Contents

1 Sufficiency and Minimal Sufficiency

1.1 Sufficiency

Model-based inference depends on a **sample space** (collection of possible outcomes) and a **model space / distribution family**: $(\mathcal{X}, \mathcal{P})$. θ denotes all the unknown parameters in the model space, and fixing θ produces a particular distribution.

In Fisher's framework, we aim to *infer* the unknown parameter, θ, using the data, X and the sample space and distn family: $(\mathcal{X}, \mathcal{P})$. Fisher's goal was to reduce the data X into a simple statistic, $T(X)$ such that no information related to inferring θ would be lost. This idea formally relied on the idea of a **sufficient statistic**.

Definition 1 (Two definitions of Sufficiency)**.** A statistic, $T(X)$, is sufficient if no information related to inferring θ is lost when converting from X to $T(X)$. Formally:

(i) **Def 1**: a statistic, $T(X)$ is sufficient if given $T(X)$, we can generate new data X^*, based only on knowing $T(X)$, such that $X^* \overset{D}{=} X$.

(ii) **Def 2**: $T(X)$ is a sufficient statistic if $X|T(X)$ does not depend on θ.

Verifying these two definitions can be somewhat cumbersome. Thankfully, there is an extremely useful theorem that helps us prove sufficiency: **the factorization theorem**.

Theorem 1 (Fisher-Neyman Factorization Theorem)**.** A statistic, $T(X)$ is a sufficient statistic for $(\mathcal{X}, \mathcal{P})$ (or θ) if and only if the pdf/pmf $f_\theta(x)$ factors:

$$\boxed{f_\theta(X) = g_\theta(T(X)) \cdot h(X)}$$

where $g_\theta(T(X))$ depends on X only through $T(X)$ and $h(X)$ does not depend on θ.

Property 1 (1-1 maps of SS, Remark 11.2 Perlman)**.** If $T(X)$ is a sufficient statistic for θ and f is a 1-1 map, $f(T(X))$ is also a sufficient statistic for θ. This is because $T(X)$ and $f(T(X))$ yield the same partitioning on the sample space $\mathcal{X}$.

Note that by this property, we conclude that sufficient statistics are *not unique*.

Definition 2 (Sufficiency of Order Statistics, Ex. 11.6 Perlman)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} f \in \mathcal{P}$ where $\mathcal{P}$ is the class of *permutation-invariant* pdfs, meaning sequence ordering of random variables is irrelevant. In other words, if you shuffle the order of input variables you won't change the resulting distribution:

$$f(X_1, \ldots, X_n) = f(X_{\pi_1}, \ldots, X_{\pi_n})$$

for all permutations π. If $\mathcal{P}$ is the class of sequence-invariant pdfs, then the **order statistics**, $(X_{(1)}, \ldots, X_{(n)})$ is a sufficient statistic for $\mathcal{P}$.

Theorem 2 (SS for subfamilies - Lemma 11.1 Perlman)**.** If $T(X)$ is a sufficient statistic with respect to $\mathcal{P}$ and if $\mathcal{P}_1 \subset \mathcal{P}$ (i.e., is a subfamily of $\mathcal{P}$), then $T(X)$ is a sufficient statistic with respect to $\mathcal{P}_1$.

Theorem 3 (SS for 2-step data reduction - Lemma 11.2 Perlman)**.** If $T(X)$ is a sufficient statistic for $(X, \mathcal{P})$ and $S(T(X))$ is a sufficient statistic for $(T, \mathcal{Q})$, $S(T(X))$ is also a sufficient statistic for $(X, \mathcal{P})$.

Sufficient statistics also have nice interpretations vis-a-vis likelihood ratios:

Property 2 (Sufficient stats and Likelihood ratios)**.** Suppose $\mathcal{P} = \{f_\theta(x) | \theta \in \Omega\}$ is a general statistical model, and we would like to choose between $\theta = \theta_1, \theta_2$. We could use the Likelihood ratio (LR) to base our decision:

$$\begin{aligned} L_{\theta_1,\theta_2}(x) &= \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} \\ &= \frac{g_{\theta_2}(T(X))}{g_{\theta_1}(T(X))} \quad \text{(By factorization thm)} \end{aligned}$$

Implying that the likelihood ratio depends on X only through the value of the sufficient statistic $T(X)$.

1.2 Minimal sufficiency

There are many sufficient statistics for θ, but which one is minimal, i.e., every other SS can be reduced to it?

Definition 3 (Minimal sufficiency)**.** $T^*(X)$ is a **minimal sufficient statistic** for $\mathcal{P}$ if for any sufficient statistic $T(X)$, there exists a function $h(\cdot)$, such that $T^*(X) = h(T(X))$, i.e., $T^*(X)$ is a reduction of $T(X)$.

Note that the minimal sufficient statistic may not be unique. There can be many minimal sufficient statistics.

Is the MSS guaranteed to exist? Yes, according to Fisher's likelihood principle.

Definition 4 (Fisher's Likelihood Principle)**.** Fisher's Likelihood Principle guarantees the existence of a minimal sufficient statistic. Fihser's likelihood principle says that

$$T^{**}(\cdot) = \{L_{\theta_1,\theta_2}(\cdot) := \frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}; \theta_1, \theta_2 \in \Theta\}$$

In other words, the set of all pairwise likelihood ratios is a minimal sufficient statistic for $\mathcal{P}$.

Pf sketch:

1. First we show $T^{**}(X)$ is a SS. Note L_{θ_1,θ_2} is a function of θ_1, θ_2. Set θ_1 to be fixed and let $\theta = \theta_2$ vary over Θ.

$$\begin{aligned} f_\theta(x) &= \underbrace{L_{\theta_1,\theta_2}(x)}_{=\frac{f_\theta(x)}{f_{\theta_1}(x)}} \cdot f_{\theta_1}(x) \\ &= g_\theta(T(X)) \cdot h(x) \implies T(X) \text{ is SS} \end{aligned}$$

2. Next, we show for all SS $T(X)$, $T^{**}(X) = h(T(X))$.

$$\begin{aligned} L_{\theta_1,\theta_2} &= \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} \\ &= \frac{g_{\theta_2}(T(X))}{g_{\theta_1}(T(X))} = \frac{g_{\theta_2}}{g_{\theta_1}}(T(X)) \end{aligned}$$

How do we find the MSS? We use the Lehmann-Scheffe theorem.

Theorem 4 (Lehmann-Scheffé theorem)**.** Suppose $X \sim \{f_\theta(x), \theta \in \Omega\}$ $T(X)$ is a minimal sufficient statistic if

$$T(X) = T(Y) \iff \frac{f_\theta(y)}{f_\theta(x)} \text{is } \theta\text{-free}$$

This usually requires an iff proof!

Note: the L-S Theorem is a sufficient criterion for MSS, but is not a necessary criterion! In other wrods, showing that Lehmann-Scheffe holds proves a statistic is minimal sufficient, but showing a statistic does not satisfy Lehmann-Scheffe does not mean it is not minimal.

Pf sketch Suppose $T(X) = T(Y) \iff \frac{f_\theta(y)}{f_\theta(x)} \text{is } \theta\text{-free}$

$$\begin{aligned} T(X) = T(Y) &\iff \frac{f_\theta(y)}{f_\theta(x)} \text{is } \theta\text{-free} \\ &\iff \forall \theta_1, \theta_2 \in \Omega, \frac{f_{\theta_1}(y)}{f_{\theta_1}(x)} = \frac{f_{\theta_2}(y)}{f_{\theta_2}(x)} \\ &\iff \forall \theta_1, \theta_2 \in \Omega, \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = \frac{f_{\theta_2}(y)}{f_{\theta_1}(y)} \\ &\iff L_{\theta_1,\theta_2}(x) = L_{\theta_1,\theta_2}(y) \; \forall \theta_1, \theta_2 \in \Omega \\ &\iff T^{**}(X) = T^{**}(Y) \end{aligned}$$

This implies $T(X) = T(Y) \iff T^{**}(X) = T^{**}(Y)$, so T must be MSS.

Here's a worked example of the Lehmann-Scheffe theorem: suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$. Show $\bar{X}_n$ is

minimal sufficient. We start by writing the joint pdf:

$$\begin{aligned} f_\theta(\mathbf{x}) &= \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_i-\theta)^2}{2}}\right] \\ &= \frac{1}{(2\pi)^{n/2}}e^{-\frac{\sum(x_i-\theta)^2}{2}} \\ &= \frac{1}{(2\pi)^{n/2}}e^{-\frac{\sum(x_i-\bar{x}_n+\bar{x}_n-\theta)^2}{2}} \\ &= \underbrace{\frac{1}{(2\pi)^{n/2}}e^{-\frac{\sum(\bar{x}_n-x_i)^2}{2}}}_{h(x)}\underbrace{e^{-\frac{n(\bar{x}_n-\theta)^2}{2}}}_{g_\theta(T(X))} \\ &\implies T(X)=\bar{X}_n \text{ is SS} \end{aligned}$$

Now Lehmann-Scheffe to show MSS:

$$\begin{aligned} \frac{f_\theta(y)}{f_\theta(x)} &= \frac{e^{-\frac{\sum(y_i-\mu)^2}{2}}}{e^{-\frac{\sum(x_i-\mu)^2}{2}}} \\ &= e^{\frac{-\sum(x_i-\mu)^2+\sum(y_i-\mu)^2}{2}} \\ &= e^{\frac{\sum y_i^2-x_i^2}{2}}\cdot e^{\mu(\sum x_i-\sum y_i)} \end{aligned}$$

If $\sum X_i = \sum Y_i$ is true, it's trivial to see that $\frac{f_\theta(y)}{f_\theta(x)}$ is μ-free. If $\frac{f_\theta(y)}{f_\theta(x)}$ is μ-free: then

$$\begin{aligned} \mu(\sum x_i - \sum y_i) &= C \quad \text{where C is const wrt } \theta \\ \implies \frac{d}{d\mu}\mu(\sum x_i - \sum y_i) &= \frac{d}{d\mu}C \\ \implies \sum x_i = \sum y_i &\implies T(X)=T(Y) \end{aligned}$$

Strategy 1 (Finding the MSS)**.**

1. Write the joint likelihood
2. Find the SS using factorization theorem
3. Use the L-S theorem to show minimal sufficiency:

Property 3 (1-D Exponential Family)**.** Suppose you have, $X_1,\ldots,X_n \overset{iid}{\sim} f_\theta$ where $\theta \in \mathbb{R}^1$ of exponential form:

$$f_\theta(x) = a(\theta)\exp[\theta T(x_i)]\cdot h(x)$$

with joint pdf:

$$f_\theta(\mathbf{x}) = [a(\theta)]^n \exp\left[\theta\sum_{i=1}^{n}T(x_i)\right]\prod_{i=1}^{n}h(x_i)$$

Then, $\sum_{i=1}^n T(X_i)$ is a minimal sufficient statistic. Thus,

Dist: $N_1(\mu, 1)$	$\theta = \mu$	MSS: $\sum_{i=1}^n X_i$
Dist: $N_1(0, \sigma^2)$	$\theta = -\frac{1}{2\sigma^2}$	MSS: $\sum_{i=1}^n X_i^2$
Dist: Binomial(n, p)	$\theta = \log\left(\frac{p}{1-p}\right)$	MSS: $\sum_{i=1}^n X_i$
Dist: Poisson(λ)	$\theta = \log(\lambda)$	MSS: $\sum_{i=1}^n X_i$
Dist: Exponential(λ)	$\theta = -\lambda$	MSS: $\sum_{i=1}^n X_i$

Property 4 (Example 11.13 Perlman - k-dimensional MSS)**.** Suppose you have, $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta$ where $\theta \in \mathbb{R}^k$ of exponential form:

$$f_\theta(x) = a(\theta_1, \ldots, \theta_k) \exp[\theta_1 T_1(x) + \ldots + \theta_k T_k(x)] \cdot h(x)$$

with joint pdf:

$$f_\theta(\mathbf{x}) = [a(\theta)]^n \exp\left[\theta_1 \sum_{i=1}^n T_1(x_i) + \theta_k \sum_{i=1}^n T_k(x_i)\right] \prod_{i=1}^n h(x_i)$$

Then, $(\sum T_1(X_i), \ldots, \sum T_k(X_i))$ is a k-dimensional minimal sufficient statistic *provided that the natural parameter space* $\Omega \subset \mathbb{R}^k$ *affinely spans* $\mathbb{R}^k$, meaning that the natural parameter space cannot be contained by a linear subspace (hyperplane) of dimension $\leq k-1$.

See the following example $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, with $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$ both unknown.

$$f_{\mu,\sigma^2}(x_1, \ldots, x_n) = \frac{e^{-\frac{n\mu^2}{2\sigma^2}}}{(2\pi)^{-n/2}\sigma^n} \cdot \exp\left(\theta \cdot \begin{pmatrix} \sum X_i \\ \sum X_i^2 \end{pmatrix}\right) \quad \text{with } \theta = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}$$

Since the natural parameter vector is not a linear subspace of $\mathbb{R}^2$, then $(\sum X_i, \sum X_i^2)$ is a MSS.

If we impose the restriction that $\sigma^2 = \mu^2 (\mu \neq 0)$ on Ω, then $X_i \sim N_1(\mu, \mu^2)$ still yields $(\sum X_i, \sum X_i^2)$ as the sufficient statistic but the natural parameter space is:

$$(\theta_1, \theta_2) = \left(\frac{1}{\mu}, -\frac{1}{2\mu^2}\right)$$
$$\implies \Omega = \left\{(\theta_1, \theta_2) \;\middle|\; \theta_2 = -\frac{\theta_1^2}{2}, \theta_1 \neq 0\right\}$$

Since Ω is a parabola in $\mathbb{R}^2$, it cannot be contained in a linear subspace of dimension ≤ 1, so the parameter space affine spans $\mathbb{R}^2$ and $(\sum X_i, \sum X_i^2)$ remains minimal sufficient.

Property 5 (1-parameter truncation family)**.** Let $X_1, \ldots, X_n$ be an iid sample from a distribution with pdf of the truncation form:

$$f_\theta(x) = \frac{b(x)\mathbb{I}(x \in [a, \theta])}{B(\theta)} \qquad x > a$$

where $b(x) > 0$ is any function, $a \in \mathbb{R}$ is known, such that for an where normalizing constant $B(\theta) = \int_a^\theta b(x)dx < \infty$. Note that $T(X) = X_{(n)}$ is a MSS with respect to this truncation family!

Similarly, $T = X_{(1)}$ is minimal sufficient if $X_1, \ldots, X_n$ is an iid sample from:

$$f_\theta(x) = \frac{b(x)\mathbb{I}(x \in [\theta, a])}{B(\theta)} \qquad x < a$$

where $b(X) > 0$ is any positive function on (∞, a), a is known, and $B(\theta) = \int_\theta^a b(x)dx < \infty$.

Property 6 (2-parameter truncation family)**.** Let $X_1, \ldots, X_n$ be an iid sample from a distribution with a pdf of the form:

$$\frac{b(x)\mathbb{I}(x \in [\theta_1, \theta_2])}{B(\theta_1, \theta_2)}$$

with finite real valued parameters, $\theta_1 < \theta_2$, $b(x) > 0$, and normalizing constant $B(\theta_1, \theta_2) = \begin{cases} \int_{\theta_1}^{\theta_2} b(x)dx < \infty \\ 0 \quad \text{else} \end{cases}$

where normalizing constant $B(\theta) = \int_a^\theta b(x)dx < \infty$.

Then $T(X) = (X_{(1)}, X_{(n)})$ is a MSS with respect to this truncation family!

Property 7 (Ancillary precision)**.** If we have a two-dimensional minimal sufficient statistic (T_1, T_2) that can be mapped via a 1-1 function to (U, V) where V is ancillary, then the joint pdf of (U, V) can be written as $f_\theta(u, v) = f_\theta(u|v) \cdot f(v)$ and the likelihood ratio:

$$\begin{aligned} L_{\theta_1,\theta_2}(u, v) &= \frac{f_{\theta_2}(u, v)}{f_{\theta_1}(u, v)} \\ &= \frac{f_{\theta_2}(u|v)}{f_{\theta_1}(u|v)} \end{aligned}$$

Suggesting that efficient inference about θ can be achieved from the conditional distribution of $U|V$.

For example, consider the distribution Uniform$[\theta, \theta + 1]$ has $(X_{(1)}, X_{(n)})$ is a minimal sufficient statistic, but can be 1-1 mapped to $(X_{(1)}, R = X_{(n)} - X_{(1)}$, where the sample range is ancillary. R provides no information about *theta*, yet it governs the accuracy of $X_{(1)}$:

$$\begin{aligned} \theta + R \leq X_{(1)} + R &\leq 1 + \theta \\ \theta \leq X_{(1)} &\leq 1 + \theta - R \\ 0 \leq X_{(1)} - \theta &\leq 1 - R \end{aligned}$$

2 Ancillarity, Completeness, & minimum variance unbiased estimation

2.1 Ancillary statistics

Definition 5 (Ancillary Statistic)**.** A statistic $V(X)$ is an *ancillary statistic* with respect to a distribution family $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ if the distribution of $V(X)$ is θ-free.

In other words, for any $A \subset \mathcal{X}$ and any integrable function $g : \mathcal{X} \to \mathbb{R}$, then $P_\theta(V(X) \in A)$ and $\mathbb{E}_\theta(g(V(X)))$ don't depend on θ.

Property 8 (Location family and ancillary statistic - Perlman Ex. 12.1)**.** A *location family* is a family of distributions that after a location shift, reduces to the same known distribution P_0: $P_\mu = P_0 + \mu$ (e.g., $N(\mu, 1)$ is a location-family with $P_0 = N(0, 1)$). In other words:

$$\text{Marginal: } \{f\mu(x) := f_0(x - \mu) \mid \mu \in \mathbb{R}\}$$
$$\text{Joint: } \left\{ f\mu(x_1, \ldots, x_n) := \prod_{i=1}^{n} f_0(x_i - \mu) \mid \mu \in \mathbb{R} \right\}$$

For any location family, **ANY** statistic, $V(\cdot)$ satisfying the **location invariant property**:

$$V(X_1, \ldots, X_n) = V(X_1 + \mu, \ldots, X_n + \mu) \text{ for all } \mu \in \Omega$$

is an ancillary statistic.

See the following example. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$.

(i) The **sample range**, $V(X) = X_{(n)} - X_{(1)}$, is ancillary because:

$$\begin{aligned} V(X) &= (X_{(n)} - \mu) - (X_{(1)} - \mu) \\ &= \max \begin{pmatrix} x_1 - \mu \\ \vdots \\ x_n - \mu \end{pmatrix} - \min \begin{pmatrix} x_1 - \mu \\ \vdots \\ x_n - \mu \end{pmatrix} \\ &\implies \text{and } (x_1 - \mu, \ldots, x_n - \mu) \sim N(0, 1) \text{ is free of } \mu \\ &\implies \max \begin{pmatrix} x_1 - \mu \\ \vdots \\ x_n - \mu \end{pmatrix} - \min \begin{pmatrix} x_1 - \mu \\ \vdots \\ x_n - \mu \end{pmatrix} \text{ is } \mu\text{-free} \\ &\implies V(X) \text{ is } \mu\text{-free} \end{aligned}$$

(ii) The **sample spacings** are ancillary:

$$V'(X) = \begin{pmatrix} X_{(2)} - X_{(1)} \\ \vdots \\ X_{(n)} - X_{(n-1)} \end{pmatrix} = \begin{pmatrix} (X_{(2)} - \mu) - (X_{(1)} - \mu) \\ \vdots \\ (X_{(n)} - \mu) - (X_{(n-1)} - \mu) \end{pmatrix} \text{ is } \theta\text{-free}$$

(iii) The **sample variance**, $V''(X) = S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is ancillary:

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^n ((X_i - \mu) - (\bar{X}_n - \mu))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^n \left((X_i - \mu) - \frac{1}{n}\sum_{j=1}^n (X_j - \mu)\right)^2 \text{ is } \theta\text{-free} \end{aligned}$$

Property 9 (Scale Family and ancillary statistic)**.** A *scale family* is a family of distribution that, after a scale shift, reduces to the same known distribution P_0: $P_\sigma = \sigma \cdot P_0$ (e.g., $N(0, \sigma^2)$ is a scale family with $P_0 = N(0,1)$). In other words:

$$\text{Marginal: } \{f\sigma(x) := \sigma^{-1} f_0\left(\sigma^{-1}x\right) \mid \sigma \in \mathbb{R}^+\}$$

$$\text{Joint: } \left\{f\sigma(x_1, \ldots, x_n) := \sigma^{-n}\prod_{i=1}^n f_0\left(\sigma^{-1}x_i\right) \mid \sigma \in \mathbb{R}^+\right\}$$

For any scale family, **ANY** statistic $V(\cdot)$ satisfying the **scale invariant property**:

$$V(X_1, \ldots, X_n) = V(\sigma X_1, \ldots, \sigma X_n) \text{ for all } \sigma \in \Omega$$

Here are some examples of ancillary statistics for scale families:

(i) The **ordered ratios** are ancillary for the scale family:

$$\begin{aligned} V(X) &= \left(\frac{X_{(1)}}{X_{(n)}}, \ldots, \frac{X_{(n-1)}}{X_{(n)}}\right) \\ &= \left(\frac{\sigma X_{(1)}}{\sigma X_{(n)}}, \ldots, \frac{\sigma X_{(n-1)}}{\sigma X_{(n)}}\right) \\ &\implies V(X) = V(\sigma X)\ \forall \mu \in \mathbb{R} \end{aligned}$$

(ii) The **t-statistic** is ancillary for the scale family:

$$\begin{aligned} V(X) &= \frac{\bar{X}_n}{s_n} \\ &= \frac{\sigma \bar{X}_n}{\sigma s_n} \\ &\implies V(X) = V(\sigma X)\ \forall \mu \in \mathbb{R} \end{aligned}$$

Property 10 (Location-scale family and ancillary statistic)**.** A *location-scale family* is a family of distribution that, after a location and scale shift, reduces to the same known distribution P_0: $P_{\mu,\sigma} = \sigma \cdot P_0 + \mu$ (e.g., $N(\mu, \sigma^2)$ is a scale family with $P_0 = N(0,1)$). In other words:

$$\text{Marginal: } \{f_{\mu,\sigma}(x) := \sigma^{-1} f_0\left(\sigma^{-1}(x-\mu)\right) \mid (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}^+\}$$
$$\text{Joint: } \left\{f_{\mu,\sigma}(x_1,\ldots,x_n) := \sigma^{-n}\prod_{i=1}^n f_0\left(\sigma^{-1}(x_i-\mu)\right) \mid (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}^+\right\}$$

Any statistic $V(\cdot)$ that is **location-scale invariant**:

$$V(X_1,\ldots,X_n) = V(\sigma X_1+\mu,\ldots,\sigma X_n+\mu)\ \forall\ (\mu,\sigma)\in\Omega$$

is ancillary for the location-scale family

Here are some examples of ancillary statistics for the location-scale family:

1. **Normalized sample spacings** are ancillary:

$$\begin{aligned} V(X) &= \left(\frac{X_{(2)}-X_{(1)}}{X_{(n)}-X_{(1)}},\ldots,\frac{X_{(n)}-X_{(n-1)}}{X_{(n)}-X_{(1)}},\right) \\ &= \left(\frac{\sigma(X_{(2)}-\mu-(X_{(1)}-\mu))}{\sigma(X_{(n)}-\mu-(X_{(1)}-\mu))},\ldots,\frac{\sigma(X_{(n)}-\mu-(X_{(n-1)}-\mu))}{\sigma(X_{(n)}-\mu-(X_{(1)}-\mu))},\right) \\ &\implies V(X) = V(\sigma X+\mu)\ \forall\ (\mu,\sigma)\in\Omega \end{aligned}$$

2. **Sample range sample standard deviation ratio** is ancillary:

$$\begin{aligned} V(X) &= \frac{X_{(n)}-X_{(1)}}{s_n} \\ &= \frac{\sigma(X_{(n)}-\mu-(X_{(1)}-\mu))}{\sigma\sqrt{\frac{1}{n-1}\sum_{i=1}^n((X_i-\mu)-(\bar{X}_n-\mu))^2}} \\ &\implies V(X) = V(\sigma X-\mu)\ \forall\ (\mu,\sigma)\in\Omega \end{aligned}$$

2.2 Complete statistics

A complete statistic is the opposite of an ancillary statistic. Note: a complete statistic may not be a sufficient statistic: e.g., $T(X)=47$.

Definition 6 (Complete statistic)**.** Given a statistical model $(\mathcal{X},\mathcal{P})$, a statistic $T(X)$ is said to be complete with respect to $\mathcal{P}$ if considering any function g, we have:

$$\mathbb{E}_\theta[g(T(X))] \text{ is } \theta\text{-free} \implies g(T(X)) \text{ is a constant function}$$

This criterion can be simplified: $T(X)$ is complete if:

$$\forall\theta\in\Omega \qquad \mathbb{E}_\theta(g(T))=0 \implies g(T)=0$$

This definition illuminates how complete and ancillary statistics are opposites. For example, an ancillary statistic yields: $\mathbb{E}_\theta(g(V(X)))$ is θ-free $\forall g$. A complete statistic yields $\mathbb{E}_\theta(g(V(X)))$ is θ-free ONLY IF g is constant.

The following are two important theorems relating complete and ancillary statistics:

Theorem 5 (Completeness versus ancillarity, Basu's theorem, relation to MSS)**.**

1. **Theorem 1**: If T is complete with respect to $\mathcal{P}$, then there exists no non-constant function of T that is ancillary.
2. **Theorem 2 (Basu's theorem)**: If T is a complete sufficient statistic with respect to $\mathcal{P}$, then we have for any P_θ, T is independent of **ANY** ancillary statistic V.
3. **Theorem 3**: If T is a complete sufficient statistic, then it is also minimally sufficient.

Here's an example proving complete sufficiency: Suppose $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$. I claim that $T = \sum_{i=1}^n X_i$ is CSS.

(i) Prove SS:

$$\begin{aligned} f_\theta(\mathbf{X}) &= \prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i} \\ &= \theta^{\sum X_i}(1-\theta)^{n-\sum X_i} \\ &\implies \sum X_i \text{ is CSS} \end{aligned}$$

(ii) Prove completeness:

$$\begin{aligned} &\text{We know } T(X) = \sum X_i \sim \text{Bin}(n, \theta) \\ &\implies \text{If } \mathbb{E}_\theta(g(T)) = 0 \\ &\implies \sum_{t=0}^n g(T) P_\theta(T = t) = 0 \\ &\implies \sum_{t=0}^n g(T) \binom{n}{t} \theta^t (1-\theta)^{n-t} = 0 \\ &\implies (1-\theta)^n \sum_{t=0}^n g(T) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t = 0 \\ &\implies \sum_{t=0}^n g(T) \binom{n}{t} x^t = 0 \quad \forall x \in (0, \infty) \\ &\implies \text{by taking derivative of polynomial wrt } X, \text{ all polynomial coefficients, } g(T), \text{ must be 0-valued} \\ &\implies g(T) = 0 \end{aligned}$$

Here's another example: Let $X \sim \text{Unif}(\{1, \dots, \theta\})$. Then $f(x) = \frac{1}{\theta} I(x \in \{1, \dots, \theta\}$.

1. If $\theta \in \mathbb{N}$, then X is CSS. Then prove completeness by induction:

$$\theta = 1: \quad \mathbb{E}(g(X)) = \sum_{X=1}^{\theta} g(X) \cdot 1/\theta = g(1) = 0$$
$$\implies g(1) = 0$$
$$\theta = 2: \quad \mathbb{E}(g(X)) = \sum_{X=1}^{\theta} g(X) \cdot 1/\theta = \frac{g(1)+g(2)}{2} = 0$$
$$\implies g(1) + g(2) = 0 \implies g(2) = 0$$
Suppose $\theta = n$ and $g(0), \ldots, g(n) = 0$
$$\theta = n+1: \quad \mathbb{E}(g(X)) = \sum_{X=1}^{\theta} g(X) \cdot 1/\theta = \frac{g(1)+\ldots+g(n+1)}{n+1} = 0$$
$$\implies g(n+1) = 0$$

   Thus, we showed that for all $\theta \in \Omega$, $E(g(X)) = 0 \implies g(X) = 0$. Thus, X is complete.

2. If $\theta \in \mathbb{N} - \{7\}$, then X is NOT CSS. Consider:

$$g(T) = \begin{cases} 1 & \text{if } \theta = 7 \\ -1 & \text{if } \theta = 8 \\ 0 & \text{else} \end{cases}$$

   Then $\mathbb{E}(g(T)) = 0$ for all $\theta \in \Omega_2$ but g is not constant. Thus, X cannot be complete.

Turns out, we have a general result for statistics with exponential family distributions.

Property 11 (CSS for 1-param/k-param exponential family - Perlman Prop 12.1)**.** (i) 1-param: Let $T(X)$ (**the statistic**) have pdf/pmf of the exponential form:

$$f_\theta(t) = a(\theta)e^{\theta t}h(t), \qquad \theta \in \Omega \subset \mathbb{R}^1$$

If Ω (the natural parameter space) contains a nondegenerate interval (a,b), then T is complete.

(ii) k-param: Let $T = (T_1, \ldots, T_k)$ have a pdf/pmf of exponential form:

$$f_\theta(t_1, \ldots, t_k) = a(\theta)e^{\theta_1 t + \cdots + \theta_k t_k}h(t_1, \ldots, t_k), \qquad \theta \in \Omega \subset \mathbb{R}^k$$

If Ω (the natural parameter space) contains a nondegenerate k-dimensional rectangle, then T is complete.

But it is not always easy to write the distribution of a statistic. Thankfully, we have a result that shows that when the *data* are distributed according to an exponential distribution, then can prove completeness:

Property 12 (CSS for general k-param exponential family - Perlman Prop 12.2)**.** Let X have pdf/pmf of exponential form:

$$f_\theta(x) = a(\theta)e^{\theta_1 T_1(x) + \cdots + \theta_k T_k(x)}h(x), \qquad \theta \in \Omega \subset \mathbb{R}^k$$

Where Ω (the natural parameter space) contains a k-dimensional open rectangle, then $T(X) = (T_1(X), \ldots, T_k(X))$ is complete.

Here are another couple good examples:

(i) Bivariate normal distribution: let $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ be iid sample from bivariate normal distribution. Note this constitutes a 4-parameter exponential family model:

$$N_2\left[\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix}\right]$$

has CSS $\bar{X}_n, s_n^2, \bar{Y}_n^2, t_n^2$ and the sample correlation r_n is ancillary.

We also have a result for data distributed according to a 1 and 2-parameter truncation families.

Property 13 (CSS for truncation families)**.**

(i) 1-parameter truncation family: Let $X_1, \ldots, X_n$ be an iid sample from the truncation pdf:

$$f_\theta(x) = \frac{b(x)\mathbb{I}(a < x \leq \theta)}{B(\theta)} \qquad x > a$$

where $a \in [-\infty, \infty)$ is specified, $\theta \in (a, \infty)$ is a real parameter, $b(x) > 0$ and $B(\theta) = \int_a^\theta b(x)dx < \infty; \forall \theta > a$. Then $T(X) = X_{(n)}$ is complete sufficient for θ. Similarly, $T(X) = X_{(1)}$ is complete sufficient if the indicator $\mathbb{I}(\theta \leq x < a)$.

(ii) 2-parameter truncation family: Let $X_1, \ldots, X_n$ be iid from truncation pdf:

$$f_{\theta_1,\theta_2}(x) = \frac{b(x)\mathbb{I}(\theta_1 \leq x \leq \theta_2)}{B(\theta_1, \theta_2)} \qquad \infty < x < \infty$$

Where $-\infty < \theta_1 < \theta_2 < \infty$, $b(x) > 0$, $B(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} b(x)dx < \infty$. Then $T(X) = (X_{(1)}, X_{(n)})$ is complete sufficient.

Property 14 (CSS Location-scale Exponential)**.** A location-scale exponential combines features of 1-parameter exponential and truncation families. Let $f_0(x) = e^{-x}\mathbb{I}(0 < x < \infty)$. Now the joint pdf of the location-scale exponential is:

$$\begin{aligned} f_{\mu,\sigma}(x_1, \ldots, x_n) &= \prod_{i=1}^n \sigma^{-1} e^{-x_i/\sigma} \left[\int_\mu^\infty \sigma^{-1} e^{-x_i/\sigma} dx_i\right]^{-1} \mathbb{I}(\mu \leq x_i < \infty) \\ &= \prod_{i=1}^n \sigma^{-1} e^{-x_i/\sigma} \left[e^{-\mu/\sigma}\right]^{-1} \mathbb{I}(\mu \leq x_i < \infty) \\ &= \prod_{i=1}^n \sigma^{-1} e^{-(x_i-\mu)/\sigma} \mathbb{I}(\mu \leq x < \infty) \\ &\implies (\sum X_i, X_{(1)}) \text{ is complete!} \end{aligned}$$

Property 15 (CSS nonparametric models - Perlman Ex. 12.9)**.** (i) $\mathcal{P} = \{$all symmetric pdfs $f(-x) = f(x)$ on $\mathbb{R}^1\}$, then $T(X) = |X|$ is complete for f_+ and $\psi := \text{sign}(X)$ is ancillary.

(ii) $\mathcal{P} = \{\text{all exchangeable pdfs } f(\pi x) = f(x) \text{ on } \mathbb{R}^n\}$, then $T(X) = |X|$ is complete and $\Pi := \text{rank}(X_1, \ldots, X_n)$ is ancillary.

(iii) $\mathcal{P} = \{\text{all radial pdfs } f(x) = g(||x||) \text{ on } \mathbb{R}^n\}$, then $T(X) = ||X||$ is complete for g and the unit vector $\vec{X} := \frac{X}{||X||}$ is ancillary.

2.3 Minimum Variance Estimation (UMVUE)

Suppose we are interested in estimating $\tau(\theta)$ based only on the data $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$. How do we measure the estimation accuracy? We do so with the mean-squared error, which has a very nice decomposition property:

Definition 7 (Mean squared error and bias-variance tradeoff)**.** For any estimator $\tilde{\tau}$ of $\tau(\theta)$, the MSE is:

$$\text{MSE}_\theta(\tilde{\tau}) = \mathbb{E}_\theta\left[(\tilde{\tau}(x) - \tau(\theta))^2\right]$$

Note that the MSE can be decomposed into a variance term and a bias term:

$$\begin{aligned}
\text{MSE}_\theta(\tilde{\tau}) &= \mathbb{E}\left[(\tilde{\tau} - \tau(\theta))^2\right] \\
&= \mathbb{E}\left[(\tilde{\tau} - \mathbb{E}(\tilde{\tau}) + \mathbb{E}(\tilde{\tau}) - \tau(\theta))^2\right] \\
&= \underbrace{(\mathbb{E}(\tilde{\tau}) - \tau(\theta))^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left((\tilde{\tau} - \mathbb{E}(\tilde{\tau}))^2\right)}_{\text{Variance}}
\end{aligned}$$

Thus, for an unbiased estimator $\tilde{\tau}(X)$, then $\text{MSE}_\theta(\tilde{\tau}) = \text{Var}(\tilde{\tau})$

Definition 8 (UMVUE)**.** An unbiased estimator $\hat{\tau}$ of $T(\theta)$ is said to be **Uniformly minimum variance unbiased estimator (UMVUE)** if:

$$\text{Var}_\theta(\hat{\tau}) \leq \text{Var}_\theta(\tilde{\tau}) \quad \forall \theta \in \Omega \text{ and all unbiased estimators } \tilde{\tau}$$

This is where *Uniformly minimum variance* gets its name: minimal variance over all θ and unbiased estimators. In other words, the UMVUE has the smallest MSE for all unbiased estimators.

This begs the question: how do we find the UMVUE? The Rao-Blackwell theorem is very useful in this regard, and guarantees the uniqueness of the UMVUE.

Theorem 6 (Rao-Blackwell Theorem)**.** Assume there exists an unbiased estimator $\tilde{\tau}(X)$ of $\tau(\theta)$ and a complete sufficient statistic: $T(X)$ for θ. Then the following admits a UMVUE:

$$\hat{\tau}(T) = \mathbb{E}[\tilde{\tau}(X)|T]$$

And $\hat{\tau}(T)$ **is unique**.

Pf sketch

(i) Define $\overset{\checkmark}{\tau}(T) = \mathbb{E}(\mathring{\tau}(X)|T)$. $\overset{\checkmark}{\tau}(T)$ is unbiased because:

$$\mathbb{E}(\overset{\checkmark}{\tau}(T)) = \mathbb{E}\left(\mathbb{E}(\mathring{\tau}(X)|T)\right) = \mathbb{E}(\mathring{\tau}(X)) = \tau(\theta)$$

And by law of total variance: $\text{Var}(\mathring{\tau}) = \text{Var}(\mathbb{E}(\mathring{\tau}|T)) + \mathbb{E}(\text{Var}(\mathring{\tau}|T))$ implying:

$$\text{Var}(\overset{\checkmark}{\tau}(T)) \leq \text{Var}(\mathring{\tau}(T))$$

(ii) Suppose $\hat{\tau}(T)$ is the UMVUE. Since $\mathbb{E}(\overset{\checkmark}{\tau}(T)) = \tau(\theta)$:

$$\begin{aligned}
&\mathbb{E}_\theta[\hat{\tau}(T) - \overset{\checkmark}{\tau}(T)] = 0 \ \forall \theta \in \Omega \\
&\qquad \implies \hat{\tau}(T) - \overset{\checkmark}{\tau}(T) = 0 \quad \text{By completeness of T} \\
&\qquad \implies \hat{\tau}(T) = \overset{\checkmark}{\tau}(T)
\end{aligned}$$

Proving that UMVUE is unique.

The Rao-Blackwell theorem importantly gives us a way to find the UMVUE and guarantees its uniqueness. However, it's not always easy to calculate the conditional expectation. The following strategy is much easier and may be the most important theorem for finding UMVUEs: **UMVUE supermarket**.

Theorem 7 (UMVUE supermarket)**.** If $\tilde{\tau}(X) = \phi(T(X))$ be a function that depends on X only through a complete sufficient statistic $T(X)$. Then $\phi(T(X))$ is the UMVUE for it's own expectation: $\mathbb{E}(\phi(T))$.

Thus, if we can find a complete statistic T and guess a function $\phi(T)$ such that $\mathbb{E}(\phi(T))$ equals our target $\tau(\theta)$, $\phi(T(X))$ will be the UMVUE for our target.

Let's explore the following example to familiarize ourselves with the power of the supermarket: suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma_0^2)$. We know that $T = \bar{X}_n$ is CSS for μ by properties of a 1-D exponential family:

(i) What if we want to find the UMVUE for μ? Let $\phi = 1$. Then $\phi(\bar{X}_n) = \bar{X}_n$ is UMVUE for $\mathbb{E}(\bar{X}_n) = \mu$. So $\bar{X}_n$ is UMVUE for μ.

(ii) What if we want to find the UMVUE for μ^2? Note that:

$$\begin{aligned}
\mathbb{E}(\bar{X}_n^2) &= \text{Var}(\bar{X}_n) + \left[\mathbb{E}(\bar{X}_n)\right]^2 \\
&= \frac{\sigma_0^2}{n} + \mu^2
\end{aligned}$$

So if we choose $\phi(\bar{X}_n) = \bar{X}_n^2 - \frac{\sigma_0^2}{n}$, then $\mathbb{E}(\phi(\bar{X}_n)) = \mu^2$, so $\phi(\bar{X}_n) = \bar{X}_n^2 - \frac{\sigma_0^2}{n}$ is UMVUE for μ^2.

Property 16 (UMVUE for Normal)**.** Consider $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. We know $\left(\bar{X}_n, s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2\right)$ is CSS for $\theta = (\mu, \sigma^2)$.

(i) $\phi(T) = \bar{X}_n$ is UMVUE for μ.

(ii) $\phi(T) = s_n^2$ is UMVUE for σ^2.

(iii) $\phi(T) = \bar{X}_n^2 - \frac{s_n^2}{n}$ is the UMVUE for μ^2.

$$\begin{aligned}\mathbb{E}\left(\phi(T)\right) &= \mathbb{E}\left(\bar{X}_n^2 - \frac{s_n^2}{n}\right) \\ &= \left(\mu^2 + \frac{\sigma^2}{n}\right) - \frac{\sigma^2}{n} = \mu^2\end{aligned}$$

(iv) $\phi(T) = \sqrt{\frac{n-1}{2}} \cdot \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} s_n$ is the UMVUE for σ (standard deviation).

Property 17 (UMVUE for Poisson distribution)**.** Consider $X_1, \ldots, X_n \overset{iid}{\sim} \text{Pois}(\lambda)$. We know $T = \sum X_i$ is CSS for λ.

(i) $\phi(T) = \bar{X}_n$ is the UMVUE for λ:

$$\begin{aligned}\mathbb{E}(\phi(T)) &= \frac{\sum \mathbb{E}(X_i)}{n} \\ &= \frac{n\lambda}{n} = \lambda\end{aligned}$$

(ii) $\phi(T) = \frac{(\sum X_i)^2 - \sum X_i}{n^2}$ is UMVUE for λ^2.

$$\begin{aligned}\mathbb{E}(\phi(T)) &= \mathbb{E}\left(\frac{\left(\sum X_i\right)^2 - \sum X_i}{n^2}\right) \\ &= \underbrace{\mathbb{E}(\bar{X}_n^2)}_{\bar{X}_n \sim N(\lambda, \frac{\lambda}{n})} - \frac{\mathbb{E}(\bar{X}_n)}{n} \\ &= \frac{\lambda}{n} + \lambda^2 - \frac{\lambda}{n} \\ &= \lambda^2\end{aligned}$$

(iii) $\phi(T) = \left(\frac{n-1}{n}\right)^{\sum X_i}$ is the UMVUE for $\tau(\lambda) = e^{-\lambda} = P_\lambda(X_1 = 0)$.

Find an unbiased estimator: $\tilde{\tau} = \mathbb{I}(X_1 = 0)$ b/c $\mathbb{E}(\mathbb{I}(X_1 = 0)) = P_\lambda(X_1 = 0)$

Use Rao-Blackwell: $\hat{\tau}(T) = \mathbb{E}(\tilde{\tau}|T)$

$$\begin{aligned}&= \mathbb{E}\left(\mathbb{I}(X_1 = 0) | \sum X_i = t\right) \\ &= P\left(X_1 = 0 | \sum X_i = t\right) \\ &= \frac{P\left(X_1 = 0, \sum X_i = t\right)}{P\left(\sum X_i = t\right)} \\ &= \frac{P\left(X_1 = 0\right) \cdot P\left(X_2 + \ldots + X_n = t\right)}{P\left(\sum X_i = t\right)} \qquad \text{Note } X_1 \sim \text{Pois}(\lambda), \sum_{i=2}^{n} X_i \sim \text{Pois}((n-1)\lambda) \\ &= \frac{P\left(\text{Pois}(\lambda) = 0\right) \cdot P\left(\text{Pois}((n-1)\lambda) = t\right)}{P(\text{Pois}(n\lambda) = t)} \\ &= \left(\frac{n-1}{n}\right)^t\end{aligned}$$

Property 18 (UMVUE for truncation family)**.** Consider $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$. I claim the UMVUE for θ is $\left(\frac{n+1}{n}\right) X_{(n)}$. We can prove it using supermarket:

$$\begin{aligned}\mathbb{E}_\theta\left(\frac{n+1}{n} X_{(n)}\right) &= \frac{n+1}{n}\left(\frac{n}{n+1}\theta\right) \\ &= \theta\end{aligned}$$

UMVUE's are great because they have the guaranteed smallest variance (MSE) among all unbiased estimators. However, this doesn't mean they have the smallest MSE! We can sometimes reduce the MSE by allowing a bit of bias: this is the rationale behind **shrinkage estimators**.

An example of a shrinkage estimator is for $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_0, \sigma^2)$. We shave shown that $\frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2$ is the UMVUE for σ^2. However, $\frac{1}{n-2}\sum_{i=1}^n (X_i - \mu_0)^2$ is a shrinkage estimator that admits some bias for lower MSE.

3 Information inequality

The Cramer-Rao lower bound provides a lower bound on the variance of any unbiased estimator in a regular (smooth) statistical model. This bound is called the information inequality because it is based on the Fisher information number. The information inequality provides an alternative approach to UMVUEs and determines the asymptotic variance of the MLE.

Definition 9 (Regular statistical family)**.** A family of pdfs $\{f_\theta(x)|\theta \in \Omega\}$ is regular if Ω is an open set and $f_\theta(x)$ is a smooth function of θ for almost every x. Note truncation families are not regular.

3.1 Fisher Information Number

Definition 10 (Score function)**. Fisher's score function** is just the derivative of the log-likelihood: $\frac{d \log f_\theta(x)}{d\theta}$.

Definition 11 (Fisher information number (FIN))**.** The Fisher Information number measures the intrinsic accuracy of a parametric statistical model. It has the following formulations:

(i) The FIN can be interpreted as the second moment of the Fisher's score function:

$$I_X(\theta) = \mathbb{E}\left[\left(\frac{d \log f_\theta(x)}{d\theta}\right)^2\right]$$

(ii) The FIN can be interpreted as the negative expected value of the second derivative of the log likelihood function:

$$I_X(\theta) = -\mathbb{E}\left[\left(\frac{d^2 \log f_\theta(x)}{d\theta^2}\right)\right]$$

(iii) The FIN can be interpreted as the variance of Fisher's score function:

$$I_X(\theta) = \text{Var}\left[\left(\frac{d \log f_\theta(x)}{d\theta}\right)\right]$$

In other words, the FIN tell us how peaked the likelihood surface is.

Example: let $f_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$:

$$\begin{aligned}
\ell(\theta) &= c + x\log(\theta) + (n-x)\log(1-\theta) \\
\frac{\partial \ell}{\partial \theta} &= \frac{x}{\theta} - \frac{n-x}{1-\theta} \\
\frac{\partial^2 \ell}{\partial \theta^2} &= -\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2} \\
\text{FIN} &= -\mathbb{E}_\theta\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] \\
&= \frac{\mathbb{E}(x)}{\theta^2} + \frac{n-\mathbb{E}(x)}{(1-\theta)^2} \\
&= \frac{n}{\theta(1-\theta)}
\end{aligned}$$

Property 19 (Properties of FIN)**.**

(i) **Non-negativity**: $I_X(\theta) \geq 0$

(ii) **Additivity**: suppose $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(\cdot)$. Then

$$\begin{aligned}
I_{\vec{X}}(\theta) &= \sum_{i=1}^n \underbrace{\text{Var}\left(\frac{d}{d\theta}\log f_\theta(X_i)\right)}_{I_{X_i}(\theta)} \\
&= \sum_{i=1}^n I_{X_i}(\theta) = nI_{X_i}(\theta)
\end{aligned}$$

3.2 Cramer-Rao Lower Bound

The Cramer-Rao lower bound links the accuracy of an estimator to the Fisher Information number.

Theorem 8 (Cramer-Rao Lower Bound)**.** Assuming that $I_X(\theta) > 0$, then:

$$\text{Var}_\theta[T(X)] \geq \frac{\left\{\frac{d}{d\theta}\mathbb{E}_\theta(T(X))\right\}^2}{I_X(\theta)}$$

Pf sketch: Relies on the *Cauchy-Schwartz Inequality*:

$$\begin{aligned}
\text{Cov}(X,Y) &= \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))) \\
&\overset{C-S}{\leq} \sqrt{\mathbb{E}[(X-\mathbb{E}(X))]^2\mathbb{E}[(Y-\mathbb{E}(Y))]^2} \\
&\leq \sqrt{\text{Var}(X)\text{Var}(Y)} \\
&\implies \text{Var}(X) \geq \frac{\text{Cov}(X,Y)^2}{\text{Var}(Y)}
\end{aligned}$$

Let $X = T$ be our estimator of $\tau(\theta)$ and let $Y = \frac{d\log f_\theta(x)}{d\theta}$ be the score function.

$$\text{Var}(T) \geq \frac{\text{Cov}(T,Y)^2}{\text{Var}(Y)}$$

Note that:

$$\begin{aligned}\text{Cov}(T,Y) &= \mathbb{E}(TY) - \mathbb{E}(T)\mathbb{E}(Y) \\ &= \mathbb{E}(TY) \quad \text{(b/c FIsher's score function has mean 0)} \\ &= \int T(X)\frac{\frac{d}{d\theta}f_\theta(x)}{f_\theta(x)}f_\theta(x)dx \\ &= \frac{d}{d\theta}\int T(x)f_\theta(x)dx \\ &= \frac{d}{d\theta}\mathbb{E}(T(X))\end{aligned}$$

And

$$\begin{aligned}\text{Var}(Y) &= \text{Var}\left(\frac{d\log f_\theta(x)}{d\theta}\right) \\ &= I_X(\theta)\end{aligned}$$

Property 20 (Properties of C-R LB)**.**

(i) The larger the FIN, the lower the variance bound will be.

(ii) If we intend to estimate $\tau(\theta)$ and we have an unbiased estimator $T(X)$, then the C-R LB reduces to:

$$\text{Var}_\theta(T(X)) \geq \frac{\{\tau'(\theta)\}^2}{I_X(\theta)}$$

(iii) If $\tau(\theta) = \theta$ and we have an unbiased estimator for θ, the bound becomes:

$$\text{Var}_\theta(T(X)) \geq \frac{1}{I_X(\theta)}$$

(iv) The C-R LB is just a lower bound, and **is not guaranteed to be achievable** (i.e., the UMVUE does not always exist).

(v) The equality on the variance bound holds iff $f_\theta(x)$ is an exponential family member of the form:

$$f_\theta(x) = \exp\{A(\theta)\}\cdot\exp\{B(\theta)T(X)\}\cdot\exp\{C(X)\}$$

Where $T(X)$ is a complete sufficient statistic. Then:

$$\text{Var}_\theta[T(X)] = \frac{\left\{\frac{d}{d\theta}\mathbb{E}_\theta(T(X))\right\}^2}{I_X(\theta)}$$

So the the CSS $T(X)$ achieves the C-R LB for $\mathbb{E}(T(X))$. In fact, any linear transformation, $a\cdot T(X)+b$ achieves the C-R LB for the for $a\cdot\mathbb{E}(T(X))+b$. However, a nonlinear transformation of the CSS will not achieve the C-R LB.

Property 21 (Parametrizations and C-R LB)**.** Let $\theta(\nu)$ be a smooth function of ν, so $g_\nu(x) = f_{\theta(\nu)}(x)$ be

a smooth reparametrization of the model.

$$\begin{aligned} I_g(\nu) &= \mathbb{E}_\nu\left(\left[\frac{d\log g_\nu(X)}{d\nu}\right]^2\right) \\ &= \mathbb{E}_\nu\left(\left[\frac{d\log f_{\theta(\nu)}(X)}{d\nu}\right]^2\right) \\ &= \mathbb{E}_\nu\left(\left[\frac{d\log f_{\theta(\nu)}(X)}{d\theta}\cdot\frac{d\theta}{d\nu}\right]^2\right) \\ &= I_f(\theta(\nu))\cdot\left(\frac{d\theta}{d\nu}\right)^2 \end{aligned}$$

For example, if $\theta = e^\nu$ then $I_g(\nu) = I_f(e^\nu)\cdot e^{2\nu}$. **NOTE**: we can swap around $\left(\frac{d\theta}{d\nu}\right)^2$ to move between the FINs. So the information number depends on the parametrization!

3.3 Multidimensional information inequality

The FIN and CR L-B can easily be extended to the multi-dimensional parameter sapce with a gradient representation:

Definition 12 (Fisher Information matrix)**.** Suppose $\theta \in \mathbb{R}^k$. The **Fisher Information Matrix** is defined as:

$$I_X(\theta) := \mathbb{E}_\theta\left[(\nabla_\theta \log f_\theta(x))\cdot(\nabla_\theta \log f_\theta(x))^T\right]$$

$$\text{where } \nabla_\theta \log f_\theta(x) = \begin{pmatrix} \frac{\partial \log f_\theta(x)}{\partial\theta_1} \\ \vdots \\ \frac{\partial \log f_\theta(x)}{\partial\theta_k} \end{pmatrix} \in \mathbb{R}^k$$

It can also be written as:

$$I_X(\theta) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) & \dots & I_{1k}(\theta) \\ \vdots & \vdots & \vdots & \vdots \\ I_{k1}(\theta) & I_{k2}(\theta) & \dots & I_{kk}(\theta) \end{pmatrix}$$

$$\begin{aligned} \text{where } I_{ij}(\theta) &= \mathbb{E}\left(\left[\frac{\partial \log f_\theta(x)}{\partial\theta_i}\right]\left[\frac{\partial \log f_\theta(x)}{\partial\theta_j}\right]\right) \\ &= -\mathbb{E}\left(\frac{\partial^2 \log f_\theta(x)}{\partial\theta_i\partial\theta_j}\right) \end{aligned}$$

Theorem 9 (Cramer-Rao Lower Bound (multidimensional))**.** Suppose $(\theta_1,\dots,\theta_k) \in \mathbb{R}^k$ are unknown and $I_X(\theta)$ is positive definite. For any real valued statistic $T(X)$:

$$\text{Var}_\theta(T(X)) = \left[\nabla_\theta \mathbb{E}(T(X))\right]^T \left[I_X(\theta)\right]^{-1} \left[\nabla_\theta \mathbb{E}(T(X))\right]$$

If we suppose $T(X)$ is an unbiased estimator of $\tau(\theta)$:

$$\text{Var}_\theta(T(X)) = [\nabla_\theta \tau(\theta)]^T [I_X(\theta)]^{-1} [\nabla_\theta \tau(\theta)]$$

Definition 13 (Information processing inequality and connection to sufficiency)**.** In short, processing your data via a function cannot increase the Fisher information:

$$I_X(\theta) \geq I_T(\theta)$$

And equality $I_X(\theta) = I_T(\theta)$ holds iff $T(X)$ is a sufficient statistic.

3.4 Nuisance parameters

For example, suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta_1, \theta_2)$. Suppose we are only interested in inferring θ_1 but wish to keep θ_2 unknown. We can treat θ_2 as a nuisance parameter.

Definition 14 (Nuisance parameter and C-R LB)**.** Given a likelihood family $(\mathcal{X}, \mathcal{P})$ with:

$$\mathcal{P} := \left\{ P_\theta | \theta = (\theta_1, \ldots, \theta_k) \in \Omega \subset \mathbb{R}^k \right\}$$

If we are only interested in estimating $\tau(\theta_1)$. Suppose we have an unbiased estimator $T(x)$ of τ. If $(\theta_2, \ldots, \theta_k)$ are known, then the 1-parameter CR bound is appropriate:

$$\text{Var}_\theta[T(X)] \geq \frac{\left(\frac{\partial \tau}{\partial \theta_1}\right)^2}{I_{11}(X)}$$

However, if we want to estimate $\tau(\theta_1)$ with $(\theta_2, \ldots, \theta_k)$ unknown, then $(\theta_2, \ldots, \theta_k)$ are nuisance parameters, and we use the k-parameter C-R LB:

$$\begin{aligned}
\nabla_\theta(\tau) &= \left(\frac{d\tau}{d\theta_1}, 0, \ldots, 0\right) \\
\text{Var}_\theta[T(X)] &\geq \left(\frac{d\tau}{d\theta_1}, 0, \ldots, 0\right) \underbrace{[I(\theta)]^{-1}}_{\text{FIM}} \left(\frac{d\tau}{d\theta_1}, 0, \ldots, 0\right)^T \\
&= \frac{\left(\frac{d\tau}{d\theta_1}\right)^2}{I_{11\cdot 2}(\theta)}
\end{aligned}$$

Where $I_{11\cdot 2}(\theta) = I_{11}(\theta) - I_{12}(\theta)[I_{22}(\theta)]^{-1} I_{21}(\theta)$.

Clearly, $I_{11\cdot 2}(\theta) \leq I_{11}(\theta)$ so:

$$\frac{\left(\frac{d\tau}{d\theta_1}\right)^2}{I_{11}(\theta)} \leq \frac{\left(\frac{d\tau}{d\theta_1}\right)^2}{I_{11\cdot 2}(\theta)}$$

Thus, nuisance parameters lead to a reduction in asymptotic efficiency, and no reduction occurs iff $I_{12}(\theta) = 0$, meaning that the covariance between the partial score function of θ_1 and the vector of other partial score functions is 0, i.e., θ_1 is orthogonal to nuisance parameters.

Pf: The resulting inequality under nuisance parameter case is because the FIM is positive semi-definite, meaning we can do a block decomposition:

$$I(\theta) = \begin{pmatrix} I_{11}(\theta) \in \mathbb{R}^1 \times \mathbb{R}^1 & I_{12}(\theta) \in \mathbb{R}^1 \times \mathbb{R}^{k-1} \\ I_{21}(\theta) \in \mathbb{R}^{k-1} \times \mathbb{R}^1 & I_{22}(\theta) \in \mathbb{R}^{k-1} \times \mathbb{R}^{k-1} \end{pmatrix}$$

And then for $\nabla_\theta(\tau) = \left(\frac{d\tau}{d\theta_1}, 0, \ldots, 0\right)$ decomposed into $(\mathbb{R}^1, \mathbb{R}^{k-1})$ components:

$$\begin{aligned} \nabla_\theta(\tau)[I(\theta)^{-1}]\nabla_\theta(\tau)^T &= \left(\frac{d\tau}{d\theta_1} - I_{12}[I_{22}]^{-1}\vec{0}\right) I_{11,2}^{-1} \left(\frac{d\tau}{d\theta_1} - I_{12}[I_{22}]^{-1}\vec{0}\right) + \vec{0} I_{22}^{-1}\vec{0} \\ &= \frac{\left(\frac{d\tau}{d\theta_1}\right)^2}{I_{11\cdot 2}(\theta)} \end{aligned}$$

Per properties of PSD matrices.

Here's a nice example: suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta_1, \theta_2)$. Then:

$$\begin{aligned} f_{\theta_1,\theta_2}(x_i) &= \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{(x_i - \theta_1)^2}{2\theta_2}\right) \\ \ell(\theta_1, \theta_2) &= c - \frac{1}{2}\log(\theta_2) - \frac{(x_i - \theta_1)^2}{2\theta_2} \\ \frac{\partial^2 \ell}{\partial \theta_1^2} &= -\frac{1}{\theta_2} \\ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} &= -\frac{x_i - \theta_1}{\theta_2^2} \\ \frac{\partial^2 \ell}{\partial \theta_2^2} = \frac{1}{2\theta_2^2} - \frac{(x_i - \theta_1)^2}{\theta_2^3} & \end{aligned}$$

The FIM is the expectation of each of these components:

$$IX_i(\theta_1, \theta_2) = \begin{pmatrix} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{pmatrix}$$

Thus, θ_1 and θ_2 are orthogonal and the CR LB for an unbiased estimator of θ_1 is:

$$\frac{1}{nI_{11}} = \frac{\theta_2}{n} = \frac{\sigma^2}{n}$$

and the CR LB for an unbiased estimator of θ_2 is:

$$\frac{1}{nI_{22}} = \frac{2\theta_2^2}{n} = \frac{2\sigma^4}{n}$$

4 Maximum Likelihood Estimation

In Fisher's framework, $X \sim P_{\theta_0}$ for some unknown θ_0. To estimate θ_0 based on the observed data, Fisher proposed locating $\hat{\theta} \in \Omega$ such that $f_{\hat{\theta}}(x) = \max_{\theta \in \Omega} f_\theta(x)$; i.e., your estimate $\hat{\theta}$ is the $\theta \in \Omega$ that maximizes the likelihood of the data.

4.1 Why is MLE such a good estimator?

Property 22 (Why is the MLE a good estimator?)**.**

1. **Strong consistency**: $\hat{\theta}$ is almost surely converging to θ_0 as $n \to \infty$.

Proof: Wald proved this using a few useful facts:

(i) Oracle inequality: $f_{\hat{\theta}}(x) \geq f_{\theta_0}(x)$ by the definition of the maximum likelihood estimator.

(ii) Strong Law of Large Numbers: a sample average convergences almost surely to its target as $n \to \infty$

(iii) The KL-divergence quantifies the difference between two pdfs:

$$KL(p, q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x) dx$$
$$KL(p, q) \geq 0 \quad \text{(By Jensen's inequal)}$$

(iv) Wald's identifiability criterion: Wald required a density such that $\forall \theta \neq \theta_0$, $KL(f_{\theta_0}, f_\theta) > 0$

By the Oracle inequality:

$$\begin{aligned} f_{\hat{\theta}}(x) &\geq f_{\theta_0}(x) \\ \implies \sum_{i=1}^n \log f_{\hat{\theta}}(x_i) &\geq \sum_{i=1}^n \log f_{\theta_0}(x_i) \\ \implies 0 &\geq \frac{1}{n} \sum_{i=1}^n \left[\log f_{\theta_0}(x_i) - \log f_{\hat{\theta}}(x_i)\right] \\ \xrightarrow{\text{SLLN}} \mathbb{E}_{\theta_0}\left[\log \frac{f_{\theta_0}(x)}{f_{\hat{\theta}}(x)}\right] &= KL(f_{\theta_0}, f_{\hat{\theta}}) \leq 0 \end{aligned}$$

However, we know that $KL(p, q) \geq 0$, but by the Oracle inequality we showed $KL(f_{\theta_0}, f_{\hat{\theta}}) \leq 0$. Under Wald's identifiability criterion, we conclude that as $n \to \infty$, $KL(f_{\theta_0}, f_{\hat{\theta}}) = 0$. Thus, $\lim_{n \to \infty} \hat{\theta} = \theta$.

2. **Asymptotically efficient**: the MLE asymptotically attains the Cramer-Rao lower bound. Fisher-Cramer Theorem for more details.

3. **Asymptotically normal**: the MLE is asymptotically normally distributed. See Fisher-Cramer Theorem for more details.

The following theorem outlines the asymptotic efficiency and limiting distribution of the MLE. This is what makes the MLE such a tantalizing target for inference.

Theorem 10 (Fisher-Cramer Theorem)**.** The Fisher-Cramer demonstrates that the MLE is asymptotically

normal and attains the Cramer-Rao lower bound:

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xRightarrow{d} N\left(0, [I_{X_i}(\theta_0)]^{-1}\right)$$

Note that $I_{X_i}(\theta_0)$ is the FIN/FIM for a single observation.

Proof: relies on Taylor expansions and Slutsky's theorem:

$$\begin{aligned}
\hat{\theta} &= \underset{\theta\in\Omega}{\text{argmax}}\left[\prod_{i=1}^{n} f_\theta(x_i)\right] \\
&= \underset{\theta\in\Omega}{\text{argmax}}\left[\sum_{i=1}^{n} \log(f_\theta(x_i))\right] \\
&\implies \hat{\theta} \text{ is the soln to } \frac{d\log(f_\theta)}{d\theta}\Big|_{\theta=\hat{\theta}} = 0 \\
&\implies \frac{d\log(f_\theta)}{d\theta}\Big|_{\theta=\hat{\theta}} \approx \frac{d\log(f_{\theta_0})}{d\theta} + (\hat{\theta}-\theta_0)\frac{d^2\log(f_\theta)}{d\theta^2} + \text{high order terms} = 0 \quad \text{(Taylor expansion at } \theta=\theta_0) \\
&\implies \frac{d\log(f_{\theta_0})}{d\theta} + (\hat{\theta}-\theta_0)\frac{d^2\log(f_\theta)}{d\theta^2} = 0 \\
&\implies -\frac{d^2\log(f_{\theta_0})}{d\theta^2}\sqrt{n}(\hat{\theta}-\theta_0) = \sqrt{n}\frac{d\log(f_{\theta_0})}{d\theta}
\end{aligned}$$

Note that:

(i) Second derivative term convergences to FIN.

$$\begin{aligned}
-\frac{1}{n}\frac{d^2\log(f_{\theta_0})}{d\theta^2} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{d^2\log(f_\theta(x_i))}{d\theta^2}\Big|_{\theta=\theta_0} \\
&\xrightarrow{\text{SLLN}} -\mathbb{E}\left(\frac{d^2\log(f_\theta)}{d\theta^2}\right) = I_{X_i}(\theta_0)
\end{aligned}$$

(ii)

$$\begin{aligned}
\frac{1}{n}\frac{d\log(f_{\theta_0})}{d\theta} &= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(\frac{d\log(f_\theta(x_i))}{d\theta}\Big|_{\theta=\theta_0}\right)}_{\text{Score function: mean=0, var=FIN}} \\
&\xrightarrow{\text{CLT}} N(0, I_{X_i}(\theta_0))
\end{aligned}$$

Using Slutsky's Theorem:

$$\begin{aligned}
-\frac{d^2\log(f_{\theta_0})}{d\theta^2}\sqrt{n}(\hat{\theta}-\theta_0) &= \sqrt{n}\frac{d\log(f_{\theta_0})}{d\theta} \\
&\implies \sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} [I_{X_i}(\theta_0)]^{-1}N(0, I_{X_i}(\theta_0)) = N(0, [I_{X_i}(\theta_0)]^{-1})
\end{aligned}$$

Another fantastic property of MLE is the invariance rule, which says the MLE of a function of θ is just the function applied to $\hat{\theta}$.

Theorem 11 (Invariance rule)**.** If $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$. The asymptotic

distribution of $\tau(\hat{\theta})$ is as follows (delta method):

$$\begin{aligned}\sqrt{n}\left(\widehat{\tau(\theta)} - \tau(\theta)\right) &\approx \sqrt{n}(\hat{\theta} - \theta)\tau'(\theta) \\ &\xrightarrow{d} \tau'(\theta) N\left(0, [I_{X_i}(\theta)]^{-1}\right) \\ &\xrightarrow{d} N\left(0, \tau'(\theta)^2 [I_{X_i}(\theta)]^{-1}\right)\end{aligned}$$

In reality, when we want to use the asymptotic distribution of $\tau(\theta)$ to construct a confidence interval, we use the observed fisher information $I_{X_i}(\hat{\theta})$ (since we don't have the true FIN) which by the CMT, $I_{X_i}(\hat{\theta}) \xrightarrow{\text{almost surely}} I_{X_i}(\theta)$.

4.2 Other interpretations and when MLE exists

Definition 15 (MLE and the LR)**.** The MLE has an interpretation via the likelihood ratio:

$$\hat{\theta} := \left\{\theta : L_{\theta,\theta'}(x) = \frac{f_\theta(x)}{f_{\theta'}(x)} \geq 1 \ \ \forall\, \theta' \in \Omega\right\}$$

$\hat{\theta}$ is therefore a function of the likelihood ratio, the MSS.

Property 23 (MLE 1-parameter exponential family)**.** When the pdf is log-concave (such as in a 1-parameter exponential family), there can exist at most one root in the parameter space, which must be the unique CANE (consistent, asymptotically normal estimator).

(i) Any exponential family $f_\theta(x) = a(\theta)e^{\theta T(x)}h(x)$ is log concave with natural parameter θ.

(ii) If $\mathbb{E}_\theta(T) = T(X)$ (i.e., function of θ = function of X) has a solution $\hat{\theta}(x)$, then this solution is unique and is the MLE of θ.

(iii) $I_X(\theta) = \text{Var}_\theta(T(X))$ (recall θ is the natural parameter).

Suppose we want to estimate the mean of a Cauchy distribution: Cauchy$(\theta, 1)$. Pearson's MoM estimator fails since there is no closed-form formula for the moments. Fisher's MLE runs into trouble as well, as there are many different roots to the log likelihood function. Here are a few workarounds:

1. Choose the root that is closest to the sample median, since we know that the Cauchy distribution is symmetric.
2. Consider a Cauchy distribution with both parameters unknown, this gives us a concave likelihood that won't be trapped by local minimizers, although we lose efficiency.
3. **Newton-Raphson method**: start with a $\sqrt{n}$-consistent (perhaps inefficient) estimator: $\theta_{(0}^{(n)}$. Then

we can write the Taylor expansion of the log-likelihood:

$$\frac{\partial \ell(\theta)}{\partial \theta} \approx \frac{\partial \ell(\theta_{(0)}^{(n)})}{\partial \theta} + \left(\theta_{(1)}^{(n)} - \theta_{(0)}^{(n)}\right) \frac{\partial^2 \ell(\theta_{(0)}^{(n)})}{\partial \theta^2} = 0$$

$$\implies \theta_{(1)}^{(n)} = \theta_{(0)}^{(n)} - \frac{\frac{\partial \ell(\theta_{(0)}^{(n)})}{\partial \theta}}{\frac{\partial^2 \ell(\theta_{(0)}^{(n)})}{\partial \theta^2}}$$

This one-step estimator $\theta_{(1}^{(n)}$ is consistent, asymptotically normal, and efficient!

5 Hypothesis testing

Estimation is only half the story. We are also concerned with testing; whether unknown parameter θ satisfies a certain property called the null hypothesis.

Definition 16 (Decision function)**.** A decision (testing procedure) is a mapping $\phi : \mathcal{X} \to [0, 1]$ (in most cases $\{0, 1\}$) which maps from the data space to a binary scale where $0 : H_0$ and $1 : H_1$.

Definition 17 (Neyman-Pearson Criterion/Level/Size/UMP)**.** The Neyman-Pearson criterion is the foundation for statistical testing. Here are the key tenants:

1. **Power function:** The power function is the probability that we reject H_0 as a function of θ (and ϕ):
$$\pi_\phi(\theta) = \mathbb{E}_\theta[\phi(x)]$$
2. **Size:** the size of a test is defined to be the highest T1 error rate among all $\theta \in \Omega_0$:
$$\text{size of } \phi := \sup_{\theta \in \Omega_0} \pi_\phi(\theta)$$
3. **Level:** a test is level-α if its T1 error rate is at most α.
$$\sup_{\theta \in \Omega_0} \pi_\phi(\theta) \leq \alpha$$
4. **UMP:** a test is UMP for $H_0 : \theta \in \Omega_0$ versus $H_1 : \theta \in \Omega_1$ at level α if:
$$\pi_\phi(\theta) \equiv \mathbb{E}_1[\phi(X)] = \sup_{\phi'} \pi_{\phi'}(\theta) \qquad \forall \theta \in \Omega_1$$
where ϕ' is all level α test for H_0. I.e., the UMP test has the lowest T2 error rate for all level-α tests.

5.1 Two-point/one-sided alternative hypothesis testing

This section pertains to testing hypotheses of the form:

$$\begin{aligned} H_0 : \theta = \theta_0 \quad &\text{vs} \quad H_1 : \theta = \theta_1 \\ H_0 : \theta = \theta_0 \quad &\text{vs} \quad H_1 : \theta > \theta_0 \\ H_0 : \theta \leq \theta_0 \quad &\text{vs} \quad H_1 : \theta > \theta_0 \end{aligned}$$

The Neyman-Pearson Lemma is an extremely valuable result for the UMP test in the case of a two-point hypothesis.

Theorem 12 (Neyman-Pearson Lemma)**.** Consider the two-point hypotheses: $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$. The test $\phi_c(x)$ based on the likelihood ratio is UMP for testing these two hypotheses at level $\alpha = \mathbb{E}_0[\phi(X)]$ (we choose our threshold c to enforce the desired α):

$$\phi_c(x) = \begin{cases} 0 & \text{if } \frac{f_1(x)}{f_0(x)} < c \\ 1 & \text{if } \frac{f_1(x)}{f_0(x)} > c \\ \gamma(x) & \text{if } \frac{f_1(x)}{f_0(x)} = c \end{cases}$$

In other words, if there exists another level-α test ϕ', then $\mathbb{E}_{\theta_1}[\phi'(x)] \leq \mathbb{E}_{\theta_1}[\phi(x)]$.

Proof: Let $\alpha_\phi = \mathbb{E}_0[\phi(X)]$

$$\int \phi(x) f_0(x) dx = \alpha_\phi$$
$$\int \phi'(x) f_0(x) dx \leq \alpha_\phi$$

Thus,

$$\mathbb{E}_1[\phi(X)] - \mathbb{E}_1[\phi'(X)] = \int (\phi(x) - \phi'(x)) f_1(x) dx$$
$$= \int_{\{\lambda(x)>c\}} \underbrace{(\phi(x) - \phi'(x))}_{\geq 0} f_1(x) dx + \int_{\{\lambda(x)<c\}} \underbrace{(\phi(x) - \phi'(x))}_{\leq 0} f_1(x) dx + \int_{\{\lambda(x)=c\}} (\phi(x) - \phi'(x)) f_1(x) dx$$
$$= \int_{\{\lambda(x)>c\}} (\phi(x) - \phi'(x)) c f_0(x) dx + \int_{\{\lambda(x)<c\}} (\phi(x) - \phi'(x)) c f_0(x) dx + \int_{\{\lambda(x)=c\}} (\phi(x) - \phi'(x)) c f_0(x) dx$$
$$= c \int (\phi - \phi') f_0 dx \geq c(\alpha - \alpha) = 0$$
$$\implies \mathbb{E}_1[\phi'(X)] < \mathbb{E}_1[\phi(X)]$$

Property 24 (Tests for 1-parameter Exp Families)**.** Suppose $X_1, \ldots, X_n$ have joint pdf:

$$f_\theta(x) = [a(\theta)]^n \exp\left(\theta \sum_{i=1}^n T(x_i)\right) \prod_{i=1}^n h(x_i)$$

Then the MP test at level α for testing $H_0 : \theta_0$, $H_1 : \theta_1$ with $\theta_0 < \theta_1$ is:

$$\phi(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^n T(x_i) < c_\alpha \\ 1 & \text{if } \sum_{i=1}^n T(x_i) > c_\alpha \\ \gamma(x) & \text{if } \sum_{i=1}^n T(x_i) = c_\alpha \end{cases}$$

If T is continuous then c_α can be chosen such that $P_{\theta_0}(T > c_\alpha) = \alpha$. If T is discrete, we can select c_α and $\gamma(x)$ to satisfy:

$$P_{\theta_0}(T > c_\alpha) + \gamma P_{\theta_0}(T = c) = \alpha$$

Note: since the test depends only on T and not on the alternative, it is UMP for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$.

Definition 18 (Monotone Likelihood Ratio (MLR))**.** $f_\theta(x)$ has a strict MLR if there exists a real-valued sufficient statistic $T(X)$ such that for each pair $\theta_1 < \theta_2 \in \Omega$, the LR is strictly increasing as a function of $T(X)$:

$$\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = g_{\theta_1,\theta_2}(T(x)) \quad \forall\, \theta_1 < \theta_2 \in \Omega$$

Theorem 13 (MLR and UMP for one-sided alternatives)**.** Suppose you are interested in one of the two sets of hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$
$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

Let $f_\theta(x)$ have MLR in T and let $\phi(T)$ be the test that:

$$\phi(t) = \begin{cases} 0 & \text{if } t < c_\alpha \\ 1 & \text{if } t > c_\alpha \\ \gamma_\alpha & \text{if } t = c_\alpha \end{cases}$$

Where c_α and γ_α chosen to satisfy:

$$P_{\theta_0}[T > c_\alpha] + \gamma_\alpha P_{\theta_0}[T = c_\alpha] = \alpha$$

Then ϕ is the UMP-level test for the two hypotheses above.

5.2 Hypothesis testing with two-sided alternatives

Suppose we are interested in testing one of the following sets of hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$
$$H_0 : a \leq \theta \leq b \quad \text{vs} \quad H_1 : \theta < a \text{ or } b < \theta$$

We turn to slight variations on the likelihood ratio test:

Definition 19 (General LRT for composite hypotheses)**.** Suppose one of the following cases:

(a) $H_0 : \theta \in \Omega_0$ and $H_1 : \theta \in \Omega_1 \equiv \Omega - \Omega_0$. If

$$\inf_{\theta \in \Omega_1} KL(\theta_0, \theta) > 0 \, , \forall \theta_0 \in \Omega_0$$
$$\inf_{\theta \in \Omega_0} KL(\theta_1, \theta) > 0 \, , \forall \theta_1 \in \Omega_1$$

Then the LRT is:

$$\phi(x_1, , x_n) = \begin{cases} 0 & \text{if } \frac{f_{\hat{\theta}_1}(x_i)}{f_{\hat{\theta}_0}(x_i)} \leq 1 \\ 1 & \text{if } \frac{f_{\hat{\theta}_1}(x_i)}{f_{\hat{\theta}_0}(x_i)} > 1 \end{cases}$$

Where $\hat{\theta}_i$ is the MLE under H_i.

(b) $H_0 : \theta \in \Omega_0$ versus $H_1 : \theta \in \Omega$ where $\Omega_0 \subset \Omega \subset \mathbb{R}^k$, $\dim(\Omega) = k - r$ and $\dim(\Omega_0) = k - r - s$ where s is the number of free parameters.

Then the LRT is:

$$\phi(x_1, , x_n) = \begin{cases} 0 & \text{if } \frac{f_{\hat{\theta}_0}(x_i)}{f_{\hat{\theta}}(x_i)} > c \\ 1 & \text{if } \frac{f_{\hat{\theta}_0}(x_i)}{f_{\hat{\theta}}(x_i)} < c \\ \gamma & \text{if } \frac{f_{\hat{\theta}_0}(x_i)}{f_{\hat{\theta}}(x_i)} = c \end{cases}$$

Where $\hat{\theta}_0$ is the MLE restricted to Ω_0 and $\hat{\theta}$ is the unrestricted MLE.

And the p-value is:

$$p = \sup_{\theta_0 \in \Omega_0} P(\lambda \leq \lambda_{\text{obs}})$$

Wilk's theorem is a very valuable way to get the distribution of the general LRT under case (b)

Theorem 14 (Wilks Theorem)**.** Let f_θ be a sample from a regular family (satisfies Cramer and Wald conditions). Let the LRT statistic, $\lambda(x_1, \ldots, x_n) = \frac{f_{\hat{\theta}_0}(x_i)}{f_{\hat{\theta}}(x_i)}$, under H_0 has asymptotic chi-square distribution with df = # of free parameters:

$$-2\log(\lambda) \xrightarrow{d} \chi^2_s \qquad \text{(Where } s = \dim(\Omega) - \dim(\Omega_0)\text{)}$$

6 Basic decision theorey

Definition 20 (Decision rule and Risk)**.** A decision rule is a map from the data space to the action space: $d(\cdot) : \mathcal{X} \to \mathcal{A}$.

Risk quantifies the reward/goodness of a decision based on a *loss function*:

$$R(d, \theta) = \mathbb{E}_\theta(L(d(X), \theta))$$

Definition 21 (Admissibility)**.** A decision $d(\cdot)$ is **admissible** if there does not exist another decision $d^{'}$ such that $R(d', \theta) \leq R(d, \theta)\ \forall \theta \in \Theta$. I.e., there does not exist another decision that achieves lower risk over all $\theta \in \Theta$.

E.g., the sample mean in $X_1, \ldots, X_n \overset{iid}{\sim} N_d(\mu, \Sigma)$ is inadmissible for estimating μ whenever $d \geq 3$ (Stein's shrinkage estimator) with squared loss.

Definition 22 (Wald's minimax principle and Bayes rule)**.** A **minimax decision** has the smallest worst risk:

$$\max_{\theta \in \Theta} R(d, \theta) = \inf_{d'} \left[\max_{\theta \in \Theta} R(d', \theta) \right]$$

The **Bayes risk** is the risk averaged over the prior distribution of θ:

$$B(d, \pi) := \int_{\theta \in \Theta} R(d, \theta) \pi(\theta) d\theta$$

A **bayes rule** minimizes the bayes risk.

7 General strategies

1. Proving sufficient:
   (i) Fisher-Neyman factorization theorem
   (ii) Show distribution of $X|T(X)$ does not depend on θ
   (iii) Is this statistic a 1-1 map of a known sufficient statistic?
   (iv) Is the distribution family you're considering a subfamily of a larger family with known sufficient statistic?
   (v) Is the pdf symmetric (abs value of X_is)? Permutation-invariant (order stats)? Radial ($||X||$)?
2. Proving not sufficient:
   (i) Show that $T(X)$ does not satisfy the factorization theorem.
   (ii) Show $X|T(X)$ depends on θ
   (iii) Show there does not exist a 1-1 map from a known SS to your proposed statistic.

3. Proving minimal sufficient:
   (i) Invoke Lehmann-Scheffe theorem (remember iff proof).
   (ii) Invoke a family-based result (exponential $(T(X) = (\sum T_1(X_i), \ldots, \sum T_k(X_i))$ when Ω affinely spans $\mathbb{R}^k$), truncation (above: $X_{(n)}$, below: $X_{(1)}$, two-sided: $(X_{(1)}, X_{(n)})$.
4. Proving not minimal sufficient:
   (i) Show the proposed statistic is not sufficient.
   (ii) Find a sufficient statistic, $\tilde{T}$ such that there exists no function such that $T = f(\tilde{T})$ (i.e., find a sufficient statistic that cannot be reduced to T).
5. Proving ancillary:
   (i) Show that the distribution of your proposed statistic is θ-free (i.e., does not depend on θ).
   (ii) Invoke a family-based result (location, scale, location-scale)
6. Proving complete:
   (i) Show that $\forall\theta \in \Omega$, $\mathbb{E}(g(T)) = 0 \implies g(T) = 0$. Use polynomial trick, induction, or other clever methods.
   (ii) Invoke a family-based result (exponential family $(T(X) = (T_1(X), \ldots, T_k(X))$ when Ω contains an open interval), truncation family $(X_(n))$, etc.)
7. Proving not complete:
   (i) Show the statistic, $T(X)$ is not minimal sufficient
   (ii) Basu's theorem: show that your proposes CSS is not independent of an ancillary statistic.
   (iii) Find a non-constant function $T(X)$ where $\mathbb{E}[T(X)]$ is 0 or constant. For multi-dimensional $T(X)$, solve for first and second moments of each $T_j(X)$ and try and find a linear combination that equals 0.
8. Proving/Finding UMVUE:
   (i) Rao-Blackwell theorem: we can improve an unbiased estimator by conditioning on a CSS. And guarantees uniqueness of the UMVUE.
   (ii) UMVUE supermarket: a function of a CSS is UMVUE for its expectation.
   (iii) If you know all unbiased estimators have a common form and you have an expression for the variance, you can minimize using calculus.
9. Minimizing MSE
   (i) If asked to minimize, write out and minimize by calculus:

$$\mathbb{E}_\theta\left[(\widehat{\tau(\theta)} - \tau(\theta))^2\right] = \mathbb{E}(\mathbb{E}((\widehat{\tau(\theta)})) - \tau(\theta))^2 + \text{Var}(\widehat{\tau(\theta)}) = \text{bias}^2 + \text{variance}$$

10. FIN/FIM and C-R LB:
    (i) FIN is the variance of the score function (telling us how peaked the likelihood is):

$$I_X(\theta) = -\mathbb{E}_\theta\left[\frac{d^2 \log f_\theta(x)}{d\theta^2}\right]$$

FIN is non-negative, additive, and is sensitive to parametrization.

(ii) C-R LB:

$$\text{General estimator } T(X)\text{: } \text{Var}(T(X)) \geq \frac{\left(\mathbb{E}_\theta[T(X)]\right)^2}{I_X(\theta)}$$

$$\text{Unbiased estimator } T(X) \text{ of } \tau(\theta)\text{: } \text{Var}(T(X)) \geq \frac{\left(\tau'(\theta)\right)^2}{I_X(\theta)}$$

$$\text{Multi-dimensional } T(X)\text{: } \text{Var}(T(X)) \geq \left(\nabla_\theta[\mathbb{E}_\theta(T(X))]\right)^T [I_\theta(X)]^{-1} \left(\nabla_\theta[\mathbb{E}_\theta(T(X))]\right)$$

(iii) Attainment: an estimator $T(X)$ achieves the C-R LB iff $f_\theta(x)$ is an exponential family of the form:

$$f_\theta(x) = \exp(A(\theta)) \exp(B(\theta)T(x)) \exp(c(x))$$

in which case $T(X)$ is an unbiased estimator for $E(T(X))$. Note that any linear function $f(T(X))$ is unbiased estimator for $\mathbb{E}(f(T(X)))$ and achieves the C-R LB. If we want to estimate $\tau(\theta)$ and τ is non-linear, our estimator will not achieve the C-R LB.

(iv) Nuisance parameters: usually increase the C-R LB, stays the same if parameters are orthogonal.

11. Maximum Likelihood estimation:

(i) If regular pdf, maximize the log-likelihood. If not, inspect and maximize the likelihood.

(ii) Check second derivative condition: $f''(x) < 0$ indicates maximum in univariate case. If in bivariate case:

$$\frac{\partial^2 f(x_0, y_0)}{\partial x^2} \frac{\partial^2 f(x_0, y_0)}{\partial y^2} - \frac{\partial^2 f(x_0, y_0)}{\partial x \partial y}^2 > 0 \text{ and } \frac{\partial^2 f(x_0, y_0)}{\partial x^2} < 0 \implies \text{local maximum}$$

(iii) Fisher Cramer: $\sqrt{n}(\hat{\theta} - \theta) \to N\left(0, [I_{X_i}(\theta)]^{-1}\right)$ or for iid samples: $\hat{\theta} \to N(\theta, [I_X(\theta)]^{-1}$

(iv) The invariance rule says that the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$. To find asymptotic distributions of MLE for a function of θ, $\tau(\theta)$, we use the delta method:

$$\sqrt{n}\left(\widehat{\tau(\theta)} - \tau(\theta)\right) \to N\left(0, \tau'(\theta)^2 [I_{X_i}(\theta)]^{-1}\right)$$

12. Hypothesis testing

(i) Power function: probability of rejecting the null hypothesis as a function of θ.

$$\pi_\phi(\theta) = \mathbb{E}_\theta[\phi]$$

(ii) Level: a test is size α fir testing $H_0 : \theta \in \Omega_0$ if:

$$\pi_\phi(\theta) \leq \alpha$$
$$\theta \in \Omega_0$$

(iii) Size: a test is size α if the worst T1 error rate for testing H_0: $\theta \in \Omega_0$ is α:

$$\pi_\phi(\theta) = \alpha$$
$$\theta \in \Omega_0$$

(iv) UMP: a test is UMP level α if the test has the smallest T2 error rate among all α level tests:

$$\pi_\phi(\theta) = \sup_{\phi' \text{ of level } \alpha} \pi_{\phi'}(\theta) \text{ for all } \theta \in \Omega_1$$

(v) Finding the MP test (2-point):

(i) For $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, the NP lemma says that the MP test is:

$$\phi(x) = \begin{cases} 0 & \lambda(x) = \frac{f_1(x)}{f_0(x)} < c \\ 1 & \lambda(x) = \frac{f_1(x)}{f_0(x)} > c \\ \gamma(x) & \lambda(x) = \frac{f_1(x)}{f_0(x)} = c \end{cases}$$

(ii) If $\lambda(x)$ is MLR with respect to a statistic $T(X)$, then the MP test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1 (> \theta_0)$ is:

$$\phi(x) = \begin{cases} 0 & T(X) < c' \\ 1 & T(X) > c' \\ \gamma(x) & T(X) = c' \end{cases}$$

Note: $T = \sum T(X_i)$ that falls out of exponential family can be used in place of the LR in LR test.

(iii) Finding c': find c' such that

$$\mathbb{E}_{\theta_0}[\phi(X)] = P_{\theta_0}(\phi(X) = 1) = P_{\theta_0}(T(X) > c') = \alpha$$

We can do this either by setting c' equal to the $(1 - \alpha)$ quantile of the distribution of $T(X)$ or by calculating the CDF.

(iv) Flipping signs: if our test depends on a MLR statistic, and our alternative hypothesis is *below* our null, $H_1 : \theta = \theta_1 (< \theta_0)$, flip the inequality directions above.

(vi) Finding the UMP (1-sided alternative):

(i) If a MP two-point test does not depend on the alternative $\theta_1 > \theta_0$, then it is UMP for the testing the one-sided alternative: $H_1 : \theta > \theta_0$.

(ii) MLR: if the likelihood ratio is an increasing function of $T(X)$, then the pdf has the monotone likelihood ratio property. For testing:

$$\begin{aligned} H_0 : \theta = \theta_0 \quad &\text{vs} \quad H_1 : \theta > \theta_0 \\ H_0 : \theta \leq \theta_0 \quad &\text{vs} \quad H_1 : \theta > \theta_0 \end{aligned}$$

the UMP level α test is:

$$\phi(t) = \begin{cases} 0 & \text{if } T < c_\alpha \\ 1 & \text{if } T > c_\alpha \\ \gamma_\alpha & \text{if } T = c_\alpha \end{cases}$$

where c_α and γ_α are chosen to satisfy:

$$P_{\theta_0}[T > c_\alpha] + \gamma_\alpha P_{\theta_0}[T = c_\alpha] = \alpha$$

(iii) Flipping signs: if our test depends on a MLR statistic, and our alternative hypothesis is *below* our null, $H_1 : \theta < \theta_0$, flip the inequality directions above.

(vii) General hypothesis tests with LR

(i) For general tests of the form:

$$\begin{aligned} H_0 : \theta \in \Omega_0 \quad &\text{vs} \quad H_1 : \theta \in \Omega - \Omega_0 \\ H_0 : \theta = \theta_0 \quad &\text{vs} \quad H_1 : \theta \neq \theta_0 \\ H_0 : \theta_1 = \theta_2 \quad &\text{vs} \quad H_1 : \theta_1 \neq \theta_2 \\ H_0 : a < \theta < b \quad &\text{vs} \quad H_1 : \theta \notin [a, b] \end{aligned}$$

Where $\Omega_0 \subset \Omega$ are subsets of $\mathbb{R}^k$, we consider the test based on the likelihood ratio statistic based on the restricted and unrestricted MLEs:

$$\lambda = \frac{\sup\limits_{\theta \in \Omega_0} f_\theta(x)}{\sup\limits_{\theta \in \Omega} f_\theta(x)} = \frac{f_{\hat{\theta}_0}(x)}{f_{\hat{\theta}}(x)}$$

The LRT form is:

$$\phi(x) = \begin{cases} 0 & \text{if } \lambda > c \quad [\text{Note: not } \lambda < c!] \\ 1 & \text{if } \lambda < c \quad [\text{Note: not } \lambda > c!] \\ \gamma_\alpha & \lambda = c_\alpha \end{cases}$$

(ii) Wilk's theorem: If we have an iid sample from a regular distribution family satisfying the Cramer-Wald conditions, the asymptotic distribution of the LR statistic is:

$$-2\log(\lambda) \xrightarrow{d} \chi^2 d - d_0$$

Where $d = \dim(\Omega)$ and $d = \dim(\Omega_0)$ (number of free parameters). Then the LRT is:

$$\phi(x) = \begin{cases} 0 & \text{if } -2\log(\lambda) < \chi^2_{d-d_0} \quad (1-\alpha)\text{-quantile} \\ 1 & \text{if } -2\log(\lambda) > \chi^2_{d-d_0} \quad (1-\alpha)\text{-quantile} \\ \gamma_\alpha & -2\log(\lambda) = \chi^2_{d-d_0} \quad (1-\alpha)\text{-quantile} \end{cases}$$

13. Useful inequalities:

(i) Cauchy-Schwarz: $[\mathbb{E}(XY)]^2 \leq \mathbb{E}(X^2)E(Y^2)$ and $[\text{Cov}(X,Y)]^2 \leq \text{Var}(X)\text{Var}(Y)$

(ii) Jensen: For convex function φ, $\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X))$