Preliminary

Theorem 1.1 (Cauchy Schwarz). Suppose X and Y are random variables, then:

$$\mathbb{C}^2(X,Y) \leq \mathbb{V}(X) \cdot \mathbb{V}(Y), \tag{1.1}$$

where equality holds $\iff$ Y and X are linearly related, $Y = aX + b$. □

1.1 Exercises

i. Prove (1.1) starting from the observation that $0 \leq \mathbb{V}(tX + Y)$ for a constant t.

Data Reduction

2.1 Notation

Suppose $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ is a random sample of size n, with realization $\boldsymbol{y} = (y_1, \cdots, y_n)$, from a distribution with joint density $f(\boldsymbol{y}|\theta) = f(y_1, \cdots, y_n|\theta)$.

2.2 Sufficiency

Definition 2.1. A statistic T is **sufficient** for θ if the conditional distribution of the sample $\boldsymbol{Y}$ given T does not depend on θ. ■

Theorem 2.1 (Factorization). A statistic $T(\boldsymbol{y})$ is sufficient for $\theta \iff$ for $\forall(\boldsymbol{y}, \theta)$ the joint density factors as:

$$f(\boldsymbol{y}|\theta) = g\big\{T(\boldsymbol{y})|\theta\big\}h(\boldsymbol{y}).$$

□

Proof. ($\implies$) If T is sufficient for θ, then $\mathbb{P}\big\{\boldsymbol{Y} = \boldsymbol{y}|T = t(\boldsymbol{y})\big\}$ does not depend on θ. Express the joint density as:

$$\begin{aligned} f(\boldsymbol{y}|\theta) &= \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}|\theta) \\ &= \mathbb{P}\big\{\boldsymbol{Y} = \boldsymbol{y} \cap T = t(\boldsymbol{y})|\theta\big\} \\ &= \mathbb{P}\big\{T = t(\boldsymbol{y})|\theta\big\} \cdot \mathbb{P}\big\{\boldsymbol{Y} = \boldsymbol{y}|T = t(\boldsymbol{y})\big\} \\ &= g(T|\theta) \cdot h(\boldsymbol{y}). \end{aligned}$$

($\impliedby$) Suppose the factorization exists. Define the subset $\mathcal{A}(\boldsymbol{y})$ of the sample space $\mathcal{Y}$:

$$\mathcal{A}(\boldsymbol{y}) = \big\{\boldsymbol{u} \in \mathcal{Y} : t(\boldsymbol{u}) = t(\boldsymbol{y})\big\}.$$

That is, $\mathcal{A}(\boldsymbol{y})$ contains those realizations of $\boldsymbol{Y}$ that lead to the same sufficient statistic as $\boldsymbol{y}$. The density of T is expressible as:

$$\mathbb{P}\big\{T = t(\boldsymbol{y})|\theta\big\} = \sum_{\boldsymbol{u}\in\mathcal{A}(\boldsymbol{y})} f(\boldsymbol{u}|\theta) = \sum_{\boldsymbol{u}\in\mathcal{A}(\boldsymbol{y})} g\big\{t(\boldsymbol{y})|\theta\big\}h(\boldsymbol{u}) = g\big\{t(\boldsymbol{y})|\theta\big\} \sum_{\boldsymbol{u}\in\mathcal{A}(\boldsymbol{y})} h(\boldsymbol{u}).$$

The distribution of the data given T is:

$$\begin{aligned}
\mathbb{P}\big\{\boldsymbol{Y} = \boldsymbol{y}|T = t(\boldsymbol{y})\big\} &= \frac{\mathbb{P}\big\{\boldsymbol{Y} = \boldsymbol{y} \cap T = t(\boldsymbol{y})\big\}}{\mathbb{P}\big\{T = t(\boldsymbol{y})\big\}} \\
&\stackrel{*}{=} \frac{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}\big\{T = t(\boldsymbol{y})\big\}} \\
&= \frac{g\big\{t(\boldsymbol{y})|\theta\big\}h(\boldsymbol{y})}{g\big\{t(\boldsymbol{y})|\theta\big\}\sum_{\boldsymbol{u}\in\mathcal{A}(\boldsymbol{y})} h(\boldsymbol{u})} \\
&= \frac{h(\boldsymbol{y})}{\sum_{\boldsymbol{u}\in\mathcal{A}(\boldsymbol{y})} h(\boldsymbol{u})}.
\end{aligned}$$

Equality $\stackrel{*}{=}$ follows since the even $\{\boldsymbol{Y} = \boldsymbol{y}\}$ is a subset of the event $\{T = t(\boldsymbol{y})\}$. That is, $\boldsymbol{Y} = \boldsymbol{y} \implies T = t(\boldsymbol{y})$, but not conversely. ■

Definition 2.2. An **exponential family** density takes the form:

$$f(y|\theta) = h(y)c(\theta)\exp\left\{\sum_{k=1}^{K} \omega_k(\theta)t_k(y)\right\}, \tag{2.1}$$

with the support of y not depending on θ.

The **canonical parameterization** of (2.1) is:

$$f(y|\eta) = h(y)c(\eta)\exp\left\{\sum_{k=1}^{K} \eta_k t_k(y)\right\}.$$

If the parameter space of η includes an open K-dimensional rectangle, then the exponential family is **full-rank**. Otherwise, it is **curved**. ■

Theorem 2.2 (Exponential Family). Suppose each Y_i follows an exponential family distribution (2.1), then the sufficient statistics for θ are:

$$\boldsymbol{T} = \left(\sum_{i=1}^{n} t_1(Y_i), \cdots, \sum_{i=1}^{n} t_K(Y_i)\right).$$

If the exponential family has full rank, then $\boldsymbol{T}$ is also complete. See Casella & Berger (2002) 6.2.10 and 6.2.25. □

2.3 Completeness

Definition 2.3. A statistic T is **complete** if $\mathbb{E}\{g(T)\} = 0$ for $\forall\theta \implies g(T) = 0$ with probability one. ■

Definition 2.4. A statistic A whose distribution does not depend on θ is **ancillary**. ■

Theorem 2.3 (Basu's). If T is a complete sufficient statistic, then T is independent of every ancillary statistic. □

Proof. Suppose A is ancillary for θ, and that T is complete and sufficient. Since A is ancillary, $\mathbb{P}(A = a)$ does not depend on θ. Define the subset $\mathcal{A}(\boldsymbol{y})$ of $\mathcal{Y}$:

$$\mathcal{A}(\boldsymbol{y}) = \big\{\boldsymbol{u} \in \mathcal{Y} : a(\boldsymbol{u}) = a(\boldsymbol{y})\big\}.$$

The distribution of A given T is expressible as:

$$\mathbb{P}\big\{A = a(\boldsymbol{y})|T = t(\boldsymbol{y})\big\} = \sum_{\boldsymbol{u}\in\mathcal{A}(\boldsymbol{y})} \mathbb{P}\big\{\boldsymbol{Y} = \boldsymbol{u}|T = t(\boldsymbol{y})\big\}.$$

Since T is *sufficient*, $\mathbb{P}\big\{\boldsymbol{Y} = \boldsymbol{u}|T = t(\boldsymbol{y})\big\}$ does not depend on θ, therefore neither does $\mathbb{P}\big\{A = a(\boldsymbol{y})|T = t(\boldsymbol{y})\big\}$. Define:

$$g(t) = \mathbb{P}(A = a|T = t) - \mathbb{P}(A = a).$$

Since neither $\mathbb{P}(A = a|T = t)$ (by sufficiency) or $\mathbb{P}(A = a)$ (by ancillarity) depend on θ, $g(T)$ is a valid statistic. By iterated expectation:

$$\mathbb{E}\big\{g(T)\big\} = \mathbb{E}\big\{\mathbb{P}(A = a|T = t)\big\} - \mathbb{P}(A = a) = \mathbb{P}(A = a) - \mathbb{P}(A = a) = 0.$$

Since T is complete, $\mathbb{P}(A = a|T = t) = \mathbb{P}(A = a)$ with probability one. Conclude that A is independent of T. ■

2.4 Exercises

i. Suppose $Y_i \sim N(\mu, \sigma^2)$. Show that $(\bar{Y}, S^2)$ are sufficient for (μ, σ^2).

ii. Suppose $Y_i \sim U(0, \theta)$. Show that $\max_i Y_i$ is complete and sufficient for θ.

iii. Suppose $Y_i \sim g(y - \theta)$. Show that $Y_{(n)} - Y_{(1)}$ is ancillary for θ.

iv. Suppose $Y_i \sim \theta^{-1}g(\theta^{-1}y)$. Show that $Y_i/\bar{Y}$ is ancillary for θ.

v. Find the complete and sufficient statistics for these distributions:

   (a) Binomial.

(b) Poisson.

(c) Gamma.

vi. (**Exponential family**):

(a) Show that for a canonical-form exponential family distribution:

$$c(\eta) = \left(\int h(\boldsymbol{y}) \exp\left\{ \sum_{k=1}^{K} \eta_k t_k(y) \right\} dy \right)^{-1}.$$

(b) Derive the moment generating function of the canonical-form exponential family distribution.

(c) Obtain expressions for $\mathbb{E}\{t_k(Y)\}$ and $\mathbb{C}\{t_k(Y), t_l(Y)\}$, $k \neq l$.

Estimation

Definition 3.1. An **estimator** is a statistic, a random function of the data, intended to estimate a parameter θ. An **estimate** is a realization of an estimator. ■

Discussion 3.1 (**Satterthwaite Approximation**). **Method of moments** is a technique for deriving estimators in which sample moments are matched with population moments to obtain a system of simultaneous equations. Suppose $Y_i \sim \chi^2_{\nu_i}(0)$. Consider approximating the distribution of $T = \sum_{i=1}^{n} \omega_i Y_i$, where the ω_i are known weights, by a $\chi^2_\nu(0)$ distribution. In particular, the problem is to find ν such that:

$$T = \sum_{i=1}^{n} \omega_i Y_i \mathrel{\dot\sim} \frac{\chi^2_\nu(0)}{\nu}.$$

Equating $\mathbb{E}(T) = \sum_{i=1}^{n} \omega_i \nu_i$ with $\mathbb{E}(\chi^2_\nu/\nu) = 1$ gives the constraint:

$$\sum_{i=1}^{n} \omega_i \nu_i = 1. \tag{3.1}$$

The second moment of the $\chi^2_\nu(0)$ distribution is $\mathbb{E}\big\{(\chi^2_\nu)^2\big\} = \nu(\nu+2)$. Equating $\mathbb{E}(T^2)$ with $\mathbb{E}\big\{(\chi^2_\nu)^2/\nu^2\big\} = 1 + 2/\nu$ and solving for ν gives:

$$\hat{\nu} = \frac{2}{\hat{\mathbb{E}}\big(T^2\big) - 1} = \frac{2}{\left(\sum_{i=1}^{n} \omega_i Y_i\right)^2 - 1}. \tag{3.2}$$

Since the estimator in (3.2) can be negative, consider instead:

$$\begin{aligned} \mathbb{E}(T^2) &= \mathbb{V}(T) + \mathbb{E}^2(T) \\ &= \mathbb{E}^2(T) \left\{ \frac{\mathbb{V}(T)}{\mathbb{E}^2(T)} + 1 \right\}. \end{aligned}$$

Setting the leading factor of $\mathbb{E}^2(T) \overset{\text{Set}}{=} 1$, since $\mathbb{E}(T) = 1$ under (3.1), and equating:

$$\mathbb{E}\left\{\frac{(\chi_\nu^2)^2}{\nu^2}\right\} = 1 + \frac{2}{\nu} \overset{\text{Set}}{=} \left\{\frac{\mathbb{V}(T)}{\mathbb{E}^2(T)} + 1\right\},$$

gives the improved estimator:

$$\hat{\nu} = \frac{2\hat{\mathbb{E}}^2(T)}{\hat{\mathbb{V}}(T)}. \tag{3.3}$$

The numerator may be approximated:

$$\hat{\mathbb{E}}(T) = \sum_{i=1}^n \omega_i Y_i.$$

Taking the variance of T analytically:

$$\mathbb{V}\left(\sum_{i=1}^n \omega_i Y_i\right) \overset{\text{IND}}{=} \sum_{i=1}^n \omega_i^2 \mathbb{V}(Y_i) = \sum_{i=1}^n \omega_i^2 \cdot 2\nu_i \overset{*}{=} 2\sum_{i=1}^n \omega_i^2 \cdot \frac{\mathbb{E}^2(Y_i)}{\nu_i},$$

where equality $\overset{*}{=}$ follows from $\mathbb{E}(Y_i) = \nu_i$. Making the approximation:

$$\hat{\mathbb{V}}(T) = 2\sum_{i=1}^n \omega_i^2 \cdot \frac{\hat{\mathbb{E}}^2(Y_i)}{\nu_i} = 2\sum_{i=1}^n \omega_i^2 \frac{Y_i^2}{\nu_i},$$

the final form of the Satterthwaite estimator in (3.3) is:

$$\hat{\nu} = \frac{(\sum_{i=1}^n \omega_i Y_i)^2}{\sum_{i=1}^n \omega_i^2 \frac{Y_i^2}{\nu_i}}. \tag{3.4}$$

(Source: Casella & Berger, 7.2.3.) ♠

3.1 Likelihood

Definition 3.2. The **likelihood** $L(\theta|\boldsymbol{y}) = f(\boldsymbol{y}|\theta)$ is the joint density of the observed data viewed as a function of θ. The *log likelihood* is denoted:

$$\ell_n(\theta) \equiv \ln f(\boldsymbol{y}|\theta).$$

The **maximum likelihood estimate** (MLE) of θ maximizes the log likelihood:

$$\hat{\theta}_n \equiv \arg\max_{\theta \in \Theta} \ell_n(\theta).$$

The sample **score** for θ is the gradient of the log likelihood with respect to θ:

$$\mathcal{U}_n(\theta) \equiv \frac{\partial \ell_n}{\partial \theta}.$$

Here the subscript n distinguishes the sample score from the unit score:

$$u_i(\theta) \equiv \frac{\partial}{\partial\theta} \ln f(y_i|\theta).$$

The MLE is often obtained by solving the score equations:

$$\mathcal{U}_n(\theta) \overset{\text{Set}}{=} 0.$$

The *Hessian* for θ is the second derivative of the log likelihood in θ:

$$\mathcal{H}_n(\theta) \equiv \frac{\partial^2 \ell_n}{\partial\theta\partial\theta'}$$

The **observed information** for θ is the negative Hessian:

$$\mathcal{J}_n(\theta) \equiv -\mathcal{H}_n(\theta).$$

The **Fisher information** is the variance of the score:

$$\mathcal{I}_n(\theta) \equiv \mathbb{V}\big\{\mathcal{U}_n(\theta)\big\}.$$

The unit Fisher information is the variance of the unit score:

$$\iota(\theta) \equiv \mathbb{V}\{u_i(\theta)\}.$$

For exponential family distributions, the Fisher information coincides with the negative expected Hessian:

$$\mathcal{I}_n(\theta) \overset{*}{=} -\mathbb{E}\big\{\mathcal{H}_n(\theta)\big\}.$$

■

Theorem 3.1 (Asymptotic Normality). For $Y_i \overset{\text{IID}}{\sim} f(y|\theta)$, suppose the following conditions are satisfies:

- θ is an interior point of the parameter space Θ.
- θ is *identified*, meaning $\theta_1 \neq \theta_2$ implies $F(y|\theta_1) \neq F(y|\theta_2)$ for at least some y.
- The first 3 partial derivatives of $\ell(\theta)$ exist for y in the support of $F(y|\theta)$.
- The 3rd derivatives of $\ell(\theta)$ is dominated element-wise by an integrable $g(y)$:

$$\left|\frac{\partial^3 \ln f(y|\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right| \leq g(y),$$

  where $\int g(y)dF(y|\theta_0) < \infty$.

- For $\theta \in \Theta$, the unit score has expectation zero $\mathbb{E}\{u_i(\theta)\} = 0$, and the unit Fisher information $\iota(\theta) = \mathbb{V}\{u_i(\theta)\}$ is positive definite.
- The solution $\hat{\theta}_n$ to the sample score equation $\mathcal{U}_n(\theta) \overset{\text{Set}}{=} 0$ is consistent for θ, meaning:

$$\lim_{n\to\infty} \mathbb{P}\big\{||\hat{\theta}_n - \theta|| > \epsilon\big\} = 0.$$

Then for $n \to \infty$:

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \rightsquigarrow N\big(0, \iota^{-1}\big), \tag{3.5}$$

where the limiting variance is the *inverse unit Fisher information*. □

Proof. The proof follows from asymptotic normality of M-estimators. See (e.g.) Boos and Stefanski (2013) theorem 7.2. ■

Lemma 3.1 (Invariance Principle). If $\hat{\theta}_n$ maximizes the log likelihood $\ell_n(\theta)$ and $\tau(\theta)$ is some function of θ, then the MLE of τ is $\hat{\tau}_n = \tau(\hat{\theta}_n)$. ■

Proof. Since τ is not necessarily bijective, the *induced log likelihood* of τ is *defined* as:

$$\ell_n^*(t) = \sup_{\{\theta:\tau(\theta)=t\}} \ell_n(\theta).$$

Since the iterated maximization is equal to unconditional maximization:

$$\sup_{t\in\mathcal{T}} \ell_n^*(t) = \sup_{t\in\mathcal{T}} \sup_{\{\theta:\tau(\theta)=t\}} \ell_n(\theta) = \sup_{\theta\in\Theta} \ell_n(\theta).$$

That is, the maximum of the induced log likelihood coincides with the maximum of the original log likelihood: $\ell_n^*(\hat{\tau}_n) = \ell_n(\hat{\theta}_n)$. Finally, since $\ell_n(\hat{\theta}_n)$ is expressible as:

$$\ell_n(\hat{\theta}_n) = \sup_{\{\theta:\tau(\theta)=\tau(\hat{\theta}_n)\}} \ell_n(\theta),$$

and by definition:

$$\sup_{\{\theta:\tau(\theta)=\tau(\hat{\theta}_n)\}} \ell_n(\theta) = \ell_n^*\big\{\tau(\hat{\theta}_n)\big\},$$

conclude that $\ell_n^*(\hat{\tau}_n) = \ell_n^*\big\{\tau(\hat{\theta}_n)\big\}$, or $\hat{\tau}_n = \tau(\hat{\theta}_n)$. ■

3.2 Evaluation of Estimators

Definition 3.3. The **mean squared error** (MSE) of an estimator $\hat{\theta}$ of θ is:

$$\text{MSE} = \mathbb{E}\big(\hat{\theta} - \theta\big)^2.$$

■

Definition 3.4. The bias of an estimator is the difference between its expectation and the true parameter:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

■

Lemma 3.2 (**Bias-Variance Decomposition**)**.** The MSE of an estimator decomposes as:

$$\text{MSE} = \mathbb{V}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}). \tag{3.6}$$

In the case of an *unbiased* estimator, the MSE is the variance. ■

Definition 3.5. $\hat{\theta}$ is the **uniform minimum variance unbiased estimator** (UMVUE) of θ if $\mathbb{E}(\hat{\theta}) = \theta$, and for any other estimator $\tilde{\theta}$ with $\mathbb{E}(\tilde{\theta}) = \theta$:

$$\mathbb{V}(\hat{\theta}) \leq \mathbb{V}(\tilde{\theta}).$$

■

Theorem 3.2 (*Uniqueness*)**.** If the UMVUE of θ exists, then it is unique. □

Proof. Suppose not. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ denote two UMVUEs of θ. Define:

$$\bar{\theta} = \frac{\hat{\theta}_1}{2} + \frac{\hat{\theta}_2}{2}.$$

Let $\varsigma^2 = \mathbb{V}(\hat{\theta}_1) = \mathbb{V}(\hat{\theta}_2)$. The variance of $\bar{\theta}$:

$$\begin{aligned} \text{Var}(\bar{\theta}) &= \frac{1}{4}\mathbb{V}(\hat{\theta}_1) + \frac{1}{4}\mathbb{V}(\hat{\theta}_2) + \frac{1}{2}\mathbb{C}(\hat{\theta}_1, \hat{\theta}_2) \\ &\overset{*}{\leq} \frac{1}{4}\varsigma^2 + \frac{1}{4}\varsigma^2 + \frac{1}{2}\sqrt{\mathbb{V}(\hat{\theta}_1)\mathbb{V}(\hat{\theta}_2)} = \varsigma^2, \end{aligned}$$

where $\overset{*}{\leq}$ is an application of the Cauchy-Schwarz inequality (1.1). If the inequality in strict, then neither $\hat{\theta}_1$ nor $\hat{\theta}_2$ is an UMVUE. Otherwise, $\hat{\theta}_2 = a\hat{\theta}_1 + b$, but then:

$$\mathbb{C}(\hat{\theta}_1, \hat{\theta}_2) = a\mathbb{V}(\hat{\theta}_1) = a\varsigma^2 \implies a = 1.$$

Moreover, to maintain unbiasedness:

$$\mathbb{E}(\hat{\theta}_2) = \mathbb{E}(\hat{\theta}_1) + b = \theta + b \implies b = 0.$$

Conclude that $\hat{\theta}_2 = \hat{\theta}_1$. ■

3.3 Cramer Rao Lower Bound

Theorem 3.3 (Cramer Rao Lower Bound). Suppose that $\boldsymbol{Y}$ is a random sample of size n, and that $\hat{\theta} = \hat{\theta}(\boldsymbol{Y})$ is an estimator satisfying:

$$\frac{d}{d\theta}\mathbb{E}(\hat{\theta}) = \int \frac{\partial}{\partial\theta}\big\{\hat{\theta}(\boldsymbol{y})f(\boldsymbol{y}|\theta)\big\}d\boldsymbol{y}, \tag{3.7}$$

and $\mathbb{V}(\hat{\theta}) < \infty$. Then:

$$\mathbb{V}(\hat{\theta}) \geq \frac{\big\{\frac{d}{d\theta}\mathbb{E}(\hat{\theta})\big\}^2}{\mathbb{E}\big\{\frac{d}{d\theta}\ln f(\boldsymbol{y}|\theta)\big\}^2}. \tag{3.8}$$

□

Proof. Applying (3.7):

$$\begin{aligned}
\frac{d}{d\theta}\mathbb{E}(\hat{\theta}) &= \int \frac{\partial}{\partial\theta}\big\{\hat{\theta}(\boldsymbol{y})f(\boldsymbol{y}|\theta)\big\}d\boldsymbol{y} = \int \hat{\theta}(\boldsymbol{y})\frac{\partial f(\boldsymbol{y}|\theta)}{\partial\theta}d\boldsymbol{y} \\
&= \int \hat{\theta}(\boldsymbol{y})\frac{\frac{\partial f(\boldsymbol{y}|\theta)}{\partial\theta}}{f(\boldsymbol{y}|\theta)} \cdot f(\boldsymbol{y}|\theta)d\boldsymbol{y} \\
&= \mathbb{E}\left\{\hat{\theta}(\boldsymbol{y}) \cdot \frac{\partial}{\partial\theta}\ln f(\boldsymbol{y}|\theta)\right\}.
\end{aligned}$$

Identify $\partial_\theta \ln f(\boldsymbol{y}|\theta)$ as the sample score for θ:

$$\frac{d}{d\theta}\mathbb{E}(\hat{\theta}) = \mathbb{E}\left\{\hat{\theta}(\boldsymbol{y}) \cdot \mathcal{U}_n(\theta)\right\}.$$

Since the score has expectation zero ($\mathbb{E}\{\mathcal{U}_n(\theta)\} = 0$):

$$\frac{d}{d\theta}\mathbb{E}(\hat{\theta}) = \mathbb{E}\left\{\hat{\theta}(\boldsymbol{y}) \cdot \mathcal{U}_n(\theta)\right\} = \mathbb{C}\big\{\hat{\theta}(\boldsymbol{y}), \mathcal{U}_n(\theta)\big\}.$$

Likewise, since the score has expectation zero:

$$\mathbb{V}\big\{\mathcal{U}_n(\theta)\big\} = \mathbb{E}\big\{\mathcal{U}_n^2(\theta)\big\}.$$

Applying the Cauchy-Schwarz inequality (1.1) to $X = \hat{\theta}$ and $Y = \mathcal{U}_n(\theta)$:

$$\left\{\frac{d}{d\theta}\mathbb{E}(\hat{\theta})\right\}^2 = \mathbb{C}^2\big\{\hat{\theta}(\boldsymbol{y}), \mathcal{U}_n(\theta)\big\} \leq \mathbb{V}(\hat{\theta}) \cdot \mathbb{V}\big\{\mathcal{U}_n(\theta)\big\} = \mathbb{V}(\hat{\theta}) \cdot \mathbb{E}\big\{\mathcal{U}_n^2(\theta)\big\}.$$

■

Remark 3.1. The denominator of the Cramer Rao lower bound (CRLB) (3.8) is the variance of the score, which is the Fisher information:

$$\mathbb{E}\left\{\frac{d}{d\theta}\ln f(\boldsymbol{y}|\theta)\right\}^2 = \mathbb{E}\big\{\mathcal{U}_n^2(\theta)\big\} = \mathbb{V}\big\{\mathcal{U}_n^2(\theta)\big\} = \mathcal{I}_n(\theta).$$

In the case of an unbiased estimator $\mathbb{E}\{\hat{\theta}\} = \theta$, the CRLB reduces to:

$$\mathbb{V}(\hat{\theta}) \geq \mathcal{I}_n^{-1}(\theta).$$

If the sample is IID, then $\mathcal{I}_n(\theta) = n\iota(\theta)$, and:

$$\mathbb{V}(\hat{\theta}) \geq \big\{n\iota(\theta)\big\}^{-1} \tag{3.9}$$

The right hand side of (3.9) is identically the limiting variance of the MLE (3.5).

The CRLB only applies when differentiation in θ commutes with integration in $\boldsymbol{y}$ (3.7). In general, this condition will fail wheno the support of $\boldsymbol{y}$ depends on θ. ♦

Lemma 3.3 (**Fisher Information**)**.** If $\boldsymbol{Y}$ is a random sample from a density $f(\boldsymbol{y}|\theta)$ that satisfies:

$$\frac{d}{d\theta}\mathbb{E}\big\{\mathcal{U}_n(\theta)\big\} = \frac{d}{d\theta}\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f(\boldsymbol{y}|\theta)\right\} = \int \frac{\partial}{\partial\theta}\left[\left\{\frac{\partial}{\partial\theta}\ln f(\boldsymbol{y}|\theta)\right\} f(\boldsymbol{y}|\theta)\right] d\boldsymbol{y},$$

then:

$$\mathcal{I}_n(\theta) = \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f(\boldsymbol{y}|\theta)\right\}^2 = -\mathbb{E}\left\{\frac{\partial^2}{\partial\theta^2}\ln f(\boldsymbol{y}|\theta)\right\}. \tag{3.10}$$

■

Remark 3.2. For exponential family densities (2.1), the sample Fisher information coincides with the negative expected Hessian of the sample log likelihood (3.10). ♦

Theorem 3.4 (**Attainment**)**.** Suppose $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ is an IID random sample, and that the CRLB condition (3.7) holds. An estimator T attains the CRLB for estimating $\tau(\theta) \iff$ the sample score is expressible as:

$$\mathcal{U}_n(\theta) = \frac{\partial}{\partial\theta}\ln f(\boldsymbol{y}|\theta) = a(\theta)\big\{T(\boldsymbol{y}) - \tau(\theta)\big\},$$

for some function $a(\theta)$ not depending on $\boldsymbol{y}$. □

Proof. Proof of the CRLB made use of the Cauchy-Schwarz inequality (1.1). Letting $X = T$ and $Y = \mathcal{U}_n(\theta)$:

$$\mathbb{C}^2\big\{T, \mathcal{U}_n(\theta)\big\} \leq \mathbb{V}(T)\cdot\mathbb{V}\big\{\mathcal{U}_n(\theta)\big\}.$$

Equality is attained if and only if:

$$\mathcal{U}_n(\theta) = aT + b.$$

Further, since the sample score must have expectation zero:

$$0 = \mathbb{E}\big\{\mathcal{U}_n(\theta)\big\} = a\tau(\theta) + b \implies b = -a\tau(\theta).$$

■

3.4 Rao Blackwell

Theorem 3.5 (**Rao Blackwell**). Suppose $\tilde{\theta}$ is unbiased for θ, and that T is sufficient for θ. Define $\hat{\theta} = \mathbb{E}(\tilde{\theta}|T)$, then $\hat{\theta}$ is also unbiased for θ and:

$$\mathbb{V}(\hat{\theta}) \leq \mathbb{V}(\tilde{\theta}).$$

That is, $\hat{\theta}$ is a uniformly better estimator than $\tilde{\theta}$. □

Proof. By the definition of sufficiency, the distribution of the data $\boldsymbol{Y}$ given T does not depend on θ. Since $\tilde{\theta} = \tilde{\theta}(\boldsymbol{y})$ is a function of $\boldsymbol{y}$ only, the expectation:

$$\hat{\theta} = \mathbb{E}\big\{\tilde{\theta}(\boldsymbol{Y})|T\big\},$$

is in fact a *statistic.*

By iterated expectation:

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}\big\{\mathbb{E}(\tilde{\theta}|T)\big\} = \mathbb{E}(\tilde{\theta}) = \theta.$$

By law of total variance:

$$\begin{aligned}\mathbb{V}(\tilde{\theta}) &= \mathbb{V}\big\{\mathbb{E}(\tilde{\theta}|T)\big\} + \mathbb{E}\big\{\mathbb{V}(\tilde{\theta}|T)\big\} \\ &= \mathbb{V}(\hat{\theta}) + \mathbb{E}\big\{\mathbb{V}(\tilde{\theta}|T)\big\} \geq \mathbb{V}(\hat{\theta}).\end{aligned}$$

■

Example 3.1. Suppose $Y_i \overset{\text{IID}}{\sim} f(\boldsymbol{y})$, continuous but not necessarily parametric. An individual observation Y_i is unbiased for the mean $\mathbb{E}(Y_i) = \mu$. The sample order statistics $\big(Y_{(1)}, \cdots, Y_{(n)}\big)$ are always sufficient.

Applying the Rao Blackwell theorem to $\tilde{\theta} = Y_i$ and $T = \big(Y_{(1)}, \cdots, Y_{(n)}\big)$:

$$\hat{\theta} = \mathbb{E}\big(Y_i|Y_{(1)}, \cdots, Y_{(n)}\big) \overset{*}{=} \frac{1}{n}\sum_{i=1}^{n} Y_{(i)} = \bar{Y}.$$

Equality $\overset{*}{=}$ holds since the distribution of Y_i given all order statistics is discrete uniform on $Y_{(1)}, \cdots, Y_{(n)}$. ♠

Theorem 3.6 (**Lehmann Scheffe**). If T is a *complete sufficient statistic*, then $h(T)$ is the UMVUE of its expectation, provided $\mathbb{V}\big\{h(T)\big\} < \infty$. □

Proof. Suppose $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are two unbiased estimators of θ. Since T is sufficient, by the Rao Blackwell theorem $\hat{\theta}_1 = \mathbb{E}(\tilde{\theta}_1|T)$ and $\hat{\theta}_2 = \mathbb{E}(\tilde{\theta}_2|T)$ are two unbiased estimators of θ with variance no greater than the original estimators. Define:

$$g(T) = \hat{\theta}_1 - \hat{\theta}_2 = \mathbb{E}(\tilde{\theta}_1|T) - \mathbb{E}(\tilde{\theta}_2|T).$$

Since $\mathbb{E}\{g(T)\} = 0$ and T is complete, conclude that $\hat{\theta}_1 = \hat{\theta}_2$. Thus, given any initially unbiased estimator $\tilde{\theta}$, the unique estimator $\hat{\theta} = \mathbb{E}(\tilde{\theta}|T)$ is also unbiased and satisfies $\mathbb{V}(\hat{\theta}) \leq \mathbb{V}(\tilde{\theta})$. If $\mathbb{V}(\hat{\theta}) < \infty$, then $\hat{\theta}$ is the UMVUE (if not, there may be multiple best estimators of θ).

Now, consider estimating $\mathbb{E}\{h(T)\}$ by $h(T)$. Observe that $h(T)$ is unbiased and that $h(T) = \mathbb{E}\{h(T)|T\}$. Provided $\mathbb{V}\{h(T)\} < \infty$, $h(T)$ is the UMVUE of its expectation. ■

3.5 Exercises

i. Find the log likelihood, score, and Fisher information for IID random samples from the following exponential family distributions:
   - (a) Binomial.
   - (b) Poisson.
   - (c) Normal.
   - (d) Gamma.

ii. Prove the bias-variance decomposition (3.6).

iii. Prove that the sample information is the negative expected Hessian of the sample log likelihood (3.10).

iv. Prove that the sample mean is the UMVUE for these distributions:
   - (a) Binomial.
   - (b) Poisson.

Hypothesis Testing

4.1 Likelihood Ratio

Definition 4.1. The likelihood ratio for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is:

$$\lambda(\boldsymbol{Y}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\boldsymbol{Y})}{\sup_{\theta \in \Theta_1} L(\theta|\boldsymbol{Y})}. \tag{4.1}$$

■

Theorem 4.1. If $T = T(\boldsymbol{Y})$ is a sufficient statistic for θ, then the likelihood ratio statistic in (4.1) is expressible as:

$$\lambda(\boldsymbol{Y}) = \frac{\sup_{\theta \in \Theta_0} L\big\{\theta|T(\boldsymbol{Y})\big\}}{\sup_{\theta \in \Theta_1} L\big\{\theta|T(\boldsymbol{Y})\big\}}. \tag{4.2}$$

□

Remark 4.1. (4.2) indicates that the likelihood ratio statistic for θ should depend on the sample $\boldsymbol{Y}$ only through a sufficient statistic for θ. ♦

Example 4.1. Suppose $Y_i \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$, and that interest lies in making inferences about μ, with σ^2 regarded as a nuisance parameter. In particular, consider evaluating the $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$. The likelihood ratio statistic is:

$$\lambda(\boldsymbol{y}) = \frac{\sup_{\sigma^2 \in (0,\infty)} L(\mu_0, \sigma^2 | \boldsymbol{y})}{\sup_{\mu \in \mathbb{R}, \sigma^2 \in (0,\infty)} L(\mu, \sigma^2 | \boldsymbol{y})} = \frac{L(\mu_0, \tilde{\sigma}_0^2 | \boldsymbol{y})}{L(\hat{\mu}, \hat{\sigma}^2 | \boldsymbol{y})},$$

where $\tilde{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n \big(y_i - \mu_0\big)^2$ and $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n y_i$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \big(y_i - \hat{\mu}\big)^2$.

♠

4.2 Power

Definition 4.2. Define the **rejection region** $\mathcal{R}$ as the subset of the sample space $\mathcal{Y}$ for which a hypothesis test ϕ rejects:

$$\mathcal{R} = \big\{\boldsymbol{y} \in \mathcal{Y} : \phi \text{ rejects}\big\}.$$

The **retention region** $\mathcal{A} = \mathcal{Y}$ is the subset of the sample space for which the hypothesis test fails to reject:

$$\mathcal{A} = \big\{\boldsymbol{y} \in \mathcal{Y} : \phi \text{ does not reject}\big\}.$$

■

Definition 4.3. The **power function** $\beta(\theta)$ is the probability that a sample falls in the rejection region as a function of the true parameter θ:

$$\beta(\theta) = \mathbb{P}_\theta\big(\boldsymbol{Y} \in \mathcal{R}\big).$$

■

Remark 4.2. The power function of the ideal test is equal to zero for $\forall \theta \in \Theta_0$, and equal to one for $\forall \theta \in \Theta_1$. ♦

Definition 4.4. A **type I error** is the probability of rejecting the null hypothesis when the null hypothesis is true:

$$\text{Type I Error} = \mathbb{P}_\theta\big(\boldsymbol{Y} \in \mathcal{R}\big) \text{ when } \theta \in \Theta_0.$$

A **type II error** is the probability of retaining the null hypothesis when the null hypothesis is false:

$$\text{Type II Error} = \mathbb{P}_\theta\big(\boldsymbol{Y} \in \mathcal{A}\big) \text{ when } \theta \in \Theta_1.$$

■

Definition 4.5. A test is described as **size** α if:

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$

By contrast, a test is **level** α if:

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$

Every size α test is also level α. ■

Example 4.2. Suppose $Y_i \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$, with σ^2 assumed known. The LRT of $H_0 : \mu \leq \mu_0$ against $H_A : \mu > \mu_0$ rejects if:

$$\frac{(\bar{Y} - \mu_0)}{\sigma/\sqrt{n}} > \zeta_{1-\alpha}.$$

The power of this test is:

$$\begin{aligned}\beta(\mu) &= \mathbb{P}\left\{\frac{(\bar{Y} - \mu_0)}{\sigma/\sqrt{n}} > \zeta_{1-\alpha}\right\} = \mathbb{P}\left\{\frac{(\bar{Y} - \mu + \mu - \mu_0)}{\sigma/\sqrt{n}} > \zeta_{1-\alpha}\right\} \\ &= \mathbb{P}\left\{\frac{(\bar{Y} - \mu)}{\sigma/\sqrt{n}} > \zeta_{1-\alpha} - \frac{(\mu - \mu_0)}{\sigma/\sqrt{n}}\right\} = 1 - \Phi\left\{\zeta_{1-\alpha} - \frac{(\mu - \mu_0)}{\sigma/\sqrt{n}}\right\}.\end{aligned}$$

♠

4.3 Neymann-Pearson

Definition 4.6. Consider a class $\mathcal{C}$ of tests for evaluating $H_0 : \theta \in \Theta_0$ against the alternative $H_A : \theta \in \Theta_A$. A test with power function $\beta(\theta)$ is **uniformly most powerful** if $\beta(\theta) \geq \tilde{\beta}(\theta)$ for every $\theta \in \Theta_A$ and every power function $\tilde{\beta}$ belonging to a test in $\mathcal{C}$. ■

Theorem 4.2 (Neymann-Pearson Lemma). Consider testing $H_0 : \theta = \theta_0$ against $H_A : \theta = \theta_1$. Suppose the rejection region takes the form:

$$\mathcal{R} = \big\{\boldsymbol{y} \in \mathcal{Y} : f(\boldsymbol{y}|\theta_1) > k_\alpha f(\boldsymbol{y}|\theta_0)\big\}, \tag{4.3}$$

where k_α is chosen such that:

$$\mathbb{P}_{\theta_0}(\boldsymbol{Y} \in \mathcal{R}) = \alpha. \tag{4.4}$$

i. Any test with rejection region (4.3) that satisfies (4.4) is a UMP level-α test.

ii. If the preceding test exists, then every UMP level-α is size-α and has a rejection region that agrees with (4.3) a.e.

□

Remark 4.3. Note that k_α is chosen such that the probability $\boldsymbol{Y}$ falls in the rejection region $\mathcal{R}$ is α *under the null hypothesis* $H_0 : \theta = \theta_0$. ♦

Proof. (i.) Define the **test function**:

$$\phi(\boldsymbol{y}) = \mathbb{I}\big(\boldsymbol{y} \in \mathcal{R}\big),$$

where $\mathcal{R}$ is defined in (4.3) and satisfies (4.4). Let $\tilde{\phi}$ denote the test function of any other level-α test; let β and $\tilde{\beta}$ denote the corresponding power functions. The power function β is related to the test function via:

$$\beta(\theta) = \mathbb{P}_\theta(\boldsymbol{Y} \in \mathcal{R}) = \mathbb{E}_\theta\big\{\mathbb{I}(\boldsymbol{Y} \in \mathcal{R})\big\} = \int \phi(\boldsymbol{y}) dF(\boldsymbol{y}; \theta).$$

The function $0 \le g(\boldsymbol{y}) = \big\{\phi(\boldsymbol{y}) - \tilde{\phi}(\boldsymbol{y})\big\}\big\{f(\boldsymbol{x}|\theta_1) - k_\alpha f(\boldsymbol{x}|\theta_0)\big\}$ is since $\tilde{\phi} \in \{0, 1\}$, $\phi(\boldsymbol{y}) = 1$ if $f(\boldsymbol{x}|\theta_1) - k_\alpha f(\boldsymbol{x}|\theta_0) > 0$ and $\phi(\boldsymbol{y}) = 0$ if $f(\boldsymbol{x}|\theta_1) - k_\alpha f(\boldsymbol{x}|\theta_0) < 0$. Now:

$$\begin{aligned} 0 \le \int g(\boldsymbol{y}) d\boldsymbol{y} &= \int \big\{\phi(\boldsymbol{y}) - \tilde{\phi}(\boldsymbol{y})\big\}\big\{f(\boldsymbol{x}|\theta_1) - k_\alpha f(\boldsymbol{x}|\theta_0)\big\} d\boldsymbol{y} \\ &= \beta(\theta_1) - \tilde{\beta}(\theta_1) - k_\alpha\big\{\beta(\theta_0) - \tilde{\beta}(\theta_0)\big\}. \end{aligned}$$

Since ϕ is size-α and $\tilde{\phi}$ is level-α, $\beta(\theta_0) - \tilde{\beta}(\theta_0) \ge 0$, hence:

$$0 \le \beta(\theta_1) - \tilde{\beta}(\theta_1) - k_\alpha\big\{\beta(\theta_0) - \tilde{\beta}(\theta_0)\big\} \le \beta(\theta_1) - \tilde{\beta}(\theta_1).$$

(ii.) Suppose ϕ is defined as previously, and $\tilde{\phi}$ is another UMP level-α test. Since ϕ and $\tilde{\phi}$ are both UMP, $\beta(\theta_1) - \tilde{\beta}(\theta_1) = 0$. From the above, conclude that:

$$0 = \beta(\theta_0) - \tilde{\beta}(\theta_0) = \alpha - \tilde{\beta}(\theta_0).$$

That is, $\tilde{\phi}$ is also size-α. Consequently,

$$\int g(\boldsymbol{y}) d\boldsymbol{y} = \int \big\{\phi(\boldsymbol{y}) - \tilde{\phi}(\boldsymbol{y})\big\}\big\{f(\boldsymbol{x}|\theta_1) - k_\alpha f(\boldsymbol{x}|\theta_0)\big\} d\boldsymbol{y} = 0.$$

Since $f(\boldsymbol{x}|\theta_1) - k_\alpha f(\boldsymbol{x}|\theta_0) \ne 0$, the integral vanishes $\iff$ $\phi(\boldsymbol{y}) = \tilde{\phi}(\boldsymbol{y})$ a.e. ■

Corollary 4.1. Consider again testing $H_0 : \theta = \theta_0$ against $H_A : \theta = \theta_1$. Suppose T is sufficient for θ, and $g(t|\theta)$ is the density of the sufficient statistic. A test based on T is UMP level-α if it satisfies:

$$\mathcal{R} = \big\{t \in \mathcal{T} : g(t|\theta_1) > k_\alpha g(t|\theta_0)\big\},$$

where k_α is chosen such that:

$$\mathbb{P}_{\theta_0}(T \in \mathcal{T}) = \alpha.$$

♣

4.4 Karlin-Rubin

Definition 4.7. A family of densities $g(t|\theta)$ for a univariate random variable T has a **monotone likelihood ratio** if for $\theta_2 > \theta_1$, the ratio:

$$\frac{g(t|\theta_2)}{g(t|\theta_1)}, \tag{4.5}$$

is **non-decreasing** *as a function of* t. ■

Proposition 4.1. Suppose T has a monotone likelihood ratio (4.5), then T is **stochastically non-decreasing** in θ. That is, for $\theta_2 > \theta_1$:

$$G(t|\theta_2) \leq G(t|\theta_1), \tag{4.6}$$

♦

Proof. Define $H(t) = G(t|\theta_2) - G(t|\theta_1)$. The derivative is:

$$\frac{d}{dt}H(t) = g(t|\theta_2) - g(t|\theta_1) = g(t|\theta_1)\left(\frac{g(t|\theta_2)}{g(t|\theta_1)} - 1\right).$$

Since $g(t|\theta_1) > 0$ and $g(t|\theta_2)/g(t|\theta_1)$ is non-decreasing, the derivative of $H(t)$ can only change sign from negative to positive. Therefore, any interior critical point of $H(t)$ is a minimum, and the maximum of $H(t)$ must occur at the boundaries, $\{-\infty, \infty\}$. By the properties of distribution functions, $H(-\infty) = 0$ and $H(\infty) = 0$. Conclude that $H(t) \leq 0 \implies G(t|\theta_2) \leq G(t|\theta_1)$. ■

Corollary 4.2. If T has a monotone likelihood ratio, the for $\theta_2 > \theta_1$:

$$\mathbb{P}_{\theta_2}(T > t) \geq \mathbb{P}_{\theta_1}(T > t). \tag{4.7}$$

♣

Theorem 4.3. Consider testing $H_0 : \theta \leq \theta_0$ against $H_A : \theta > \theta_0$. Suppose T is sufficient for θ, and that $g(t|\theta)$ has a monotone likelihood ratio. Then, for any t_0, the test with rejection region:

$$\mathcal{R} = \big\{t \in \mathcal{T} : t > t_0\big\},$$

is a UMP level-α test, where $\alpha = \mathbb{P}_{\theta_0}(T > t_0)$. □

Proof. Let $\beta(\theta) = \mathbb{P}_\theta(T > t_0)$ denote the power function. By (4.7), $\beta(\theta)$ is non-decreasing. Therefore:

$$\sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = \mathbb{P}_{\theta_0}(T > t_0),$$

demonstrating this is a level $\alpha \equiv \mathbb{P}_{\theta_0}(T > t_0)$ test.

Fix $\theta_1 > \theta_0$, and define:

$$k = \inf_{t \in \mathcal{U}} \frac{g(t|\theta_1)}{g(t|\theta_0)},$$

where $\mathcal{U}$ is the set:

$$\mathcal{U} = \big\{t \in \mathcal{T} : t > t_0 \text{ and either } g(t|\theta_1) > 0 \text{ or } g(t|theta_0) > 0\big\}.$$

Now $t > t_0 \iff g(t|\theta_1) > kg(t|\theta_0)$. By the corollary to the Neymann-Pearson lemma, the test with rejection region:

$$\mathcal{R} = \big\{t \in \mathcal{T} : t > t_0\big\} = \big\{t \in \mathcal{T} : g(t|\theta_1) > kg(t|\theta_0)\big\},$$

is UMP for testing $H_0 : \theta = \theta_0$ against $H_A : \theta = \theta_1$. Since $\theta_1 > \theta_0$ was arbitrary, the test is UMP for $\forall \theta > \theta_0$. ■

Corollary 4.3. If T is sufficient for θ and $g(t|\theta)$ has a monotone likelihood ratio, then the rejection region of the UMP test of $H_0 : \theta \geq \theta_0$ against $H_A : \theta < \theta_0$ takes the form:

$$\mathcal{R} = \big\{t \in \mathcal{T} : t < t_0\big\},$$

where $\alpha = \mathbb{P}_{\theta_0}(T < t_0)$. ♣

4.5 p-Values

Definition 4.8. A **p-value** is a *statistic* $p(\boldsymbol{y}) \in [0, 1]$ such that p approaching zero provides increasing evidence against H_0. A p-value is **valid** if:

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta\big\{p(\boldsymbol{y}) \leq \alpha\big\} \leq \alpha.$$

■

4.6 Exercises

i. Prove (4.2).

ii. Prove the corollary to the Neymann-Pearson Lemma.

iii. Verify (4.7).

iv. Suppose $Y_i \sim N(\mu, \sigma^2)$ with σ^2 known.

   (a) Find the UMP, size α test of $H_0 : \mu \leq \mu_0$ against $H_A : \mu > \mu_0$.

(b) Show that a UMP test of $H_0 : \mu = \mu_0$ v. $H_0 : \mu \neq \mu_0$ DNE.

v. Suppose $Y_i \sim \text{Weibull}(\alpha, \lambda)$, with the shape α and rate λ parameters both unknown. Find the likelihood ratio test of $H_0 : \alpha = 1$ against $H_A : \alpha \neq 1$ in the presence of the nuisance parameter λ.

Confidence Intervals

5.1 Interval Estimators

Definition 5.1. An **interval estimator** of a scalar parameter θ is a pair of statistics $L(\boldsymbol{y})$ and $U(\boldsymbol{y})$, with $L(\boldsymbol{y}) \leq U(\boldsymbol{y})$, such that $L(\boldsymbol{y}) \leq \theta \leq U(\boldsymbol{y})$. ■

Definition 5.2. The **coverage probability** of an interval estimator is the probability the interval covers the true parameter θ:

$$\text{Coverage}(\theta) = \mathbb{P}_\theta\big\{(L \leq \theta) \cap (U \geq \theta)\big\}.$$

The **confidence coefficient** is the infimum of the coverage probability:

$$\gamma = \inf_{\theta \in \Theta} \mathbb{P}_\theta\big\{(L \leq \theta) \cap (U \geq \theta)\big\}.$$

■

Remark 5.1. In defining the coverage probability, the interval $[L, U]$, not the parameter, is random. In general, the coverage probability can depend on θ. When θ is unknown, we can only guarantee that the coverage probability is at least the confidence coefficient γ. ♦

5.2 Test Inversion

Example 5.1. Suppose $Y_i \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$ with σ^2 known. Consider testing $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$. The sample mean $\hat{\mu} = \bar{Y}$ is sufficient for μ. The rejection region of the UMP, unbiased, level-α test of $H_0 : \mu = \mu_0$ is:

$$\mathcal{R}(\mu_0) = \left(\boldsymbol{y} \in \mathcal{Y} : \frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{n}} > z_{1-\alpha/2}\right).$$

Under H_0, the probability that $\boldsymbol{Y}$ falls in the rejection region is:

$$\mathbb{P}_{\mu_0}\big\{\boldsymbol{Y} \in \mathcal{R}(\mu_0)\big\} = \mathbb{P}_{\mu_0}\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{n}} > z_{1-\alpha/2}\right) = \alpha.$$

Equivalently, the probability that $\boldsymbol{Y}$ falls in the retention region is:

$$\mathbb{P}_{\mu_0}\big\{\boldsymbol{Y} \in \mathcal{A}(\mu_0)\big\} = \mathbb{P}_{\mu_0}\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}\right) = 1 - \alpha.$$

Rearranging gives:

$$\mathbb{P}_{\mu_0}\left(\hat{\mu} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \hat{\mu} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Define:

$$L = \hat{\mu} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad U = \hat{\mu} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Now, $\mathbb{P}_{\mu_0}(L \le \mu_0 \le U) = 1 - \alpha$. Finally, observe that the last probability statement holds for every μ_0. Thus (L, U) provides a $(1 - \alpha)$ confidence interval for μ. ♠

Discussion 5.1. Recall that the *rejection region* $\mathcal{R}$ is defined as the subset of the sample space $\mathcal{Y}$ for which a test ϕ rejects, and the *retention region* $\mathcal{A}$ is the subset of the sample space for which ϕ fails to reject. In general, the retention region $\mathcal{A}$ depend on the value θ_0 of the parameter under H_0. In the previous example:

$$\mathcal{A}(\mu_0) = \left\{\boldsymbol{y} \in \mathcal{Y} : \hat{\mu} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \hat{\mu} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

The corresponding **confidence set** is the subset of parameter space for which θ is a plausible value, given the data:

$$\mathcal{C}(\boldsymbol{y}) = \left\{\mu \in \mathbb{R} : \hat{\mu} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \hat{\mu} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

The retention region and the confidence set are linked by the duality:

$$\boldsymbol{y} \in \mathcal{A}(\theta_0) \iff \theta_0 \in \mathcal{C}(\boldsymbol{y}).$$

♠

Theorem 5.1 (Duality). For each $\theta \in \Theta$, let $\mathcal{A}(\theta_0)$ denote the retention region of a level-α test of $H_0 : \theta = \theta_0$. Now, for each $\boldsymbol{y} \in \mathcal{Y}$, define the set:

$$\mathcal{C}(\boldsymbol{y}) = \big\{\theta \in \Theta : \boldsymbol{y} \in \mathcal{A}(\theta)\big\}.$$

Then $\mathcal{C}(\boldsymbol{y})$ is a $(1 - \alpha)$ confidence set for θ. Conversely, suppose $\mathcal{C}(\boldsymbol{y})$ is a $(1 - \alpha)$ confidence set for θ. For each $\theta \in \Theta$, define the set:

$$\mathcal{A}(\theta_0) = \big\{\boldsymbol{y} \in \mathcal{Y} : \theta_0 \in \mathcal{C}(\boldsymbol{y})\big\}.$$

Then $\mathcal{A}(\theta_0)$ is the retention region of a level-α test of $H_0 : \theta = \theta_0$. □

Example 5.2 (Inverting the LRT)**.** Suppose $Y_i \overset{\text{IID}}{\sim} F_\theta$. The rejection region for a likelihood ratio test of $H_0 : \theta = \theta_0$ is:

$$\mathcal{R}(\theta_0) = \big[\boldsymbol{y} \in \mathcal{Y} : -2\{\ell_n(\theta_0) - \ell_n(\hat{\theta})\} > \chi^2_{1,1-\alpha}\big],$$

where $\hat{\theta}$ is the MLE of θ, and $\chi^2_{1,1-\alpha}$ is the critical value of the $\chi^2_1(0)$ distribution.

The retention region is:

$$\mathcal{A}(\theta_0) = \big[\boldsymbol{y} \in \mathcal{Y} : -2\{\ell_n(\theta_0) - \ell_n(\hat{\theta})\} \leq \chi^2_{1,1-\alpha}\big].$$

Viewing the sample as fixed and the retention region as a function of the parameter gives the corresponding confidence set:

$$\mathcal{C}(\boldsymbol{y}) = \big[\theta \in \Theta : -2\{\ell_n(\theta_0) - \ell_n(\hat{\theta})\} \leq \chi^2_{1,1-\alpha}\big].$$

♠

Example 5.3 (P-Inversion)**.** Suppose $Y_i \overset{\text{IID}}{\sim} F_\theta$. Let $P(\boldsymbol{y};\theta_0)$ denote a p-value assessing $H_0 : \theta = \theta_0$ based on $\boldsymbol{y}$. For a level α test, the rejection region is:

$$\mathcal{R}(\theta_0) = \big\{\boldsymbol{y} \in \mathcal{Y} : P(\boldsymbol{y};\theta_0) \leq \alpha\big\}.$$

The retention region is:

$$\mathcal{A}(\theta_0) = \big\{\boldsymbol{y} \in \mathcal{Y} : P(\boldsymbol{y};\theta_0) > \alpha\big\}.$$

Viewing the sample as fixed and θ as variable, the confidence set is:

$$\mathcal{C}(\boldsymbol{y}) = \big\{\theta \in \Theta : P(\boldsymbol{y};\theta) > \alpha\big\}.$$

♠

Example 5.4 (**Clopper-Pearson Interval**)**.** Suppose $Y_i \overset{\text{IID}}{\sim} \text{Bern}(\pi)$. The total number of successes $T = \sum_{i=1}^n Y_i$ is sufficient for π. Define the function:

$$u(\theta) = \mathbb{P}\big\{\text{Binom}(n,\theta) \leq t_{\text{obs}}\big\}.$$

$u(\theta)$ is the p-value for testing $H_0 : \pi \geq \theta$ against $H_A : \pi < \theta$, and $u(\theta)$ is a *decreasing* function of θ. An upper confidence bound is given by:

$$U = \sup\left\{\theta \in (0,1) : u(\theta) > \frac{\alpha}{2}\right\}.$$

U is the largest value of θ for which $H_0 : \pi \geq \theta$ fails to reject at level $(\alpha/2)$.

Reciprocally, define the function:

$$l(\theta) = \mathbb{P}\big\{\text{Binom}(n, \theta) \geq t_{\text{obs}}\big\}.$$

$l(\theta)$ is the p-value for testing $H_0 : \pi \leq \theta$ against $H_A : \pi > \theta$, and $l(\theta)$ is an *increasing* function of θ. A lower confidence bound is given by:

$$L = \inf\left\{\theta \in (0,1) : l(\theta) > \frac{\alpha}{2}\right\}.$$

L is the smallest value of θ for which $H_0 : \pi \leq \theta$ fails to reject at level $(\alpha/2)$.

♠

5.3 Pivots

Definition 5.3. A **pivot** is a function $Q(\boldsymbol{Y}, \theta)$ of the data $\boldsymbol{Y}$ and parameter θ whose distribution no longer depend on θ. ■

Example 5.5. Suppose $Y_i \overset{\text{IID}}{\sim} U(0, \theta)$. A sufficient statistic for θ is the sample maximum $Y_{(n)} = \max(Y_1, \cdots, Y_n)$. Recall that, $Y_i \overset{d}{=} \theta X_i$, $Y_{(n)} \overset{d}{=} \theta X_{(n)}$, and $X_{(n)} \sim \text{Beta}(n, 1)$. The quantity $X_{(n)} = \theta^{-1} Y_{(n)}$ is pivotal for θ. To construct a confidence interval, we seek constants a and b such that:

$$\mathbb{P}(a \leq \theta^{-1} Y_{(n)} \leq b) = \mathbb{P}(a \leq X_{(n)} \leq b) = \int_a^b n t^{n-1} dt = t^n \Big|_a^b = b^n - a^n \overset{\text{Set}}{=} 1 - \alpha.$$

Having obtained $a < b$ numerically, a confidence interval for θ is given by:

$$\mathbb{P}\left(\frac{Y_{(n)}}{b} \leq \theta \leq \frac{Y_{(n)}}{a}\right) = 1 - \alpha.$$

♠

Example 5.6. Suppose $Y_i \overset{\text{IID}}{\sim} \text{Exp}(\lambda)$. A sufficient statistic for λ is the sum total $S = \sum_{i=1}^n Y_i$. Recall that, $Y_i \overset{d}{=} \lambda^{-1} X_i$, $S = \lambda^{-1} T$, where $T = \sum_{i=1}^n X_i$, and that $T \sim \text{Gamma}(n, 1)$. The quantity $T = \lambda S$ is pivotal for λ. To construct a confidence interval, we seek constants a and b such that:

$$\mathbb{P}(a \leq \lambda S \leq b) = \mathbb{P}(a \leq T \leq b) = \frac{1}{\Gamma(n)} \int_a^b t^{n-1} e^{-t} dt \overset{\text{Set}}{=} 1 - \alpha.$$

Having obtained $a < b$ numerically, a confidence interval for λ is given by:

$$\mathbb{P}\left(\frac{a}{S} \leq \lambda \leq \frac{b}{S}\right) = 1 - \alpha.$$

♠

5.4 Exercises

i. Prove the duality of confidence sets and hypothesis tests.

ii. Construct a confidence set for the rate λ of an exponential distribution by inverting the likelihood ratio test.

iii. Suppose $Y_i \overset{\text{IID}}{\sim} \text{Poi}(\lambda)$. Construct a Clopper-Pearson type confidence interval for λ.

iv. Suppose $Y_i \overset{\text{IID}}{\sim} \text{Bern}(\pi)$.

   (a) Find the variance stabilizing transformation $g(\cdot)$ of Y_i.

   (b) Use the variance stabilized random variable $Z_i = g(Y_i)$ to construct an asymptotic confidence interval for π.