracting and cancelling the common terms we see that for any θ.

$$L(\theta; C) - L(\theta; A) \;=\; L(\theta; A^c \cap C) - L(\theta; A \cap C^c)$$

If A is any other critical region of size α we want to show

$$L(\theta_1; C) - L(\theta_1; A) \geq 0.$$

Because of the assumptions about C and C^c, we have $L(\theta_1; A^c \cap C) \geq k^{-1} L(\theta_0; A^c \cap C)$ because for $\mathbf{x} \in C$, $L(\theta_0; \mathbf{x}) \leq k L(\theta_1; \mathbf{x})$.

Also, $-L(\theta_1; A \cap C^c) \geq -k^{-1} L(\theta_0; A \cap C^c)$ because for $\mathbf{x} \in C^c$, $L(\theta_0; \mathbf{x}) \geq k L(\theta_1; \mathbf{x})$

Thus, we have:

$$L(\theta_1; A^c \cap C) \geq k^{-1} L(\theta_0; A^c \cap C),$$
$$-L(\theta_1; A \cap C^c) \geq -k^{-1} L(\theta_0; A \cap C^c).$$

$$\begin{aligned} L(\theta_1; C) - L(\theta_1; A) \; &= \\ L(\theta_1; A^c \cap C) - L(\theta_1; A \cap C^c) \; &\geq \\ k^{-1}[L(\theta_0; A^c \cap C) - L(\theta_0; A \cap C^c)] \\ = \; k^{-1}[L(\theta_0; C) - L(\theta_0; A)]. \end{aligned}$$

Because C and A are critical regions of size α, $L(\theta_0; C) = L(\theta_0; A) = \alpha$.

So $L(\theta_1; C) - L(\theta_1; A) \geq 0$.

The Neyman-Pearson Theorem works with the same proof with C and A randomized tests. From that we get

HMC Corollary 8.1.1: If C is a best critical region of size α for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, then $P_{\theta_1}(C) \geq \alpha$. That is, the power is greater than or equal to the size.

Proof: We compare C with the trivial randomized test which uses $Y \sim Bern(\alpha)$, and we use $A = \{Y = 1\}$. So the power equals the size for this test because $P_\theta(A) = \alpha$ for all θ.

By the Neyman-Pearson Theorem $P_{\theta_1}(C) \geq P_{\theta_1}(A) = \alpha$.

Here θ_0 and θ_1 are known parameter values. So $\Lambda(\theta_0, \theta_1; \mathbf{x})$ is a statistic and so for any k we can define the critical region $C_k = \{\Lambda(\theta_0, \theta_1; \mathbf{x}) \leq k\}$. The Neyman-Pearson Theorem then says that C_k is a best critical region of its size

If we start with size α, then we choose k so that $P_{\theta_0}(C_k) = \alpha$.

We can often express Λ in terms of a single statistic $T(\mathbf{X})$ separate from θ_0 and θ_1. In the next section we will see how this is done in general.

Example 8.1.2 and 8.2.3: Let $X \sim \mathcal{N}(\theta, 1)$ so that $f(x;\theta) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(x-\theta)^2}{2})$. We test for θ_0 against $\theta_1 > \theta_0$. Notice first that

$$(-\frac{\sum_{i=1}^n (x_i - \theta_0)^2}{2}) - (-\frac{\sum_{i=1}^n (x_i - \theta_1)^2}{2}) =$$

$$-(\theta_1 - \theta_0)\sum_{i=1}^n x_i + \frac{n}{2}(\theta_1^2 - \theta_0^2).$$

So

$$\Lambda(\theta_0, \theta_1; \mathbf{X}) = exp(-(\theta_1 - \theta_0)\sum_{i=1}^n X_i + \frac{n}{2}(\theta_1^2 - \theta_0^2))$$

$$= exp(n(\theta_1 - \theta_0)[-\bar{X} + \frac{\theta_1 + \theta_0}{2}]),$$

and we use

$$C_k = \{\Lambda(\theta_0, \theta_1; \mathbf{X}) \leq k\} = \{\bar{X} \geq \frac{\theta_1 + \theta_0}{2} - \frac{\ln k}{n(\theta_1 - \theta_0)}\}.$$

For any θ, $\sqrt{n}(\bar{X} - \theta) \sim \mathcal{N}(0,1)$. So if $1 - \Phi(z_\alpha) = \alpha$, C_k has size α with $\theta = \theta_0$ when $C_k = \{\bar{X} \geq \theta_0 + \frac{z_\alpha}{\sqrt{n}}\}$. We can solve this for k.

$$k = exp[(\theta_1 - \theta_0)[\frac{(\theta_1 - \theta_0)}{2} - \frac{z_\alpha}{\sqrt{n}}].$$

However, we don't need to determine k.
If $\theta = \theta_1$, then $X - \theta_1 \sim \mathcal{N}(0,1)$ and so
$\sqrt{n}(\bar{X} - \theta_1) \sim \mathcal{N}(0,1)$. The power is given by

$$\begin{aligned}\gamma_{C_k}(\theta_1) = P_{\theta_1}(C_k) &= P(\sqrt{n}(\bar{X} - \theta_1) \geq z_\alpha - \sqrt{n}(\theta_1 - \theta_0)) \\ &= 1 - \Phi(z_\alpha - \sqrt{n}(\theta_1 - \theta_0)) \geq 1 - \Phi(z_\alpha) = \alpha.\end{aligned}$$

As HMC remark, although we have been assuming that the pdf's or pmf's come from a parameterized family, this need not so For the Neyman-Pearson result. All that is needed is that the two distributions of the two simple hypotheses have the same range.

Example 8.1.3: Here the authors test the Poiss(1) pmf $H_0 : f_0(x) = e^{-1}/x!$ against the Geom($\frac{1}{2}$) $H_1 : (\frac{1}{2})^{x+1}$ for $x = 0, 1, \ldots$.

$$\Lambda(\mathbf{X}) = (e^{-n}/x_1! \ldots x_n!) \div ((1/2)^n (1/2)^{x_1 + \ldots x_n}) = \frac{(2e^{-1})^n 2^{\sum x_i}}{\prod x_i!}.$$

For any k, $C_k = \{\Lambda \leq k\}$ defines a best critical region for $\alpha = P_0(C_k)$. However, as the book illustrates, computing what the set C is can be complicated enough that this is really only of theoretical interest.

> *In theory, theory and practice are the same thing, but in practice, they are really not.*

Remark 8.1.2: Recall that the size $\alpha = P_{\theta_0}(C)$ is the probability of a Type I error and $\beta = P_{\theta_1}(C^c) = 1 - P_{\theta_1}(C)$ is the probability of a Type II error. With $d_0, d_1 > 0$, suppose that we want to minimize $d_0\alpha + d_1\beta$. In the notation of the proof of the Neyman-Pearson Theorem, this is

$$d_0 \int_C L(\theta_0) + d_1 \int_{C^c} L(\theta_1) = d_1 + \int_C [d_0 L(\theta_0) - d_1 L(\theta_1)].$$

Clearly we minimize this by choosing

$$C = \{d_0 L(\theta_0) - d_1 L(\theta_1) < 0\} = \{\Lambda < \frac{d_1}{d_0}\}.$$

Given α this is the same as minimizing β, ie. maximizing the power $\gamma_C(\theta_1)$. That is, choosing a best critical region.

Sec. 8.2: Uniformly Most Powerful Tests

We continue to test the simple hypothesis $H_0 : \theta = \theta_0$, but now against a composite set of alternatives $H_1 : \theta \in \omega_1$.

HMC Definition 8.2.1: A set $C \subset \mathcal{S}^n$ is a *uniformly most powerful critical region (UMP)* for the simple hypothesis $H_0 : \theta = \theta_0$, the set of alternatives $H_1 : \theta \in \omega_1$ when it is a best critical region of size α for θ_0 against each $\theta \in \omega_1$. A test using such a critical region is called a UMP test.

UMP tests occur in the following situation:

HMC Definition 8.2.2: The likelihood function $L(\theta; \mathbf{x})$ has the *monotone likelihood ratio property (mlr)* in the statistic $T(\mathbf{X})$ if the likelihood ratio function $\Lambda(\theta_1, \theta_2, \mathbf{x})$ is either a monotone increasing function of $T(\mathbf{x})$ for all $\theta_1 < \theta_2$ or a monotone decreasing function of $T(\mathbf{x})$ for all $\theta_1 < \theta_2$.

Let us pause to see what this means. For two parameter values θ_a, θ_b,

$$\Lambda(\theta_a, \theta_b, \mathbf{x}) \;=\; \frac{L(\theta_a, \mathbf{x})}{L(\theta_b, \mathbf{x})}.$$

To say that this is a monotone increasing function of $T(\mathbf{x})$ is to say that there is a positive, increasing function of a real variable t, $g(\theta_a, \theta_b, t)$ such that $\Lambda(\theta_a, \theta_b, \mathbf{x}) = g(\theta_a, \theta_b, T(\mathbf{x}))$.

Notice that the function g of t depends, like Λ, on the parameters θ_a and θ_b. To say that $g(\theta_a, \theta_b, t)$ is an increasing function of t is to say that it preserves inequalities: if $t_1 < t_2$ then $g(\theta_a, \theta_b, t_1) < g(\theta_a, \theta_b, t_2)$. In particular, for any c,

$$t < c \quad \Longleftrightarrow \quad g(\theta_a, \theta_b, t) < g(\theta_a, \theta_b, c).$$

Notice what happens when we reverse the parameters. $\Lambda(\theta_b, \theta_a, \mathbf{x})$ is the reciprocal of $\Lambda(\theta_a, \theta_b, \mathbf{x})$ and taking the reciprocal reverses inequalities between positive numbers. So

$$\Lambda(\theta_b, \theta_a, \mathbf{x}) \;=\; g(\theta_b, \theta_a, T(\mathbf{x})) \;=\; 1/g(\theta_a, \theta_b, T(\mathbf{x})).$$

So $g(\theta_b, \theta_a, t)$ is a decreasing function of t. It reverses inequalities: if $t_1 < t_2$ then $g(\theta_b, \theta_a, t_1) > g(\theta_b, \theta_a, t_2)$. In particular, for any c,

$$t > c \quad \Longleftrightarrow \quad g(\theta_b, \theta_a, t) < g(\theta_b, \theta_a, c).$$

Suppose $\Lambda(\theta_a, \theta_b, \mathbf{x}) = g(\theta_a, \theta_b, T(\mathbf{x}))$ and whenever $\theta_a < \theta_b$, $g(\theta_a, \theta_b, t)$ is a monotone increasing function of t Whenever $\theta_a < \theta_b$ we define for any real number m the critical region

$$C_m = \{T \leq m\} = \{\Lambda(\theta_a, \theta_b, \mathbf{x}) \leq g(\theta_a, \theta_b, m)\}, \quad \text{so that}$$
$$\mathbf{X} \in C_m \Leftrightarrow T(\mathbf{X}) \leq m \Leftrightarrow \Lambda(\theta_a, \theta_b, \mathbf{X}) \leq g(\theta_a, \theta_b, m),$$
$$\mathbf{X} \in C_m^c \Leftrightarrow T(\mathbf{X}) > m \Leftrightarrow \Lambda(\theta_a, \theta_b, \mathbf{X}) > g(\theta_a, \theta_b, m).$$

As a test of the null hypothesis $H_0 : \theta = \theta_a$ the size is $P_{\theta_a}(C_m) = P_{\theta_a}(T \leq m)$. The Neyman-Pearson Theorem says that C_m is a best critical region of this size against any alternative $H_1 : \theta = \theta_b$ with $\theta_a < \theta_b$. The power of the test is $P_{\theta_b}(C_m) = P_{\theta_b}(T \leq m)$.

When we reverse the parameters, $\Lambda(\theta_b, \theta_a, \mathbf{x}) = g(\theta_b, \theta_a, T(\mathbf{x}))$ and whenever $\theta_b > \theta_a$, $g(\theta_b, \theta_a, t) = 1/g(\theta_a, \theta_b, t)$ is a monotone decreasing function of t Whenever $\theta_b > \theta_a$ we define for any real number M the critical region

$$C_M = \{T \geq M\} = \{\Lambda(\theta_b, \theta_a, \mathbf{x}) \leq g(\theta_b, \theta_a, M)\}, \quad \text{so that}$$
$$\mathbf{X} \in C_M \Leftrightarrow T(\mathbf{X}) \geq M \Leftrightarrow \Lambda(\theta_b, \theta_a, \mathbf{X}) \leq g(\theta_b, \theta_a, M),$$
$$\mathbf{X} \in C_M^c \Leftrightarrow T(\mathbf{X}) < M \Leftrightarrow \Lambda(\theta_b, \theta_a, \mathbf{X}) > g(\theta_b, \theta_a, M).$$

As a test of the null hypothesis $H_0 : \theta = \theta_b$ the size is $P_{\theta_b}(C_M) = P_{\theta_b}(T \geq M)$. The Neyman-Pearson Theorem says that C_M is a best critical region of this size against any alternative $H_1 : \theta = \theta_a$ with $\theta_b > \theta_a$. The power of the test is $P_{\theta_a}(C_M) = P_{\theta_a}(T \geq M)$.

When $L(\theta; \mathbf{x})$ has the mlr, we can use the Neyman-Pearson Theorem to obtained a UMP test for either of the *one-sided alternatives*: $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$ or else $H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$.

Suppose that the likelihood ratio function $\Lambda(\theta_1, \theta_2, \mathbf{x})$ is a monotone decreasing function of $T(\mathbf{x})$ for all $\theta_1 < \theta_2$.

To get a UMP test of size α for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. We choose M so that $P_{\theta_0}(T \geq M) = \alpha$ and use as our critical region $C_M = \{\mathbf{X} : T(\mathbf{X}) \geq M\}$, so that C_M is a critical region of size α.

For any $\theta > \theta_0$, $\Lambda(\theta_0, \theta, \mathbf{x})$ is a decreasing function of $T(\mathbf{x})$ so there is a k (which may depend on θ) such that $C_M = \{\mathbf{X} : \Lambda(\theta_0, \theta, \mathbf{X}) \leq k\}$.

The Neyman-Pearson Theorem says that C_M is a best critical region testing the null hypothesis $\theta = \theta_0$ against any $\theta > \theta_0$.

Corollary: With $C_M = \{T \geq M\}$ the power function $\gamma_C(\theta) = P_\theta(C_M)$ is a nondecreasing function of θ.

Proof: Fix $\theta_1 < \theta_2$ and let $\beta = P_{\theta_1}(C_M)$. The above argument shows that C_M provides a best critical region of size β for $H_0 : \theta = \theta_1$ against $H_1 : \theta = \theta_2$. Corollary 8.1.1 then implies

$$P_{\theta_2}(C_M) \geq \beta = P_{\theta_1}(C_M).$$

Notice that the Corollary is true for all $\theta_1 < \theta_2$ on both sides of θ_0. Therefore for $C_M = \{T \geq M\}$, it follows that

$$\alpha \;=\; \max_{\theta \leq \theta_0} \; P_\theta(C_M).$$

This means that C_M provides a UMP critical region for $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.

To get a UMP test of size α for $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$. We choose m so that $P_{\theta_0}(T \leq m) = \alpha$ and use as our critical region $C_m = \{\mathbf{X} : T(\mathbf{X}) \leq m\}$, so that C_m is a critical region of size α.

For any $\theta < \theta_0$, $\Lambda(\theta_0, \theta, \mathbf{x}) = 1/\Lambda(\theta, \theta_0, \mathbf{x})$ is an increasing function of $T(\mathbf{x})$ so there is a k (which may depend on θ) such that $C_m = \{\mathbf{X} : \Lambda(\theta_0, \theta, \mathbf{X}) \leq k\}$.

The Neyman-Pearson Theorem says that C_m is a best critical region testing the null hypothesis $\theta = \theta_0$ against any $\theta < \theta_0$.

Corollary: With $C_2 = \{T \leq m\}$ the power function $\gamma_C(\theta) = P_\theta(C_m)$ is a nonincreasing function of θ.

Proof: Again fix $\theta_1 < \theta_2$ and this time let $\beta = P_{\theta_2}(C_m)$. The above argument shows that C_m provides a best critical region of size β for $H_0 : \theta = \theta_2$ against $H_1 : \theta = \theta_1 < \theta_2$. Again Corollary 8.1.1 then implies $P_{\theta_1}(C_m) \geq \beta = P_{\theta_2}(C_m)$.

We look back at example 8.1.2 and at examples 8.2.2 and 8.2.4.

Example 8.2.2: Our family is $\mathcal{N}(0, \theta)$, ie. normal rv's with mean 0 and variance the unknown positive parameter θ. So the likelihood function is

$$L(\theta; \mathbf{x}) = (\frac{1}{2\pi\theta})^{n/2} exp[-\frac{1}{2\theta}\sum_{i=1}^{n} x_i^2].$$

and the likelihood ratio is

$$\Lambda(\theta_0, \theta_1; \mathbf{x}) = (\frac{\theta_1}{\theta_0})^{n/2} exp[-(\frac{\theta_1 - \theta_0}{2\theta_0\theta_1})\sum_{i=1}^{n} x_i^2].$$

When $\theta_1 > \theta_0$, Λ is a decreasing function of $t = \sum_{i=1}^{n} x_i^2$. So we use as the critical region

$$C = \{\sum_{i=1}^{n} X_i^2 \geq c\}.$$

Since $X/\sqrt{\theta} \sim \mathcal{N}(0,1)$, it follows that $\frac{1}{\theta}\sum_{i=1}^{n} X_i^2$ is a χ^2 rv with n degrees of freedom. If Ψ_n is the cdf of such a distribution, and Q_α is chosen so that $1 - \Psi_n(Q_\alpha) = \alpha$, then for $H_0 : \theta = \theta_0$ the critical region has size α when $c/\theta_0 = Q_\alpha$, or $c = \theta_0 Q_\alpha$. We thus and obtain a UMP test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

The power function is then given by

$$\gamma_C(\theta) = 1 - \Psi_n(\frac{\theta_0}{\theta} Q_\alpha).$$

This is an increasing function of θ.

Example 8.2.4: Suppose the family is Poiss(θ) with $f(\theta; x) = e^{-\theta}\theta^x/x!$ for $x = 0, 1, \ldots$.

$$\Lambda(\theta_0, \theta_1; \mathbf{X}) = e^{n(\theta_1 - \theta_0)}(\frac{\theta_0}{\theta_1})^{\sum_i x_i}.$$

(Notice that the $x_1! \ldots x_n!$ factors cancel out).
With $\theta_1 > \theta_0$ this is a decreasing function of $\sum_i X_i$. So we can use as our critical region $C = \{\sum_{i=1}^n X_i \geq k\}$.
From the properties of the Poisson distribution, $\sum_{i=1}^n X_i \sim Poiss(n\theta)$. We can use this to choose k so that the size is approximately α and to compute the power of the test.
When n is large we can use the normal approximation from the CLT to choose c with approximate size α.

Sec. 1.10: Inequalities

HMC Theorem 1.10.2 (**Markov's Inequality**): Assume X is a nonnegative rv with mean μ and that $c > 0$.

$$P(X \geq c) \leq \frac{\mu}{c}.$$

Proof: Let $Y = \frac{X}{c}$ and let I be the indicator function of the event $\{Y \geq 1\}$. Recall that $E(I) = P(Y \geq 1)$. Observe that $Y \geq I$. So $E(Y) \geq E(I)$.

HMC Theorem 1.10.3 (**Chebyshev's Inequality**): Assume X is an rv with mean μ and variance σ^2. For every $k > 0$

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof: $Z = (X - \mu)^2$ is a nonnegative rv with mean σ^2. By Markov's Inequality $P(Z \geq (k\sigma)^2) \leq \frac{\sigma^2}{(k\sigma)^2}$.

Let $\phi : (a, b) \to \mathbb{R}$ be a differentiable function. We call ϕ (strictly) *concave upwards* or *convex* when the derivative ϕ' is a (strictly) increasing function. For any $c \in (a, b)$ this means that the graph of ϕ is above the tangent line at c. That is, for all $x \in (a, b)$

$$\phi(x) \geq \phi(c) + \phi'(c)(x - c).$$

Proof: The function $x \mapsto \phi(x) - [\phi(c) + \phi'(c)(x - c)]$ has a minimum at $x = c$.

It suffices that the second derivative ϕ'' be non-negative and for strict convexity that ϕ'' be positive.

HMC Theorem 1.10.3 (**Jensen's Inequality**): Assume that ϕ is convex on (a, b) and X is an rv with mean μ and range contained in (a, b).

$$E(\phi(X)) \ \geq \ \phi(E(X)),$$

and if ϕ is strictly convex, then the inequality is strict unless X is a constant rv.

Proof: With $c = \mu$ in the above inequality, we have $\phi(X) \geq \phi(\mu) + \phi'(\mu)(X - \mu)$. Take the expected value.

We look at HMC Examples 1.10.1 and 1.10.4.

Example 1.10.1: Suppose $f(x) = \frac{1}{2\theta}, \quad -\theta < x < \theta$ and $= 0$ otherwise. Thus $\mu = 0$.

$$Var(X) = \frac{1}{2\theta}\int_{-\theta}^{+\theta} x^2 dx = \frac{1}{\theta}\int_{0}^{+\theta} x^2 dx = \frac{\theta^2}{3}.$$

Therefore, $\sigma = \frac{\theta}{\sqrt{3}}$.

$$P(|X-\mu| \geq k\sigma) = 1 - \frac{1}{2\theta}\int_{-k\theta/\sqrt{3}}^{+k\theta/\sqrt{3}} dx = 1 - \frac{k}{\sqrt{3}}.$$

Chebyshev's Inequality says that this probability is bounded by $1/k^2$ which is always larger. One can show using calculus that the function $q(t) = \frac{t}{\sqrt{3}} + \frac{1}{t^2}$ has a minimum where $t_m^3 = \frac{\sqrt{3}}{2}$ and

$$q(t_m) = t_m \cdot \sqrt{3} > t_m^3 \cdot \sqrt{3} = 3/2 > 1.$$

Example 1.10.4: A finite rv X takes a finite number of values which we can list as $X_1, \ldots, X_n$. The pmf of X, is given by $f_X(i) = p_i = P(X = X_i)$.

The mean $E(X) = \sum_{i=1}^{n} p_i X_i$ is the weighted average of the values of X. It is the *arithmetic mean* with weights $p_1, \ldots, p_n$.

For any positive rv X the *geometric mean* is $G(X) = exp(E(\ln(X))$. In the finite case

$$G(X) \;=\; exp(\sum_{i=1}^{n} p_i \ln(X_i)) \;=\; \prod_{i=1}^{n} X_i^{p_i}.$$

Because the function $t \mapsto -\ln(t)$ is concave upward, Jensen's Inequalty implies $E(-\ln(X)) > -\ln(E(X))$. Multiplying by (-1) reverses the inequality. We then exponentiate and get for any positive rv X:

$$E(X) \;>\; exp\; E(\ln(X)) \;=\; G(X).$$

For any positive rv X the *harmonic mean* is $H(X) = 1/E(1/X)$. In the finite case $H(X) = 1/\sum_{i=1}^{n} \frac{p_i}{x_i}$.

For any positive rv X we have $E(X) > G(X) > H(X)$.

Proof: We saw that $E(X) > G(X)$. Notice for any real exponent a

$$\begin{aligned} G(X^a) &= exp(E(\ln(X^a))) = exp(aE(\ln(X))) \\ &= [exp(E(\ln(X)))]^a = G(X)^a. \end{aligned}$$

In particular, with $a = -1$ if we let $Y = 1/X$, then $G(Y) = G(1/X) = 1/G(X)$.

But $E(Y) > G(Y)$. So $E(1/X) > 1/G(X)$. Taking the reciprocals reverses the inequality.

We can see directly that $E(X) > H(X)$ by applying Jensen's Inequality to the function $t \mapsto 1/t$.

Sec. 5.1: Pointwise Convergence and Convergence in Probability

Throughout Chapter 5 we have a sequence of rv's $\{X_n\}$ and a target rv X. We consider different meanings for the phrase $\{X_n\}$ *converges to* X.

A sequence of numbers $\{a_n\}$ converges to a number a when

> *For every* $\epsilon > 0$, *eventually (i.e. for* $n \geq N$ *for some* N*)* $\quad |a_n - a| < \epsilon$.

So the sequence $\{a_n\}$ does <u>not</u> converge to a when

> *For some* $\epsilon > 0$, *infinitely often (i.e. for infinitely many* n *)* $\quad |a_n - a| \geq \epsilon$.

The sequence $\{X_n\}$ converges to X pointwise when for every point s of the sample space S the sequence of numbers $\{X_n(s)\}$ converges to the number $X(s)$.

In probability theory we must allow a set of exceptions of probability zero. So we define:

Definition: $\{X_n\}$ *converges to* X *almost everywhere* (written ae) when the set of $s \in S$ such that $\{X_n(s)\}$ does not converge to the number $X(s)$ has probability zero.

So $\{X_n\}$ converges to X, when only with probability zero does it happen for $s \in S$ that there exist some $\epsilon > 0$ (which ϵ may depend on s) so that infinitely often $\quad |X_n(s) - X(s)| \geq \epsilon$.

Define for $\epsilon > 0$ the event, that is the subset of the sample space, $E_n(\epsilon) = \{|X_n - X| \geq \epsilon\}$.
So $\{X_n(s)\}$ does not converge to $X(s)$ when for some $\epsilon > 0$ s is in the set

$$\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} E_n(\epsilon).$$

This means that $\{X_n\}$ converges to X ae when for every $\epsilon > 0$,

$$P(\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} E_n(\epsilon)) = 0$$

Because the sequence $\{\bigcup_{n=N}^{\infty} E_n(\epsilon)\}$ is a decreasing sequence in N, we obtain

Theorem: $\{X_n\}$ converges to X ae if and only if for every $\epsilon > 0$

$$Lim_{N\to\infty} P(\bigcup_{n=N}^{\infty} E_n(\epsilon)) = 0.$$

Since $E_N(\epsilon) \subset \bigcup_{n=N}^{\infty} E_n(\epsilon)$. we get

Corollary: If $\{X_n\}$ converges to X ae then for every $\epsilon > 0$

$$Lim_{N\to\infty} P(E_N(\epsilon)) = 0.$$

From this we define a weaker notion of convergence.

HMC Definition 5.1.1: $\{X_n\}$ *converges to* X *in probability* (written (P)) when for every $\epsilon > 0$

$$Lim_{n\to\infty} P(E_n(\epsilon)) = 0.$$

That is,

$$Lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0.$$

Both limit ideas satisfy the usual properties that we expect for limits. These can be collected by the following

Theorem: Let $g(x, y)$ be a continuous function defined on a subset D of $\mathbb{R}^2$. Let $\{X_n\}, \{Y_n\}$ be sequences of rv's and X, Y be rv's such that the pairs (X_n, Y_n) and (X, Y) all have range in the subset D.

If $\{X_n\}$ converges to X ae and $\{Y_n\}$ converges to Y ae, then $\{g(X_n, Y_n)\}$ converges to $g(X, Y)$ ae.

If $\{X_n\}$ converges to X (P) and $\{Y_n\}$ converges to Y (P), then $\{g(X_n, Y_n)\}$ converges to $g(X, Y)$ (P).

Since continuity preserves limits, this is easy to prove for convergence ae. The proof is a bit trickier for convergence in probability. We will omit both proofs and just use the results.

Of great importance are the Laws of Large Numbers. We will state both but only give a proof of the weak law.
HMC Theorem 5.1.1: Let $\{X_n\}$ be an iid sequence of rv's with common mean μ. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Strong Law of Large Numbers: The sequence $\{\bar{X}_n\}$ converges ae to the constant μ.

Weak Law of Large Numbers: The sequence $\{\bar{X}_n\}$ converges (P) to the constant μ.

For the proof of the weak law we assume that the rv's have a finite variance σ^2 so that $Var(\bar{X}_n) = \frac{\sigma^2}{n}$. We apply Chebyshev's Inequality:

$$\begin{aligned} P(|\bar{X}_n - \mu| \geq \epsilon) &= P(|\bar{X}_n - \mu| \geq \frac{\epsilon\sqrt{n}}{\sigma} \cdot \frac{\sigma}{\sqrt{n}}) \\ &\leq \frac{\sigma^2}{n\epsilon^2} \quad \rightarrow 0. \end{aligned}$$

HMC Definition 5.1.2 (**Consistency**) Let X be an rv with cdf $F(x, \theta), \theta \in \Omega$. Let $\{X_n\}$ be an iid sequence with distribution that of X so the $X_1, \ldots, X_n$ is a sample of size n. Let $T_n(X_1, \ldots, X_n)$ be a statistic. The sequence $\{T_n\}$ is called a *consistent estimator* of θ if $\{T_n\} \to \theta$ (P).

If $\{T_n\}$ is a consistent estimator of θ and $\{a_n\}$ is a sequence of numbers converging to 1, then $\{a_n \cdot T_n\}$ is a consistent estimator of θ as well.

Theorem: Assume that each T_n is an unbiased estimator and that each T_n has finite variance σ_n^2. If $\sigma_n^2 \to 0$, then $\{T_n\}$ is a consistent estimator of θ.

Proof: $E(T_n) = \theta$ for all n because the estimator is unbiased. Apply Chebyshev's Inequality:

$$P(|T_n - \theta| \geq \epsilon) \;=\; P(|T_n - \theta| \geq \frac{\epsilon}{\sigma_n} \cdot \sigma_n) \;\leq\; \frac{\sigma_n^2}{\epsilon^2} \;\to 0.$$

Example 5.1.1: For an infinite sequence $X_1, X_2, \ldots$ of iid's with mean μ and variance σ^2, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of μ. That is, $E(\bar{X}_n) = \mu$. Recall that $Var(\bar{X}_n) = \sigma^2/n$. The Law of Law Numbers implies that the sequence $\{\bar{X}_n\}$ provides a consistent estimate of μ as well. Notice that this is a special case of the above theorem.

Now consider the sample variance.

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 =$$

$$\frac{n}{n-1}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2] = \frac{n}{n-1}[(\frac{1}{n}\sum_{i=1}^{n} X_i^2) - (\bar{X}_n)^2].$$

By the Law of Large Numbers applied to $\{X_i^2\}$ the sequence $\{\frac{1}{n}\sum_{i=1}^{n} X_i^2\}$ converges to $E(X_i^2)$ and $\{(\bar{X}_n)^2\}$ converges to $E(X_i)^2 = \mu^2$.

It follows that $\{S_n^2\}$ is a consistent estimate for σ^2.

Hence, $\{S_n = \sqrt{S_n^2}\}$ is a consistent estimate for σ.

It is not unbiased. Jensen's Inequality applied to the concave function $t \mapsto -\sqrt{t}$ implies that

$$E(\sqrt{S_n^2}) \; < \; \sqrt{E(S_n^2)} \; = \; \sigma.$$

Example 5.1.2: Let $X_1, \ldots$ be an iid sequence of the uniform distribution on $(0, \theta)$. Recall that using the sample $X_1, \ldots, X_n$ the maximum $Y_n = \max(X_1, \ldots, X_n)$ is the MLE estimator for θ.

The cdf of Y_n is $F_{Y_n}(t) = (t/\theta)^n$ for $0 \leq t \leq \theta$ and so the density $f_{Y_n}(t) = \frac{n}{\theta^n} t^{n-1}$ for $0 \leq t \leq \theta$. Therefore,

$$E(Y_n) = \frac{n}{\theta^n} \int_0^\theta t^n dt = \frac{n}{n+1}\theta,$$

$$E(Y_n^2) = \frac{n}{\theta^n} \int_0^\theta t^{n+1} dt = \frac{n}{n+2}\theta^2,$$

$$\text{so} \quad Var(Y_n) = \frac{n}{n+2}\theta^2 - (\frac{n}{n+1}\theta)^2 = \frac{n}{(n+1)^2(n+2)}\theta^2.$$

So $\frac{n+1}{n} Y_n$ is an unbiased estimate with variance $\frac{1}{n(n+2)}\theta^2$ and so it is a consistent estimator. Hence, the biased estimate Y_n is consistent as well.

We can see directly that $\{Y_n\} \to \theta$ (P) because

$$P(|Y_n - \theta| \geq \epsilon) = P(Y_n \leq \theta - \epsilon) = F_{Y_n}(\theta - \epsilon) = (\frac{\theta - \epsilon}{\theta})^n \to 0.$$

Recall that $\bar{X}_n$ is an unbiased and consistent estimator for the mean $\mu = \theta/2$. It follows that $2\bar{X}_n$ is an unbiased and consistent estimator for θ.

$Var(\bar{X}_n) = Var(X_i)/n = \frac{1}{12n}\theta^2$ and so $Var(2\bar{X}_n) = \frac{1}{3n}\theta^2$. This is $\frac{n+2}{3}$ times the variance of $\frac{n+1}{n}Y_n$.

Thus, while both $2\bar{X}_n$ and $\frac{n+1}{n}Y_n$ are unbiased and consistent estimators of θ, the latter is a better choice because its variance is much smaller.

Sec. 5.2: Convergence in Distribution

Now we turn to a notion of convergence which is even weaker but nonetheless of great importance. For it we refer to the cdf's F_{X_n} and F_X of the random variables. We let $C(F_X)$ denote the set of points $x \in \mathbb{R}$ at which the distribution function F_X is continuous.

HMC Definition 5.2.1: $\{X_n\}$ *converges to* X *in distribution* (written $\xrightarrow{(D)}$) when

$$Lim_{n\to\infty} F_{X_n}(x) \;=\; F_X(x) \quad \text{for all } x \in C(F_X).$$

Example 5.2.4: As in Example 5.1.2 we let $X_1, \ldots$ be an iid sequence of the uniform distribution on $(0, \theta)$ and consider $Y_n = \max(X_1, \ldots, X_n)$. Let $Z_n = n(\theta - Y_n)$.

$$E(Z_n) = n(\theta - E(Y_n)) = n(\theta - \frac{n}{n+1}\theta) = \frac{n}{n+1}\theta,$$

$$Var(Z_n) = n^2 Var(Y_n) = \frac{n^3}{(n+1)^2(n+2)}\theta^2.$$

$$F_{Z_n}(t) = P(Z_n \leq t) = P(Y_n \geq \theta - \frac{t}{n}) =$$

$$1 - (\frac{\theta - (t/n)}{\theta})^n = 1 - (1 - \frac{t/\theta}{n})^n \to 1 - exp(-(t/\theta)).$$

That is, $Z_n \to Z$ where $Z \sim \Gamma(1, \theta)$, i.e. it has an exponential distribution with mean θ.

In order to compare convergence in probability with convergence in distribution, we remember that, for example, $F_X(x) = P(X \leq x)$.

$$F_X(x - \epsilon) \leq F_{X_n}(x) + P(|X - X_n| \geq \epsilon),$$

$$F_{X_n}(x) \leq F_X(x + \epsilon) + P(|X - X_n| \geq \epsilon).$$

Therefore,

$$\begin{aligned} & F_X(x) - F_{X_n}(x) - [F_X(x) - F_X(x - \epsilon)] \\ = \; & F_X(x - \epsilon) - F_{X_n}(x) \;\leq\; P(|X - X_n| \geq \epsilon), \end{aligned}$$

$$\begin{aligned} & F_{X_n}(x) - F_X(x) - [F_X(x + \epsilon) - F_X(x)] \\ = \; & F_{X_n}(x) - F_X(x + \epsilon) \;\leq\; P(|X - X_n| \geq \epsilon). \end{aligned}$$

Because $|a - b| = \max(a - b, b - a)$ we put these two together to get:

$$|F_{X_n}(x) - F_X(x)| \leq \max([F_X(x) - F_X(x-\epsilon)], [F_X(x+\epsilon) - F_X(x)]) + P(|X - X_n| \geq \epsilon).$$

Notice that if X is a continuous with density $f_X(x)$ bounded by some constant M, then

$$\max([F_X(x) - F_X(x-\epsilon)], [F_X(x+\epsilon) - F_X(x)]) \leq M\epsilon.$$

Let $\delta > 0$. If x is a continuity point we can choose $\epsilon > 0$ so that $\max([F_X(x) - F_X(x-\epsilon)], [F_X(x+\epsilon) - F_X(x)]) \leq \delta/2$. If X has a density bounded by M then we can choose $\epsilon = \delta/2M$ which will work for all x.

Now assume that $X_n \to X$ (P). Having chosen ϵ so that $\max([F_X(x) - F_X(x-\epsilon)], [F_X(x+\epsilon) - F_X(x)]) \leq \delta/2$ we can now choose N so that when $n \geq N, \quad P(|X - X_n| \geq \epsilon) \leq \delta/2$.

It follows that when $n \geq N, \quad |F_{X_n}(x) - F_X(x)| \leq \delta$.

Thus, we have proved:

HMC Theorem 5.2.1: If $\{X_n\}$ converges to X in probability, then $\{X_n\}$ converges to X in distribution.

Furthermore, if X is a continuous rv with a bounded density function, then

$$Lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_{X_n}(x) - F_X(x)| \;=\; 0.$$

The converse is not true in general, but it is true if the limit rv is a constant.

HMC Theorem 5.2.2: If $\{X_n\}$ converges to b in distribution, then it converges to b in probability.

Proof:
$P(|X_n - b| \leq \epsilon) = F_{X_n}(b + \epsilon) - F_{X_n}((b - \epsilon) - 0) \to 1 - 0.$

We state two results from the book which we will use without proof.

HMC Theorem 5.2.3: If g is a function continuous on $D \subset \mathbb{R}$ and the rv's have range in D then $X_n \to X$ (D) implies $\{g(X_n)\} \to g(X)$ (D).

HMC Theorem 5.2.4 (**Slutsky's Theorem**): If $X_n \to X$ (D), $A_n \to a$ (P) and $B_n \to b$ (P), then $A_n + B_n X_n \to a + bX$ (D).

If $X_n \to X$ (D), we would like to conclude that $E(X_n) \to E(X)$, but this need not be true. Among other things, an rv with an unbounded range need not have an expected value. What is true is the following:

Theorem: $X_n \to X$ (D) if and only if for every bounded continuous function $g : \mathbb{R} \to \mathbb{R}$, $E(g(X_n)) \to E(g(X)$.

If there is a bounded subset of $\mathbb{R}$ which contains all the ranges of the rv's then any continuous function is bounded on the range and $E(g(X_n)) \to E(g(X))$ follows. However, the function $g(x) = x$ is not bounded on all of $\mathbb{R}$. The most useful result for us is:

Theorem: If $X_n \to X$ (D) and there exists $M > 0$ such that $E(X_n^2) \leq M$ for all n, then $E(X_n) \to E(X)$.

Corollary: If $X_n \to X$ (D) and there exists $M > 0$ and $h > 0$ such that $E(e^{2h|X_n|}) \leq M$, then $E(e^{tX_n}) \to E(e^{tX})$ for all t with $|t| < h$.

That is, convergence in distribution together with some boundedness conditions implies convergence near 0 of the mgf's. Most important for us is the deep converse result.

HMC Theorem 5.2.10: Assume that for some $h > 0$ the mgf's $M_{X_n}(t), M_X(t)$ exist for $|t| \leq h$. If $M_{X_n}(t) \to M_X(t)$ for every t with $|t| \leq h$, then $X_n \to X$ (D).

That is, we can recognize convergence in distribution by using convergence of the moment generating functions.

Squeeze Theorem

Theorem (**Squeeze Theorem**) If $X_n^1 \geq Y_n \geq X_n^2$ and $X_n^1, X_n^2 \to X$ (P) (or (D)) then $Y_n \to X$ (P) (respectively, (D)).

Convergence in probability follows because

$$\{|Y_n - X| > \epsilon\} \subset \{|X_n^1 - X| > \epsilon\} \cup \{|X_n^2 - X| > \epsilon\}.$$

Convergence in distribution follows because for every x

$$F_{X_n^1}(x) \leq F_{Y_n}(x) \leq F_{X_n^2}(x).$$

Sec. 5.3: Central Limit Theorem

Let $\{X_n\}$ be an iid sequence of rv's with mean μ and variance σ^2. So $\{Y_n = \frac{X_n - \mu}{\sigma}\}$ is an iid sequence of rv's with mean 0 and variance 1.

The Law of Large Numbers says that the sample means satisfy $\bar{X}_n \to \mu$ (P) or, equivalently, $\bar{Y}_n \to 0$ (P). In fact convergence holds ae by the Strong Law. The averaged rv's

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$$

all have mean 0 and variance 1. Let $Z \sim \mathcal{N}(0,1)$.

HMC Theorem 5.3.1 (**Central Limit Theorem**):

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to Z \ (D).$$

We will prove the result assuming that for some $h > 0$ X has an mgf defined for $|t| \leq h\sigma$. So the mgf of Y, $M(t) = E(e^{tY}) = e^{-\mu t/\sigma} E(e^{(t/\sigma)X})$ is defined for $|t| \leq h\sigma$.

$$E(e^{t(Y_1+\cdots+Y_n)/\sqrt{n}}) = E(e^{tY_1/\sqrt{n}}) \ldots E(e^{tY_n/\sqrt{n}}) = M(\frac{t}{\sqrt{n}})^n.$$

To compute the limit as $n \to \infty$ we take the log and use L'Hôpital's Rule after the change of variable $y = 1/\sqrt{n}$

$$Lim_{n\to\infty}\ n \ln M(\frac{t}{\sqrt{n}}) = Lim_{y\to 0}\ \frac{\ln M(yt)}{y^2} =$$

$$Lim_{y\to 0}\ \frac{tM'(yt)}{2yM(yt)} = \frac{t}{2} Lim_{y\to 0}\ \frac{1}{M(yt)} Lim_{y\to 0}\ \frac{M'(yt)}{y}$$

$$= \frac{t^2}{2} Lim_{y\to 0}\ \frac{1}{M(yt)} Lim_{y\to 0}\ \frac{M''(yt)}{1} = \frac{t^2}{2}.$$

Recall that for $Z \sim \mathcal{N}(0,1)$, $E(e^{tZ}) = e^{t^2/2}$.

Thus, for each t with $|t| \leq h$, $E(e^{t(Y_1+\cdots+Y_n)/\sqrt{n}}) \to E(e^{tZ})$.

Since the mgf's converge it follows from HMC Theorem 5.2.10 that

$$\frac{\sum_{i=1}^n Y_i}{\sqrt{n}} \to Z \ (D)$$

as $n \to \infty$.

Since $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \to \sigma^2 \quad (P)$, HMC Theorems 5.2.3 and 5.2.4 imply that

$$\frac{\sum_{i=1}^n X_i - n\mu}{S\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S} \to Z \ (D).$$

HMC Theorem 5.2.9: If $g(x)$ is continuously differentiable at θ with $g'(\theta) \neq 0$ and $\{\sqrt{n}(X_n - \theta)\} \rightarrow Z\ (D)$ with $Z \sim \mathcal{N}(0, \sigma^2)$, then $\{\sqrt{n}(g(X_n) - g(\theta))\} \rightarrow Z^1\ (D)$ with $Z^1 \sim \mathcal{N}(0, \sigma^2(g'(\theta))^2)$.

Proof: By the Mean Value Theorem $\sqrt{n}(g(X_n) - g(\theta)) = g'(c_n)[\sqrt{n}(X_n - \theta)]$ with c_n between θ and X_n.

$X_n \rightarrow \theta\ (D)$ and so (P). So by the Squeeze Theorem $c_n \rightarrow \theta\ (P)$.

Because g' is continuous $g'(c_n) \rightarrow g'(\theta) \quad (P)$. So by Slutsky's Theorem

$$\{\sqrt{n}(g(X_n) - g(\theta))\} \rightarrow g'(\theta)Z\ (D).$$

$Z^1 = g'(\theta)Z \sim \mathcal{N}(0, \sigma^2(g'(\theta))^2)$.

Kolmogorov-Smirnov Test

We consider a continuous rv X with support (a, b) with $-\infty \leq a < b \leq \infty$. That is, the pdf $f_X(x)$ is positive for $x \in (a, b)$ and is zero elsewhere. It follows that $F_X : (a, b) \to (0, 1)$ is a strictly increasing continuous function.

When we looked at Sec 4.1.2[4.1.1], I mentioned that the best estimate for the cdf F_X of a continuous rv X is the sample cdf. We now look at what this means.

Let $X_1, \ldots, X_n$ be a sample. The *empirical distribution* $X_{[n]}$ is the discrete rv taking the values $X_1, \ldots, X_n$ with each value equally likely and so with probability $\frac{1}{n}$. That is, we let $[n]$ be an independent rv with values $\{1, \ldots, n\}$, so that for $i = 1, \ldots, n$ $P(X_{[n]} = X_i) = \frac{1}{n}$.

For a real number x, $F_{X_{[n]}}(x)$ counts the number of X_i's with value at most x and divides by n.

We let I_x denote the indicator function of $(-\infty, x] \subset \mathbb{R}$ so that $I_x(t) = 1$ if $t \leq x$ and $= 0$ otherwise. Thus, $E(I_x(X)) = P(X \leq x) = F_X(x)$.

$$F_{X_{[n]}}(x) = \frac{1}{n} \sum_{i=1}^{n} I_x(X_i).$$

It follows from the Strong Law of Large Numbers that as

$$Lim_{n \to \infty} \; F_{X_{[n]}}(x) = F_X(x) \quad \text{ae}$$

.

This says that for almost every realization $x_1, x_2, \ldots$ of the sequence $X_1, X_2, \ldots$ the sequence of discrete realized rv's $\{x_{[n]}\}$ converges in distribution to X.

However, a stronger result is true.

We can define the statistic $D_n = \sup_{x\in(a,b)} |F_{X_{[n]}}(x) - F_X(x)|$.

Theorem (**Glivenko-Cantelli Theorem**) $Lim_{n\to\infty} D_n = 0$ ae.

Amazingly, the statistic does not depend on the choice of continuous distribution F_X. To see this, recall that $U = F_X(X) \sim Unif(0,1)$, So that $U_1 = F_X(X_1), U_2 = F_X(X_2), \ldots$ is an iid sequence all with the same uniform distribution and with empirical distribution $U_{[n]} = F_X(X_{[n]})$.

We use the change of variable $u = F_X(x)$. So that

$$\{X_i \leq x\} = \{F_X(X_i) \leq F_X(x)\} = \{U_i \leq u\}$$

From which we obtain $F_{X_{[n]}}(x) = F_{U_{[n]}}(u)$ and $F_X(x) = F_U(u) = u$. Therefore,

$$D_n = \sup_{x\in(a,b)} |F_{X_{[n]}}(x) - F_X(x)| = \sup_{u\in(0,1)} |F_{U_{[n]}}(u) - u|.$$

While D_n is the supremum over an infinite set, it really requires only $2n$ computations.

Let $Y_1 < \cdots < Y_n$ be the order statistics for $X_1, \ldots, X_n$. Since X is continuous, we may assume that the n values are distinct. It follows that

$$F_{X_{[n]}}(x) = \begin{cases} 0 & x < Y_1, \\ \frac{i}{n} & Y_i \leq x < Y_{i+1}, \\ 1 & Y_n \leq x. \end{cases}$$

Note that if G is a continuous increasing function on a interval and c is a constant, then the maximum value of $|c - G|$ on the interval occurs at one of the endpoints. It easily follows that

$$D_n = \max_{i=1}^{n} \{|\frac{i}{n} - F_X(Y_i)|, |\frac{i-1}{n} - F_X(Y_i)|\}.$$

The Kolmogorov-Smirnov Test uses the computed values of this statistic to test whether an iid sample with realization $x_1, \ldots, x_n$ and associated order sequence $y_1 < \cdots < y_n$ comes from an rv with cdf F_X. We compute

$$D_n = \max_{i=1}^{n} \{|\frac{i}{n} - F_X(y_i)|, |\frac{i-1}{n} - F_X(y_i)|\}.$$

Under the null hypothesis that F_X is the true distribution, it is known that, for example, with $n = 50$ the probability that D_{50} is greater than .23 has probability .01 (there are tables for this).

Therefore, if $n = 50$ and the value of D_n computed above is greater than .23, then we reject the null hypothesis with confidence level .01.

Sec. 6.1: Maximum Likelihood Estimation: Existence and Consistency

We will be considering a family of pdf's $f(x;\theta)$ parametrized by θ in an interval of the real line. Among the parameter values is the true value θ_0 which we desire to estimate.

For a sample $X_1, \ldots, X_n$ the *likelihood*,

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta),$$

is the joint density of $\mathbf{X} = (X_1, \ldots, X_n)$.

In Section 4.1, the *maximum likelihood estimator* $\hat{\theta}$ was defined. For a realization $\mathbf{x} = (x_1, \ldots, x_n)$ of the sample $\hat{\theta}(x_1, \ldots, x_n)$ is a parameter value at which the likelihood $L(\theta; x_1, \ldots, x_n)$ achieves its maximum. The statistic $\hat{\theta}(\mathbf{X})$ is the maximum likelihood estimator.

We make various *regularity assumptions*

HMC Assumptions 6.1.1:

R0 The pdf's are distinct: If $\theta_1 \neq \theta_2$, then $f(\cdot, \theta_1) \neq f(\cdot, \theta_2)$.

R1 The pdf's have common support for all θ.

HMC Theorem 6.1.1: If $\theta_1 \neq \theta_0$, then
$Lim_{n\to\infty} P_{\theta_0}[L(\theta_0, \mathbf{X}) > L(\theta_1, \mathbf{X})] = 1.$

Proof: $L(\theta_0, \mathbf{X}) > L(\theta_1, \mathbf{X})$ if and only if

$$\frac{1}{n}\sum_{i=1}^{n} \ln[\frac{f(x_i;\theta_1)}{f(x_i;\theta_0)}] < 0.$$

$$\frac{1}{n}\sum_{i=1}^{n} \ln[\frac{f(x_i;\theta_1)}{f(x_i;\theta_0)}] \ \rightarrow \ E_{\theta_0}[\ln[\frac{f(X,\theta_1)}{f(X,\theta_0)}]] \quad (P)$$

by the Law of Large Numbers.

By Assumption (R0) the rv $\frac{f(X,\theta_1)}{f(X,\theta_0)}$ is a nonconstant rv, which is positive on the common support of the pdf's (R1). So Jensen's Inequality implies:

$$E_{\theta_0}[\ln[\frac{f(X,\theta_1)}{f(X,\theta_0)}]] \ < \ln \ E_{\theta_0}[\frac{f(X,\theta_1)}{f(X,\theta_0)}].$$

But,

$$E_{\theta_0}[\frac{f(X,\theta_1)}{f(X,\theta_0)}] \ = \ \int \ \frac{f(x,\theta_1)}{f(x,\theta_0)} \cdot f(x,\theta_0) \ dx = 1.$$

So there is a positive number k so that $E_{\theta_0}[\ln[\frac{f(X,\theta_1)}{f(X,\theta_0)}]] = -k$ and so

$$\frac{1}{n}\sum_{i=1}^{n} \ln[\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}] \ \rightarrow \ -k \quad (P).$$

$$P(|\frac{1}{n}\sum_{i=1}^{n} \ln[\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}] - (-k)| < k) \ \rightarrow \ 1.$$

$$P(\frac{1}{n}\sum_{i=1}^{n} \ln[\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}] < 0) \ \rightarrow \ 1.$$

HMC Example 6.1.1: $f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}$ (Laplace or double exponential, shifted by θ.

$$\ell(\theta;\mathbf{x}) \;=\; -n\ln(2) - \sum_{i=1}^{n}|x_i-\theta|, \qquad \ell'(\theta) = \sum_{i=1}^{n} sgn(x_i-\theta).$$

where $sgn(t) = \begin{cases} +1 & t > 0, \\ -1 & t < 0 \end{cases}$ and is undefined if $t = 0$.

So $\ell'(\theta) > 0$ if $\#\{x_i > \theta\} > \#\{x_i < \theta\}$ and $\ell'(\theta) < 0$ if $\#\{x_i > \theta\} < \#\{x_i < \theta\}$.

Let $Y_1, \ldots, Y_n$ be the order statistics for the sample $X_1, \ldots, X_n$.

If $n = 2k+1$ then $\hat{\theta} = Y_{k+1}$, the median.

If $n = 2k$, we can use $\hat{\theta} = (Y_k + Y_{k+1})/2$ but the maximum value for the likelihood is achieved at any point y with $Y_k < y < Y_{k+1}$.

Example 6.1.3: With the Bernoulli mass function $f(x;\theta) = \theta^x(1-\theta)^{1-x}$ for $x = 0, 1$ we saw in Example 4.1.4 that the mle $\hat{\theta}$ for $\mathbf{X}$ is the sample mean $\bar{X}$.

But if the range is restricted to $[0, \frac{1}{3}]$ then we are looking for the maximum of $L(\theta)$ on this closed interval. When the critical point $\bar{x}$ lies in the interval then it is the mle. If $\bar{x} > \frac{1}{3}$ then $L(\theta; \mathbf{x})$ is increasing on the interval and so the mle is the right hand endpoint.

$$\hat{\theta} \;=\; \min(\bar{X}, \frac{1}{3}).$$

HMC Theorem 6.1.2: If g is a one-to-one function on the parameter interval and $\eta = g(\theta)$, then the mle estimate $\hat{\eta}$ equals $g(\hat{\theta})$.

Proof: For any $\eta = g(\theta)$, $L(\eta; \mathbf{x}) = L(\theta; \mathbf{x})$. That is, we are just relabeling.

If we let $\tilde{L}(\eta, \mathbf{x})$ be the likelihood in terms of η, then $\tilde{L}(\eta, \mathbf{x})$ just means $L(\theta, \mathbf{x}) = L(g^{-1}(\eta), \mathbf{x})$.

We are assuming that $L(\hat{\theta}; \mathbf{x}) > L(\theta; \mathbf{x})$ with $\theta \neq \hat{\theta}$. Let $\hat{\eta} = g(\hat{\theta})$ and let $\eta = g(\theta)$ be some other value of the variable.

$$\tilde{L}(\hat{\eta}, \mathbf{x}) = L(\hat{\theta}, \mathbf{x}) > L(\theta, \mathbf{x}) = \tilde{L}(\eta; \mathbf{x}).$$

HMC Assumptions 6.1.1:

R2 The true value θ_0 is in the interior of the parameter interval.

R3 The pdf $f(x, \theta)$ is twice continuously differentiable as a function of θ.

HMC Theorem 6.1.3: There is a sequence of statistics $\{\tilde{\theta}_n\}$ in the interior of the parameter interval, which converges ae to θ_0 and such that

$$P_{\theta_0}(L(\theta; X_1, \ldots, X_n) \text{ has a local maximum at } \tilde{\theta}_n) \to 1.$$

Notice that at an interior local maximum $\frac{\partial L}{\partial \theta} = 0$.

Proof: Let k_0 be the smallest positive integer such that the closed interval $[\theta_0 - \frac{1}{k_0}, \theta_0 + \frac{1}{k_0}]$ is contained in the interior of the parameter interval. For each $k \geq k_0$ let

$$\theta_k^- = \theta_0 - \frac{1}{k}, \quad \theta_k^+ = \theta_0 + \frac{1}{k}.$$

so that the intervals $[\theta_k^-, \theta_k^+]$ each have mid-point θ_0 and they are closing in on θ_0.

For each $k \geq k_0$ define the event

$$S_{n,k} = \{\mathbf{X} : L(\theta_0; \mathbf{X}) > L(\theta_k^-; \mathbf{X})\} \cap \{\mathbf{X} : L(\theta_0; \mathbf{X}) > L(\theta_k^+; \mathbf{X})\}.$$

Looking ahead, notice that if $\mathbf{X} \in S_{n,k}$ then on the interval $[\theta_k^-, \theta_k^+]$ the function of θ $L(\theta; \mathbf{X})$ takes its maximum at a point $\tilde{\theta}$ in the interior (θ_k^-, θ_k^+) rather than one of the endpoints and so $\tilde{\theta}$ is a local maximum for $L(\theta; \mathbf{X})$ on the entire parameter interval.

Theorem 6.1.1 says that for each $k \geq k_0$ the probability $P_{\theta_0}(S_{n,k}) \to 1$ as $n \to \infty$.

This means we can choose a positive integer N_k so that $P_{\theta_0}(S_{n,k}) > 1 - \frac{1}{k}$ for all $n \geq N_k$ and so that $N_{k+1} > N_k$.

For $n < N_{k_0}$ let $\tilde{\theta}_n(\mathbf{X}) = \theta_0$.

For $k \geq k_0$ and $N_k \leq n < N_{k+1}$ we restrict $L(\theta; \mathbf{X})$ to the parameter interval $[\theta_k^-, \theta_k^+]$ and let $\tilde{\theta}_n(\mathbf{X})$ be a parameter value in this interval at which the restricted function achieves its maximum.

So for $N_k \leq n < N_{k+1}$

(i) $\theta_k^- \leq \tilde{\theta}_n(\mathbf{X}) \leq \theta_k^+$.

(ii) If $\mathbf{X} \in S_{n,k}$, then $\tilde{\theta}_n(\mathbf{X}) \in (\theta_k^-, \theta_k^+)$ is a local maximum for the unrestricted function $L(\theta; \mathbf{X})$.

(iii) The θ_0 probability that $\tilde{\theta}_n(\mathbf{X})$ is a local maximum for $L(\theta; \mathbf{X})$ is greater than $1 - \frac{1}{k}$.

From (i) it follows that the sequence $\{\tilde{\theta}_n\}$ converges to θ_0.
From (iii) it follows that

$$P_{\theta_0}(L(\theta; \mathbf{X}) \text{ has a local maximum at } \tilde{\theta}_n) \to 1.$$

Notice that this is a pure existence theorem. We don't know what θ_0 is and so none of this can be computed.

Let us summarize the proof.

(a) For θ_0 in the interior of the parameter interval there is a k_0 so that with $k \geq k_0$ the interval $[\theta_k^-, \theta_k^+]$ with midpoint θ_0 and of length $2/k$ is contained in the interior of the parameter interval.

(b) Using Theorem 6.1.1 we get an increasing sequence of positive integers $\{N_k\}$ so that for $n \geq N_k$

$$P_{\theta_0}[L(\theta_0, \mathbf{X}) > L(\theta_k^-, \mathbf{X}) \ \& \ L(\theta_0, \mathbf{X}) > L(\theta_k^+, \mathbf{X})] \ \geq \ 1 - (1/k).$$

(c) For n with $N_k \leq n < N_{k+1}$ let $\tilde{\theta}_n(\mathbf{X})$ be a point in $[\theta_k^-, \theta_k^+]$ at which $L(\theta, \mathbf{X})$ has its maximum on the interval. When the condition in (b) holds, $\tilde{\theta}_n(\mathbf{X})$ is not an endpoint and so is a local maximum for $L(\theta, \mathbf{X})$ as θ varies over the whole parameter interval. From (a) the sequence $\{\tilde{\theta}_n(\mathbf{X})\}$ converges to θ_0 for every $\mathbf{X}$.

HMC Corollary 6.1.1: If the likelihood equation $\frac{\partial L}{\partial \theta} = 0$ has a unique solution $\hat{\theta}(\mathbf{X})$, then $P(\hat{\theta}_n = \tilde{\theta}_n) \rightarrow 1$ and so $\hat{\theta}$ is a consistent estimator for θ_0.

Proof: When $\tilde{\theta}_n(\mathbf{X})$ is a local maximum for $L(\theta; \mathbf{X})$ then it is a solution of the likelihood equation and so equals $\hat{\theta}$.

It then follows from (iii) above that for
$N_k \leq n < N_{k+1}, \quad P(\hat{\theta}_n = \tilde{\theta}_n) > 1 - \frac{1}{k}$.

From (i) it then follows that $P(|\hat{\theta}_n - \theta_0| > \frac{1}{k}) < \frac{1}{k}$, and so $\hat{\theta}_n \rightarrow \theta_0$ (P).

Sec. 6.2: Information and Efficiency

HMC Assumptions 6.2.1:

R4 The integrals can be differentiated under the integral sign as functions of θ.

Because $1 = \int_{-\infty}^{\infty} f(x;\theta)\, dx$ we can differentiate to get:

$$0 = \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial \theta}\, dx = \int_{-\infty}^{\infty} \frac{\partial \ln f(x;\theta)}{\partial \theta} \cdot f(x;\theta)\, dx$$

That is, the rv $\frac{\partial \ln f(X;\theta)}{\partial \theta}$ has expectation 0. That is,

$$E_\theta[\frac{\partial \ln f(X;\theta)}{\partial \theta}] = 0.$$

Differentiating again we have

$$0 = \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2} \cdot f(x;\theta)\, dx + \int_{-\infty}^{\infty} \left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2 \cdot f(x;\theta)\, dx$$

The second integral, denoted $I(\theta)$, is a variance of $\frac{\partial \ln f(X;\theta)}{\partial \theta}$ and is called the *Fisher information*. Therefore,

$$I(\theta) \;=\; Var_\theta\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right] \;=\; -E_\theta\left[\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2}\right].$$

The function $\frac{\partial \ln f(x;\theta)}{\partial \theta}$ is called the *score function*.

Note that

$$\ell(\theta; \mathbf{X}) = \sum_{i=1}^{n} \ln f(X_i; \theta)$$

$$\ell'(\theta; \mathbf{X}) = \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(X_i; \theta)}{\partial \theta}.$$

Examples 6.2.1- If $X \sim Bern(\theta)$ then the pmf is $f(\theta, x) = \theta^x(1-\theta)^{1-x}$ with $x = 0, 1$ and $E(X) = \theta$. So

$$\ell(\theta) = x \ln \theta + (1-x) \ln(1-\theta)$$

$$\ell'(\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

$$\ell''(\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

$$I(\theta) = -E_\theta(\ell''(\theta)) = \frac{1}{\theta(1-\theta)}.$$

Examples 6.2.2- For a *location model* we assume that the iid sequence $X_1, X_2, \ldots$ are such that $X_i - \theta$ are rv's with density $f_X(x)dx$, not depending on θ. For example, $X_i \sim \mathcal{N}(\theta, 1)$ for all i is a location model. Hence, the common pdf of X_i's is $f_X(x-\theta)dx$.
The information is given by:

$$I(\theta) = E_\theta(\ell'(\theta)^2) = \int_{-\infty}^{\infty} (\frac{f'(x-\theta)}{f(x-\theta)})^2 f(x-\theta)dx = \int_{-\infty}^{\infty} (\frac{f'(z)}{f(z)})^2 f(z)dz$$

via the change of variables $z = x - \theta$.

Thus, the information in this case does not depend on θ.

For a sample $\mathbf{X} = (X_1, \ldots, X_n)$, $\frac{\partial \ln L(\theta;\mathbf{X})}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta}$ and so $Var(\frac{\partial \ln L(\theta;\mathbf{X})}{\partial \theta}) = nI(\theta)$.

HMC Theorem 6.2.1 (**Rao-Cramer Lower Bound**): For an rv $Y = u(\mathbf{X})$, if $k(\theta) = E_\theta(Y)$, then

$$Var_\theta(Y) \geq \frac{k'(\theta)^2}{nI(\theta)}.$$

In particular, if Y is an unbiased estimator of θ, so that $k(\theta) = \theta$, then

$$Var_\theta(Y) \geq \frac{1}{nI(\theta)}.$$

Proof:

$$k(\theta) = \int_{\mathbb{R}^n} u(\mathbf{x}) L(\theta; \mathbf{x}) d\mathbf{x},$$

$$k'(\theta) = \int_{\mathbb{R}^n} u(\mathbf{x}) \frac{\partial \ln L(\theta; \mathbf{x})}{\partial \theta} L(\theta; \mathbf{x}) d\mathbf{x}.$$

With $Z = \frac{\partial \ln L(\theta; \mathbf{X})}{\partial \theta}$, we have $E(Z) = 0$, $Var(Z) = nI(\theta)$. So

$$k'(\theta) = E(Y \cdot Z) = Cov(Y, Z) = \rho \sigma_Y \sqrt{nI(\theta)}.$$

Because the correlation coefficient ρ satisfies $\rho^2 \leq 1$, it follows that

$$k'(\theta)^2 \leq \sigma_Y^2 \cdot nI(\theta).$$

For an unbiased estimator Y of θ we define its *efficiency* to be $1/\sigma_Y^2 nI(\theta)$ which is at most 1 and we call Y *efficient* when it is unbiased and with efficiency equal to 1. That is, its variance is as small as possible, namely equal to $1/nI(\theta)$.

Example 6.2.3- Consider the case with $X_i \sim Poiss(\theta)$ so that the mean and variance equal θ.

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i; \theta) = (\sum_i x_i) \ln \theta - n\theta - \sum_i \ln(x_i!)$$

$$\ell'(\theta; \mathbf{x}) = \frac{\sum_i x_i}{\theta} - n.$$

So the mle is $\hat{\theta} = \bar{X}$ with mean θ and variance θ/n. In particular, it is unbiased.

With $n = 1$

$$I(\theta) = E(\ell'(\theta)^2) = E((\frac{X-\theta}{\theta})^2) = \frac{\theta}{\theta^2} = \frac{1}{\theta}.$$

The Rao-Cramer lower bound for the variance is $\frac{1}{nI(\theta)} = \frac{\theta}{n} = Var(\bar{X})$.

Thus, $\hat{\theta}$ is an efficient estimator.

HMC Assumptions 6.2.2:

R5 There exists a continuous function $M(x)$ with $E_{\theta_0}(M(X)) < \infty$ such that for θ_1, θ_2 in the parameter interval

$$\left| \frac{\partial^2 \ln f(x;\theta_1)}{\partial \theta^2} - \frac{\partial^2 \ln f(x;\theta_2)}{\partial \theta^2} \right| \leq M(x)|\theta_1 - \theta_2|.$$

We will use the notation

$$\ell(\theta; \mathbf{X}) \;=\; \ln L(\theta; \mathbf{X}) \;=\; \sum_{i=1}^{n} \ln f(x_i; \theta),$$

$$\ell'(\theta; \mathbf{X}) \;=\; \frac{\partial \ln L(\theta; \mathbf{X})}{\partial \theta}, \quad \ell''(\theta; \mathbf{X}) \;=\; \frac{\partial^2 \ln L(\theta; \mathbf{X})}{\partial \theta^2}.$$

As we have seen above

$$E_\theta(\ell'(\theta; \mathbf{X})) \;=\; 0, \quad nI(\theta) = Var_\theta(\ell'(\theta; \mathbf{X})) = -E_\theta(\ell''(\theta; \mathbf{X})).$$

From the Central Limit Theorem and the Law of Large Numbers, we then get

$$\frac{1}{\sqrt{n}}\ell'(\theta_0; \mathbf{X}) \;\to\; \mathcal{N}(0, I(\theta_0))\,(D), \quad -\frac{1}{n}\ell''(\theta_0; \mathbf{X}) \;\to\; I(\theta_0)\,(P).$$

We refer to $\ell'(\theta; \mathbf{X}) = 0$ as the *MLE equation* which is satisfied by an interior maximum of $\ell(\theta; \mathbf{X})$.

HMC Theorem 6.2.2: Assume that $0 < I(\theta_0) < \infty$. If $\{\hat{\theta}_n(\mathbf{X})\}$ is a consistent sequence of solutions of the MLE equations, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \ \rightarrow \ \mathcal{N}(0, \frac{1}{I(\theta_0)}).$$

Proof: We apply the Mean Value Theorem to the function $\ell'(\theta; \mathbf{X})$.

$$0 = \ell'(\hat{\theta}) = \ell'(\theta_0) + (\hat{\theta} - \theta_0)\ell''(\theta^*)$$

with $\theta^*(\mathbf{X})$ between $\hat{\theta}$ and θ_0.

By Assumption R5:

$$n^{-1}|\ell''(\theta_0) - \ell''(\theta^*)| \leq \bar{M}(\mathbf{X})|\theta_0 - \theta^*|, \text{ with } \bar{M}(\mathbf{X}) = n^{-1}\sum_{i=1}^{n} M(X_i).$$

By the Law of Large Numbers, $\bar{M}(\mathbf{X}) \to E_{\theta_0}(M(X))$, and by consistency and the Squeeze Theorem $|\theta_0 - \theta^*| \to 0$ (P).

It follows that $-n^{-1}\ell''(\theta^*) \to I(\theta_0)$ (P).

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{n^{-1/2}\ell'(\theta_0)}{-n^{-1}\ell''(\theta^*)} \to \frac{1}{I(\theta_0)}\mathcal{N}(0, I(\theta_0)).$$

Thus,

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, 1).$$

HMC Corollary 6.2.2: $\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}I(\theta_0)}\ell'(\theta_0) + R_n$ with $R_n \to 0$ (P).

Proof: From the previous equation,

$$R_n = [n^{-1/2}\ell'(\theta_0)] \cdot [\frac{1}{-n^{-1}\ell''(\theta^*)} - \frac{1}{I(\theta_0)}].$$

The first factor tends to $\mathcal{N}(0, I(\theta_0))$ in distribution and the second tends to 0 in probability and so the product tends to 0 in probability.

Because $I(\theta)$ is a continuous function and $\hat{\theta}$ is consistent, $I(\hat{\theta}) \to I(\theta_0)$ (P).

Consequently,

$$\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, 1).$$

So we can use

$$(\hat{\theta}_n - z_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta}_n)}}, \hat{\theta}_n + z_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta}_n)}})$$

as an approximate $(1 - \alpha)100\%$ confidence interval for θ_0.

Sec. 6.3: Maximum Likelihood Tests

We now use the mle to test the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$.

NOTICE an annoying change in notation (I follow the book). Previously, θ_0 denoted the unknown true value of θ. But now θ_0 is a known value and we are testing to see whether it equals the true value.

We will restrict attention to cases where the likelihood $L(\theta; \mathbf{X})$ has a unique interior local maximum $\hat{\theta}$ at which, of course, the likelihood equation, $L' = 0$, holds. That is, the likelihood equation has a unique solution at which $L(\theta)$ has its maximum value.

We use the ratio $\Lambda = L(\theta_0)/L(\hat{\theta}) \leq 1$. If H_0 is true then since $\hat{\theta}$ is a consistent estimate, and so tends to the true value, Λ should be close to 1.

So for significance level α we Reject H_0 if $\Lambda \leq c$ with c chosen so that $P_{\theta_0}(\Lambda \leq c) = \alpha$. Notice that we use the fact that θ_0, as opposed to the true value, is known.

For Example 6.3.1: $f(x;\theta) = \theta^{-1}exp(-x/\theta)$. So with a sample of size n the likelihood function is $L(\theta) = \theta^{-n}exp(-n\bar{X}/\theta)$ and the MLE is $\bar{X}$. So

$$\Lambda = \frac{L(\theta_0)}{L(\bar{X})} = e^n(\frac{\bar{X}}{\theta_0})^n exp(-n\bar{X}/\theta_0) = g(\bar{X}/\theta_0)$$

with $g(t) = e^n t^n exp(-nt)$.

For the function $g(t)$ it is easier to look at the log:

$$(\ln g)(t) = n(1+\ln(t)-t);\ (\ln g)'(t) = n(\frac{1}{t}-1);\ (\ln g)''(t) = -\frac{n}{t^2}.$$

The maximum occurs at $t = 1$ with $g(1) = 1$. That is, the max of Λ is 1 when $\theta_0 = \hat{\theta} = \bar{X}$.

For $0 < c < 1$ there are two values c_1, c_2 with $g(c_1) = c = g(c_2)$ and $0 < c_1 < 1 < c_2$. Thus, $\Lambda \leq c$ if and only if $\bar{X}/\theta_0 \leq c_1$ or $\bar{X}/\theta_0 \geq c_2$.

Observe that $X/\theta \sim Exp(1) = \Gamma(1,1)$ and so $2X/\theta \sim \Gamma(1,2)$.

Therefore, $(2/\theta)\sum_{i=1}^{n} X_i \sim \Gamma(n,2) = \chi^2(2n)$.

Example 6.3.2: We consider the family $\{\mathcal{N}(\theta, \sigma^2)\}$ with σ^2 known. Observe that

$$\sum_i (x_i - \theta)^2 = \sum_i [(x_i - \bar{x}) + (\bar{x} - \theta)]^2 = \sum_i (x_i - \bar{x})^2 + \sum_i (\bar{x} - \theta)^2.$$

The cross term vanishes as usual because $\sum_i (x_i - \bar{x}) = 0$.

$$L(\theta) = (\frac{1}{2\pi\sigma^2})^{n/2} exp(-(2\sigma^2)^{-1} \sum_{i=1}^{n} (x_i - \theta)^2) = (\frac{1}{2\pi\sigma^2})^{n/2} \times$$

$$exp(-(2\sigma^2)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2) exp(-(2\sigma^2)^{-1} \sum_{i=1}^{n} n(\bar{x} - \theta)^2).$$

Since $\hat{\theta} = \bar{X}$,

$$\Lambda = L(\theta_0)/L(\hat{\theta}) = exp(-(2\sigma^2)^{-1} \sum_{i=1}^{n} n(\bar{X} - \theta_0)^2).$$

The condition $\Lambda \leq c$ is equivalent to $-2 \ln \Lambda \geq -2 \ln c$.

$$-2 \ln \Lambda = (\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}})^2 = W$$

which has a $\chi^2(1)$ distribution, as it is the square of a standard normal. Given α we choose Q_α so that $P(W \geq Q_\alpha) = \alpha$. Then with $-2 \ln c = Q_\alpha$ we obtain our test of size α.

HMC Theorem 6.3.1: Under the null hypothesis that θ_0 is the true value of θ,

$$-2 \ln \Lambda \ \rightarrow \ \chi^2(1) \ (D).$$

Proof: We use Taylor's Theorem (the second order version of the Mean Value Theorem to get:

$$\ell(\hat{\theta}) = \ell(\theta_0) + (\hat{\theta} - \theta_0)\ell'(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^2\ell''(\theta_n^*),$$

with θ_n^* between $\hat{\theta}$ and θ_0.

We are assuming that θ_0 is the true value. So as in the proof of Theorem 6.2.2, we use Assumption R5 and the Squeeze Theorem to get $-n^{-1}\ell''(\theta_n^*) \rightarrow I(\theta_0)$ (P). Next, Corollary 6.2.3 says that

$$n^{-1/2}\ell'(\theta_0) = n^{1/2}(\hat{\theta} - \theta_0)I(\theta_0) + R_n$$

with $R_n \rightarrow 0$ (P) Therefore

$$-2\ln\Lambda \;=\; 2(\ell(\hat{\theta}) - \ell(\theta_0)) \;=$$

$$2[\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)]^2 \cdot [1 - (\frac{1}{2}\ell''(\theta_n^*)/nI(\theta_0)] \;+\; R_n^*$$

$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \;\to \mathcal{N}(0,1)$ (D) and so
$[\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)]^2 \;\to\; \chi^2(1)$ (D).

$1 - (\frac{1}{2}\ell'(\theta_n^*)/nI(\theta_0)) \;\to\; \frac{1}{2}$ (P). Finally,

$R_n^* = \sqrt{n}(\hat{\theta} - \theta_0) \cdot R_n$. Because
$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, 1/I(\theta_0)$ (D) and $R_n \to 0$ (P), the product R_n^* tends to 0 in probability as well.

With χ^2_L defined to be $-2 \ln \Lambda$, we have under the null hypothesis that $\chi^2_L \rightarrow \chi^2(1)$ (D) and so we can use the test: Reject H_0 if $\chi^2_L \geq \chi^2_\alpha(1)$. Here $c = \chi^2_\alpha(1)$ is the value such that with $Z^2 \sim \chi^2(1)$.

$$F_{Z^2}(c) = P(Z^2 \leq c) = 1 - \alpha.$$

Under the null hypothesis $I(\hat{\theta}) \rightarrow I(\theta_0)$ and so with χ^2_W defined to be $[\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0)]^2$ we also have that $\chi^2_W \rightarrow \chi^2(1)$ (D) and so we can use the test: Reject H_0 if $\chi^2_W \geq \chi^2_\alpha(1)$.

We can also define $\chi^2_R = [\ell'(\theta_0)/\sqrt{nI(\theta_0)}]^2$.

The difference between any two of the statistics χ^2_L, χ^2_W and χ^2_R tends to 0 in probability under the null hypothesis.