STAT 713
MATHEMATICAL STATISTICS II

Spring 2018

Lecture Notes

Joshua M. Tebbs
Department of Statistics
University of South Carolina

Contents

6 Principles of Data Reduction

Complementary reading: Chapter 6 (CB). Sections 6.1-6.2.

6.1 Introduction

Recall: We begin by recalling the definition of a statistic. Suppose that $X_1, X_2, ..., X_n$ is an iid sample. A **statistic** $T = T(\mathbf{X}) = T(X_1, X_2, ..., X_n)$ is a function of the sample $\mathbf{X} = (X_1, X_2, ..., X_n)$. The only restriction is that T cannot depend on unknown parameters. For example,

$$T(\mathbf{X}) = \overline{X} \qquad T(\mathbf{X}) = X_{(n)} \qquad T(\mathbf{X}) = \left(\prod_{i=1}^{n} X_i\right)^{1/n} \qquad \mathbf{T}(\mathbf{X}) = \mathbf{X}.$$

Recall: We can think of $\mathbf{X}$ and T as functions:

- $(S, \mathcal{B}, P)$: probability space for random experiment
- $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_{\mathbf{X}}) \longrightarrow$ range space of $\mathbf{X}$
  - **Recall:** $\mathbf{X} : S \to \mathbb{R}^n$ is a random vector if $\mathbf{X}^{-1}(B) \equiv \{\omega \in S : \mathbf{X}(\omega) \in B\} \in \mathcal{B}$, for all $B \in \mathcal{B}(\mathbb{R}^n)$
  - $P_{\mathbf{X}}$: induced probability measure of $\mathbf{X}$; one-to-one correspondence with $F_{\mathbf{X}}(\mathbf{x})$
  - $\mathcal{X}$ = support of $\mathbf{X}$; $\mathcal{X} \subseteq \mathbb{R}^n$
- $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_T) \longrightarrow$ range space of T
  - $T : \mathbb{R}^n \to \mathbb{R}$, if T is a scalar statistic
  - P_T describes the (sampling) distribution of T; Chapters 4-5 (CB)
  - $\mathcal{T}$ = support of T; $\mathcal{T} \subseteq \mathbb{R}$; $\mathcal{T} = \{t : t = T(\mathbf{x}),\ \mathbf{x} \in \mathcal{X}\}$, the image set of $\mathcal{X}$ under T.

Conceptualization: A statistic T forms a **partition** of $\mathcal{X}$, the support of $\mathbf{X}$. Specifically, T partitions $\mathcal{X} \subseteq \mathbb{R}^n$ into sets

$$A_t = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) = t\},$$

for $t \in \mathcal{T}$. The statistic T **summarizes** the data $\mathbf{x}$ in that one can report

$$T(\mathbf{x}) = t \iff \mathbf{x} \in A_t$$

instead of reporting $\mathbf{x}$ itself. This is the idea behind **data reduction**. We reduce the data $\mathbf{x}$ so that they can be more easily understood without losing the meaning associated with the set of observations.

Example 6.1. Suppose X_1, X_2, X_3 are iid Bernoulli(θ), where $0 < \theta < 1$. The support of $\mathbf{X} = (X_1, X_2, X_3)$ is

$$\mathcal{X} = \{(0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\}.$$

Define the statistic

$$T = T(\mathbf{X}) = X_1 + X_2 + X_3.$$

The statistic T partitions $\mathcal{X} \subset \mathbb{R}^3$ into the following sets:

$$\begin{aligned} A_0 &= \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) = 0\} = \{(0,0,0)\} \\ A_1 &= \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) = 1\} = \{(1,0,0), (0,1,0), (0,0,1)\} \\ A_2 &= \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) = 2\} = \{(1,1,0), (1,0,1), (0,1,1)\} \\ A_3 &= \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) = 3\} = \{(1,1,1)\}. \end{aligned}$$

The image of $\mathcal{X}$ under T is

$$\mathcal{T} = \{t : t = T(\mathbf{x}),\ \mathbf{x} \in \mathcal{X}\} = \{0,1,2,3\},$$

the support of T. The statistic T summarizes the data in that it reports only the value $T(\mathbf{x}) = t$. It does not report which $\mathbf{x} \in \mathcal{X}$ produced $T(\mathbf{x}) = t$.

Connection: Data reduction plays an important role in **statistical inference**. Suppose $X_1, X_2, ..., X_n$ is an iid sample from $f_X(x|\theta)$, where $\theta \in \Theta$. We would like to use the sample $\mathbf{X}$ to learn about which member (or members) of this family might be reasonable. We also do not want to be burdened by having to work with the entire sample $\mathbf{X}$. Therefore, we are interested in statistics T that reduce the data $\mathbf{X}$ (for convenience) while still not compromising our ability to learn about θ.

Preview: Chapter 6 (CB) discusses three methods of data reduction:

- Section 6.2: Sufficiency Principle
- Section 6.3: Likelihood Principle
- Section 6.4: Equivariance Principle

We will focus exclusively on Section 6.2.

6.2 The Sufficiency Principle

6.2.1 Sufficient statistics

Informal Definition: A statistic $T = T(\mathbf{X})$ is a **sufficient statistic** for a parameter θ if it contains "all of the information" about θ that is available in the sample. In other words, we do not lose any information about θ by reducing the sample $\mathbf{X}$ to the statistic T.

Sufficiency Principle: If $T = T(\mathbf{X})$ is a sufficient statistic for θ, then any inference regarding θ should depend on $\mathbf{X}$ only through the value of $T(\mathbf{X})$.

- In other words, if $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{X}$, and $T(\mathbf{x}) = T(\mathbf{y})$, then inference for θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.
- For example, in Example 6.1, suppose

$$\begin{aligned} \mathbf{x} &= (1,0,0) \\ \mathbf{y} &= (0,0,1) \end{aligned}$$

so that $t = T(\mathbf{x}) = T(\mathbf{y}) = 1$. The Sufficiency Principle says that inference for θ depends only on the value of $t = 1$ and not on whether $\mathbf{x}$ or $\mathbf{y}$ was observed.

Definition 6.2.1/Theorem 6.2.2. A statistic $T = T(\mathbf{X})$ is a sufficient statistic for θ if the conditional distribution of $\mathbf{X}$ given T is free of θ; i.e., if the ratio

$$f_{\mathbf{X}|T}(\mathbf{x}|t) = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_T(t|\theta)}$$

is free of θ, for all $\mathbf{x} \in \mathcal{X}$. In other words, after conditioning on T, we have removed all information about θ from the sample $\mathbf{X}$.

Discussion: Note that in the **discrete** case, all distributions above can be interpreted as probabilities. From the definition of a conditional distribution,

$$f_{\mathbf{X}|T}(\mathbf{x}|t) = \frac{f_{\mathbf{X},T}(\mathbf{x},t|\theta)}{f_T(t|\theta)} = \frac{P_\theta(\mathbf{X} = \mathbf{x}, T = t)}{P_\theta(T = t)}.$$

Because $\{\mathbf{X} = \mathbf{x}\} \subset \{T = t\}$, we have

$$P_\theta(\mathbf{X} = \mathbf{x}, T = t) = P_\theta(\mathbf{X} = \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta).$$

Therefore,

$$f_{\mathbf{X}|T}(\mathbf{x}|t) = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_T(t|\theta)}$$

as claimed. If T is **continuous**, then $f_T(t|\theta) \neq P_\theta(T = t)$ and $f_{\mathbf{X}|T}(\mathbf{x}|t)$ cannot be interpreted as a conditional probability. Fortunately, the criterion above; i.e.,

$$f_{\mathbf{X}|T}(\mathbf{x}|t) = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_T(t|\theta)}$$

being free of θ, still applies in the continuous case (although a more rigorous explanation would be needed to see why).

Example 6.2. Suppose $X_1, X_2, ..., X_n$ are iid Poisson(θ), where $\theta > 0$. Use Definition 6.2.1/Theorem 6.2.2 to show that

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a sufficient statistic.

Proof. The pmf of $\mathbf{X}$, for $x_i = 0, 1, 2, ...$, is given by

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!}.$$

Recall that $T \sim \text{Poisson}(n\theta)$, shown by using mgfs. Therefore, the pmf of T, for $t = 0, 1, 2, ...$, is

$$f_T(t|\theta) = \frac{(n\theta)^t e^{-n\theta}}{t!}.$$

With $t = \sum_{i=1}^n x_i$, the conditional distribution

$$f_{\mathbf{X}|T}(\mathbf{x}|t) = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_T(t|\theta)} = \frac{\frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!}}{\frac{(n\theta)^t e^{-n\theta}}{t!}} = \frac{t!}{n^t \prod_{i=1}^n x_i!},$$

which is free of θ. From the definition of sufficiency and from Theorem 6.2.2, we have shown that $T = T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic. □

Example 6.3. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from

$$f_X(x|\theta) = \frac{1}{\theta} e^{-x/\theta} I(x > 0),$$

an exponential distribution with mean $\theta > 0$. Show that

$$T = T(\mathbf{X}) = \overline{X}$$

is a sufficient statistic.

Proof. The pdf of $\mathbf{X}$, for $x_i > 0$, is given by

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = \frac{1}{\theta^n} e^{-\sum_{i=1}^n x_i/\theta}.$$

Recall that if $X_1, X_2, ..., X_n$ are iid exponential(θ), then

$$\overline{X} \sim \text{gamma}(n, \theta/n).$$

Therefore, the pdf of $T = T(\mathbf{X}) = \overline{X}$, for $t > 0$, is

$$f_T(t|\theta) = \frac{1}{\Gamma(n)\left(\frac{\theta}{n}\right)^n} t^{n-1} e^{-nt/\theta}.$$

With $t = \overline{x}$ (i.e., $nt = \sum_{i=1}^n x_i$), the conditional distribution

$$f_{\mathbf{X}|T}(\mathbf{x}|t) = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_T(t|\theta)} = \frac{\frac{1}{\theta^n} e^{-\sum_{i=1}^n x_i/\theta}}{\frac{1}{\Gamma(n)\left(\frac{\theta}{n}\right)^n} t^{n-1} e^{-nt/\theta}} = \frac{\Gamma(n)}{n^n t^{n-1}},$$

which is free of θ. From the definition of sufficiency and from Theorem 6.2.2, we have shown that $T = T(\mathbf{X}) = \overline{X}$ is a sufficient statistic. □

Example 6.4. Suppose $X_1, X_2, ..., X_n$ is an iid sample from a **continuous** distribution with pdf $f_X(x|\theta)$, where $\theta \in \Theta$. Show that $\mathbf{T} = \mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(2)}, ..., X_{(n)})$, the vector of order statistics, is always sufficient.

Proof. Recall from Section 5.4 (CB) that the joint distribution of the n order statistics is

$$\begin{aligned} f_{X_{(1)},X_{(2)},...,X_{(n)}}(x_1, x_2, ..., x_n|\theta) &= n! f_X(x_1|\theta) f_X(x_2|\theta) \cdots f_X(x_n|\theta) \\ &= n! f_{\mathbf{X}}(\mathbf{x}|\theta), \end{aligned}$$

for $-\infty < x_1 < x_2 < \cdots < x_n < \infty$. Therefore, the ratio

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{T}}(\mathbf{x}|\theta)} = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{n! f_{\mathbf{X}}(\mathbf{x}|\theta)} = \frac{1}{n!},$$

which is free of θ. From the definition of sufficiency and from Theorem 6.2.2, we have shown that $\mathbf{T} = \mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(2)}, ..., X_{(n)})$ is a sufficient statistic. □

Discussion: Example 6.4 shows that (with continuous distributions), the order statistics are always sufficient.

- Of course, reducing the sample $\mathbf{X} = (X_1, X_2, ..., X_n)$ to $\mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(2)}, ..., X_{(n)})$ is not that much of a reduction. However, in some parametric families, it is not possible to reduce $\mathbf{X}$ any further without losing information about θ (e.g., Cauchy, logistic, etc.); see pp 275 (CB).
- In some instances, it may be that the parametric form of $f_X(x|\theta)$ is not specified. With so little information provided about the population, we should not be surprised that the only available reduction of $\mathbf{X}$ is to the order statistics.

Remark: The approach we have outlined to show that a statistic T is sufficient appeals to Definition 6.2.1 and Theorem 6.2.2; i.e., we are using the definition of sufficiency directly by showing that the conditional distribution of $\mathbf{X}$ given T is free of θ.

- If I ask you to show that T is sufficient by appealing to the definition of sufficiency, this is the approach I want you to take.
- What if we need to find a sufficient statistic? Then the approach we have just outlined is not practical to implement (i.e., imagine trying different statistics T and for each one attempting to show that $f_{\mathbf{X}|T}(\mathbf{x}|t)$ is free of θ). This might involve a large amount of trial and error and you would have to derive the sampling distribution of T each time (which for many statistics can be difficult or even intractable).
- The Factorization Theorem makes getting sufficient statistics much easier.

Theorem 6.2.6 (**Factorization Theorem**). A statistic $T = T(\mathbf{X})$ is **sufficient** for θ if and only if there exists functions $g(t|\theta)$ and $h(\mathbf{x})$ such that

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = g(t|\theta)h(\mathbf{x}),$$

for all support points $\mathbf{x} \in \mathcal{X}$ and for all $\theta \in \Theta$.

Proof. We prove the result for the discrete case only; the continuous case is beyond the scope of this course.
Necessity ($\Longrightarrow$): Suppose T is sufficient. It suffices to show there exists functions $g(t|\theta)$ and $h(\mathbf{x})$ such that the factorization holds. Because T is sufficient, we know

$$f_{\mathbf{X}|T}(\mathbf{x}|t) = P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t)$$

is free of θ (this is the definition of sufficiency). Therefore, take

$$\begin{aligned} g(t|\theta) &= P_\theta(T(\mathbf{X}) = t) \\ h(\mathbf{x}) &= P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t). \end{aligned}$$

Because $\{\mathbf{X} = \mathbf{x}\} \subset \{T(\mathbf{X}) = t\}$,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= P_\theta(\mathbf{X} = \mathbf{x}) \\ &= P_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t) \\ &= P_\theta(T(\mathbf{X}) = t)P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t) = g(t|\theta)h(\mathbf{x}). \end{aligned}$$

Sufficiency ($\Longleftarrow$): Suppose the factorization holds. To establish that $T = T(\mathbf{X})$ is sufficient, it suffices to show that

$$f_{\mathbf{X}|T}(\mathbf{x}|t) = P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t)$$

is free of θ. Denoting $T(\mathbf{x}) = t$, we have

$$\begin{aligned} f_{\mathbf{X}|T}(\mathbf{x}|t) = P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t) &= \frac{P_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t)}{P_\theta(T(\mathbf{X}) = t)} \\ &= \frac{P_\theta(\mathbf{X} = \mathbf{x})\ I(T(\mathbf{X}) = t)}{P_\theta(T(\mathbf{X}) = t)} \\ &= \frac{g(t|\theta)h(\mathbf{x})\ I(T(\mathbf{X}) = t)}{P_\theta(T(\mathbf{X}) = t)}, \end{aligned}$$

because the factorization holds by assumption. Now write

$$P_\theta(T(\mathbf{X}) = t) = P_\theta(\mathbf{X} \in A_t),$$

where recall $A_t = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) = t\}$ is a set over $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_{\mathbf{X}})$. Note that

$$\begin{aligned} P_\theta(\mathbf{X} \in A_t) &= \sum_{\mathbf{x} \in \mathcal{X}:\ T(\mathbf{x}) = t} P_\theta(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x} \in \mathcal{X}:\ T(\mathbf{x}) = t} g(t|\theta)h(\mathbf{x}) \\ &= g(t|\theta) \sum_{\mathbf{x} \in \mathcal{X}:\ T(\mathbf{x}) = t} h(\mathbf{x}). \end{aligned}$$

Therefore,

$$f_{\mathbf{X}|T}(\mathbf{x}|t) = \frac{g(t|\theta)h(\mathbf{x})\ I(T(\mathbf{X})=t)}{g(t|\theta)\sum_{\mathbf{x}\in\mathcal{X}:\ T(\mathbf{x})=t} h(\mathbf{x})} = \frac{h(\mathbf{x})\ I(T(\mathbf{X})=t)}{\sum_{\mathbf{x}\in\mathcal{X}:\ T(\mathbf{x})=t} h(\mathbf{x})},$$

which is free of θ. $\square$

Example 6.2 (continued). Suppose $X_1, X_2, ..., X_n$ are iid Poisson(θ), where $\theta > 0$. We have already shown that

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a sufficient statistic (using the definition of sufficiency). We now show this using the Factorization Theorem. For $x_i = 0, 1, 2, ...$, the pmf of $\mathbf{X}$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \\ &= \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!} \\ &= \underbrace{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}_{=\ g(t|\theta)}\ \underbrace{\frac{1}{\prod_{i=1}^n x_i!}}_{=\ h(\mathbf{x})}, \end{aligned}$$

where $t = \sum_{i=1}^n x_i$. By the Factorization Theorem, $T = T(\mathbf{X}) = \sum_{i=1}^n X_i$ is sufficient.

Example 6.5. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0, \theta)$, where $\theta > 0$. Find a sufficient statistic.

Solution. The pdf of $\mathbf{X}$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= \prod_{i=1}^{n} \frac{1}{\theta}\ I(0 < x_i < \theta) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 < x_i < \theta) \\ &= \underbrace{\frac{1}{\theta^n}\ I(x_{(n)} < \theta)}_{=\ g(t|\theta)}\ \underbrace{\prod_{i=1}^{n} I(x_i > 0)}_{=\ h(\mathbf{x})}, \end{aligned}$$

where $t = x_{(n)}$. By the Factorization Theorem, $T = T(\mathbf{X}) = X_{(n)}$ is sufficient.

Example 6.6. Suppose $X_1, X_2, ..., X_n$ are iid gamma(α, β), where $\alpha > 0$ and $\beta > 0$. Note that in this family, the parameter $\boldsymbol{\theta} = (\alpha, \beta)$ is two-dimensional. The pdf of $\mathbf{X}$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^\alpha}\ x_i^{\alpha-1} e^{-x_i/\beta} I(x_i > 0) \\ &= \underbrace{\left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\right]^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha} e^{-\sum_{i=1}^n x_i/\beta}}_{=\ g(t_1,t_2|\boldsymbol{\theta})}\ \underbrace{\prod_{i=1}^{n} \frac{I(x_i > 0)}{x_i}}_{=\ h(\mathbf{x})}, \end{aligned}$$

where $t_1 = \prod_{i=1}^n x_i$ and $t_2 = \sum_{i=1}^n x_i$. By the Factorization Theorem,

$$\mathbf{T} = \mathbf{T}(\mathbf{X}) = \begin{pmatrix} \prod_{i=1}^n X_i \\ \sum_{i=1}^n X_i \end{pmatrix}$$

is sufficient.

Remark: In previous examples, we have seen that the dimension of a sufficient statistic T often equals the dimension of the parameter θ:

- Example 6.2: Poisson(θ). $T = \sum_{i=1}^n X_i$; $\dim(T) = \dim(\theta) = 1$
- Example 6.3: exponential(θ). $T = \overline{X}$; $\dim(T) = \dim(\theta) = 1$
- Example 6.5: $\mathcal{U}(0, \theta)$. $T = X_{(n)}$; $\dim(T) = \dim(\theta) = 1$
- Example 6.6: gamma(α, β). $\mathbf{T} = (\prod_{i=1}^n X_i, \sum_{i=1}^n X_i)$; $\dim(\mathbf{T}) = \dim(\boldsymbol{\theta}) = 2$.

Sometimes the dimension of a sufficient statistic is **larger** than that of the parameter. We have already seen this in Example 6.4 where $\mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(2)}, ..., X_{(n)})$, the vector of order statistics, was sufficient; i.e., $\dim(\mathbf{T}) = n$. In some parametric families (e.g., Cauchy, etc.), this statistic is sufficient and no further reduction is possible.

Example 6.7. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(\theta, \theta + 1)$, where $-\infty < \theta < \infty$. This is a one-parameter family; i.e., $\dim(\theta) = 1$. The pdf of $\mathbf{X}$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= \prod_{i=1}^n I(\theta < x_i < \theta + 1) \\ &= \prod_{i=1}^n I(x_i > \theta) \prod_{i=1}^n I(x_i - 1 < \theta) \\ &= \underbrace{I(x_{(1)} > \theta) I(x_{(n)} - 1 < \theta)}_{= g(t_1, t_2|\theta)} \underbrace{\prod_{i=1}^n I(x_i \in \mathbb{R})}_{= h(\mathbf{x})}, \end{aligned}$$

where $t_1 = x_{(1)}$ and $t_2 = x_{(n)}$. By the Factorization Theorem,

$$\mathbf{T} = \mathbf{T}(\mathbf{X}) = \begin{pmatrix} X_{(1)} \\ X_{(n)} \end{pmatrix}$$

is sufficient. In this family, $2 = \dim(\mathbf{T}) > \dim(\theta) = 1$.

Remark: Sufficiency also extends to non-iid situations.

Example 6.8. Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$ and the x_i's are fixed constants (i.e., not random). In this model, it is easy to show that

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2),$$

so that $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$. Note that $Y_1, Y_2, ..., Y_n$ are independent random variables (functions of independent random variables are independent); however, $Y_1, Y_2, ..., Y_n$ are not identically distributed because $E(Y_i) = \beta_0 + \beta_1 x_i$ changes as i does.

For $\mathbf{y} \in \mathbb{R}^n$, the pdf of $\mathbf{Y}$ is

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}. \end{aligned}$$

It is easy to show that

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \underbrace{\sum_{i=1}^{n} y_i^2 - 2\beta_0 \sum_{i=1}^{n} y_i - 2\beta_1 \sum_{i=1}^{n} x_i y_i + n\beta_0^2 + 2\beta_0\beta_1 \sum_{i=1}^{n} x_i + \beta_1^2 \sum_{i=1}^{n} x_i^2}_{= g(t_1, t_2, t_3|\boldsymbol{\theta})},$$

where $t_1 = \sum_{i=1}^{n} y_i^2$, $t_2 = \sum_{i=1}^{n} y_i$, and $t_3 = \sum_{i=1}^{n} x_i y_i$. Taking $h(\mathbf{y}) = 1$, the Factorization Theorem shows that

$$\mathbf{T} = \mathbf{T}(\mathbf{Y}) = \begin{pmatrix} \sum_{i=1}^{n} Y_i^2 \\ \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix}$$

is sufficient. Note that $\dim(\mathbf{T}) = \dim(\boldsymbol{\theta}) = 3$.

Sufficient statistics in the exponential family:

Theorem 6.2.10. Suppose $X_1, X_2, ..., X_n$ are iid from the exponential family

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right\},$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$, $d \leq k$. Then

$$\mathbf{T} = \mathbf{T}(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \sum_{j=1}^{n} t_2(X_j), ..., \sum_{j=1}^{n} t_k(X_j)\right)$$

is sufficient.

Proof. Use the Factorization Theorem. The pdf of $\mathbf{X}$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) &= \prod_{j=1}^{n} h(x_j) c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x_j)\right\} \\ &= \underbrace{\prod_{j=1}^{n} h(x_j)}_{=\, h^*(\mathbf{x})} \underbrace{[c(\boldsymbol{\theta})]^n \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) \sum_{j=1}^{n} t_i(x_j)\right\}}_{=\, g(t_1, t_2, \ldots, t_k|\boldsymbol{\theta})}, \end{aligned}$$

where $t_i = \sum_{j=1}^{n} t_i(x_j)$, for $i = 1, 2, ..., k$. □

Example 6.9. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(θ), where $0 < \theta < 1$. For $x = 0, 1$, the pmf of X is

$$\begin{aligned} f_X(x|\theta) &= \theta^x (1-\theta)^{1-x} \\ &= (1-\theta)\left(\frac{\theta}{1-\theta}\right)^x \\ &= (1-\theta)\exp\left\{\ln\left(\frac{\theta}{1-\theta}\right) x\right\} \\ &= h(x)c(\theta)\exp\{w_1(\theta)t_1(x)\}, \end{aligned}$$

where $h(x) = 1$, $c(\theta) = 1 - \theta$, $w_1(\theta) = \ln\{\theta/(1-\theta)\}$, and $t_1(x) = x$. By Theorem 6.2.10,

$$T = T(\mathbf{X}) = \sum_{j=1}^{n} t_1(X_j) = \sum_{j=1}^{n} X_j$$

is sufficient.

Result: Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$, where $\theta \in \Theta$, and suppose $T = T(\mathbf{X})$ is sufficient. If r is a one-to-one function, then $r(T(\mathbf{X}))$ is also sufficient.
Proof. Let $T^*(\mathbf{X}) = r(T(\mathbf{X}))$ so that $T(\mathbf{X}) = r^{-1}(T^*(\mathbf{X}))$. We have

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= g(T(\mathbf{x})|\theta)h(\mathbf{x}) \\ &= g(r^{-1}(T^*(\mathbf{x}))|\theta)h(\mathbf{x}) \\ &= g^*(T^*(\mathbf{x})|\theta)h(\mathbf{x}), \end{aligned}$$

where $g^* = g \circ r^{-1}$; i.e., g^* is the composition of g and r^{-1}. By the Factorization Theorem, $T^*(\mathbf{X})$ is sufficient. □

Applications:

- In Example 6.9, we showed that
$$T = \sum_{i=1}^{n} X_i$$
is a sufficient statistic for the Bernoulli family. By the previous result, we know that $T_1^*(\mathbf{X}) = \overline{X}$ and $T_2^*(\mathbf{X}) = e^{\sum_{i=1}^{n} X_i}$ are also sufficient. Note that $r_1(t) = t/n$ and $r_2(t) = e^t$ are one-to-one functions over $\mathcal{T} = \{t : t = 0, 1, 2, ..., n\}$.

- In the $\mathcal{N}(\mu, \sigma^2)$ family where both parameters are unknown, it is easy to show that

$$\mathbf{T} = \mathbf{T}(\mathbf{X}) = \begin{pmatrix} \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i^2 \end{pmatrix}$$

is sufficient (just apply the Factorization Theorem directly or use our result dealing with exponential families). Define the function

$$r(\mathbf{t}) = r(t_1, t_2) = \begin{pmatrix} t_1/n \\ \frac{1}{n-1}(t_2 - t_1^2/n) \end{pmatrix},$$

and note that $r(\mathbf{t})$ is one-to-one over $\mathcal{T} = \{(t_1, t_2) : -\infty < t_1 < \infty,\ t_2 \geq 0\}$. Therefore,

$$r(\mathbf{T}(\mathbf{X})) = r\begin{pmatrix} \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i^2 \end{pmatrix} = \begin{pmatrix} \overline{X} \\ S^2 \end{pmatrix}$$

is also sufficient in the $\mathcal{N}(\mu, \sigma^2)$ family.

Remark: In the $\mathcal{N}(\mu, \sigma^2)$ family where both parameters are unknown, the statistic $\mathbf{T}(\mathbf{X}) = (\overline{X}, S^2)$ is sufficient.

- In the $\mathcal{N}(\mu, \sigma_0^2)$ subfamily where σ_0^2 is known, $T(\mathbf{X}) = \overline{X}$ is sufficient.
- In the $\mathcal{N}(\mu_0, \sigma^2)$ subfamily where μ_0 is known,

$$T(\mathbf{X}) = \sum_{i=1}^{n} (X_i - \mu_0)^2$$

is sufficient. Interestingly, S^2 is not sufficient in this subfamily. It is easy to show that $f_{\mathbf{X}|S^2}(\mathbf{x}|s^2)$ depends on σ^2.

6.2.2 Minimal sufficient statistics

Example 6.10. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and σ_0^2 is known. Each of the following statistics is sufficient:

$$T_1(\mathbf{X}) = \overline{X}, \quad \mathbf{T}_2(\mathbf{X}) = \left(X_1, \sum_{i=2}^{n} X_i\right), \quad \mathbf{T}_3(\mathbf{X}) = (X_{(1)}, X_{(2)}, ..., X_{(n)}), \quad \mathbf{T}_4(\mathbf{X}) = \mathbf{X}.$$

Q: How much data reduction is possible?

Definition: A sufficient statistic $T = T(\mathbf{X})$ is called a **minimal sufficient statistic** if, for any other sufficient statistic $T^*(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T^*(\mathbf{x})$.

Remark: A minimal sufficient statistic is a sufficient statistic that offers the most data reduction. Note that "$T(\mathbf{x})$ is a function of $T^*(\mathbf{x})$" means

$$T^*(\mathbf{x}) = T^*(\mathbf{y}) \implies T(\mathbf{x}) = T(\mathbf{y}).$$

Informally, if you know $T^*(\mathbf{x})$, you can calculate $T(\mathbf{x})$, but not necessarily vice versa.

Remark: You can also characterize minimality of a sufficient statistic using the partitioning concept described at the beginning of this chapter. Consider the collection of sufficient statistics. A minimal sufficient statistic $T = T(\mathbf{X})$ admits the **coarsest** possible partition in the collection.

Consider the following table:

Statistic	Description	Partition of $\mathcal{X}$
$T(\mathbf{x})$	Minimal sufficient	A_t, $t = 1, 2, ...,$
$T^*(\mathbf{x})$	Sufficient	B_t, $t = 1, 2, ...,$

By "coarsest possible partition," we mean that $\mathcal{X}$ (the support of $\mathbf{X}$) cannot be split up further and still be a **sufficient partition** (i.e., a partition for a sufficient statistic). This means that $\{B_t, t = 1, 2, ...,\}$ must be a sub-partition of $\{A_t, t = 1, 2, ...,\}$; i.e., each B_t set associated with $T^*(\mathbf{X})$ is a subset of some A_t associated with $T(\mathbf{X})$.

Theorem 6.2.13. Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$, where $\theta \in \Theta$. Suppose there is a function $T(\mathbf{x})$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{y}|\theta)} \text{ is free of } \theta \iff T(\mathbf{x}) = T(\mathbf{y}).$$

Then $T(\mathbf{X})$ is a minimal sufficient statistic.

Example 6.10 (continued). Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and σ_0^2 is known. For $\mathbf{x} \in \mathbb{R}^n$, the pdf of $\mathbf{X}$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-(x_i-\mu)^2/2\sigma_0^2} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma_0^2}. \end{aligned}$$

Now write

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2$$

and form the ratio

$$\begin{aligned}\frac{f_{\mathbf{X}}(\mathbf{x}|\mu)}{f_{\mathbf{X}}(\mathbf{y}|\mu)} &= \frac{\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n \exp\{-\sum_{i=1}^n (x_i-\mu)^2/2\sigma_0^2\}}{\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n \exp\{-\sum_{i=1}^n (y_i-\mu)^2/2\sigma_0^2\}} \\ &= \frac{\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n \exp\{-[\sum_{i=1}^n (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2]/2\sigma_0^2\}}{\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n \exp\{-[\sum_{i=1}^n (y_i-\overline{y})^2 + n(\overline{y}-\mu)^2]/2\sigma_0^2\}}.\end{aligned}$$

Clearly, this ratio is free of μ if and only if $\overline{x} = \overline{y}$. By Theorem 6.2.13, we know that $T(\mathbf{X}) = \overline{X}$ is a minimal sufficient statistic.

Result: Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$, where $\theta \in \Theta$, and suppose $T = T(\mathbf{X})$ is minimal sufficient. If r is a one-to-one function, then $r(T(\mathbf{X}))$ is also minimal sufficient.

Example 6.7 (continued). Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(\theta, \theta+1)$, where $-\infty < \theta < \infty$. We have already shown the pdf of $\mathbf{X}$ is

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = I(x_{(1)} > \theta) I(x_{(n)} - 1 < \theta) \prod_{i=1}^n I(x_i \in \mathbb{R}).$$

Clearly, the ratio

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{y}|\theta)} = \frac{I(x_{(1)} > \theta) I(x_{(n)} - 1 < \theta) \prod_{i=1}^n I(x_i \in \mathbb{R})}{I(y_{(1)} > \theta) I(y_{(n)} - 1 < \theta) \prod_{i=1}^n I(y_i \in \mathbb{R})}$$

is free of θ if and only if $(x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)})$. By Theorem 6.2.13, we know that $\mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic. Note that in this family, the dimension of a minimal sufficient statistic does not match the dimension of the parameter. Note also that a one-to-one function of $\mathbf{T}(\mathbf{X})$ is

$$\begin{pmatrix} X_{(n)} - X_{(1)} \\ (X_{(1)} + X_{(n)})/2 \end{pmatrix}$$

which is also minimal sufficient.

6.2.3 Ancillary statistics

Definition: A statistic $S = S(\mathbf{X})$ is an **ancillary statistic** if the distribution of S does not depend on the model parameter θ.

Remark: You might characterize an ancillary statistic as being "unrelated" to a sufficient statistic. After all, sufficient statistics contain all the information about the parameter θ and ancillary statistics have distributions that are free of θ.

Example 6.11. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 > 0$. Note that

$$\overline{X} \sim \mathcal{N}(0, \sigma^2/n),$$

so $\overline{X}$ is not ancillary (its distribution depends on σ^2). However, the statistic

$$S(\mathbf{X}) = \frac{\overline{X}}{S/\sqrt{n}} \sim t_{n-1}$$

is ancillary because its distribution, t_{n-1}, does not depend on σ^2. Also, it is easy to show that

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i^2$$

is a (minimal) sufficient statistic for σ^2.

Recap:

- $T(\mathbf{X}) = \sum_{i=1}^{n} X_i^2$ contains all the information about σ^2.
- The distribution of $S(\mathbf{X})$ does not depend on σ^2.
- Might we conclude that $T(\mathbf{X}) \perp\!\!\!\perp S(\mathbf{X})$?
- I used R to generate $B = 1000$ draws from the bivariate distribution of $(T(\mathbf{X}), S(\mathbf{X}))$, when $n = 10$ and $\sigma^2 = 100$; see Figure 6.1.

Remark: Finding ancillary statistics is easy when you are dealing with location or scale families.

Location-invariance: For any $c \in \mathbb{R}$, suppose the statistic $S(\mathbf{X})$ satisfies

$$S(x_1 + c, x_2 + c, ..., x_n + c) = S(x_1, x_2, ..., x_n)$$

for all $\mathbf{x} \in \mathcal{X}$. We say that $S(\mathbf{X})$ is a **location-invariant statistic**. In other words, the value of $S(\mathbf{x})$ is unaffected by location shifts.

Result: Suppose $X_1, X_2, ..., X_n$ are iid from

$$f_X(x|\mu) = f_Z(x - \mu),$$

a location family with standard pdf $f_Z(\cdot)$ and location parameter $-\infty < \mu < \infty$. If $S(\mathbf{X})$ is location invariant, then it is ancillary.

Proof. Define $W_i = X_i - \mu$, for $i = 1, 2, ..., n$. We perform an n-variate transformation to find the distribution of $\mathbf{W} = (W_1, W_2, ..., W_n)$. The inverse transformation is described by

$$x_i = g_i^{-1}(w_1, w_2, ..., w_n) = w_i + \mu,$$

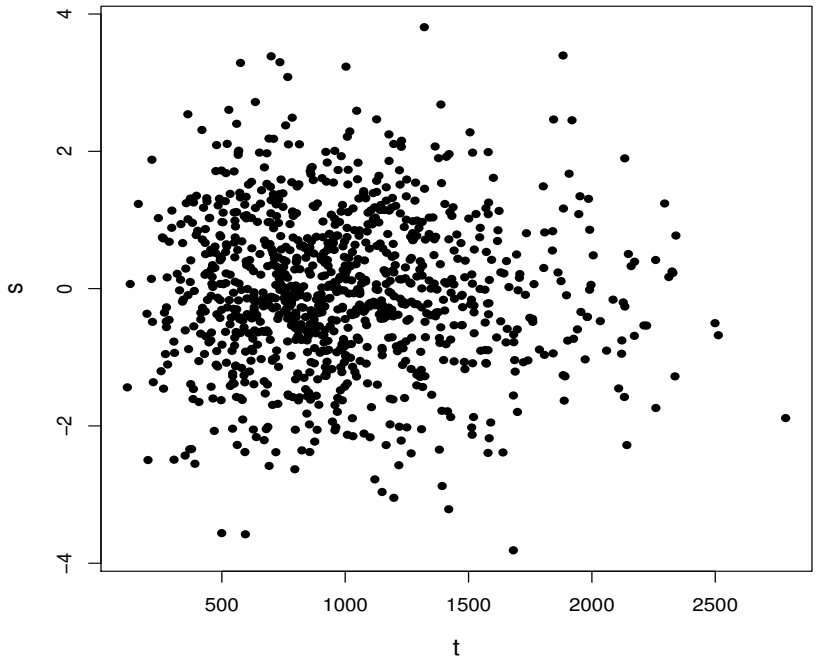

Figure 6.1: Scatterplot of $B = 1000$ pairs of $T(\mathbf{x})$ and $S(\mathbf{x})$ in Example 6.11. Each point was calculated based on an iid sample of size $n = 10$ with $\sigma^2 = 100$.

for $i = 1, 2, ..., n$. It is easy to see that the Jacobian of the inverse transformation is 1 and therefore

$$\begin{aligned} f_{\mathbf{W}}(\mathbf{w}) &= f_{\mathbf{X}}(w_1 + \mu, w_2 + \mu, ..., w_n + \mu) \\ &= \prod_{i=1}^{n} f_X(w_i + \mu) \\ &= \prod_{i=1}^{n} f_Z(w_i + \mu - \mu) = \prod_{i=1}^{n} f_Z(w_i), \end{aligned}$$

which does not depend on μ. Because $S(\mathbf{X})$ is location invariant, we know

$$\begin{aligned} S(\mathbf{X}) &= S(X_1, X_2, ..., X_n) \\ &= S(W_1 + \mu, W_2 + \mu, ..., W_n + \mu) \\ &= S(W_1, W_2, ..., W_n) \\ &= S(\mathbf{W}). \end{aligned}$$

Because the distribution of $\mathbf{W}$ does not depend on μ, the distribution of the statistic $S(\mathbf{W})$ cannot depend on μ either. But $S(\mathbf{W}) = S(\mathbf{X})$, so we are done. □

Examples: Each of the following is a location-invariant statistic (and hence is ancillary when sampling from a location family):

$$S(\mathbf{X}) = \overline{X} - M, \quad S(\mathbf{X}) = X_{(n)} - X_{(1)}, \quad S(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}|, \quad S(\mathbf{X}) = S^2.$$

Note: Above M denotes the sample median of $X_1, X_2, ..., X_n$.

Example 6.12. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and σ_0^2 is known. Show that the sample variance S^2 is ancillary.
Proof. First note that

$$f_X(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-(x-\mu)^2/2\sigma_0^2}\ I(x \in \mathbb{R}) = f_Z(x-\mu),$$

where

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-z^2/2\sigma_0^2}\ I(z \in \mathbb{R}),$$

the $\mathcal{N}(0, \sigma_0^2)$ pdf. Therefore, the $\mathcal{N}(\mu, \sigma_0^2)$ family is a location family. We now show that $S(\mathbf{X}) = S^2$ is location invariant. Let $W_i = X_i + c$, for $i = 1, 2, ..., n$. Clearly, $\overline{W} = \overline{X} + c$ and

$$\begin{aligned} S(\mathbf{W}) &= \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \overline{W})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}[(X_i + c) - (\overline{X} + c)]^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = S(\mathbf{X}). \end{aligned}$$

This shows that $S(\mathbf{X}) = S^2$ is location invariant and hence is ancillary.

Remark: The preceding argument only shows that the distribution of S^2 does not depend on μ. However, in this example, it is easy to find the distribution of S^2 directly. Recall that

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1} \stackrel{d}{=} \text{ gamma}\left(\frac{n-1}{2}, 2\right) \implies S^2 \sim \text{ gamma}\left(\frac{n-1}{2}, \frac{2\sigma_0^2}{n-1}\right),$$

which, of course, does not depend on μ.

Scale-invariance: For any $d > 0$, suppose the statistic $S(\mathbf{X})$ satisfies

$$S(dx_1, dx_2, ..., dx_n) = S(x_1, x_2, ..., x_n)$$

for all $\mathbf{x} \in \mathcal{X}$. We say that $S(\mathbf{X})$ is a **scale-invariant statistic**. In other words, the value of $S(\mathbf{x})$ is unaffected by changes in scale.

Result: Suppose $X_1, X_2, ..., X_n$ are iid from

$$f_X(x|\sigma) = \frac{1}{\sigma} f_Z\left(\frac{x}{\sigma}\right),$$

a scale family with standard pdf $f_Z(\cdot)$ and scale parameter $\sigma^2 > 0$. If $S(\mathbf{X})$ is scale invariant, then it is ancillary.

Proof. Define $W_i = X_i/\sigma$, for $i = 1, 2, ..., n$. We perform an n-variate transformation to find the distribution of $\mathbf{W} = (W_1, W_2, ..., W_n)$. The inverse transformation is described by

$$x_i = g_i^{-1}(w_1, w_2, ..., w_n) = \sigma w_i,$$

for $i = 1, 2, ..., n$. It is easy to see that the Jacobian of the inverse transformation is σ^n and therefore

$$\begin{aligned} f_{\mathbf{W}}(\mathbf{w}) &= f_{\mathbf{X}}(\sigma w_1, \sigma w_2, ..., \sigma w_n) \times \sigma^n \\ &= \sigma^n \prod_{i=1}^{n} f_X(\sigma w_i) \\ &= \sigma^n \prod_{i=1}^{n} \frac{1}{\sigma} f_Z\left(\frac{\sigma w_i}{\sigma}\right) = \prod_{i=1}^{n} f_Z(w_i), \end{aligned}$$

which does not depend on σ. Because $S(\mathbf{X})$ is scale invariant, we know

$$\begin{aligned} S(\mathbf{X}) &= S(X_1, X_2, ..., X_n) \\ &= S(\sigma W_1, \sigma W_2, ..., \sigma W_n) \\ &= S(W_1, W_2, ..., W_n) \\ &= S(\mathbf{W}). \end{aligned}$$

Because the distribution of $\mathbf{W}$ does not depend on σ, the distribution of the statistic $S(\mathbf{W})$ cannot depend on σ either. But $S(\mathbf{W}) = S(\mathbf{X})$, so we are done. □

Examples: Each of the following is a scale-invariant statistic (and hence is ancillary when sampling from a scale family):

$$S(\mathbf{X}) = \frac{S}{\overline{X}}, \quad S(\mathbf{X}) = \frac{X_{(n)}}{X_{(1)}}, \quad S(\mathbf{X}) = \frac{\sum_{i=1}^{k} X_i^2}{\sum_{i=1}^{n} X_i^2}.$$

Example 6.13. Suppose $X_1, X_2, ..., X_n$ is an iid sample from

$$f_X(x|\sigma) = \frac{1}{2\sigma} e^{-|x|/\sigma} I(x \in \mathbb{R}).$$

Show that

$$S(\mathbf{X}) = \frac{\sum_{i=1}^{k} |X_i|}{\sum_{i=1}^{n} |X_i|}$$

is ancillary.

Proof. First note that

$$f_X(x|\sigma) = \frac{1}{\sigma} f_Z\left(\frac{x}{\sigma}\right),$$

where

$$f_Z(z) = \frac{1}{2} e^{-|z|} I(z \in \mathbb{R}),$$

a standard LaPlace pdf. Therefore, $\{f_X(x|\sigma),\ \sigma > 0\}$ is a scale family. We now show that $S(\mathbf{X})$ is scale invariant. For $d > 0$, let $W_i = dX_i$, for $i = 1, 2, ..., n$. We have

$$\begin{aligned} S(\mathbf{W}) &= \frac{\sum_{i=1}^{k} |W_i|}{\sum_{i=1}^{n} |W_i|} \\ &= \frac{\sum_{i=1}^{k} |dX_i|}{\sum_{i=1}^{n} |dX_i|} \\ &= \frac{d\sum_{i=1}^{k} |X_i|}{d\sum_{i=1}^{n} |X_i|} = S(\mathbf{X}). \end{aligned}$$

This shows that $S(\mathbf{X})$ is scale invariant and hence is ancillary.

Remark: The preceding argument only shows that the distribution of $S(\mathbf{X})$ does not depend on σ. It can be shown (verify!) that

$$S(\mathbf{X}) = \frac{\sum_{i=1}^{k} |X_i|}{\sum_{i=1}^{n} |X_i|} \sim \text{beta}(k, n-k),$$

which, of course, does not depend on σ.

Remark: It is straightforward to extend our previous discussions to location-scale families.

6.2.4 Sufficient, ancillary, and complete statistics

Definition: Let $\{f_T(t|\theta); \theta \in \Theta\}$ be a family of pdfs (or pmfs) for a statistic $T = T(\mathbf{X})$. We say that this family is a **complete family** if the following condition holds:

$$E_\theta[g(T)] = 0 \ \ \forall \theta \in \Theta \implies P_\theta(g(T) = 0) = 1 \ \ \forall \theta \in \Theta;$$

i.e., $g(T) = 0$ almost surely for all $\theta \in \Theta$. We call $T = T(\mathbf{X})$ a **complete statistic**.

Remark: This condition basically says that the only function of T that is an unbiased estimator of zero is the function that is zero itself (with probability 1).

Example 6.14. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(θ), where $0 < \theta < 1$. Show that

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a complete statistic.

Proof. We know that $T \sim b(n, \theta)$, so it suffices to show that this family of distributions is a complete family. Suppose

$$E_\theta[g(T)] = 0 \ \ \forall \theta \in (0,1).$$

It suffices to show that $P_\theta(g(T) = 0) = 1$ for all $\theta \in (0,1)$. Note that

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

where $r = \theta/(1-\theta)$. Because $(1-\theta)^n$ is never zero, it must be that

$$\sum_{t=0}^{n} g(t)\binom{n}{t} r^t = 0.$$

The LHS of this equation is a polynomial (in r) of degree n. The only way this polynomial can be zero **for all** $\theta \in (0,1)$; i.e., **for all** $r > 0$, is for the coefficients

$$g(t)\binom{n}{t} = 0, \ \text{for } t = 0, 1, 2, ..., n.$$

Because $\binom{n}{t} \neq 0$, this can only happen when $g(t) = 0$, for $t = 0, 1, 2, ..., n$. We have shown that $P_\theta(g(T) = 0) = 1$ for all θ. Therefore, $T(\mathbf{X})$ is complete. □

Remark: To show that a statistic $T = T(\mathbf{X})$ is not complete, all we have to do is find one **nonzero** function $g(T)$ that satisfies $E_\theta[g(T)] = 0$, for all θ.

Example 6.15. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\theta, \theta^2)$, where $\theta \in \Theta = (-\infty, 0) \cup (0, \infty)$. Putting

$$f_X(x|\theta) = \frac{1}{\sqrt{2\pi\theta^2}} e^{-(x-\theta)^2/2\theta^2} I(x \in \mathbb{R})$$

into exponential family form shows that

$$\mathbf{T} = \mathbf{T}(\mathbf{X}) = \begin{pmatrix} \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i^2 \end{pmatrix}$$

is sufficient. However, $\mathbf{T}$ is not complete. Consider

$$g(\mathbf{T}) = 2\left(\sum_{i=1}^{n} X_i\right)^2 - (n+1)\sum_{i=1}^{n} X_i^2.$$

It is straightforward to show that

$$E_\theta[g(\mathbf{T})] = E_\theta\left[2\left(\sum_{i=1}^n X_i\right)^2 - (n+1)\sum_{i=1}^n X_i^2\right] = 0.$$

We have found a nonzero function $g(\mathbf{T})$ that has zero expectation. Therefore $\mathbf{T}$ cannot be complete.

Theorem 6.2.24 (**Basu's Theorem**). Suppose T is a sufficient statistic. If T is also complete, then T is independent of every ancillary statistic S.

Proof. Suppose S is ancillary. Let ϕ and ψ be any functions. Using iterated expectation,

$$\begin{aligned} E_\theta[\phi(S)\psi(T)] &= E_\theta\{E[\phi(S)\psi(T)|T]\} \\ &= E_\theta\{\psi(T)E[\phi(S)|T]\}, \end{aligned} \tag{6.1}$$

the last step following because once we condition on $T = t$; i.e., we write $E[\phi(S)\psi(t)|T = t]$, then $\psi(t)$ is constant. Now, consider the quantity

$$E_\theta\{E[\phi(S)|T]\} = E_\theta[\phi(S)] = k,$$

where k is a constant free of θ (because S is ancillary by assumption). Define

$$g(T) = E[\phi(S)|T] - k.$$

From our argument above, we have

$$E_\theta[g(T)] = E_\theta\{E[\phi(S)|T] - k\} = E_\theta[\phi(S)] - k = 0$$

for all θ. However, because T is complete by assumption, we know that

$$g(T) \overset{a.s.}{=} 0 \ \forall\theta \implies E[\phi(S)|T] \overset{a.s.}{=} k \ \forall\theta.$$

Because T is sufficient by assumption, we know that $E[\phi(S)|T]$ does not depend on θ either. From Equation (6.1), we have

$$\begin{aligned} E_\theta[\phi(S)\psi(T)] &= E_\theta\{\psi(T)E[\phi(S)|T]\} \\ &= kE_\theta\{\psi(T)\} \\ &= E_\theta[\phi(S)]E_\theta[\psi(T)]. \end{aligned}$$

Because $E_\theta[\phi(S)\psi(T)] = E_\theta[\phi(S)]E_\theta\{\psi(T)\}$ holds for all functions (ϕ and ψ were arbitrarily chosen), then equality holds when

$$\begin{aligned} \phi(S) &= I(S \leq s) \\ \psi(T) &= I(T \leq t), \end{aligned}$$

for $s, t \in \mathbb{R}$. Using this choice of ϕ and ψ, the joint cdf of (S, T)

$$\begin{aligned} F_{T,S}(t, s) &= P_\theta(S \leq s, T \leq t) \\ &= E_\theta[\phi(S)\psi(T)] \\ &= E_\theta[\phi(S)]E_\theta[\psi(T)] \\ &= P_\theta(S \leq s)P_\theta(T \leq t) = F_S(s)F_T(t). \end{aligned}$$

We have shown that the joint cdf of (S, T) factors into the product of the marginal cdfs. Because s and t are arbitrary, we are done. □

Example 6.16. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0, \theta)$, where $\theta > 0$. Show that $X_{(n)}$ and $X_{(1)}/X_{(n)}$ are independent.
Proof. We will show that

- $T(\mathbf{X}) = X_{(n)}$ is complete and sufficient.
- $S(\mathbf{X}) = X_{(1)}/X_{(n)}$ is ancillary.

The result will then follow from Basu's Theorem. First, note that

$$\begin{aligned} f_X(x|\theta) &= \frac{1}{\theta} I(0 < x < \theta) \\ &= \frac{1}{\theta} f_Z\left(\frac{x}{\theta}\right), \end{aligned}$$

where $f_Z(z) = I(0 < z < 1)$ is the standard uniform pdf. Therefore, the $\mathcal{U}(0, \theta)$ family is a scale family. We now show that $S(\mathbf{X})$ is scale invariant. For $d > 0$, let $W_i = dX_i$, for $i = 1, 2, ..., n$. We have

$$S(\mathbf{W}) = \frac{W_{(1)}}{W_{(n)}} = \frac{dX_{(1)}}{dX_{(n)}} = \frac{X_{(1)}}{X_{(n)}} = S(\mathbf{X}).$$

This shows that $S(\mathbf{X})$ is scale invariant and hence is ancillary.

We have already shown that $T = T(\mathbf{X}) = X_{(n)}$ is sufficient; see Example 6.5 (notes). We now show T is complete. We first find the distribution of T. The pdf of T, the maximum order statistic, is given by

$$\begin{aligned} f_T(t) &= n f_X(t)[F_X(t)]^{n-1} \\ &= n \, \frac{1}{\theta} \, I(0 < t < \theta) \, \left(\frac{t}{\theta}\right)^{n-1} \\ &= \frac{n t^{n-1}}{\theta^n} \, I(0 < t < \theta). \end{aligned}$$

Suppose $E_\theta[g(T)] = 0$ for all $\theta > 0$. This implies that

$$\begin{aligned} \int_0^\theta g(t) \, \frac{n t^{n-1}}{\theta^n} dt = 0 \;\; \forall \theta > 0 &\implies \int_0^\theta g(t) t^{n-1} dt = 0 \;\; \forall \theta > 0 \\ &\implies \frac{d}{d\theta} \int_0^\theta g(t) t^{n-1} dt = 0 \;\; \forall \theta > 0 \\ &\implies g(\theta)\theta^{n-1} = 0 \;\; \forall \theta > 0, \end{aligned}$$

the last step following from the Fundamental Theorem of Calculus, provided that g is Riemann-integrable. Because $\theta^{n-1} \neq 0$, it must be true that $g(\theta) = 0$ for all $\theta > 0$. We have

therefore shown that the only function g satisfying $E_\theta[g(T)] = 0$ for all $\theta > 0$ is the function that is itself zero; i.e., we have shown

$$P_\theta(g(T) = 0) = 1, \text{ for all } \theta > 0.$$

Therefore $T = T(\mathbf{X}) = X_{(n)}$ is complete. $\square$

Remark: Our completeness argument in Example 6.16 is not entirely convincing. We have basically established that

$$E_\theta[g(T)] = 0 \ \ \forall\theta > 0 \implies P_\theta(g(T) = 0) = 1 \ \ \forall\theta > 0$$

for the class of functions g which are Riemann-integrable. There are many functions g that are not Riemann-integrable. CB note that "this distinction is not of concern." This is another way of saying that the authors do not want to present completeness from a more general point of view (for good reason; this would involve a heavy dose of measure theory).

Extension: Suppose that, in Example 6.16, I asked you to find

$$E\left(\frac{X_{(1)}}{X_{(n)}}\right).$$

At first glance, this appears to be an extremely challenging expectation to calculate. From first principles, we could find the joint distribution of $(X_{(1)}, X_{(n)})$ and then calculate the first moment of the ratio. Another approach is to use Basu's Theorem. Note that

$$\begin{aligned} E(X_{(1)}) &= E\left(X_{(n)}\,\frac{X_{(1)}}{X_{(n)}}\right) \\ &= E(X_{(n)})E\left(\frac{X_{(1)}}{X_{(n)}}\right), \end{aligned}$$

the last step following because $X_{(n)}$ and $X_{(1)}/X_{(n)}$ are independent. Therefore, we can calculate the desired expectation by instead calculating $E(X_{(1)})$ and $E(X_{(n)})$. These are easier to calculate:

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

Therefore, we have

$$\frac{\theta}{n+1} = \left(\frac{n}{n+1}\right)\theta\ E\left(\frac{X_{(1)}}{X_{(n)}}\right) \implies E\left(\frac{X_{(1)}}{X_{(n)}}\right) = \frac{1}{n}.$$

It makes sense that this expectation would not depend on θ; recall that $S(\mathbf{X}) = X_{(1)}/X_{(n)}$ is ancillary.

Completeness in the exponential family:

Recall: Suppose $X_1, X_2, ..., X_n$ are iid from the exponential family

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right\},$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$, $d \leq k$. In Theorem 6.2.10, we showed that

$$\mathbf{T} = \mathbf{T}(\mathbf{X}) = \begin{pmatrix} \sum_{j=1}^{n} t_1(X_j) \\ \sum_{j=1}^{n} t_2(X_j) \\ \vdots \\ \sum_{j=1}^{n} t_k(X_j) \end{pmatrix}$$

is a sufficient statistic.

New result (Theorem 6.2.25): In the exponential family, the statistic $\mathbf{T} = \mathbf{T}(\mathbf{X})$ is **complete** if the natural parameter space

$$\{\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_k) : \eta_i = w_i(\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta\}$$

contains an open set in $\mathbb{R}^k$. For the most part, this means:

- $\mathbf{T} = \mathbf{T}(\mathbf{X})$ is complete if $d = k$ (**full** exponential family)
- $\mathbf{T} = \mathbf{T}(\mathbf{X})$ is not complete if $d < k$ (**curved** exponential family).

Example 6.17. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from a gamma$(\alpha, 1/\alpha^2)$ distribution. The pdf of X is

$$\begin{aligned} f_X(x|\alpha) &= \frac{1}{\Gamma(\alpha)\left(\frac{1}{\alpha^2}\right)^{\alpha}} x^{\alpha-1}e^{-x/(1/\alpha^2)}I(x>0) \\ &= \frac{I(x>0)}{x}\frac{\alpha^{2\alpha}}{\Gamma(\alpha)}\, e^{\alpha \ln x}\, e^{-\alpha^2 x} \\ &= \frac{I(x>0)}{x}\frac{\alpha^{2\alpha}}{\Gamma(\alpha)} \exp\left(\alpha \ln x - \alpha^2 x\right) \\ &= h(x)c(\alpha)\exp\{w_1(\alpha)t_1(x) + w_2(\alpha)t_2(x)\}, \end{aligned}$$

where $h(x) = I(x > 0)/x$, $c(\alpha) = \alpha^{2\alpha}/\Gamma(\alpha)$, $w_1(\alpha) = \alpha$, $t_1(x) = \ln x$, $w_2(\alpha) = -\alpha^2$, and $t_2(x) = x$. Theorem 6.2.10 tells us that

$$\mathbf{T} = \mathbf{T}(\mathbf{X}) = \begin{pmatrix} \sum_{i=1}^{n} \ln X_i \\ \sum_{i=1}^{n} X_i \end{pmatrix}$$

is a sufficient statistic. However, Theorem 6.2.25 tells us that $\mathbf{T}$ is not complete because $\{f_X(x|\alpha), \alpha > 0\}$ is an exponential family with $d = 1$ and $k = 2$. Note also that

$$\{\boldsymbol{\eta} = (\eta_1, \eta_2) : (\alpha, -\alpha^2);\ \alpha > 0\}$$

is a half-parabola (which opens downward); this set does not contain an open set in $\mathbb{R}^2$.

Example 6.18. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Prove that $\overline{X} \perp\!\!\!\perp S^2$.

Proof. We use Basu's Theorem, but we have to use it carefully. Fix $\sigma^2 = \sigma_0^2$ and consider first the $\mathcal{N}(\mu, \sigma_0^2)$ subfamily. The pdf of $X \sim \mathcal{N}(\mu, \sigma_0^2)$ is

$$\begin{aligned} f_X(x|\mu) &= \frac{1}{\sqrt{2\pi}\sigma_0} e^{-(x-\mu)^2/2\sigma_0^2} I(x \in \mathbb{R}) \\ &= \frac{I(x \in \mathbb{R}) e^{-x^2/2\sigma_0^2}}{\sqrt{2\pi}\sigma_0}\ e^{-\mu^2/2\sigma_0^2}\ e^{(\mu/\sigma_0^2)x} \\ &= h(x)c(\mu)\exp\{w_1(\mu)t_1(x)\}. \end{aligned}$$

Theorem 6.2.10 tells us that

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a sufficient statistic. Because $d = k = 1$ (remember, this is for the $\mathcal{N}(\mu, \sigma_0^2)$ subfamily), Theorem 6.2.25 tells us that T is complete. In Example 6.12 (notes), we have already showed that

- the $\mathcal{N}(\mu, \sigma_0^2)$ subfamily is a location family
- $S(\mathbf{X}) = S^2$ is location invariant and hence ancillary for this subfamily.

Therefore, by Basu's Theorem, we have proven that, **in the $\mathcal{N}(\mu, \sigma_0^2)$ subfamily**,

$$\sum_{i=1}^{n} X_i \perp\!\!\!\perp S^2 \implies \overline{X} \perp\!\!\!\perp S^2,$$

the last implication being true because $\overline{X}$ is a function of $T = T(\mathbf{X}) = \sum_{i=1}^n X_i$ and functions of independent statistics are independent. Finally, because we fixed $\sigma^2 = \sigma_0^2$ arbitrarily, this same argument holds for all σ_0^2 fixed. Therefore, this independence result holds for any choice of σ^2 and hence for the full $\mathcal{N}(\mu, \sigma^2)$ family. □

Remark: It is important to see that in the preceding proof, we cannot work directly with the $\mathcal{N}(\mu, \sigma^2)$ family and claim that

- $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is complete and sufficient
- $S(\mathbf{X}) = S^2$ is ancillary

for this family. In fact, neither statement is true in the full family.

Remark: Outside the exponential family, Basu's Theorem can be useful in showing that a sufficient statistic $T(\mathbf{X})$ is not complete.

Basu's Theorem (Contrapositive version): Suppose $T(\mathbf{X})$ is sufficient and $S(\mathbf{X})$ is ancillary. If $T(\mathbf{X})$ and $S(\mathbf{X})$ are not independent, then $T(\mathbf{X})$ is not complete.

Example 6.19. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from

$$f_X(x|\theta) = \frac{1}{\pi[1+(x-\theta)^2]}\ I(x \in \mathbb{R}),$$

where $-\infty < \theta < \infty$. It is easy to see that $\{f_X(x|\theta) : -\infty < \theta < \infty\}$ is a location family; i.e., $f_X(x|\theta) = f_Z(x-\theta)$, where

$$f_Z(z) = \frac{1}{\pi(1+z^2)}\ I(z \in \mathbb{R})$$

is the standard Cauchy pdf. We now prove the sample range $S(\mathbf{X}) = X_{(n)} - X_{(1)}$ is location invariant. Let $W_i = X_i + c$, for $i = 1, 2, ..., n$, and note

$$S(\mathbf{W}) = W_{(n)} - W_{(1)} = (X_{(n)} + c) - (X_{(1)} + c) = X_{(n)} - X_{(1)} = S(\mathbf{X}).$$

This shows that $S(\mathbf{X})$ is ancillary in this family. Finally, we know from Example 6.4 (notes) that the order statistics

$$\mathbf{T} = \mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(2)}, ..., X_{(n)})$$

are sufficient for this family (in fact, $\mathbf{T}$ is minimal sufficient; see Exercise 6.9, CB, pp 301). However, clearly $S(\mathbf{X})$ and $\mathbf{T}(\mathbf{X})$ are not independent; e.g., if you know $\mathbf{T}(\mathbf{x})$, you can calculate $S(\mathbf{x})$. By Basu's Theorem (the contrapositive version), we know that $\mathbf{T}(\mathbf{X})$ cannot be complete.

Theorem 6.2.28. Suppose that $T(\mathbf{X})$ is sufficient. If $T(\mathbf{X})$ is complete, then $T(\mathbf{X})$ is minimal sufficient.

Remark: Example 6.19 shows that the converse to Theorem 6.2.28 is not true; i.e.,

$$T(\mathbf{X}) \text{ minimal sufficient } \not\Longrightarrow\ T(\mathbf{X}) \text{ complete.}$$

Example 6.7 provides another counterexample. We showed that if $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(\theta, \theta+1)$, then $\mathbf{T} = \mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic. However, $\mathbf{T}$ cannot be complete because $\mathbf{T}$ and the sample range $X_{(n)} - X_{(1)}$ (which is location invariant and hence ancillary in this model) are not independent. This implies that there exists a nonzero function $g(\mathbf{T})$ that has zero expectation for all $\theta \in \mathbb{R}$. In fact, it is easy to show that

$$E_\theta(X_{(n)} - X_{(1)}) = \frac{n-1}{n+1}.$$

Therefore,

$$g(\mathbf{T}) = X_{(n)} - X_{(1)} - \frac{n-1}{n+1}$$

satisfies $E_\theta[g(\mathbf{T})] = 0$ for all θ.

7 Point Estimation

Complementary reading: Chapter 7 (CB).

7.1 Introduction

Remark: We will approach "the point estimation problem" from the following point of view. We have a parametric model for $\mathbf{X} = (X_1, X_2, ..., X_n)$:

$$\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}), \text{ where } \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k,$$

and the model parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ is unknown. We will assume that $\boldsymbol{\theta}$ is fixed (except when we discuss Bayesian estimation). Possible goals include

1. Estimating $\boldsymbol{\theta}$
2. Estimating a function of $\boldsymbol{\theta}$, say $\tau(\boldsymbol{\theta})$, where $\tau : \mathbb{R}^k \to \mathbb{R}^q$, $q \leq k$ (often, $q = 1$; i.e., $\tau(\boldsymbol{\theta})$ is a scalar parameter).

Remark: For most of the situations we will encounter in this course, the random vector $\mathbf{X}$ will consist of $X_1, X_2, ..., X_n$, an **iid sample** from the population $f_X(x|\boldsymbol{\theta})$. However, our discussion is also relevant when the independence assumption is relaxed, the identically distributed assumption is relaxed, or both.

Definition: A **point estimator**

$$W(\mathbf{X}) = W(X_1, X_2, ..., X_n)$$

is any function of the sample $\mathbf{X}$. Therefore, any statistic is a point estimator. We call $W(\mathbf{x}) = W(x_1, x_2, ..., x_n)$ a **point estimate**. $W(\mathbf{x})$ is a realization of $W(\mathbf{X})$.

Preview: This chapter is split into two parts. In this first part (Section 7.2), we present different approaches of **finding** point estimators. These approaches are:

- Section 7.2.1: Method of Moments (MOM)
- Section 7.2.2: Maximum Likelihood Estimation (MLE)
- Section 7.2.3: Bayesian Estimation
- Section 7.2.4: EM Algorithm (we will skip).

The second part (Section 7.3) focuses on **evaluating** point estimators; e.g., which estimators are good/bad? What constitutes a "good" estimator? Is it possible to find the best one? For that matter, how should we even define "best?"

7.2 Methods of Finding Estimators

7.2.1 Method of moments

Strategy: Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k) \in \Theta \subseteq \mathbb{R}^k$. The method of moments (MOM) approach says to equate the first k sample moments to the first k population moments and then to solve for $\boldsymbol{\theta}$.

Recall: The jth **sample moment** (uncentered) is given by

$$m_j' = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

If $X_1, X_2, ..., X_n$ is an iid sample, the jth **population moment** (uncentered) is

$$\mu_j' = E(X^j).$$

Intuition: The first k sample moments depend on the sample $\mathbf{X}$. The first k population moments will generally depend on $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$. Therefore, the system of equations

$$\begin{aligned} m_1' &\stackrel{\text{set}}{=} E(X) \\ m_2' &\stackrel{\text{set}}{=} E(X^2) \\ &\vdots \\ m_k' &\stackrel{\text{set}}{=} E(X^k) \end{aligned}$$

can (at least in theory) be solved for $\theta_1, \theta_2, ..., \theta_k$. A solution to this system of equations is called a **method of moments (MOM) estimator**.

Example 7.1. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0, \theta)$, where $\theta > 0$. The first sample moment is

$$m_1' = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

The first population moment is

$$\mu_1' = E(X) = \frac{\theta}{2}.$$

We set these moments equal to each other; i.e.,

$$\overline{X} \stackrel{\text{set}}{=} \frac{\theta}{2}$$

and solve for θ. The solution

$$\widehat{\theta} = 2\overline{X}$$

is a method of moments estimator for θ.

Example 7.2. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(-\theta, \theta)$, where $\theta > 0$. For this population, $E(X) = 0$ so this will not help us. Moving to second moments, we have

$$m_2' = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

and

$$\mu_2' = E(X^2) = \text{var}(X) = \frac{\theta^2}{3}.$$

Therefore, we can set

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \stackrel{\text{set}}{=} \frac{\theta^2}{3}$$

and solve for θ. The solution

$$\widehat{\theta} = +\sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}$$

is a method of moments estimator for θ. We keep the positive solution because $\theta > 0$ (although, technically, the negative solution is still a MOM estimator).

Example 7.3. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. The first two population moments are $E(X) = \mu$ and $E(X^2) = \text{var}(X) + [E(X)]^2 = \sigma^2 + \mu^2$. Therefore, method of moments estimators for μ and σ^2 are found by solving

$$\begin{aligned} \overline{X} &\stackrel{\text{set}}{=} \mu \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &\stackrel{\text{set}}{=} \sigma^2 + \mu^2. \end{aligned}$$

We have $\widehat{\mu} = \overline{X}$ and

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Note that the method of moments estimator for σ^2 is not our "usual" sample variance (with denominator $n - 1$).

Remarks:

- I think of MOM estimation as a "quick and dirty" approach. All we are doing is matching moments. We are attempting to learn about a population $f_X(x|\theta)$ by using moments only.
- Sometimes MOM estimators have good finite-sample properties (e.g., unbiasedness, small variance, etc.). Sometimes they do not.
- MOM estimators generally do have desirable large-sample properties (e.g., large-sample normality, etc.) but are usually less (asymptotically) efficient than other estimators.

- MOM estimators can be nonsensical. In fact, sometimes MOM estimators fall outside the parameter space Θ. For example, in linear models with random effects, variance components estimated via MOM can be negative.

7.2.2 Maximum likelihood estimation

Note: We first formally define a likelihood function; see also Section 6.3 (CB).

Definition: Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. Given that $\mathbf{X} = \mathbf{x}$ is observed, the function

$$L(\boldsymbol{\theta}|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$$

is called the **likelihood function**.

Note: The likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ is the same function as the joint pdf/pmf $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$. The only difference is in how we interpret each one.

- The function $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$ is a model that describes the random behavior of $\mathbf{X}$ when $\boldsymbol{\theta}$ is fixed.
- The function $L(\boldsymbol{\theta}|\mathbf{x})$ is viewed as a function of $\boldsymbol{\theta}$ with the data $\mathbf{X} = \mathbf{x}$ held fixed.

Interpretation: When $\mathbf{X}$ is discrete,

$$L(\boldsymbol{\theta}|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}).$$

That is, when $\mathbf{X}$ is discrete, we can interpret the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ literally as a joint probability.

- Suppose that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two possible values of $\boldsymbol{\theta}$. Suppose $\mathbf{X}$ is discrete and

  $$L(\boldsymbol{\theta}_1|\mathbf{x}) = P_{\boldsymbol{\theta}_1}(\mathbf{X} = \mathbf{x}) > P_{\boldsymbol{\theta}_2}(\mathbf{X} = \mathbf{x}) = L(\boldsymbol{\theta}_2|\mathbf{x}).$$

  This suggests the sample $\mathbf{x}$ is more likely to have occurred with $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ rather than if $\boldsymbol{\theta} = \boldsymbol{\theta}_2$. Therefore, in the discrete case, we can interpret $L(\boldsymbol{\theta}|\mathbf{x})$ as "the probability of the data $\mathbf{x}$."
- Of course, this interpretation of $L(\boldsymbol{\theta}|\mathbf{x})$ is not appropriate when $\mathbf{X}$ is continuous because $P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}) = 0$. However, this description is still used informally when describing the likelihood function with continuous data. An attempt to make this description mathematical is given on pp 290 (CB).
- Section 6.3 (CB) describes how the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ can be viewed as a **data reduction** device.

Definition: Any maximizer $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$ of the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ is called a **maximum likelihood estimate**.

- With our previous interpretation, we can think of $\widehat{\boldsymbol{\theta}}$ as "the value of $\boldsymbol{\theta}$ that maximizes the probability of the data $\mathbf{x}$."

We call $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ a **maximum likelihood estimator** (MLE).

Remarks:

1. Finding the MLE $\widehat{\boldsymbol{\theta}}$ is essentially a maximization problem. The estimate $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ must fall in the parameter space Θ because we are maximizing $L(\boldsymbol{\theta}|\mathbf{x})$ over Θ; i.e.,
$$\widehat{\boldsymbol{\theta}}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\mathbf{x}).$$
There is no guarantee that an MLE $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ will be unique (although it often is).
2. Under certain conditions (so-called "regularity conditions"), maximum likelihood estimators $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ have very nice large-sample properties (Chapter 10, CB).
3. In most "real" problems, the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ must be maximized numerically to calculate $\widehat{\boldsymbol{\theta}}(\mathbf{x})$.

Example 7.4. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{U}[0,\theta]$, where $\theta > 0$. Find the MLE of θ.
Solution. The likelihood function is
$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\theta}\, I(0 \le x_i \le \theta) = \underbrace{\frac{1}{\theta^n}\, I(x_{(n)} \le \theta) \prod_{i=1}^{n} I(x_i \ge 0)}_{\text{view this as a function of } \theta \text{ with } \mathbf{x} \text{ fixed}}.$$
Note that

- For $\theta \ge x_{(n)}$, $L(\theta|\mathbf{x}) = 1/\theta^n$, which decreases as θ increases.
- For $\theta < x_{(n)}$, $L(\theta|\mathbf{x}) = 0$.

Clearly, the MLE of θ is $\widehat{\theta} = X_{(n)}$.

Remark: Note that in this example, we "closed the endpoints" on the support of X; i.e., the pdf of X is
$$f_X(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{otherwise.} \end{cases}$$
Mathematically, this model is no different than had we "opened the endpoints." However, if we used open endpoints, note that
$$x_{(n)} < \arg\max_{\theta>0} L(\theta|\mathbf{x}) < x_{(n)} + \epsilon$$
for all $\epsilon > 0$, and therefore the maximizer of $L(\theta|\mathbf{x})$; i.e., the MLE, would not exist.

Curiosity: In this uniform example, we derived the MOM estimator to be $\widehat{\theta} = 2\overline{X}$ in Example 7.1. The MLE is $\widehat{\theta} = X_{(n)}$. Which estimator is "better?"

Note: In general, when the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ is a differentiable function of $\boldsymbol{\theta}$, we can use calculus to maximize $L(\boldsymbol{\theta}|\mathbf{x})$. If an MLE $\widehat{\boldsymbol{\theta}}$ exists, it must satisfy

$$\frac{\partial}{\partial \theta_j} L(\widehat{\boldsymbol{\theta}}|\mathbf{x}) = 0, \quad j = 1, 2, ..., k.$$

Of course, second-order conditions must be verified to ensure that $\widehat{\boldsymbol{\theta}}$ is a maximizer (and not a minimizer or some other value).

Example 7.5. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\theta, 1)$, where $-\infty < \theta < \infty$. The likelihood function is

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}. \end{aligned}$$

The derivative

$$\begin{aligned} \frac{\partial}{\partial \theta} L(\theta|\mathbf{x}) &= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}}_{\text{this can never be zero}} \sum_{i=1}^{n}(x_i - \theta) \stackrel{\text{set}}{=} 0 \\ \implies \sum_{i=1}^{n}(x_i - \theta) &= 0. \end{aligned}$$

Therefore, $\widehat{\theta} = \overline{x}$ is a first-order critical point of $L(\theta|\mathbf{x})$. Is $\widehat{\theta} = \overline{x}$ a maximizer? I calculated

$$\frac{\partial^2}{\partial \theta^2} L(\theta|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \left\{ \left[\sum_{i=1}^{n}(x_i - \theta)\right]^2 - n \right\}.$$

Because

$$\left.\frac{\partial^2}{\partial \theta^2} L(\theta|\mathbf{x})\right|_{\theta=\overline{x}} = -n\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2} < 0,$$

the function $L(\theta|\mathbf{x})$ is concave down when $\theta = \overline{x}$; i.e., $\widehat{\theta} = \overline{x}$ maximizes $L(\theta|\mathbf{x})$. Therefore,

$$\widehat{\theta} = \widehat{\theta}(\mathbf{X}) = \overline{X}$$

is the MLE of θ.

Illustration: Under the $\mathcal{N}(\theta, 1)$ model assumption, I graphed in Figure 7.1 the likelihood function $L(\theta|\mathbf{x})$ after observing $x_1 = 2.437$, $x_2 = 0.993$, $x_3 = 1.123$, $x_4 = 1.900$, and $x_5 = 3.794$ (an iid sample of size $n = 5$). The sample mean $\overline{x} = 2.049$ is our ML estimate of θ based on this sample $\mathbf{x}$.

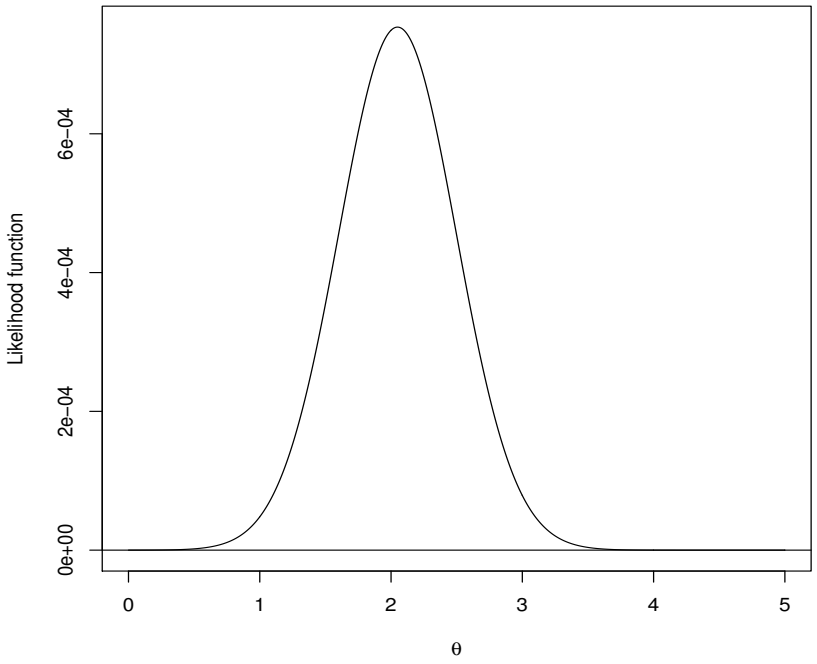

Figure 7.1: Plot of $L(\theta|\mathbf{x})$ versus θ in Example 7.5. The data $\mathbf{x}$ were generated from a $\mathcal{N}(\theta = 1.5, 1)$ distribution with $n = 5$. The sample mean (MLE) is $\overline{x} = 2.049$.

Q: What if, in Example 7.5, we constrained the parameter space to be $\Theta_0 = \{\theta : \theta \geq 0\}$? What is the MLE over Θ_0?
A: We simply maximize $L(\theta|\mathbf{x})$ over Θ_0 instead. It is easy to see the restricted MLE is

$$\widehat{\theta}^* = \widehat{\theta}^*(\mathbf{X}) = \begin{cases} \overline{X}, & \overline{X} \geq 0 \\ 0, & \overline{X} < 0. \end{cases}$$

Important: Suppose that $L(\boldsymbol{\theta}|\mathbf{x})$ is a likelihood function. Then

$$\begin{aligned} \widehat{\boldsymbol{\theta}}(\mathbf{x}) &= \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\mathbf{x}) \\ &= \arg\max_{\boldsymbol{\theta}\in\Theta} \ln L(\boldsymbol{\theta}|\mathbf{x}). \end{aligned}$$

The function $\ln L(\boldsymbol{\theta}|\mathbf{x})$ is called the **log-likelihood function**. Analytically, it is usually easier to work with $\ln L(\boldsymbol{\theta}|\mathbf{x})$ than with the likelihood function directly. The equations

$$\frac{\partial}{\partial\theta_j} \ln L(\boldsymbol{\theta}|\mathbf{x}) = 0, \quad j = 1, 2, ..., k,$$

are called the **score equations**.

Example 7.6. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Set $\boldsymbol{\theta} = (\mu, \sigma^2)$. The likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

The log-likelihood function is

$$\ln L(\boldsymbol{\theta}|\mathbf{x}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

The score equations are

$$\begin{aligned} \frac{\partial}{\partial\mu}\ln L(\boldsymbol{\theta}|\mathbf{x}) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \stackrel{\text{set}}{=} 0 \\ \frac{\partial}{\partial\sigma^2}\ln L(\boldsymbol{\theta}|\mathbf{x}) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \stackrel{\text{set}}{=} 0. \end{aligned}$$

Clearly $\widehat{\mu} = \overline{x}$ solves the first equation; inserting $\widehat{\mu} = \overline{x}$ into the second equation and solving for σ^2 gives $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. A first-order critical point is $(\overline{x}, n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2)$.

Q: How can we verify this solution is a maximizer?
A: In general, for a k-dimensional maximization problem, we can calculate the Hessian matrix

$$\mathbf{H} = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ln L(\boldsymbol{\theta}|\mathbf{x}),$$

a $k \times k$ matrix of second-order partial derivatives, and show this matrix is **negative definite** when we evaluate it at the first-order critical point $\widehat{\boldsymbol{\theta}}$. This is a sufficient condition. Recall a $k \times k$ matrix $\mathbf{H}$ is negative definite if $\mathbf{a}'\mathbf{H}\mathbf{a} < 0$ for all $\mathbf{a} \in \mathbb{R}^k$, $\mathbf{a} \neq \mathbf{0}$.

For the $\mathcal{N}(\mu, \sigma^2)$ example, I calculated

$$\mathbf{H} = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) & \frac{n}{\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 \end{pmatrix}.$$

With $\mathbf{a}' = (a_1, a_2)$, it follows that

$$\mathbf{a}'\mathbf{H}\mathbf{a}\Big|_{\mu=\widehat{\mu},\sigma^2=\widehat{\sigma}^2} = -\frac{na_1^2}{\widehat{\sigma}^2} < 0.$$

This shows that

$$\widehat{\boldsymbol{\theta}}(\mathbf{X}) = \begin{pmatrix} \overline{X} \\ \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \end{pmatrix}$$

is the MLE of $\boldsymbol{\theta}$ in the $\mathcal{N}(\mu, \sigma^2)$ model.

Exercise: Find the MLEs of μ and σ^2 in the respective sub-families:

- $\mathcal{N}(\mu, \sigma_0^2)$, where σ_0^2 is known
- $\mathcal{N}(\mu_0, \sigma^2)$, where μ_0 is known.

Example 7.7. *ML estimation under parameter constraints.* Suppose X_1, X_2 are independent random variables where

$$\begin{aligned} X_1 &\sim b(n_1, p_1) \\ X_2 &\sim b(n_2, p_2), \end{aligned}$$

where $0 < p_1 < 1$ and $0 < p_2 < 1$. The likelihood function of $\boldsymbol{\theta} = (p_1, p_2)$ is

$$\begin{aligned} L(\boldsymbol{\theta}|x_1, x_2) &= f_{X_1}(x_1|p_1) f_{X_2}(x_2|p_2) \\ &= \binom{n_1}{x_1} p_1^{x_1}(1-p_1)^{n_1-x_1} \binom{n_2}{x_2} p_2^{x_2}(1-p_2)^{n_2-x_2}. \end{aligned}$$

The log-likelihood function is

$$\ln L(\boldsymbol{\theta}|x_1, x_2) = c + x_1 \ln p_1 + (n_1 - x_1)\ln(1-p_1) + x_2 \ln p_2 + (n_2 - x_2)\ln(1-p_2),$$

where $c = \ln \binom{n_1}{x_1} + \ln \binom{n_2}{x_2}$ is free of $\boldsymbol{\theta}$. Over

$$\Theta = \{\boldsymbol{\theta} = (p_1, p_2) : 0 < p_1 < 1, 0 < p_2 < 1\},$$

it is easy to show that $\ln L(\boldsymbol{\theta}|x_1, x_2)$ is maximized at

$$\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(X_1, X_2) = \begin{pmatrix} \widehat{p}_1 \\ \widehat{p}_2 \end{pmatrix} = \begin{pmatrix} \dfrac{X_1}{n_1} \\ \dfrac{X_2}{n_2} \end{pmatrix},$$

the vector of sample proportions. Because this is the maximizer over the entire parameter space Θ, we call $\widehat{\boldsymbol{\theta}}$ the **unrestricted MLE** of $\boldsymbol{\theta}$.

Q: How do we find the MLE of $\boldsymbol{\theta}$ subject to the constraint that $p_1 = p_2$?
A: We would now like to maximize $\ln L(\boldsymbol{\theta}|x_1, x_2)$ over

$$\Theta_0 = \{\boldsymbol{\theta} = (p_1, p_2) : 0 < p_1 < 1, 0 < p_2 < 1,\ p_1 = p_2\}.$$

We can use Lagrange multipliers to maximize $\ln L(\boldsymbol{\theta}|x_1, x_2)$ subject to the constraint that

$$g(\boldsymbol{\theta}) = g(p_1, p_2) = p_1 - p_2 = 0.$$

We are left to solve

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta}|x_1, x_2) &= \lambda \frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta}) \\ g(\boldsymbol{\theta}) &= 0. \end{aligned}$$

This system becomes

$$\begin{aligned} \frac{x_1}{p_1} - \frac{n_1 - x_1}{1 - p_1} &= \lambda \\ \frac{x_2}{p_2} - \frac{n_2 - x_2}{1 - p_2} &= -\lambda \\ p_1 - p_2 &= 0. \end{aligned}$$

Solving this system for p_1 and p_2, we get

$$\widehat{\boldsymbol{\theta}}^* = \widehat{\boldsymbol{\theta}}^*(X_1, X_2) = \begin{pmatrix} \widehat{p}_1^* \\ \widehat{p}_2^* \end{pmatrix} = \begin{pmatrix} \dfrac{X_1 + X_2}{n_1 + n_2} \\ \dfrac{X_1 + X_2}{n_1 + n_2} \end{pmatrix}.$$

Because this is the maximizer over the subspace Θ_0, we call $\widehat{\boldsymbol{\theta}}^*$ the **restricted MLE**; i.e., the MLE of $\boldsymbol{\theta}$ subject to the $p_1 = p_2$ restriction.

Discussion: The parameter constraint $p_1 = p_2$ might arise in a **hypothesis test**; e.g., $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$. If H_0 is true, then we would expect $\widehat{\boldsymbol{\theta}}^*$ and $\widehat{\boldsymbol{\theta}}$ to be "close" and the ratio

$$\lambda(x_1, x_2) = \frac{L(\widehat{\boldsymbol{\theta}}^*|x_1, x_2)}{L(\widehat{\boldsymbol{\theta}}|x_1, x_2)} \approx 1.$$

The farther $\widehat{\boldsymbol{\theta}}^*$ is from $\widehat{\boldsymbol{\theta}}$, the smaller $\lambda(x_1, x_2)$ becomes. Therefore, it would make sense to reject H_0 when $\lambda(x_1, x_2)$ is small. This is the idea behind **likelihood ratio tests**.

Example 7.8. *Logistic regression.* In practice, finding maximum likelihood estimates usually requires numerical methods. Suppose $Y_1, Y_2, ..., Y_n$ are independent Bernoulli random variables; specifically, $Y_i \sim \text{Bernoulli}(p_i)$, where

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i \iff p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

In this model, the x_i's are fixed constants. The likelihood function of $\boldsymbol{\theta} = (\beta_0, \beta_1)$ is

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) &= \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left[\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]^{y_i} \left[1 - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]^{1-y_i}. \end{aligned}$$

Taking logarithms and simplifying gives

$$\ln L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 x_i) - \ln(1 + e^{\beta_0 + \beta_1 x_i})\right].$$

Closed-form expressions for the maximizers $\widehat{\beta}_0$ and $\widehat{\beta}_1$ do not exist except in very simple situations. Numerical methods are needed to maximize $\ln L(\boldsymbol{\theta}|\mathbf{y})$; e.g., iteratively re-weighted least squares (the default method in R's `glm` function).

Theorem 7.2.10 (Invariance property of MLEs). Suppose $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. For any function $\tau(\boldsymbol{\theta})$, the MLE of $\tau(\boldsymbol{\theta})$ is $\tau(\widehat{\boldsymbol{\theta}})$.

Proof. For simplicity, suppose θ is a scalar parameter and that $\tau : \mathbb{R} \to \mathbb{R}$ is one-to-one (over Θ). In this case,

$$\eta = \tau(\theta) \iff \theta = \tau^{-1}(\eta).$$

The likelihood function of interest is $L^*(\eta)$. It suffices to show that $L^*(\eta)$ is maximized when $\eta = \tau(\widehat{\theta})$, where $\widehat{\theta}$ is the maximizer of $L(\theta)$. For simplicity in notation, I drop emphasis of a likelihood function's dependence on $\mathbf{x}$. Let $\widehat{\eta}$ be a maximizer of $L^*(\eta)$. Then

$$\begin{aligned} L^*(\widehat{\eta}) &= \sup_{\eta} L^*(\eta) \\ &= \sup_{\eta} L(\tau^{-1}(\eta)) \\ &= \sup_{\theta} L(\theta). \end{aligned}$$

Therefore, the maximizer $\widehat{\eta}$ satisfies $\tau^{-1}(\widehat{\eta}) = \widehat{\theta}$. Because τ is one-to-one, $\widehat{\eta} = \tau(\widehat{\theta})$. □

Remark: Our proof assumes that τ is a one-to-one function. However, Theorem 7.2.10 is true for any function; see pp 319-320 (CB).

Example 7.8 (continued). In the logistic regression model

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \iff p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \tau(\beta_0, \beta_1), \text{ say,}$$

the MLE of p_i is

$$\widehat{p}_i = \tau(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)}.$$

Example 7.9. Suppose $X_1, X_2, ..., X_n$ are iid exponential(β), where $\beta > 0$. The likelihood function is

$$L(\beta|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum_{i=1}^n x_i/\beta}.$$

The log-likelihood function is

$$\ln L(\beta|\mathbf{x}) = -n \ln \beta - \frac{\sum_{i=1}^n x_i}{\beta}$$

The score equation becomes

$$\frac{\partial}{\partial \beta} \ln L(\beta|\mathbf{x}) = -\frac{n}{\beta} + \frac{\sum_{i=1}^n x_i}{\beta^2} \stackrel{\text{set}}{=} 0.$$

Solving the score equation for β gives $\widehat{\beta} = \overline{x}$. It is easy to show that this value maximizes $\ln L(\beta|\mathbf{x})$. Therefore,

$$\widehat{\beta} = \widehat{\beta}(\mathbf{X}) = \overline{X}$$

is the MLE of β.

Applications of invariance: In Example 7.9,

- $\overline{X}^2$ is the MLE of β^2
- $1/\overline{X}$ is the MLE of $1/\beta$
- For t fixed, $e^{-t/\overline{X}}$ is the MLE of $S_X(t|\beta) = e^{-t/\beta}$, the **survivor function** of X at t.

7.2.3 Bayesian estimation

Remark: Non-Bayesians think of inference in the following way:

$$\text{Observe } \mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta) \quad \longrightarrow \quad \text{Use } \mathbf{x} \text{ to make statement about } \theta.$$

In this paradigm, the model parameter θ is fixed (and unknown). I have taken θ to be a scalar here for ease of exposition.

Bayesians do not consider the parameter θ to be fixed. They regard θ as random, having its own probability distribution. Therefore, Bayesians think of inference in this way:

$$\text{Model } \theta \sim \pi(\theta) \quad \longrightarrow \quad \text{Observe } \mathbf{X}|\theta \sim f_{\mathbf{X}}(\mathbf{x}|\theta) \quad \longrightarrow \quad \text{Update with } \pi(\theta|\mathbf{x}).$$

The model for θ on the front end is called the **prior distribution**. The model on the back end is called the **posterior distribution**. The posterior distribution **combines** prior information (supplied through the prior model) and the observed data $\mathbf{x}$. For a Bayesian, all inference flows from the posterior distribution.

Important: Here are the relevant probability distributions that arise in a Bayesian context. These are given "in order" as to how the Bayesian uses them. Continue to assume that θ is a scalar.

1. **Prior distribution:** $\theta \sim \pi(\theta)$. This distribution incorporates the information available about θ before any data are observed.
2. **Conditional distribution:** $\mathbf{X}|\theta \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$. This is the distribution of $\mathbf{X}$, but now viewed conditionally on θ:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= L(\theta|\mathbf{x}) \\ &\overset{\text{iid}}{=} \prod_{i=1}^{n} f_{X|\theta}(x_i|\theta). \end{aligned}$$

   Mathematically, the conditional distribution is the same as the likelihood function.

3. **Joint distribution:** This distribution describes how $\mathbf{X}$ and θ vary jointly. From the definition of a conditional distribution,

$$f_{\mathbf{X},\theta}(\mathbf{x},\theta) = \underbrace{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)}_{\text{likelihood}} \; \underbrace{\pi(\theta)}_{\text{prior}}.$$

4. **Marginal distribution.** This describes how $\mathbf{X}$ is distributed marginally. From the definition of a marginal distribution,

$$\begin{aligned} m_{\mathbf{X}}(\mathbf{x}) &= \int_{\Theta} f_{\mathbf{X},\theta}(\mathbf{x},\theta)d\theta \\ &= \int_{\Theta} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)d\theta, \end{aligned}$$

where Θ is the "support" of θ (remember, we are now treating θ as a random variable).

5. **Posterior distribution.** This is the Bayesian's "updated" distribution of θ, given that the data $\mathbf{X} = \mathbf{x}$ have been observed. From the definition of a conditional distribution,

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{f_{\mathbf{X},\theta}(\mathbf{x},\theta)}{m_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)}{\int_{\Theta} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)d\theta}. \end{aligned}$$

Remark: The process of starting with $\pi(\theta)$ and performing the necessary calculations to end up with $\pi(\theta|\mathbf{x})$ is informally known as "turning the Bayesian crank." The distributions above can be viewed as steps in a "recipe" for posterior construction (i.e., start with the prior and the conditional, calculate the joint, calculate the marginal, calculate the posterior). We will see momentarily that not all steps are needed. In fact, in practice, computational techniques are used to essentially bypass Step 4 altogether. You can see that this might be desirable, especially if θ is a vector (and perhaps high-dimensional).

Example 7.10. Suppose that, conditional on θ, $X_1, X_2, ..., X_n$ are iid Poisson(θ), where the prior distribution for $\theta \sim$ gamma(a, b), a, b known. We now turn the Bayesian crank.

1. **Prior distribution.**

$$\pi(\theta) = \frac{1}{\Gamma(a)b^a}\ \theta^{a-1}e^{-\theta/b}I(\theta > 0).$$

2. **Conditional distribution.** For $x_i = 0, 1, 2, ...$,

$$f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}e^{-\theta}}{x_i!} = \frac{\theta^{\sum_{i=1}^n x_i}e^{-n\theta}}{\prod_{i=1}^n x_i!}.$$

Recall that this is the same function as the likelihood function.

3. **Joint distribution.** For $x_i = 0, 1, 2, ...$, and $\theta > 0$,

$$\begin{aligned} f_{\mathbf{X},\theta}(\mathbf{x},\theta) &= f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta) \\ &= \frac{\theta^{\sum_{i=1}^n x_i}e^{-n\theta}}{\prod_{i=1}^n x_i!}\ \frac{1}{\Gamma(a)b^a}\ \theta^{a-1}e^{-\theta/b} \\ &= \underbrace{\frac{1}{\prod_{i=1}^n x_i!\ \Gamma(a)b^a}}_{\text{does not depend on } \theta}\ \theta^{\sum_{i=1}^n x_i + a - 1}e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}. \end{aligned}$$

4. **Marginal distribution.** For $x_i = 0, 1, 2, ...,$

$$\begin{aligned} m_{\mathbf{X}}(\mathbf{x}) &= \int_{\Theta} f_{\mathbf{X},\theta}(\mathbf{x}, \theta) d\theta \\ &= \frac{1}{\prod_{i=1}^{n} x_i!\ \Gamma(a) b^a} \int_0^{\infty} \underbrace{\theta^{\sum_{i=1}^{n} x_i + a - 1} e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}}_{\text{gamma}(a^*, b^*)\text{ kernel}} d\theta, \end{aligned}$$

where

$$a^* = \sum_{i=1}^{n} x_i + a \quad \text{and} \quad b^* = \frac{1}{n + \frac{1}{b}}.$$

Therefore,

$$m_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\prod_{i=1}^{n} x_i!\ \Gamma(a) b^a}\ \Gamma\left(\sum_{i=1}^{n} x_i + a\right) \left(\frac{1}{n + \frac{1}{b}}\right)^{\sum_{i=1}^{n} x_i + a}.$$

5. **Posterior distribution.** For $\theta > 0$,

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{f_{\mathbf{X},\theta}(\mathbf{x}, \theta)}{m_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{\frac{1}{\prod_{i=1}^{n} x_i!\ \Gamma(a) b^a}\ \theta^{\sum_{i=1}^{n} x_i + a - 1} e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}}{\frac{1}{\prod_{i=1}^{n} x_i!\ \Gamma(a) b^a}\ \Gamma\left(\sum_{i=1}^{n} x_i + a\right) \left(\frac{1}{n + \frac{1}{b}}\right)^{\sum_{i=1}^{n} x_i + a}} \\ &= \frac{1}{\Gamma\left(\sum_{i=1}^{n} x_i + a\right) \left(\frac{1}{n+\frac{1}{b}}\right)^{\sum_{i=1}^{n} x_i + a}}\ \theta^{\sum_{i=1}^{n} x_i + a - 1} e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}, \end{aligned}$$

which we recognize as the gamma pdf with parameters

$$\begin{aligned} a^* &= \sum_{i=1}^{n} x_i + a \\ b^* &= \frac{1}{n + \frac{1}{b}}. \end{aligned}$$

That is, the posterior distribution

$$\theta|\mathbf{X} = \mathbf{x} \sim \text{gamma}\left(\sum_{i=1}^{n} x_i + a,\ \frac{1}{n + \frac{1}{b}}\right).$$

Remark: Note that the shape and scale parameters of the posterior distribution $\pi(\theta|\mathbf{x})$ depend on

- a and b, the prior distribution parameters (i.e., the "hyperparameters")
- the data $\mathbf{x}$ through the sufficient statistic $t(\mathbf{x}) = \sum_{i=1}^{n} x_i$.

In this sense, the posterior distribution combines information from the prior and the data.

Q: In general, which functional of $\pi(\theta|\mathbf{x})$ should we use as a point estimator?
A: Answering this question technically would require us to discuss **loss functions** (see Section 7.3.4, CB). In practice, it is common to use one of

$$\begin{aligned}
\widehat{\theta}_B &= E(\theta|\mathbf{X}=\mathbf{x}) \longrightarrow \text{posterior mean} \\
\widetilde{\theta}_B &= \text{med}(\theta|\mathbf{X}=\mathbf{x}) \longrightarrow \text{posterior median} \\
\widehat{\theta}^*_B &= \text{mode}(\theta|\mathbf{X}=\mathbf{x}) \longrightarrow \text{posterior mode.}
\end{aligned}$$

Note that in Example 7.10 (the Poisson-gamma example), the posterior mean equals

$$\begin{aligned}
\widehat{\theta}_B = E(\theta|\mathbf{X}=\mathbf{x}) &= \frac{\sum_{i=1}^n x_i + a}{n + \frac{1}{b}} \\
&= \left(\frac{nb}{nb+1}\right)\overline{x} + \left(\frac{1}{nb+1}\right)ab.
\end{aligned}$$

That is, the posterior mean is a **weighted average** of the sample mean $\overline{x}$ and the prior mean ab. Note also that as the sample size n increases, more weight is given to the data (through $\overline{x}$) and less weight is given to the the prior (through the prior mean).

Remark: In Example 7.10, we wrote the joint distribution (in **Step 3**) as

$$\begin{aligned}
f_{\mathbf{X},\theta}(\mathbf{x},\theta) &= f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta) \\
&= \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!} \; \frac{1}{\Gamma(a)b^a} \; \theta^{a-1}e^{-\theta/b} \\
&= \underbrace{\frac{1}{\prod_{i=1}^n x_i! \; \Gamma(a)b^a}}_{\text{does not depend on } \theta} \; \underbrace{\theta^{\sum_{i=1}^n x_i + a - 1} e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}}_{\text{gamma}(a^*,b^*) \text{ kernel}}.
\end{aligned}$$

At this step, we can clearly identify the kernel of the posterior distribution. We can therefore skip calculating the marginal distribution $m_{\mathbf{X}}(\mathbf{x})$ in Step 4, because we know $m_{\mathbf{X}}(\mathbf{x})$ does not depend on θ. Because of this, it is common to write, in general,

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta) \\
&= L(\theta|\mathbf{x})\pi(\theta).
\end{aligned}$$

The posterior distribution is proportional to the likelihood function times the prior distribution. A (classical) Bayesian analysis requires these two functions $L(\theta|\mathbf{x})$ and $\pi(\theta)$ only.

Remark: Suppose $\mathbf{X}|\theta \sim f_{\mathbf{X}|\theta}(x|\theta)$. If $T = T(\mathbf{X})$ is sufficient, we can write

$$f_{\mathbf{X}|\theta}(x|\theta) = g(t|\theta)h(\mathbf{x}),$$

by the Factorization Theorem. Therefore, the posterior distribution

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta) \\
&\propto g(t|\theta)\pi(\theta).
\end{aligned}$$

This shows that the posterior distribution will depend on the data $\mathbf{x}$ through the value of the sufficient statistic $t = T(\mathbf{x})$. We can therefore write the posterior distribution as depending on t only; i.e.,

$$\pi(\theta|t) \quad \propto \quad f_{T|\theta}(t|\theta)\pi(\theta),$$

and restrict attention to the (sampling) distribution of $T = T(\mathbf{X})$ from the beginning.

Example 7.11. Suppose that $X_1, X_2, ..., X_n$ are iid Bernoulli(θ), where the prior distribution for $\theta \sim \text{beta}(a, b)$, a, b known. We know that

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a sufficient statistic for the Bernoulli family and that $T \sim b(n, \theta)$. Therefore, for $t = 0, 1, 2, ..., n$ and $0 < \theta < 1$, the posterior distribution

$$\begin{aligned}
\pi(\theta|t) &\propto f_{T|\theta}(t|\theta)\pi(\theta) \\
&= \binom{n}{t}\theta^t(1-\theta)^{n-t}\,\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1} \\
&= \underbrace{\binom{n}{t}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{does not depend on } \theta}\;\underbrace{\theta^{t+a-1}(1-\theta)^{n-t+b-1}}_{\text{beta}(a^*, b^*) \text{ kernel}},
\end{aligned}$$

where $a^* = t + a$ and $b^* = n - t + b$. From here, we can immediately conclude that the posterior distribution

$$\theta|T = t \sim \text{beta}(t + a, n - t + b),$$

where $t = T(\mathbf{x}) = \sum_{i=1}^{n} x_i$.

Discussion: In Examples 7.10 and 7.11, we observed the following occurrence:

- Example 7.10. $\theta \sim$ gamma (prior) $\longrightarrow$ $\theta|\mathbf{X} = \mathbf{x} \sim$ gamma (posterior).
- Example 7.11. $\theta \sim$ beta (prior) $\longrightarrow$ $\theta|T = t \sim$ beta (posterior).

Definition: Let $\mathcal{F} = \{f_X(x|\theta) : \theta \in \Theta\}$ denote a class of pdfs or pmfs. A class Π of prior distributions is said to be a **conjugate prior family** for $\mathcal{F}$ if the posterior distribution also belongs to Π.

As we have already seen in Examples 7.10 and 7.11,

- The gamma family is conjugate for the Poisson family.
- The beta family is conjugate for the binomial family.

Example 7.12. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

- If σ^2 is known, a conjugate prior for μ is

$$\mu \sim \mathcal{N}(\xi, \tau^2), \quad \xi, \tau^2 \text{ known.}$$

- If μ is known, a conjugate prior for σ^2 is

$$\sigma^2 \sim \text{IG}(a, b) \quad a, b \text{ known.}$$

7.3 Methods of Evaluating Estimators

7.3.1 Bias, variance, and MSE

Definition: Suppose $W = W(\mathbf{X})$ is a point estimator. We call W an **unbiased estimator** of θ if

$$E_\theta(W) = \theta \quad \text{for all } \theta \in \Theta.$$

More generally, we call W an unbiased estimator of $\tau(\theta)$ if

$$E_\theta(W) = \tau(\theta) \quad \text{for all } \theta \in \Theta.$$

Definition: The **mean-squared error** (**MSE**) of a point estimator $W = W(\mathbf{X})$ is

$$\begin{aligned} \text{MSE}_\theta(W) &= E_\theta[(W-\theta)^2] \\ &= \text{var}_\theta(W) + [E_\theta(W) - \theta]^2 \\ &= \text{var}_\theta(W) + \text{Bias}_\theta^2(W), \end{aligned}$$

where $\text{Bias}_\theta(W) = E_\theta(W) - \theta$ is the **bias** of W as an estimator of θ. Note that if W is an unbiased estimator of θ, then for all $\theta \in \Theta$,

$$E_\theta(W) = \theta \implies \text{Bias}_\theta(W) = E_\theta(W) - \theta = 0.$$

In this case,

$$\text{MSE}_\theta(W) = \text{var}_\theta(W).$$

Remark: In general, the MSE incorporates two components:

- $\text{var}_\theta(W)$; this measures **precision**
- $\text{Bias}_\theta(W)$; this measures **accuracy**.

Obviously, we prefer estimators with small MSE because these estimators have small bias (i.e., high accuracy) and small variance (i.e., high precision).

Example 7.13. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters unknown. Set $\boldsymbol{\theta} = (\mu, \sigma^2)$. Recall that our "usual" sample variance estimator is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

and for all $\boldsymbol{\theta}$,

$$\begin{aligned} E_{\boldsymbol{\theta}}(S^2) &= \sigma^2 \\ \text{var}_{\boldsymbol{\theta}}(S^2) &= \frac{2\sigma^4}{n-1}. \end{aligned}$$

Consider the "competing estimator:"

$$S_b^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

which recall is the MOM and MLE of σ^2.

Note that

$$S_b^2 = \left(\frac{n-1}{n}\right) S^2 \implies E_{\boldsymbol{\theta}}(S_b^2) = E_{\boldsymbol{\theta}}\left[\left(\frac{n-1}{n}\right) S^2\right] = \left(\frac{n-1}{n}\right) E_{\boldsymbol{\theta}}(S^2) = \left(\frac{n-1}{n}\right)\sigma^2.$$

That is, the estimator S_b^2 is biased; it **underestimates** σ^2 on average.

Comparison: Let's compare S^2 and S_b^2 on the basis of MSE. Because S^2 is an unbiased estimator of σ^2,

$$\text{MSE}_{\boldsymbol{\theta}}(S^2) = \text{var}_{\boldsymbol{\theta}}(S^2) = \frac{2\sigma^4}{n-1}.$$

The MSE of S_b^2 is

$$\text{MSE}_{\boldsymbol{\theta}}(S_b^2) = \text{var}_{\boldsymbol{\theta}}(S_b^2) + \text{Bias}_{\boldsymbol{\theta}}^2(S_b^2).$$

The variance of S_b^2 is

$$\begin{aligned} \text{var}_{\boldsymbol{\theta}}(S_b^2) &= \text{var}_{\boldsymbol{\theta}}\left[\left(\frac{n-1}{n}\right) S^2\right] \\ &= \left(\frac{n-1}{n}\right)^2 \text{var}_{\boldsymbol{\theta}}(S^2) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}. \end{aligned}$$

The bias of S_b^2 is

$$\text{E}_{\boldsymbol{\theta}}(S_b^2 - \sigma^2) = \text{E}_{\boldsymbol{\theta}}(S_b^2) - \sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2.$$

Therefore,

$$\text{MSE}_{\boldsymbol{\theta}}(S_b^2) = \underbrace{\frac{2(n-1)\sigma^4}{n^2}}_{\text{var}_{\boldsymbol{\theta}}(S_b^2)} + \underbrace{\left[\left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2\right]^2}_{\text{Bias}_{\boldsymbol{\theta}}^2(S_b^2)} = \left(\frac{2n-1}{n^2}\right)\sigma^4.$$

Finally, to compare $\text{MSE}_{\boldsymbol{\theta}}(S^2)$ with $\text{MSE}_{\boldsymbol{\theta}}(S_b^2)$, we are left to compare the constants

$$\frac{2}{n-1} \quad \text{and} \quad \frac{2n-1}{n^2}.$$

Note that the ratio

$$\frac{\frac{2n-1}{n^2}}{\frac{2}{n-1}} = \frac{2n^2-3n+1}{2n^2} < 1,$$

for all $n \geq 2$. Therefore,

$$\text{MSE}_{\boldsymbol{\theta}}(S_b^2) < \text{MSE}_{\boldsymbol{\theta}}(S^2),$$

showing that S_b^2 is a "better" estimator than S^2 on the basis of MSE.

Discussion: In general, how should we **compare** two competing estimators W_1 and W_2?

- If both W_1 and W_2 are unbiased, we prefer the estimator with the smaller variance.
- If either W_1 or W_2 is biased (or perhaps both are biased), we prefer the estimator with the smaller MSE.

There is no guarantee that one estimator, say W_1, will **always** beat the other for all $\theta \in \Theta$ (i.e., for all values of θ in the parameter space). For example, it may be that W_1 has smaller MSE for some values of $\theta \in \Theta$, but larger MSE for other values.

Remark: In some situations, we might have a biased estimator, but we can calculate its bias. We can then "adjust" the (biased) estimator to make it unbiased. I like to call this "making biased estimators unbiased." The following example illustrates this.

Example 7.14. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}[0, \theta]$, where $\theta > 0$. We know (from Example 7.4) that the MLE of θ is $X_{(n)}$, the maximum order statistic. It is easy to show that

$$E_\theta(X_{(n)}) = \left(\frac{n}{n+1}\right)\theta.$$

The MLE is biased because $E_\theta(X_{(n)}) \neq \theta$. However, the estimator

$$\left(\frac{n+1}{n}\right)X_{(n)},$$

an "adjusted version" of $X_{(n)}$, is unbiased.

Remark: In the previous example, we might compare the following estimators:

$$\begin{aligned} W_1 = W_1(\mathbf{X}) &= \left(\frac{n+1}{n}\right)X_{(n)} \\ W_2 = W_2(\mathbf{X}) &= 2\overline{X}. \end{aligned}$$

The estimator W_1 is an unbiased version of the MLE. The estimator W_2 is the MOM (which is also unbiased). I have calculated

$$\text{var}_\theta(W_1) = \frac{\theta^2}{n(n+2)} \qquad \text{and} \qquad \text{var}_\theta(W_2) = \frac{\theta^2}{3n}.$$

It is easy to see that $\text{var}_\theta(W_1) \leq \text{var}_\theta(W_2)$, for all $n \geq 2$. Therefore, W_1 is a "better" estimator on the basis of this variance comparison. Are you surprised?

Curiosity: Might there be another unbiased estimator, say $W_3 = W_3(\mathbf{X})$ that is "better" than both W_1 and W_2? If a better (unbiased) estimator does exist, how do we find it?

7.3.2 Best unbiased estimators

Goal: Consider the class of estimators

$$\mathcal{C}_\tau = \{W = W(\mathbf{X}) : E_\theta(W) = \tau(\theta) \ \ \forall \theta \in \Theta\}.$$

That is, $\mathcal{C}_\tau$ is the collection of all unbiased estimators of $\tau(\theta)$. Our goal is to find the (unbiased) estimator $W^* \in \mathcal{C}_\tau$ that has the smallest variance.

Remark: On the surface, this task seems somewhat insurmountable because $\mathcal{C}_\tau$ is a very large class. In Example 7.14, for example, both $W_1 = \left(\frac{n+1}{n}\right) X_{(n)}$ and $W_2 = 2\overline{X}$ are unbiased estimators of θ. However, so is the convex combination

$$W_a = W_a(\mathbf{X}) = a\left(\frac{n+1}{n}\right) X_{(n)} + (1-a)2\overline{X},$$

for all $a \in (0, 1)$.

Remark: It seems that our discussion of "best" estimators starts with the restriction that we will consider only those that are unbiased. If we did not make a restriction like this, then we would have to deal with too many estimators, many of which are nonsensical. For example, suppose $X_1, X_2, ..., X_n$ are iid Poisson(θ), where $\theta > 0$.

- The estimators $\overline{X}$ and S^2 emerge as candidate estimators because they are unbiased.
- However, suppose we widen our search to consider all possible estimators and then try to find the one with the smallest MSE. Consider the estimator $\widehat{\theta} = 17$.
  - If $\theta = 17$, then $\widehat{\theta}$ can never be beaten in terms of MSE; its MSE $= 0$.
  - If $\theta \neq 17$, then $\widehat{\theta}$ may be a terrible estimator; its MSE $= (17 - \theta)^2$.
- We want to exclude nonsensical estimators like this. Our solution is to restrict attention to estimators that are unbiased.

Definition: An estimator $W^* = W^*(\mathbf{X})$ is a **uniformly minimum variance unbiased estimator** (**UMVUE**) of $\tau(\theta)$ if

1. $E_\theta(W^*) = \tau(\theta)$ for all $\theta \in \Theta$
2. $\text{var}_\theta(W^*) \leq \text{var}_\theta(W)$, for all $\theta \in \Theta$, where W is any other unbiased estimator of $\tau(\theta)$.

Note: This definition is stated in full generality. Most of the time (but certainly not always), we will be interested in estimating θ itself; i.e., $\tau(\theta) = \theta$. Also, as the notation suggests, we assume that $\tau(\theta)$ is a scalar parameter and that estimators are also scalar.

Discussion/Preview: How do we find UMVUEs? We start by noting the following:

- UMVUEs may not exist.
- If a UMVUE does exist, it is unique (we'll prove this later).

We present **two approaches** to find UMVUEs:

Approach 1: Determine a **lower bound**, say $B(\theta)$, on the variance of any unbiased estimator of $\tau(\theta)$. Then, if we can find an unbiased estimator W^* whose variance attains this lower bound, that is,

$$\text{var}_\theta(W^*) = B(\theta),$$

for all $\theta \in \Theta$, then we know that W^* is UMVUE.

Approach 2: Link the notion of being "best" with that of sufficiency and completeness.

Theorem 7.3.9 (Cramér-Rao Inequality). Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$, where

1. the support of $\mathbf{X}$ is free of all unknown parameters
2. for any function $h(\mathbf{x})$ such that $E_\theta[h(\mathbf{X})] < \infty$ for all $\theta \in \Theta$, the interchange

$$\frac{d}{d\theta}\int_{\mathbb{R}^n} h(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta) d\mathbf{x} = \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta} h(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta) d\mathbf{x}$$

is justified; i.e., we can interchange the derivative and integral (derivative and sum if $\mathbf{X}$ is discrete).

For any estimator $W(\mathbf{X})$ with $\text{var}_\theta[W(\mathbf{X})] < \infty$, the following inequality holds:

$$\text{var}_\theta[W(\mathbf{X})] \geq \frac{\left\{\frac{d}{d\theta} E_\theta[W(\mathbf{X})]\right\}^2}{E_\theta\left\{\left[\frac{\partial}{\partial\theta} \ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\}}.$$

The quantity on the RHS is called the **Cramér-Rao Lower Bound** (**CRLB**) on the variance of the estimator $W(\mathbf{X})$.

Remark: Note that in the statement of the CRLB in Theorem 7.3.9, we haven't said exactly what $W(\mathbf{X})$ is an estimator for. This is to preserve the generality of the result; Theorem 7.3.9 holds for any estimator with finite variance. However, given our desire to restrict attention to unbiased estimators, we will usually consider one of these cases:

- If $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$, then the numerator becomes

$$\left[\frac{d}{d\theta}\,\tau(\theta)\right]^2 = [\tau'(\theta)]^2.$$

- If $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta) = \theta$, then the numerator equals 1.

Important special case (Corollary 7.3.10): When $\mathbf{X}$ consists of $X_1, X_2, ..., X_n$ which are **iid** from the population $f_X(x|\theta)$, then the denominator in Theorem 7.3.9

$$E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\} = nE_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\},$$

or, using other notation,

$$I_n(\theta) = nI_1(\theta).$$

We call $I_n(\theta)$ the **Fisher information** based on the sample $\mathbf{X}$. We call $I_1(\theta)$ the **Fisher information** based on one observation X.

Lemma 7.3.11 (Information Equality): Under fairly mild assumptions (which hold for exponential families, for example), the Fisher information based on one observation

$$I_1(\theta) = E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\} = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right].$$

The second expectation is often easier to calculate.

Preview: In Chapter 10, we will investigate the large-sample properties of MLEs. Under certain regularity conditions, we will show an MLE $\widehat{\theta}$ satisfies

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\widehat{\theta}}),$$

where the asymptotic variance

$$\sigma^2_{\widehat{\theta}} = \frac{1}{I_1(\theta)}.$$

This is an extremely useful (large-sample) result; e.g., it makes getting large-sample CIs and performing large-sample tests straightforward. Furthermore, an analogous large-sample result holds for vector-valued MLEs. If $\widehat{\boldsymbol{\theta}}$ is the MLE of a $k \times 1$ dimensional parameter $\boldsymbol{\theta}$, then

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \text{mvn}_k(\mathbf{0}, \boldsymbol{\Sigma}),$$

where the asymptotic variance-covariance matrix (now, $k \times k$)

$$\boldsymbol{\Sigma} = [I_1(\boldsymbol{\theta})]^{-1}$$

is the inverse of the $k \times k$ Fisher information matrix $I_1(\boldsymbol{\theta})$.

Example 7.15. Suppose $X_1, X_2, ..., X_n$ are iid Poisson(θ), where $\theta > 0$. Find the CRLB on the variance of unbiased estimators of $\tau(\theta) = \theta$.
Solution. We know that the CRLB is

$$\frac{1}{I_n(\theta)} = \frac{1}{nI_1(\theta)},$$

where

$$I_1(\theta) = E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\} = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right].$$

For $x = 0, 1, 2, ...$,

$$\ln f_X(x|\theta) = \ln\left(\frac{\theta^x e^{-\theta}}{x!}\right) = x\ln\theta - \theta - \ln x!.$$

Therefore,

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln f_X(x|\theta) &= \frac{x}{\theta} - 1\\ \frac{\partial^2}{\partial\theta^2}\ln f_X(x|\theta) &= -\frac{x}{\theta^2}.\end{aligned}$$

The Fisher information based on one observation is

$$\begin{aligned}I_1(\theta) &= -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right]\\ &= -E_\theta\left(-\frac{X}{\theta^2}\right) = \frac{1}{\theta}.\end{aligned}$$

Therefore, the CRLB on the variance of all unbiased estimators of $\tau(\theta) = \theta$ is

$$\text{CRLB} = \frac{1}{nI_1(\theta)} = \frac{\theta}{n}.$$

Observation: Because $W(\mathbf{X}) = \overline{X}$ is an unbiased estimator of $\tau(\theta) = \theta$ in the Poisson(θ) model and because

$$\text{var}_\theta(\overline{X}) = \frac{\theta}{n},$$

we see that $\text{var}_\theta(\overline{X})$ does attain the CRLB. This means that $W(\mathbf{X}) = \overline{X}$ is the UMVUE for $\tau(\theta) = \theta$.

Example 7.16. Suppose $X_1, X_2, ..., X_n$ are iid gamma(α_0, β), where α_0 is known and $\beta > 0$. Find the CRLB on the variance of unbiased estimators of β.
Solution. We know that the CRLB is

$$\frac{1}{I_n(\beta)} = \frac{1}{nI_1(\beta)},$$

where

$$I_1(\beta) = E_\beta\left\{\left[\frac{\partial}{\partial\beta}\ln f_X(X|\beta)\right]^2\right\} = -E_\beta\left[\frac{\partial^2}{\partial\beta^2}\ln f_X(X|\beta)\right].$$

For $x > 0$,

$$\begin{aligned}\ln f_X(x|\beta) &= \ln\left[\frac{1}{\Gamma(\alpha_0)\beta^{\alpha_0}}x^{\alpha_0-1}e^{-x/\beta}\right] \\ &= -\ln\Gamma(\alpha_0) - \alpha_0\ln\beta + (\alpha_0-1)\ln x - \frac{x}{\beta}.\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\partial}{\partial\beta}\ln f_X(x|\beta) &= -\frac{\alpha_0}{\beta} + \frac{x}{\beta^2} \\ \frac{\partial^2}{\partial\beta^2}\ln f_X(x|\beta) &= \frac{\alpha_0}{\beta^2} - \frac{2x}{\beta^3}.\end{aligned}$$

The Fisher information based on one observation is

$$\begin{aligned}I_1(\beta) &= -E_\beta\left[\frac{\partial^2}{\partial\beta^2}\ln f_X(X|\beta)\right] \\ &= -E_\beta\left(\frac{\alpha_0}{\beta^2} - \frac{2X}{\beta^3}\right) = \frac{\alpha_0}{\beta^2}.\end{aligned}$$

Therefore, the CRLB on the variance of all unbiased estimators of β is

$$\text{CRLB} = \frac{1}{nI_1(\beta)} = \frac{\beta^2}{n\alpha_0}.$$

Observation: Consider the estimator

$$W(\mathbf{X}) = \frac{\overline{X}}{\alpha_0}.$$

Note that

$$E_\beta[W(\mathbf{X})] = E_\beta\left(\frac{\overline{X}}{\alpha_0}\right) = \frac{\alpha_0\beta}{\alpha_0} = \beta$$

and

$$\text{var}_\beta[W(\mathbf{X})] = \text{var}_\beta\left(\frac{\overline{X}}{\alpha_0}\right) = \frac{\alpha_0\beta^2}{n\alpha_0^2} = \frac{\beta^2}{n\alpha_0}.$$

We see that $W(\mathbf{X}) = \overline{X}/\alpha_0$ is an unbiased estimator for β and $\text{var}_\beta(\overline{X}/\alpha_0)$ attains the CRLB. This means that $W(\mathbf{X}) = \overline{X}/\alpha_0$ is the UMVUE for β.

Discussion: Instead of estimating β in Example 7.16, suppose that we were interested in estimating $\tau(\beta) = 1/\beta$ instead.

1. Show that
$$W(\mathbf{X}) = \frac{n\alpha_0 - 1}{n\overline{X}}$$
is an unbiased estimator of $\tau(\beta) = 1/\beta$.

2. Derive the CRLB for the variance of unbiased estimators of $\tau(\beta) = 1/\beta$.

3. Calculate $\text{var}_\beta[W(\mathbf{X})]$ and show that it is strictly larger than the CRLB (i.e., the variance does not attain the CRLB).

Q: Does this necessarily imply that $W(\mathbf{X})$ cannot be the UMVUE of $\tau(\beta) = 1/\beta$?

Remark: In general, the CRLB offers a **lower bound** on the variance of any unbiased estimator of $\tau(\theta)$. However, this lower bound may be unattainable. That is, the CRLB may be strictly smaller than the variance of any unbiased estimator. If this is the case, then our "CRLB approach" to finding an UMVUE will not be helpful.

Corollary 7.3.15 (Attainment). Suppose $X_1, X_2, ..., X_n$ is an iid sample from $f_X(x|\theta)$, where $\theta \in \Theta$, a family that satisfies the regularity conditions stated for the Cramér-Rao Inequality. If $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$, then $\text{var}_\theta[W(\mathbf{X})]$ attains the CRLB if and only if the score function

$$S(\theta|\mathbf{x}) = a(\theta)[W(\mathbf{x}) - \tau(\theta)]$$

is a linear function of $W(\mathbf{x})$.

Recall: The **score function** is given by

$$\begin{aligned} S(\theta|\mathbf{x}) &= \frac{\partial}{\partial\theta} \ln L(\theta|\mathbf{x}) \\ &= \frac{\partial}{\partial\theta} \ln f_{\mathbf{X}}(\mathbf{x}|\theta). \end{aligned}$$

Example 7.16 (continued). Suppose $X_1, X_2, ..., X_n$ are iid gamma(α_0, β), where α_0 is known and $\beta > 0$. The likelihood function is

$$\begin{aligned} L(\beta|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_0)\beta^{\alpha_0}} x_i^{\alpha_0 - 1} e^{-x_i/\beta} \\ &= \left[\frac{1}{\Gamma(\alpha_0)\beta^{\alpha_0}}\right]^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha_0 - 1} e^{-\sum_{i=1}^{n} x_i/\beta}. \end{aligned}$$

The log-likelihood function is

$$\ln L(\beta|\mathbf{x}) = -n \ln \Gamma(\alpha_0) - n\alpha_0 \ln \beta + (\alpha_0 - 1) \sum_{i=1}^{n} \ln x_i - \frac{\sum_{i=1}^{n} x_i}{\beta}.$$

The score function is

$$\begin{aligned} S(\beta|\mathbf{x}) = \frac{\partial}{\partial\beta}\ln L(\beta|\mathbf{x}) &= -\frac{n\alpha_0}{\beta} + \frac{\sum_{i=1}^n x_i}{\beta^2} \\ &= \frac{n\alpha_0}{\beta^2}\left(\frac{\sum_{i=1}^n x_i}{n\alpha_0} - \beta\right) \\ &= a(\beta)[W(\mathbf{x}) - \tau(\beta)], \end{aligned}$$

where

$$W(\mathbf{x}) = \frac{\sum_{i=1}^n x_i}{n\alpha_0} = \frac{\overline{x}}{\alpha_0}.$$

We have written the score function $S(\beta|\mathbf{x})$ as a linear function of $W(\mathbf{x}) = \overline{x}/\alpha_0$. Because $W(\mathbf{X}) = \overline{X}/\alpha_0$ is an unbiased estimator of $\tau(\beta) = \beta$ (shown previously), the variance $\text{var}_\beta[W(\mathbf{X})]$ attains the CRLB for the variance of unbiased estimators of $\tau(\beta) = \beta$.

Remark: The attainment result is interesting, but I have found that its usefulness may be limited if you want to find the UMVUE. Even if we can write

$$S(\theta|\mathbf{x}) = a(\theta)[W(\mathbf{x}) - \tau(\theta)]$$

where $E_\theta[W(\mathbf{X})] = \tau(\theta)$, the RHS might involve a function $\tau(\theta)$ for which there is no desire to estimate. To illustrate this, suppose $X_1, X_2, ..., X_n$ are iid beta$(\theta, 1)$, where $\theta > 0$. The score function is

$$\begin{aligned} S(\theta|\mathbf{x}) &= \frac{n}{\theta} + \sum_{i=1}^n \ln x_i \\ &= n\left[\frac{\sum_{i=1}^n \ln x_i}{n} - \left(-\frac{1}{\theta}\right)\right] \\ &= a(\theta)[W(\mathbf{x}) - \tau(\theta)]. \end{aligned}$$

It turns out that

$$E_\theta[W(\mathbf{X})] = E_\theta\left(\frac{1}{n}\sum_{i=1}^n \ln X_i\right) = -\frac{1}{\theta}.$$

We have shown that $\text{var}_\theta[W(\mathbf{X})]$ attains the CRLB on the variance of unbiased estimators of $\tau(\theta) = -1/\theta$, a parameter we likely have no desire to estimate.

Unresolved issues:

1. What if $f_X(x|\theta)$ does not satisfy the regularity conditions needed for the Cramér-Rao Inequality to apply? For example, $X \sim \mathcal{U}(0, \theta)$.

2. What if the CRLB is unattainable? Can we still find the UMVUE?

7.3.3 Sufficiency and completeness

Remark: We now move to our "second approach" on how to find UMVUEs. This approach involves sufficiency and completeness−two topics we discussed in the last chapter. We can also address the unresolved issues on the previous page.

Theorem 7.3.17 (**Rao-Blackwell**). Let $W = W(\mathbf{X})$ be an unbiased estimator of $\tau(\theta)$. Let $T = T(\mathbf{X})$ be a sufficient statistic for θ. Define

$$\phi(T) = E(W|T).$$

Then

1. $E_\theta[\phi(T)] = \tau(\theta)$ for all $\theta \in \Theta$
2. $\text{var}_\theta[\phi(T)] \leq \text{var}_\theta(W)$ for all $\theta \in \Theta$.

That is, $\phi(T) = E(W|T)$ is a uniformly better unbiased estimator than W.
Proof. This result follows from the iterated rules for means and variances. First,

$$E_\theta[\phi(T)] = E_\theta[E(W|T)] = E_\theta(W) = \tau(\theta).$$

Second,

$$\begin{aligned} \text{var}_\theta(W) &= E_\theta[\text{var}(W|T)] + \text{var}_\theta[E(W|T)] \\ &= E_\theta[\text{var}(W|T)] + \text{var}_\theta[\phi(T)] \\ &\geq \text{var}_\theta[\phi(T)], \end{aligned}$$

because $\text{var}(W|T) \geq 0$ (a.s.) and hence $E_\theta[\text{var}(W|T)] \geq 0$. □

Implication: We can always "improve" the unbiased estimator W by conditioning on a sufficient statistic.

Remark: To use the Rao-Blackwell Theorem, some students think they have to

1. Find an unbiased estimator W.
2. Find a sufficient statistic T.
3. Derive the conditional distribution $f_{W|T}(w|t)$.
4. Find the mean $E(W|T)$ of this conditional distribution.

This is not the case at all! Because $\phi(T) = E(W|T)$ is a function of the sufficient statistic T, the Rao-Blackwell result simply convinces us that in our search for the UMVUE, we can restrict attention to those estimators that are functions of a sufficient statistic.

Q: In the proof of the Rao-Blackwell Theorem, where did we use the fact that T was sufficient?
A: Nowhere. Thus, it would seem that conditioning on any statistic, sufficient or not, will result in an improvement over the unbiased W. However, there is a catch:

- If T is not sufficient, then there is no guarantee that $\phi(T) = E(W|T)$ will be an estimator; i.e., it could depend on θ. See Example 7.3.18 (CB, pp 343).

Remark: To understand how we can use the Rao-Blackwell result in our quest to find a UMVUE, we need two additional results. One deals with uniqueness; the other describes an interesting characterization of a UMVUE itself.

Theorem 7.3.19 (**Uniqueness**). If W is UMVUE for $\tau(\theta)$, then it is unique.
Proof. Suppose that W' is also UMVUE. It suffices to show that $W = W'$ with probability one. Define

$$W^* = \frac{1}{2}(W + W').$$

Note that

$$E_\theta(W^*) = \frac{1}{2}[E_\theta(W) + E_\theta(W')] = \tau(\theta), \quad \text{for all } \theta \in \Theta,$$

showing that W^* is an unbiased estimator of $\tau(\theta)$. The variance of W^* is

$$\begin{aligned}
\text{var}_\theta(W^*) &= \text{var}_\theta\left[\frac{1}{2}(W + W')\right] \\
&= \frac{1}{4}\text{var}_\theta(W) + \frac{1}{4}\text{var}_\theta(W') + \frac{1}{2}\text{cov}_\theta(W, W') \\
&\leq \frac{1}{4}\text{var}_\theta(W) + \frac{1}{4}\text{var}_\theta(W') + \frac{1}{2}\left[\text{var}_\theta(W)\text{var}_\theta(W')\right]^{1/2} \\
&= \text{var}_\theta(W),
\end{aligned}$$

where the inequality arises from the covariance inequality (CB, pp 188, application of Cauchy-Schwarz) and the final equality holds because both W and W' are UMVUE by assumption (so their variances must be equal). Therefore, we have shown that

1. W^* is unbiased for $\tau(\theta)$
2. $\text{var}_\theta(W^*) \leq \text{var}_\theta(W)$.

Because W is UMVUE (by assumption), the inequality in (2) can not be strict (or else it would contradict the fact that W is UMVUE). Therefore, it must be true that

$$\text{var}_\theta(W^*) = \text{var}_\theta(W).$$

This implies that the inequality above (arising from the covariance inequality) is an equality; therefore,

$$\text{cov}_\theta(W, W') = \left[\text{var}_\theta(W)\text{var}_\theta(W')\right]^{1/2}.$$

Therefore,

$$\text{corr}_\theta(W, W') = \pm 1 \implies W' = \underbrace{a(\theta)W + b(\theta)}_{\text{linear function of } W}, \text{ with probability 1,}$$

by Theorem 4.5.7 (CB, pp 172), where $a(\theta)$ and $b(\theta)$ are constants. It therefore suffices to show that $a(\theta) = 1$ and $b(\theta) = 0$. Note that

$$\begin{aligned} \text{cov}_\theta(W, W') = \text{cov}_\theta[W, a(\theta)W + b(\theta)] &= a(\theta)\text{cov}_\theta(W, W) \\ &= a(\theta)\text{var}_\theta(W). \end{aligned}$$

However, we have previously shown that

$$\begin{aligned} \text{cov}_\theta(W, W') = [\text{var}_\theta(W)\text{var}_\theta(W')]^{1/2} &= [\text{var}_\theta(W)\text{var}_\theta(W)]^{1/2} \\ &= \text{var}_\theta(W). \end{aligned}$$

This implies $a(\theta) = 1$. Finally,

$$\begin{aligned} E_\theta(W') = E_\theta[a(\theta)W + b(\theta)] &= E_\theta[W + b(\theta)] \\ &= E_\theta(W) + b(\theta). \end{aligned}$$

Because both W and W' are unbiased, this implies $b(\theta) = 0$. □

Theorem 7.3.20. Suppose $E_\theta(W) = \tau(\theta)$ for all $\theta \in \Theta$. W is UMVUE of $\tau(\theta)$ if and only if W is uncorrelated with all unbiased estimators of 0.

Proof. Necessity ($\implies$): Suppose $E_\theta(W) = \tau(\theta)$ for all $\theta \in \Theta$. Suppose W is UMVUE of $\tau(\theta)$. Suppose $E_\theta(U) = 0$ for all $\theta \in \Theta$. It suffices to show $\text{cov}_\theta(W, U) = 0$ for all $\theta \in \Theta$. Define

$$\phi_a = W + aU,$$

where a is a constant. It is easy to see that ϕ_a is an unbiased estimator of $\tau(\theta)$; for all $\theta \in \Theta$,

$$E_\theta(\phi_a) = E_\theta(W + aU) = E_\theta(W) + a\underbrace{E_\theta(U)}_{=\,0} = \tau(\theta).$$

Also,

$$\begin{aligned} \text{var}_\theta(\phi_a) &= \text{var}_\theta(W + aU) \\ &= \text{var}_\theta(W) + \underbrace{a^2\text{var}_\theta(U) + 2a\ \text{cov}_\theta(W, U)}_{\text{Key question: Can this be negative?}}. \end{aligned}$$

- **Case 1:** Suppose $\exists\ \theta_0 \in \Theta$ such that $\text{cov}_{\theta_0}(W, U) < 0$. Then

$$\begin{aligned} a^2\text{var}_{\theta_0}(U) + 2a\ \text{cov}_{\theta_0}(W, U) < 0 &\iff a^2\text{var}_{\theta_0}(U) < -2a\ \text{cov}_{\theta_0}(W, U) \\ &\iff a^2 < -\frac{2a\ \text{cov}_{\theta_0}(W, U)}{\text{var}_{\theta_0}(U)}. \end{aligned}$$

I can make this true by picking

$$0 < a < -\frac{2\ \text{cov}_{\theta_0}(W,U)}{\text{var}_{\theta_0}(U)}$$

and therefore I have shown that

$$\text{var}_{\theta_0}(\phi_a) < \text{var}_{\theta_0}(W).$$

However, this contradicts the assumption that W is UMVUE. Therefore, it must be true that $\text{cov}_\theta(W,U) \geq 0$.

- **Case 2:** Suppose $\exists\ \theta_0 \in \Theta$ such that $\text{cov}_{\theta_0}(W,U) > 0$. Then

$$\begin{aligned} a^2\text{var}_{\theta_0}(U) + 2a\ \text{cov}_{\theta_0}(W,U) < 0 &\iff a^2\text{var}_{\theta_0}(U) < -2a\ \text{cov}_{\theta_0}(W,U) \\ &\iff a^2 < -\frac{2a\ \text{cov}_{\theta_0}(W,U)}{\text{var}_{\theta_0}(U)}. \end{aligned}$$

I can make this true by picking

$$-\frac{2\ \text{cov}_{\theta_0}(W,U)}{\text{var}_{\theta_0}(U)} < a < 0$$

and therefore I have shown that

$$\text{var}_{\theta_0}(\phi_a) < \text{var}_{\theta_0}(W).$$

However, this again contradicts the assumption that W is UMVUE. Therefore, it must be true that $\text{cov}_\theta(W,U) \leq 0$.

Combining Case 1 and Case 2, we are forced to conclude that $\text{cov}_\theta(W,U) = 0$. This proves the necessity.

Sufficiency ($\Longleftarrow$): Suppose $E_\theta(W) = \tau(\theta)$ for all $\theta \in \Theta$. Suppose $\text{cov}_\theta(W,U) = 0$ for all $\theta \in \Theta$ where U is any unbiased estimator of zero; i.e., $E_\theta(U) = 0$ for all $\theta \in \Theta$. Let W' be any other unbiased estimator of $\tau(\theta)$. It suffices to show that $\text{var}_\theta(W) \leq \text{var}_\theta(W')$. Write

$$W' = W + (W' - W)$$

and calculate

$$\text{var}_\theta(W') = \text{var}_\theta(W) + \text{var}_\theta(W' - W) + 2\text{cov}_\theta(W, W' - W).$$

However, $\text{cov}_\theta(W, W' - W) = 0$ because $W' - W$ is an unbiased estimator of 0. Therefore,

$$\text{var}_\theta(W') = \text{var}_\theta(W) + \underbrace{\text{var}_\theta(W' - W)}_{\geq 0} \ \geq\ \text{var}_\theta(W).$$

This proves the sufficiency. $\square$

Summary: We are now ready to put Theorem 7.3.17 (Rao-Blackwell), Theorem 7.3.19 (UMVUE uniqueness) and Theorem 7.3.20 together. Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$, where $\theta \in \Theta$. Our goal is to find the UMVUE of $\tau(\theta)$.

- Theorem 7.3.17 (Rao-Blackwell) assures us that we can restrict attention to functions of sufficient statistics.

Therefore, suppose T is a sufficient statistic for θ. Suppose that $\phi(T)$, a function of T, is an unbiased estimator of $\tau(\theta)$; i.e.,

$$E_\theta[\phi(T)] = \tau(\theta), \quad \text{for all } \theta \in \Theta.$$

- Theorem 7.3.20 assures us that $\phi(T)$ is UMVUE if and only if $\phi(T)$ is uncorrelated with all unbiased estimators of 0.

Add the assumption that T is a complete statistic. *The only unbiased estimator of 0 in complete families is the zero function itself.* Because $\text{cov}_\theta[\phi(T), 0] = 0$ holds trivially, we have shown that $\phi(T)$ is uncorrelated with "all" unbiased estimators of 0. Theorem 7.3.20 says that $\phi(T)$ must be UMVUE; Theorem 7.3.19 guarantees that $\phi(T)$ is unique.

Recipe for finding UMVUEs: Suppose we want to find the UMVUE for $\tau(\theta)$.

1. Start by finding a statistic T that is both sufficient and complete.
2. Find a function of T, say $\phi(T)$, that satisfies

$$E_\theta[\phi(T)] = \tau(\theta), \quad \text{for all } \theta \in \Theta.$$

Then $\phi(T)$ is the UMVUE for $\tau(\theta)$. This is essentially what is summarized in Theorem 7.3.23 (CB, pp 347).

Example 7.17. Suppose $X_1, X_2, ..., X_n$ are iid Poisson(θ), where $\theta > 0$.

- We already know that $\overline{X}$ is UMVUE for θ; we proved this by showing that $\overline{X}$ is unbiased and that $\text{var}_\theta(\overline{X})$ attains the CRLB on the variance of all unbiased estimators of θ.
- We now show $\overline{X}$ is UMVUE for θ by using sufficiency and completeness.

The pmf of X is

$$\begin{aligned} f_X(x|\theta) &= \frac{\theta^x e^{-\theta}}{x!} I(x = 0, 1, 2, ...,) \\ &= \frac{I(x = 0, 1, 2, ...,)}{x!} e^{-\theta} e^{(\ln \theta)x} \\ &= h(x)c(\theta) \exp\{w_1(\theta)t_1(x)\}. \end{aligned}$$

Therefore X has pmf in the exponential family. Theorem 6.2.10 says that

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a sufficient statistic. Because $d = k = 1$ (i.e., a full family), Theorem 6.2.25 says that T is complete. Now,

$$E_\theta(T) = E_\theta\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E_\theta(X_i) = n\theta.$$

Therefore,

$$E_\theta\left(\frac{T}{n}\right) = E_\theta(\overline{X}) = \theta.$$

Because $\overline{X}$ is unbiased and is a function of T, a complete and sufficient statistic, we know that $\overline{X}$ is the UMVUE.

Example 7.18. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0, \theta)$, where $\theta > 0$. We have previously shown that

$$T = T(\mathbf{X}) = X_{(n)}$$

is sufficient and complete (see Example 6.5 and Example 6.16, respectively, in the notes). It follows that

$$E_\theta(T) = E_\theta(X_{(n)}) = \left(\frac{n}{n+1}\right)\theta$$

for all $\theta > 0$. Therefore,

$$E_\theta\left[\left(\frac{n+1}{n}\right)X_{(n)}\right] = \theta.$$

Because $(n+1)X_{(n)}/n$ is unbiased and is a function of $X_{(n)}$, a complete and sufficient statistic, it must be the UMVUE.

Example 7.19. Suppose $X_1, X_2, ..., X_n$ are iid gamma(α_0, β), where α_0 is known and $\beta > 0$. Find the UMVUE of $\tau(\beta) = 1/\beta$.
Solution. The pdf of X is

$$\begin{aligned} f_X(x|\beta) &= \frac{1}{\Gamma(\alpha_0)\beta^{\alpha_0}} x^{\alpha_0 - 1} e^{-x/\beta} I(x > 0) \\ &= \frac{x^{\alpha_0 - 1} I(x > 0)}{\Gamma(\alpha_0)} \, \frac{1}{\beta^{\alpha_0}} \, e^{(-1/\beta)x} \\ &= h(x)c(\beta)\exp\{w_1(\beta)t_1(x)\} \end{aligned}$$

a one-parameter exponential family with $d = k = 1$ (a full family). Theorem 6.2.10 and Theorem 6.2.25 assure that

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a sufficient and complete statistic, respectively. In Example 7.16 (notes), we saw that

$$\phi(T) = \frac{n\alpha_0 - 1}{T}$$

is an unbiased estimator of $\tau(\beta) = 1/\beta$. Therefore, $\phi(T)$ must be the UMVUE.

Remark: In Example 7.16, recall that the CRLB on the variance of unbiased estimators of $\tau(\beta) = 1/\beta$ was unattainable.

Example 7.20. Suppose $X_1, X_2, ..., X_n$ are iid Poisson(θ), where $\theta > 0$. Find the UMVUE for

$$\tau(\theta) = P_\theta(X = 0) = e^{-\theta}.$$

Solution. We use an approach known as "direct conditioning." We start with

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i,$$

which is sufficient and complete. We know that the UMVUE therefore is a function of T. Consider forming

$$\phi(T) = E(W|T),$$

where W is any unbiased estimator of $\tau(\theta) = e^{-\theta}$. We know that $\phi(T)$ by this construction is the UMVUE; clearly $\phi(T) = E(W|T)$ is a function of T and

$$E_\theta[\phi(T)] = E_\theta[E(W|T)] = E_\theta(W) = e^{-\theta}.$$

How should we choose W? Any unbiased W will "work," so let's keep our choice simple, say

$$W = W(\mathbf{X}) = I(X_1 = 0).$$

Note that

$$E_\theta(W) = E_\theta[I(X_1 = 0)] = P_\theta(X_1 = 0) = e^{-\theta},$$

showing that W is an unbiased estimator. Now, we just calculate $\phi(T) = E(W|T)$ directly. For t fixed, we have

$$\begin{aligned}
\phi(t) = E(W|T = t) &= E[I(X_1 = 0)|T = t] \\
&= P(X_1 = 0|T = t) \\
&= \frac{P_\theta(X_1 = 0, T = t)}{P_\theta(T = t)} \\
&= \frac{P_\theta\left(X_1 = 0, \sum_{i=2}^{n} X_i = t\right)}{P_\theta(T = t)} \\
&\overset{\text{indep}}{=} \frac{P_\theta(X_1 = 0) P_\theta\left(\sum_{i=2}^{n} X_i = t\right)}{P_\theta(T = t)}.
\end{aligned}$$

We can now calculate each of these probabilities. Recall that $X_1 \sim \text{Poisson}(\theta)$, $\sum_{i=2}^{n} X_i \sim \text{Poisson}((n-1)\theta)$, and $T \sim \text{Poisson}(n\theta)$. Therefore,

$$\begin{aligned}\phi(t) &= \frac{P_\theta(X_1 = 0) P_\theta\left(\sum_{i=2}^{n} X_i = t\right)}{P_\theta(T = t)} \\ &= \frac{e^{-\theta}\,\dfrac{[(n-1)\theta]^t e^{-(n-1)\theta}}{t!}}{\dfrac{(n\theta)^t e^{-n\theta}}{t!}} = \left(\frac{n-1}{n}\right)^t.\end{aligned}$$

Therefore,

$$\phi(T) = \left(\frac{n-1}{n}\right)^T$$

is the UMVUE of $\tau(\theta) = e^{-\theta}$.

Remark: It is interesting to note that in this example

$$\phi(t) = \left(\frac{n-1}{n}\right)^t = \left[\left(\frac{n-1}{n}\right)^n\right]^{\overline{x}} = \left[\left(1 - \frac{1}{n}\right)^n\right]^{\overline{x}} \approx e^{-\overline{x}},$$

for n large. Recall that $e^{-\overline{X}}$ is the MLE of $\tau(\theta) = e^{-\theta}$ by invariance.

Remark: The last subsection in CB (Section 7.3.4) is on loss-function optimality. This material will be covered in STAT 822.

7.4 Appendix: CRLB Theory

Remark: In this section, we provide the proofs that pertain to the CRLB approach to finding UMVUEs. These proofs are also relevant for later discussions on MLEs and their large-sample characteristics.

Remark: We start by reviewing the Cauchy-Schwarz Inequality. Essentially, the main Cramér-Rao inequality result (Theorem 7.3.9) follows as an application of this inequality.

Recall: Suppose X and Y are random variables. Then

$$|E(XY)| \leq E(|XY|) \leq [E(X^2)]^{1/2}[E(Y^2)]^{1/2}.$$

This is called the **Cauchy-Schwarz Inequality**. In this inequality, if we replace X with $X - \mu_X$ and Y with $Y - \mu_Y$, we get

$$|E[(X-\mu_X)(Y-\mu_Y)]| \leq \{E[(X-\mu_X)^2]\}^{1/2}\{E[(Y-\mu_Y)^2]\}^{1/2}.$$

Squaring both sides, we get

$$[\text{cov}(X,Y)]^2 \leq \sigma_X^2 \sigma_Y^2.$$

This is called the **covariance inequality**.

Theorem 7.3.9 (Cramér-Rao Inequality). Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$, where

1. the support of $\mathbf{X}$ is free of all unknown parameters
2. for any function $h(\mathbf{x})$ such that $E_\theta[h(\mathbf{X})] < \infty$ for all $\theta \in \Theta$, the interchange

$$\frac{d}{d\theta}\int_{\mathbb{R}^n} h(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} = \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta} h(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x}$$

is justified; i.e., we can interchange the derivative and integral (derivative and sum if $\mathbf{X}$ is discrete).

For any estimator $W(\mathbf{X})$ with $\text{var}_\theta[W(\mathbf{X})] < \infty$, the following inequality holds:

$$\text{var}_\theta[W(\mathbf{X})] \geq \frac{\left\{\frac{d}{d\theta}E_\theta[W(\mathbf{X})]\right\}^2}{E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\}}.$$

Proof. First we state and prove a lemma.

Lemma. Let

$$S(\theta|\mathbf{X}) = \frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)$$

denote the **score function**. The score function is a zero-mean random variable; that is,

$$E_\theta[S(\theta|\mathbf{X})] = E_\theta\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right] = 0.$$

Proof of Lemma: Note that

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right] = \int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{x}|\theta) f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} &= \int_{\mathbb{R}^n}\frac{\frac{\partial}{\partial\theta}f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{x}|\theta)} f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} \\ &= \int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} \\ &= \frac{d}{d\theta}\underbrace{\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x}}_{=1} = 0. \end{aligned}$$

The interchange of derivative and integral above is justified based on the assumptions stated in Theorem 7.3.9. Therefore, the lemma is proven. □

Note: Because the score function is a zero-mean random variable,

$$\text{var}_\theta[S(\theta|\mathbf{X})] = E_\theta\{[S(\theta|\mathbf{X})]^2\};$$

that is,

$$\text{var}_\theta\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right] = E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\}.$$

We now return to the CRLB proof. Consider

$$\begin{aligned}
\text{cov}_\theta\left[W(\mathbf{X}), \frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right] &= E_\theta\left[W(\mathbf{X})\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right] - E_\theta[W(\mathbf{X})]\underbrace{E_\theta\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]}_{=0} \\
&= E_\theta\left[W(\mathbf{X})\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right] \\
&= \int_{\mathbb{R}^n} W(\mathbf{x})\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{x}|\theta) f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} \\
&= \int_{\mathbb{R}^n} W(\mathbf{x})\frac{\frac{\partial}{\partial\theta}f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{x}|\theta)} f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} \\
&= \int_{\mathbb{R}^n} W(\mathbf{x})\frac{\partial}{\partial\theta}f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} \\
&= \frac{d}{d\theta}\int_{\mathbb{R}^n} W(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} \\
&= \frac{d}{d\theta}E_\theta[W(\mathbf{X})].
\end{aligned}$$

Now, write the covariance inequality with

1. $W(\mathbf{X})$ playing the role of "X"
2. $S(\theta|\mathbf{X}) = \frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)$ playing the role of "Y."

We get

$$\left\{\text{cov}_\theta\left[W(\mathbf{X}), \frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]\right\}^2 \leq \text{var}_\theta[W(\mathbf{X})]\ \text{var}_\theta\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right],$$

that is,

$$\left\{\frac{d}{d\theta}E_\theta[W(\mathbf{X})]\right\}^2 \leq \text{var}_\theta[W(\mathbf{X})]\ E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\}.$$

Dividing both sides by $E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\}$ gives the result. □

Corollary 7.3.10 (Cramér-Rao Inequality−iid case). With the same regularity conditions stated in Theorem 7.3.9, in the iid case,

$$\text{var}_\theta[W(\mathbf{X})] \geq \frac{\left\{\frac{d}{d\theta}E_\theta[W(\mathbf{X})]\right\}^2}{nE_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\}}.$$

Proof. It suffices to show

$$E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\} = nE_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\}.$$

Because $X_1, X_2, ..., X_n$ are iid,

$$\begin{aligned}
\text{LHS} &= E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f_X(X_i|\theta)\right]^2\right\} \\
&= E_\theta\left\{\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\ln f_X(X_i|\theta)\right]^2\right\} \\
&= E_\theta\left\{\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta)\right]^2\right\} \\
&= \sum_{i=1}^{n} E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta)\right]^2\right\} + \sum\sum_{i\neq j} E_\theta\left[\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta)\frac{\partial}{\partial\theta}\ln f_X(X_j|\theta)\right] \\
&\overset{\text{indep}}{=} \sum_{i=1}^{n} E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta)\right]^2\right\} + \sum\sum_{i\neq j} \underbrace{E_\theta\left[\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta)\right]}_{=\,0}\underbrace{E_\theta\left[\frac{\partial}{\partial\theta}\ln f_X(X_j|\theta)\right]}_{=\,0}.
\end{aligned}$$

Therefore, all cross product expectations are zero and thus

$$\text{LHS} = \sum_{i=1}^{n} E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta)\right]^2\right\} \overset{\text{ident}}{=} nE_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\}.$$

This proves the iid case. □

Remark: Recall our notation:

$$\begin{aligned}
I_n(\theta) &= E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\} \\
I_1(\theta) &= E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\}.
\end{aligned}$$

In the iid case, we have just proven that $I_n(\theta) = nI(\theta)$. Therefore, in the iid case,

- If $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$, then

$$\text{CRLB} = \frac{[\tau'(\theta)]^2}{nI_1(\theta)}.$$

- If $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta) = \theta$, then

$$\text{CRLB} = \frac{1}{nI_1(\theta)}.$$

Lemma 7.3.11 (Information Equality). Under regularity conditions,

$$I_1(\theta) = E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\} = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right].$$

Proof. From the definition of mathematical expectation,

$$E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right] = \int_{\mathbb{R}}\frac{\partial^2}{\partial\theta^2}\ln f_X(x|\theta)f_X(x|\theta)dx = \int_{\mathbb{R}}\underbrace{\frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta}f_X(x|\theta)}{f_X(x|\theta)}\right]}_{\text{use quotient rule here}}f_X(x|\theta)dx$$

Note: A sum replaces the integral above if X is discrete. The derivative

$$\begin{aligned}\frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta}f_X(x|\theta)}{f_X(x|\theta)}\right] &= \frac{\frac{\partial^2}{\partial\theta^2}f_X(x|\theta)f_X(x|\theta) - \frac{\partial}{\partial\theta}f_X(x|\theta)\frac{\partial}{\partial\theta}f_X(x|\theta)}{[f_X(x|\theta)]^2}\\ &= \frac{\frac{\partial^2}{\partial\theta^2}f_X(x|\theta)}{f_X(x|\theta)} - \frac{\left[\frac{\partial}{\partial\theta}f_X(x|\theta)\right]^2}{[f_X(x|\theta)]^2}.\end{aligned}$$

Therefore, the last integral becomes

$$\begin{aligned}\int_{\mathbb{R}}\left\{\frac{\frac{\partial^2}{\partial\theta^2}f_X(x|\theta)}{f_X(x|\theta)} - \frac{\left[\frac{\partial}{\partial\theta}f_X(x|\theta)\right]^2}{[f_X(x|\theta)]^2}\right\}f_X(x|\theta)dx &= \int_{\mathbb{R}}\left\{\frac{\partial^2}{\partial\theta^2}f_X(x|\theta) - \frac{\left[\frac{\partial}{\partial\theta}f_X(x|\theta)\right]^2}{f_X(x|\theta)}\right\}dx\\ &= \int_{\mathbb{R}}\frac{\partial^2}{\partial\theta^2}f_X(x|\theta)dx - \int_{\mathbb{R}}\frac{\left[\frac{\partial}{\partial\theta}f_X(x|\theta)\right]^2}{f_X(x|\theta)}\,dx\\ &= \frac{d^2}{d\theta^2}\underbrace{\int_{\mathbb{R}}f_X(x|\theta)dx}_{=1} - \int_{\mathbb{R}}\left[\frac{\partial}{\partial\theta}\ln f_X(x|\theta)\right]^2 f_X(x|\theta)dx\\ &= -E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\}.\end{aligned}$$

We have shown

$$E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right] = -E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\}.$$

Multiplying both sides by -1 gives the information equality. □

Remark: We now finish by proving the attainment result.

Corollary 7.3.15. Suppose $X_1, X_2, ..., X_n$ is an iid sample from $f_X(x|\theta)$, where $\theta \in \Theta$, a family that satisfies the regularity conditions stated for the Cramér-Rao Inequality. If $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$, then $\text{var}_\theta[W(\mathbf{X})]$ attains the CRLB if and only if the score function

$$S(\theta|\mathbf{x}) = a(\theta)[W(\mathbf{x}) - \tau(\theta)]$$

is a linear function of $W(\mathbf{x})$.

Proof. From the CRLB proof, recall that we had

1. $W(\mathbf{X})$ playing the role of "X"
2. $\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)$ playing the role of "Y"

in applying the covariance inequality, which yields

$$\begin{aligned}\text{var}_\theta[W(\mathbf{X})] &\geq \frac{[\tau'(\theta)]^2}{E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\}} \\ &\overset{\text{iid}}{=} \frac{[\tau'(\theta)]^2}{E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f_X(X_i|\theta)\right]^2\right\}}.\end{aligned}$$

Now, in the covariance inequality, we have *equality* when the correlation of $W(\mathbf{X})$ and $\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)$ equals ± 1, which in turn implies

$$c(X-\mu_X) = Y-\mu_Y \ \text{ a.s.},$$

or restated,

$$c[W(\mathbf{X})-\tau(\theta)] = \frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta) - 0 \ \text{ a.s.}$$

This is an application of Theorem 4.5.7 (CB, pp 172); i.e., two random variables are perfectly correlated if and only if the random variables are perfectly linearly related. In these equations, c is a constant. Also, I have written "-0" on the RHS of the last equation to emphasize that

$$E_\theta\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right] = E_\theta\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f_X(X_i|\theta)\right] = 0.$$

Also, $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$ by assumption. Therefore, we have

$$\begin{aligned}c[W(\mathbf{X})-\tau(\theta)] &= \frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta) \\ &= \frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f_X(X_i|\theta) \\ &= \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{X}) \\ &= S(\theta|\mathbf{X}),\end{aligned}$$

where $S(\theta|\mathbf{X})$ is the score function. The constant c cannot depend on $W(\mathbf{X})$ nor on $\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)$, but it can depend on θ. To emphasize this, we write

$$S(\theta|\mathbf{X}) = a(\theta)[W(\mathbf{X})-\tau(\theta)].$$

Thus, $\text{var}_\theta[W(\mathbf{X})]$ attains the CRLB when the score function $S(\theta|\mathbf{X})$ can be written as a linear function of the unbiased estimator $W(\mathbf{X})$. □

8 Hypothesis Testing

Complementary reading: Chapter 8 (CB).

8.1 Introduction

Setting: We observe $\mathbf{X} = (X_1, X_2, ..., X_n) \sim f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. For example, $X_1, X_2, ..., X_n$ might constitute a random sample (iid sample) from a population $f_X(x|\boldsymbol{\theta})$. We regard $\boldsymbol{\theta}$ as fixed and unknown.

Definition: A **statistical hypothesis** is a statement about $\boldsymbol{\theta}$. This statement specifies a collection of distributions that $\mathbf{X}$ can possibly have. Two complementary hypotheses in a testing problem are the **null hypothesis**

$$H_0 : \boldsymbol{\theta} \in \Theta_0$$

and the **alternative hypothesis**

$$H_1 : \boldsymbol{\theta} \in \Theta_0^c,$$

where $\Theta_0^c = \Theta \setminus \Theta_0$. We call Θ_0 the **null parameter space** and Θ_0^c the **alternative parameter space**.

Example 8.1. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\theta, \sigma_0^2)$, where $-\infty < \theta < \infty$ and σ_0^2 is known. Consider testing

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_1 : \theta \neq \theta_0, \end{gathered}$$

where θ_0 is a specified value of θ. The null parameter space $\Theta_0 = \{\theta_0\}$, a singleton. The alternative parameter space $\Theta_0^c = \mathbb{R} \setminus \{\theta_0\}$.

Terminology: In Example 8.1, we call $H_0 : \theta = \theta_0$ a **simple** (or **sharp**) hypothesis. Note that H_0 specifies exactly one distribution, namely, $\mathcal{N}(\theta_0, \sigma_0^2)$. A simple hypothesis specifies a single distribution.

Terminology: In Example 8.1, suppose we wanted to test

$$\begin{gathered} H_0 : \theta \leq \theta_0 \\ \text{versus} \\ H_1 : \theta > \theta_0. \end{gathered}$$

We call H_0 a **composite** (or **compound**) hypothesis. Note that H_0 specifies a family of distributions, namely, $\{\mathcal{N}(\theta, \sigma_0^2) : \theta \leq \theta_0\}$.

Goal: In a statistical hypothesis testing problem, we decide between the two complementary hypotheses H_0 and H_1 on the basis of observing $\mathbf{X} = \mathbf{x}$. In essence, a hypothesis test is a specification of the **test function**

$$\phi(\mathbf{x}) = P(\text{Reject } H_0 | \mathbf{X} = \mathbf{x}).$$

Terminology: Let $\mathcal{X}$ denote the support of $\mathbf{X}$.

- The subset of $\mathcal{X}$ for which H_0 is rejected is called the **rejection region**, denoted by R.
- The subset of $\mathcal{X}$ for which H_0 is not rejected is called the **acceptance region**, denoted by R^c.

If

$$\phi(\mathbf{x}) = I(\mathbf{x} \in R) = \begin{cases} 1, & \mathbf{x} \in R \\ 0, & \mathbf{x} \in R^c, \end{cases}$$

the test is said to be **non-randomized**.

Example 8.2. Suppose $X \sim b(10, \theta)$, where $0 < \theta < 1$, and consider testing

$$H_0 : \theta \geq 0.35$$
$$\text{versus}$$
$$H_1 : \theta < 0.35.$$

Here is an example of a **randomized test function**:

$$\phi(x) = \begin{cases} 1, & x \leq 2 \\ \frac{1}{5}, & x = 3 \\ 0, & x \geq 4. \end{cases}$$

Using this test function, we would reject H_0 if $x = 0, 1$, or 2. If $x = 3$, we would reject H_0 with probability 1/5. If $x \geq 4$, we would not reject H_0.

- If we observed $x = 3$, we could then subsequently generate $U \sim \mathcal{U}(0, 1)$.
  - If $u \leq 0.2$, then reject H_0.
  - If $u > 0.2$, then do not reject H_0.

Remark: In most problems, a test function ϕ depends on $\mathbf{X}$ through a one-dimensional **test statistic**, say

$$W = W(\mathbf{X}) = W(X_1, X_2, ..., X_n).$$

1. We would like to work with test statistics that are sensible and confer tests with nice statistical properties (does sufficiency play a role?)
2. We would like to find the **sampling distribution** of W under H_0 and H_1.

Example 8.3. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Consider testing

$$H_0 : \sigma^2 = 40$$
versus
$$H_1 : \sigma^2 \neq 40.$$

In this problem, both

$$\begin{aligned} W_1 = W_1(\mathbf{X}) &= |S^2 - 40| \\ W_2 = W_2(\mathbf{X}) &= \frac{(n-1)S^2}{40} \end{aligned}$$

are reasonable test statistics.

- Because S^2 is an unbiased estimator of σ^2, large values of W_1 (intuitively) are evidence against H_0. However, what is W_1's sampling distribution?
- The advantage of working with W_2 is that we know its sampling distribution when H_0 is true; i.e., $W_2 \sim \chi^2_{n-1}$. It is also easy to calculate the sampling distribution of W_2 when H_0 is not true; i.e., for values of $\sigma^2 \neq 40$.

Example 8.4. McCann and Tebbs (2009) summarize a study examining perceived unmet need for dental health care for people with HIV infection. Baseline in-person interviews were conducted with 2,864 HIV infected individuals (aged 18 years and older) as part of the HIV Cost and Services Utilization Study. Define

$$\begin{aligned} X_1 &= \text{number of patients with private insurance} \\ X_2 &= \text{number of patients with medicare and private insurance} \\ X_3 &= \text{number of patients without insurance} \\ X_4 &= \text{number of patients with medicare but no private insurance.} \end{aligned}$$

Set $\mathbf{X} = (X_1, X_2, X_3, X_4)$ and model $\mathbf{X} \sim \text{mult}(2864, p_1, p_2, p_3, p_4; \sum_{i=1}^4 p_i = 1)$. Under this assumption, consider testing

$$H_0 : p_1 = p_2 = p_3 = p_4 = \tfrac{1}{4}$$
versus
$$H_1 : H_0 \text{ not true.}$$

Note that an observation like $\mathbf{x} = (0, 0, 0, 2864)$ should lead to a rejection of H_0. An observation like $\mathbf{x} = (716, 716, 716, 716)$ should not. What about $\mathbf{x} = (658, 839, 811, 556)$? Can we find a reasonable one-dimensional test statistic?

8.2 Methods of Finding Tests

Preview: The authors present three methods of finding tests:

1. Likelihood ratio tests (LRTs)
2. Bayesian tests
3. Union-Intersection and Intersection-Union tests (UIT/IUT)

We will focus largely on LRTs. We will discuss Bayesian tests briefly.

8.2.1 Likelihood ratio tests

Recall: Suppose $\mathbf{X} = (X_1, X_2, ..., X_n) \sim f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. The **likelihood function** is

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}) &= f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) \\ &\stackrel{\text{iid}}{=} \prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}), \end{aligned}$$

where $f_X(x|\boldsymbol{\theta})$ is the common population distribution (in the iid case). Recall that Θ is the **parameter space**.

Definition: The **likelihood ratio test** (**LRT**) **statistic** for testing

$$\begin{gathered} H_0 : \boldsymbol{\theta} \in \Theta_0 \\ \text{versus} \\ H_1 : \boldsymbol{\theta} \in \Theta \setminus \Theta_0 \end{gathered}$$

is defined by

$$\lambda(\mathbf{x}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}|\mathbf{x})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}|\mathbf{x})}.$$

A LRT is a test that has a rejection region of the form

$$R = \{\mathbf{x} \in \mathcal{X} : \lambda(\mathbf{x}) \leq c\},$$

where $0 \leq c \leq 1$.

Intuition: The numerator of $\lambda(\mathbf{x})$ is the largest the likelihood function can be over the null parameter space Θ_0. The denominator is the largest the likelihood function can be over the entire parameter space Θ. Clearly,

$$0 \leq \lambda(\mathbf{x}) \leq 1.$$

The form of the rejection region above says to "reject H_0 when $\lambda(\mathbf{x})$ is too small." When $\lambda(\mathbf{x})$ is small, the data $\mathbf{x}$ are not consistent with the collection of models under H_0.

Connection with MLEs:

- The numerator of $\lambda(\mathbf{x})$ is
$$\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}|\mathbf{x}) = L(\widehat{\boldsymbol{\theta}}_0|\mathbf{x}),$$
where $\widehat{\boldsymbol{\theta}}_0$ is the MLE of $\boldsymbol{\theta}$ subject to the constraint that $\boldsymbol{\theta} \in \Theta_0$. That is, $\widehat{\boldsymbol{\theta}}_0$ is the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta}|\mathbf{x})$ over the null parameter space Θ_0. We call $\widehat{\boldsymbol{\theta}}_0$ the **restricted MLE**.

- The denominator of $\lambda(\mathbf{x})$ is
$$\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}|\mathbf{x}) = L(\widehat{\boldsymbol{\theta}}|\mathbf{x}),$$
where $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. That is, $\widehat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta}|\mathbf{x})$ over the entire parameter space Θ. We call $\widehat{\boldsymbol{\theta}}$ the **unrestricted MLE**.

- For notational simplicity, we often write
$$\lambda(\mathbf{x}) = \frac{L(\widehat{\boldsymbol{\theta}}_0|\mathbf{x})}{L(\widehat{\boldsymbol{\theta}}|\mathbf{x})}.$$
This notation is easier and emphasizes how the definition of $\lambda(\mathbf{x})$ is tied to maximum likelihood estimation.

Special case: When H_0 is a **simple hypothesis**; i.e.,
$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0,$$
the null parameter space is $\Theta_0 = \{\boldsymbol{\theta}_0\}$, a singleton. Clearly, in this case,
$$\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}|\mathbf{x}) = L(\widehat{\boldsymbol{\theta}}_0|\mathbf{x}) = L(\boldsymbol{\theta}_0|\mathbf{x}).$$
That is, there is only one value of $\boldsymbol{\theta}$ "allowed" under H_0. We are therefore maximizing the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ over a single point in Θ.

Large-sample intuition: We will learn in Chapter 10 that (under suitable regularity conditions), an MLE
$$\widehat{\theta} \xrightarrow{p} \theta, \quad \text{as } n \to \infty,$$
i.e., "MLEs are consistent" (I have switched to the scalar case here only for convenience). In the light of this asymptotic result, consider each of the following cases:

- Suppose that H_0 is **true**; i.e., $\theta \in \Theta_0$. Then

$$\begin{aligned} \widehat{\theta}_0 &\xrightarrow{p} \theta \in \Theta_0 \\ \widehat{\theta} &\xrightarrow{p} \theta \in \Theta_0. \end{aligned}$$

  The MLEs $\widehat{\theta}_0$ and $\widehat{\theta}$ are converging to the same quantity (in probability) so they should be close to each other in large samples. Therefore, we would expect

$$\lambda(\mathbf{x}) = \frac{L(\widehat{\theta}_0|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})}$$

  to be "close" to 1.

- Suppose H_0 is **not true**; i.e., $\theta \in \Theta \setminus \Theta_0$. Then

$$\widehat{\theta} \xrightarrow{p} \theta \in \Theta \setminus \Theta_0,$$

  but $\widehat{\theta}_0 \in \Theta_0$ because $\widehat{\theta}_0$ is calculated by maximizing $L(\theta|\mathbf{x})$ over Θ_0 (i.e., $\widehat{\theta}_0$ can never "escape from" Θ_0). Therefore, there is no guarantee that $\widehat{\theta}_0$ and $\widehat{\theta}$ will be close to each other in large samples, and, in fact, the ratio

$$\lambda(\mathbf{x}) = \frac{L(\widehat{\theta}_0|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})}$$

  could be much smaller than 1.

- This is why (at least by appealing to large-sample intuition) it makes sense to reject H_0 when $\lambda(\mathbf{x})$ is small.

Example 8.5. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and $\sigma_0^2 = 1$. Consider testing

$$\begin{gathered} H_0 : \mu = \mu_0 \\ \text{versus} \\ H_1 : \mu \neq \mu_0. \end{gathered}$$

The likelihood function is

$$L(\mu|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

The relevant parameter spaces are

$$\begin{aligned} \Theta_0 &= \{\mu_0\}, \text{ a singleton} \\ \Theta &= \{\mu : -\infty < \mu < \infty\}. \end{aligned}$$

Clearly,

$$\sup_{\mu\in\Theta_0} L(\mu|\mathbf{x}) = L(\mu_0|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2}.$$

Over the entire parameter space Θ, the MLE is $\widehat{\mu} = \overline{X}$; see Example 7.5 (notes, pp 31). Therefore,

$$\sup_{\mu\in\Theta} L(\mu|\mathbf{x}) = L(\overline{x}|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\overline{x})^2}.$$

The LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\mu_0|\mathbf{x})}{L(\overline{x}|\mathbf{x})} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\overline{x})^2}} = e^{-\frac{1}{2}\left[\sum_{i=1}^n (x_i-\mu_0)^2 - \sum_{i=1}^n (x_i-\overline{x})^2\right]}.$$

Recall the algebraic identity

$$\sum_{i=1}^n (x_i-\mu_0)^2 = \sum_{i=1}^n (x_i-\overline{x})^2 + n(\overline{x}-\mu_0)^2.$$

Therefore, $\lambda(\mathbf{x})$ reduces to

$$\lambda(\mathbf{x}) = e^{-\frac{n}{2}(\overline{x}-\mu_0)^2}.$$

An LRT rejects H_0 when $\lambda(\mathbf{x})$ is "too small," say, $\lambda(\mathbf{x}) \leq c$.

Goal: Write the rejection rule

$$\lambda(\mathbf{x}) \leq c$$

as a statement in terms of an easily-identified statistic. Note that

$$\begin{aligned}
\lambda(\mathbf{x}) = e^{-\frac{n}{2}(\overline{x}-\mu_0)^2} \leq c &\iff -\frac{n}{2}(\overline{x}-\mu_0)^2 \leq \ln c \\
&\iff (\overline{x}-\mu_0)^2 \geq -\frac{2\ln c}{n} \\
&\iff |\overline{x}-\mu_0| \geq \sqrt{-\frac{2\ln c}{n}} = c', \text{ say.}
\end{aligned}$$

Therefore, the LRT rejection region can be written as

$$R = \{\mathbf{x}\in\mathcal{X} : \lambda(\mathbf{x}) \leq c\} \;=\; \{\mathbf{x}\in\mathcal{X} : |\overline{x}-\mu_0| \geq c'\}.$$

Rejecting H_0 when $\lambda(\mathbf{x})$ is "too small" is the same as rejecting H_0 when $|\overline{x}-\mu_0|$ is "too large." The latter decision rule makes sense intuitively. Note that we have written our LRT rejection region and the corresponding test function

$$\phi(\mathbf{x}) = I(\mathbf{x}\in R) = I(|\overline{x}-\mu_0| \geq c') = \begin{cases} 1, & |\overline{x}-\mu_0| \geq c' \\ 0, & |\overline{x}-\mu_0| < c' \end{cases}$$

in terms of the one-dimensional statistic $W(\mathbf{X}) = \overline{X}$. Recall that $W(\mathbf{X}) = \overline{X}$ is a sufficient statistic for the $\mathcal{N}(\mu, 1)$ family.

Example 8.6. Suppose $X_1, X_2, ..., X_n$ are iid with population pdf

$$f_X(x|\theta) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta \\ 0, & x < \theta, \end{cases}$$

where $-\infty < \theta < \infty$. Note that this is a location exponential population pdf; the location parameter is θ. Consider testing

$$H_0 : \theta \leq \theta_0$$
versus
$$H_1 : \theta > \theta_0.$$

The likelihood function is

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} e^{-(x_i-\theta)} I(x_i \geq \theta) \\ &= e^{-\sum_{i=1}^{n} x_i + n\theta} I(x_{(1)} \geq \theta) \prod_{i=1}^{n} I(x_i \in \mathbb{R}) \\ &= \underbrace{e^{n\theta} I(x_{(1)} \geq \theta)}_{g(x_{(1)}|\theta)} \underbrace{e^{-\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} I(x_i \in \mathbb{R})}_{h(\mathbf{x})}. \end{aligned}$$

Note that $W(\mathbf{X}) = X_{(1)}$ is a sufficient statistic by the Factorization Theorem. The relevant parameter spaces are

$$\begin{aligned} \Theta_0 &= \{\theta : -\infty < \theta \leq \theta_0\} \\ \Theta &= \{\theta : -\infty < \theta < \infty\}. \end{aligned}$$

We need to find the unrestricted MLE

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|\mathbf{x})$$

and the restricted MLE

$$\widehat{\theta}_0 = \arg\max_{\theta \in \Theta_0} L(\theta|\mathbf{x}).$$

Unrestricted MLE: Note that

- When $\theta \leq x_{(1)}$, $L(\theta|\mathbf{x}) = e^{-\sum_{i=1}^{n} x_i + n\theta}$, which increases as θ increases.
  - For graphing purposes, it is helpful to note that

$$\frac{\partial^2}{\partial \theta^2} L(\theta|\mathbf{x}) = n^2 e^{-\sum_{i=1}^{n} x_i + n\theta} > 0,$$

    i.e., $L(\theta|\mathbf{x})$ is convex.

- When $\theta > x_{(1)}$, $L(\theta|\mathbf{x}) = 0$.
- Therefore, $L(\theta|\mathbf{x})$ is an increasing function when θ is less than or equal to the minimum order statistic $x_{(1)}$; when θ is larger than $x_{(1)}$, the likelihood function drops to zero.
- Clearly, the unrestricted MLE of θ is $\widehat{\theta} = X_{(1)}$ and hence the denominator of $\lambda(\mathbf{x})$ is

$$\sup_{\theta \in \Theta} L(\theta|\mathbf{x}) = L(\widehat{\theta}|\mathbf{x}) = L(x_{(1)}|\mathbf{x}).$$

Restricted MLE: By "restricted," we mean "subject to the constraint that the estimate fall in $\Theta_0 = \{\theta : -\infty < \theta \leq \theta_0\}$."

- **Case 1:** If $\theta_0 < x_{(1)}$, then the largest $L(\theta|\mathbf{x})$ can be is $L(\theta_0|\mathbf{x})$. Therefore, the restricted MLE is $\widehat{\theta}_0 = \theta_0$.
- **Case 2:** If $\theta_0 \geq x_{(1)}$, then the restricted MLE $\widehat{\theta}_0$ coincides with the unrestricted MLE $\widehat{\theta} = X_{(1)}$.
- Therefore,

$$\widehat{\theta}_0 = \begin{cases} \theta_0, & \theta_0 < X_{(1)} \\ X_{(1)}, & \theta_0 \geq X_{(1)}. \end{cases}$$

The LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\widehat{\theta}_0|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})} = \begin{cases} \dfrac{L(\theta_0|\mathbf{x})}{L(x_{(1)}|\mathbf{x})}, & \theta_0 < x_{(1)} \\ \dfrac{L(x_{(1)}|\mathbf{x})}{L(x_{(1)}|\mathbf{x})} = 1, & \theta_0 \geq x_{(1)}. \end{cases}$$

That $\lambda(\mathbf{x}) = 1$ when $\theta_0 \geq x_{(1)}$ makes perfect sense in testing

$$H_0 : \theta \leq \theta_0$$
$$\text{versus}$$
$$H_1 : \theta > \theta_0.$$

- If $x_{(1)} \leq \theta_0$, we certainly don't want to reject H_0 and conclude that $\theta > \theta_0$.
- It is only when $x_{(1)} > \theta_0$ do we have evidence that θ might be larger than θ_0. The larger $x_{(1)}$ is ($x_{(1)} > \theta_0$), the smaller $\lambda(\mathbf{x})$ becomes; see Figure 8.2.1 (CB, pp 377). That is,

$$\text{larger } x_{(1)} \iff \text{smaller } \lambda(\mathbf{x}) \iff \text{more evidence against } H_0.$$

Not surprisingly, we can write our LRT rejection region in terms of $W(\mathbf{X}) = X_{(1)}$. When $\theta_0 < x_{(1)}$, the LRT statistic

$$\lambda(\mathbf{x}) = \frac{L(\theta_0|\mathbf{x})}{L(x_{(1)}|\mathbf{x})} = \frac{e^{-\sum_{i=1}^n x_i + n\theta_0}}{e^{-\sum_{i=1}^n x_i + nx_{(1)}}} = e^{-n(x_{(1)}-\theta_0)}.$$

Note that

$$\begin{aligned} \lambda(\mathbf{x}) = e^{-n(x_{(1)}-\theta_0)} \leq c &\iff -n(x_{(1)} - \theta_0) \leq \ln c \\ &\iff x_{(1)} \geq \theta_0 - \frac{\ln c}{n} = c', \text{ say.} \end{aligned}$$

Therefore, the LRT rejection region can be written as

$$R = \{\mathbf{x} \in \mathcal{X} : \lambda(\mathbf{x}) \leq c\} = \{\mathbf{x} \in \mathcal{X} : x_{(1)} \geq c'\}.$$

Rejecting H_0 when $\lambda(\mathbf{x})$ is "too small" is the same as rejecting H_0 when $x_{(1)}$ is "too large." As noted earlier, the latter decision rule makes sense intuitively. Note that we have written our LRT rejection region and the corresponding test function

$$\phi(\mathbf{x}) = I(\mathbf{x} \in R) = I(x_{(1)} \geq c') = \begin{cases} 1, & x_{(1)} \geq c' \\ 0, & x_{(1)} < c' \end{cases}$$

in terms of the one-dimensional statistic $W(\mathbf{X}) = X_{(1)}$, which is sufficient for the location exponential family.

Theorem 8.2.4. Suppose $T = T(\mathbf{X})$ is a sufficient statistic for θ. If $\lambda^*(T(\mathbf{x})) = \lambda^*(t)$ is the LRT statistic based on T and if $\lambda(\mathbf{x})$ is the LRT statistic based on $\mathbf{X}$, then $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

Proof. Because $T = T(\mathbf{X})$ is sufficient, we can write (by the Factorization Theorem)

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = g_T(t|\theta)h(\mathbf{x}),$$

where $g_T(t|\theta)$ is the pdf (pmf) of T and $h(\mathbf{x})$ is free of θ. Therefore,

$$\begin{aligned} \lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Theta_0} L(\theta|\mathbf{x})}{\sup_{\theta\in\Theta} L(\theta|\mathbf{x})} &= \frac{\sup_{\theta\in\Theta_0} g_T(t|\theta)h(\mathbf{x})}{\sup_{\theta\in\Theta} g_T(t|\theta)h(\mathbf{x})} \\ &= \frac{\sup_{\theta\in\Theta_0} g_T(t|\theta)}{\sup_{\theta\in\Theta} g_T(t|\theta)} \\ &= \frac{\sup_{\theta\in\Theta_0} L^*(\theta|t)}{\sup_{\theta\in\Theta} L^*(\theta|t)}, \end{aligned}$$

where $L^*(\theta|t)$ is the likelihood function based on observing $T = t$. □

Implication: If a sufficient statistic T exists, we can immediately restrict attention to its distribution when deriving an LRT.

Example 8.7. Suppose $X_1, X_2, ..., X_n$ are iid exponential(θ), where $\theta > 0$. Consider testing

$$H_0 : \theta = \theta_0$$
versus
$$H_1 : \theta \neq \theta_0.$$

(a) Show that the LRT statistic based on $\mathbf{X} = \mathbf{x}$ is

$$\lambda(\mathbf{x}) = \left(\frac{e}{n\theta_0}\right)^n \left(\sum_{i=1}^n x_i\right)^n e^{-\sum_{i=1}^n x_i/\theta_0}.$$

(b) Show that the LRT statistic based on $T = T(\mathbf{X}) = \sum_{i=1}^n X_i$ is

$$\lambda^*(t) = \left(\frac{e}{n\theta_0}\right)^n t^n e^{-t/\theta_0},$$

establishing that $\lambda^*(t) = \lambda(\mathbf{x})$, as stated in Theorem 8.2.4.
(c) Show that

$$\lambda^*(t) \leq c \iff t \leq c_1 \text{ or } t \geq c_2,$$

for some c_1 and c_2 satisfying $c_1 < c_2$.

Example 8.8. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Set $\boldsymbol{\theta} = (\mu, \sigma^2)$. Consider testing

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu \neq \mu_0.$$

The null hypothesis H_0 above looks simple, but it is not. The relevant parameter spaces are

$$\begin{aligned} \Theta_0 &= \{\boldsymbol{\theta} = (\mu, \sigma^2) : \ \mu = \mu_0, \ \sigma^2 > 0\} \\ \Theta &= \{\boldsymbol{\theta} = (\mu, \sigma^2) : -\infty < \mu < \infty, \ \sigma^2 > 0\}. \end{aligned}$$

In this problem, we call σ^2 a **nuisance parameter**, because it is not the parameter that is of interest in H_0 and H_1. The likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}. \end{aligned}$$

Unrestricted MLE: In Example 7.6 (notes, pp 33), we showed that

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \overline{X} \\ S_b^2 \end{pmatrix} = \begin{pmatrix} \overline{X} \\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{pmatrix}$$

maximizes $L(\boldsymbol{\theta}|\mathbf{x})$ over Θ.

Restricted MLE: It is easy to show that

$$\widehat{\boldsymbol{\theta}}_0 = \begin{pmatrix} \mu_0 \\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2 \end{pmatrix}$$

maximizes $L(\boldsymbol{\theta}|\mathbf{x})$ over Θ_0.

(a) Show that

$$\lambda(\mathbf{x}) = \frac{L(\widehat{\boldsymbol{\theta}}_0|\mathbf{x})}{L(\widehat{\boldsymbol{\theta}}|\mathbf{x})} = \left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2}\right]^{n/2}.$$

(b) Show that

$$\lambda(\mathbf{x}) \leq c \iff \left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right| \geq c'.$$

This demonstrates that the "one-sample t test" is a LRT under normality.

Exercise: In Example 7.7 (notes, pp 34-35), derive the LRT statistic to test

$$H_0 : p_1 = p_2$$
versus
$$H_1 : p_1 \neq p_2.$$

Exercise: In Example 8.4 (notes, pp 67), show that the LRT statistic is

$$\lambda(\mathbf{x}) = \lambda(x_1, x_2, x_3, x_4) = \prod_{i=1}^{4}\left(\frac{2864}{4x_i}\right)^{x_i}.$$

Also, show that

$$\lambda(\mathbf{x}) \leq c \iff -2\ln\lambda(\mathbf{x}) \geq c'.$$

Under $H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$, we will learn later that $-2\ln\lambda(\mathbf{X})$ is distributed approximately as χ^2_3. This suggests a "large-sample" LRT, namely, to reject H_0 if $-2\ln\lambda(\mathbf{x})$ is "too large." We can use the χ^2_3 distribution to specify what "too large" actually means.

8.2.2 Bayesian tests

Remark: Hypothesis tests of the form

$$H_0 : \theta \in \Theta_0$$
versus
$$H_1 : \theta \in \Theta_0^c,$$

where $\Theta_0^c = \Theta \setminus \Theta_0$, can also be carried out within the Bayesian paradigm, but they are performed differently. Recall that, for a Bayesian, all inference is carried out using the posterior distribution $\pi(\theta|\mathbf{x})$.

Realization: The posterior distribution $\pi(\theta|\mathbf{x})$ is a valid probability distribution. It is the distribution that describes the behavior of the random variable θ, updated after observing the data $\mathbf{x}$. In this light, the probabilities

$$\begin{aligned} P(H_0 \text{ true}|\mathbf{x}) &= P(\theta \in \Theta_0|\mathbf{x}) = \int_{\Theta_0} \pi(\theta|\mathbf{x})d\theta \\ P(H_1 \text{ true}|\mathbf{x}) &= P(\theta \in \Theta_0^c|\mathbf{x}) = \int_{\Theta_0^c} \pi(\theta|\mathbf{x})d\theta \end{aligned}$$

make perfect sense and be calculated (or approximated) "exactly." Note that these probabilities make no sense to the non-Bayesian. S/he regards θ as fixed, so that $\{\theta \in \Theta_0\}$ and $\{\theta \in \Theta_0^c\}$ are not random events. We do not assign probabilities to events that are not random.

Example 8.9. Suppose that $X_1, X_2, ..., X_n$ are iid Poisson(θ), where the prior distribution for $\theta \sim$ gamma(a, b), a, b known. In Example 7.10 (notes, pp 38-39), we showed that the posterior distribution

$$\theta|\mathbf{X} = \mathbf{x} \sim \text{gamma}\left(\sum_{i=1}^{n} x_i + a,\ \frac{1}{n + \frac{1}{b}}\right).$$

As an application, consider the following data, which summarize the number of goals per game in the 2013-2014 English Premier League season:

Goals	0	1	2	3	4	5	6	7	8	9	10+
Frequency	27	73	80	72	65	39	17	4	1	2	0

There were $n = 380$ games total. I modeled the number of goals per game X as a Poisson random variable and assumed that $X_1, X_2, ..., X_{380}$ are iid Poisson(θ). Before the season started, I modeled the mean number of goals per game as $\theta \sim$ gamma$(1.5, 2)$, which is a fairly diffuse prior distribution.

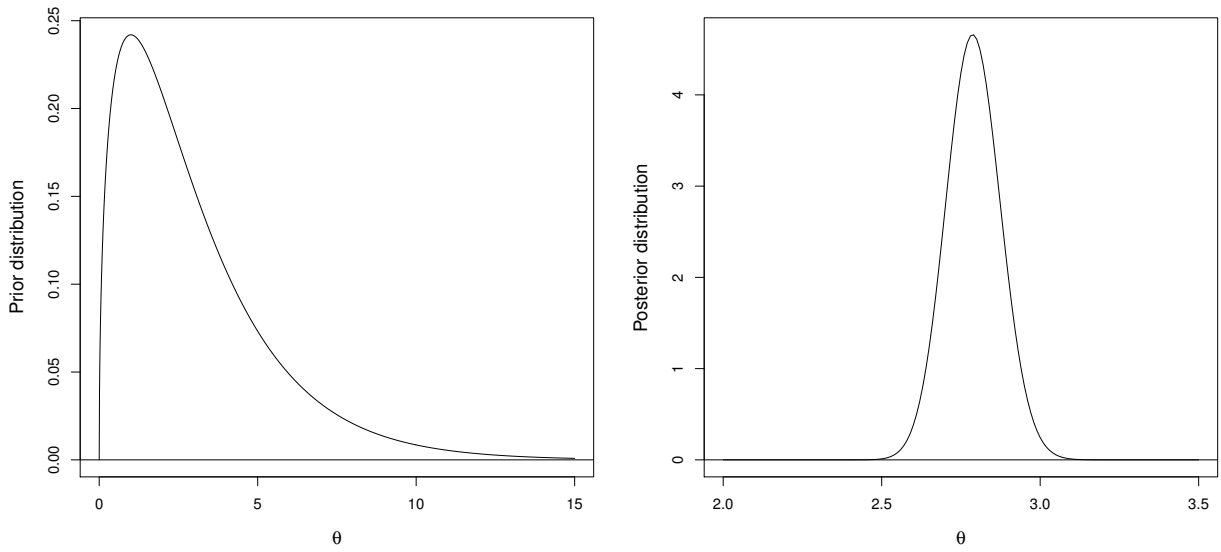

Figure 8.1: 2013-2014 English Premier League data. Prior distribution (left) and posterior distribution (right) for θ, the mean number of goals scored per game. Note that the horizontal axes are different in the two figures.

Based on the observed data, I used R to calculate

```
> sum(goals)
[1] 1060
```

The posterior distribution is therefore

$$\theta|\mathbf{X} = \mathbf{x} \sim \text{gamma}\left(1060 + 1.5,\ \frac{1}{380 + \frac{1}{2}}\right) \stackrel{d}{=} \text{gamma}(1061.5, 0.002628).$$

I have depicted the prior distribution $\pi(\theta)$ and the posterior distribution $\pi(\theta|\mathbf{x})$ in Figure 8.1. Suppose that I wanted to test $H_0 : \theta \geq 3$ versus $H_1 : \theta < 3$ on the basis of the assumed Bayesian model and the observed data $\mathbf{x}$. The probability that H_0 is true is

$$P(\theta \geq 3|\mathbf{x}) = \int_3^\infty \pi(\theta|\mathbf{x})d\theta \approx 0.008,$$

which I calculated in R using

```
> 1-pgamma(3,1061.5,1/0.002628)
[1] 0.008019202
```

Therefore, it is far more likely that H_1 is true, in fact, with probability over 0.99.

8.3 Methods of Evaluating Tests

Setting: Suppose $\mathbf{X} = (X_1, X_2, ..., X_n) \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$ and consider testing

$$H_0 : \theta \in \Theta_0$$
versus
$$H_1 : \theta \in \Theta_0^c,$$

where $\Theta_0^c = \Theta \setminus \Theta_0$. I will henceforth assume that θ is a scalar parameter (for simplicity only).

8.3.1 Error probabilities and the power function

Definition: For a test (with test function)

$$\phi(\mathbf{x}) = I(\mathbf{x} \in R),$$

we can make one of two mistakes:

1. Type I Error: Rejecting H_0 when H_0 is true
2. Type II Error: Not rejecting H_0 when H_1 is true.

Therefore, for any test that we perform, there are four possible scenarios, described in the following table:

		Decision	
		Reject H_0	Do not reject H_0
Truth	H_0	Type I Error	☺
	H_1	☺	Type II Error

Calculations:

1. Suppose $H_0 : \theta \in \Theta_0$ is true. For $\theta \in \Theta_0$,

$$P(\text{Type I Error}|\theta) = P_\theta(\mathbf{X} \in R) = E_\theta[I(\mathbf{X} \in R)] = E_\theta[\phi(\mathbf{X})].$$

2. Suppose $H_1 : \theta \in \Theta_0^c$ is true. For $\theta \in \Theta_0^c$,

$$P(\text{Type II Error}|\theta) = P_\theta(\mathbf{X} \in R^c) = 1 - P_\theta(\mathbf{X} \in R) = 1 - E_\theta[\phi(\mathbf{X})] = E_\theta[1 - \phi(\mathbf{X})].$$

It is very important to note that both of these probabilities depend on θ. This is why we emphasize this in the notation.

Definition: The **power function** of a test $\phi(\mathbf{x})$ is the function of θ given by

$$\beta(\theta) = P_\theta(\mathbf{X} \in R) = E_\theta[\phi(\mathbf{X})].$$

In other words, the power function gives the probability of rejecting H_0 **for all** $\theta \in \Theta$. Note that if H_1 is true, so that $\theta \in \Theta_0^c$,

$$\beta(\theta) = P_\theta(\mathbf{X} \in R) = 1 - P_\theta(\mathbf{X} \in R^c) = 1 - P(\text{Type II Error}|\theta).$$

Example 8.10. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and σ_0^2 is known. Consider testing

$$H_0 : \mu \leq \mu_0$$
$$\text{versus}$$
$$H_1 : \mu > \mu_0.$$

The LRT of H_0 versus H_1 uses the test function

$$\phi(\mathbf{x}) = \begin{cases} 1, & \dfrac{\overline{x} - \mu_0}{\sigma_0/\sqrt{n}} \geq c \\ 0, & \text{otherwise.} \end{cases}$$

The power function for this test is given by

$$\begin{aligned} \beta(\mu) = P_\mu(\mathbf{X} \in R) &= P_\mu\left(\frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \geq c\right) \\ &= P_\mu\left(\overline{X} \geq \frac{c\sigma_0}{\sqrt{n}} + \mu_0\right) \\ &= P_\mu\left(\frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} \geq \frac{\frac{c\sigma_0}{\sqrt{n}} + \mu_0 - \mu}{\sigma_0/\sqrt{n}}\right) = 1 - F_Z\left(c + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right), \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ and $F_Z(\cdot)$ is the standard normal cdf.

Exercise: Determine n and c such that

$$\begin{aligned} \sup_{\mu \leq \mu_0} \beta(\mu) &= 0.10 \\ \inf_{\mu \geq \mu_0 + \sigma_0} \beta(\mu) &= 0.80. \end{aligned}$$

- The first requirement implies that $P(\text{Type I Error}|\mu)$ will not exceed 0.10 for all $\mu \leq \mu_0$ (H_0 true).
- The second requirement implies that $P(\text{Type II Error}|\mu)$ will not exceed 0.20 for all $\mu \geq \mu_0 + \sigma_0$ (these are values of μ that make H_1 true).

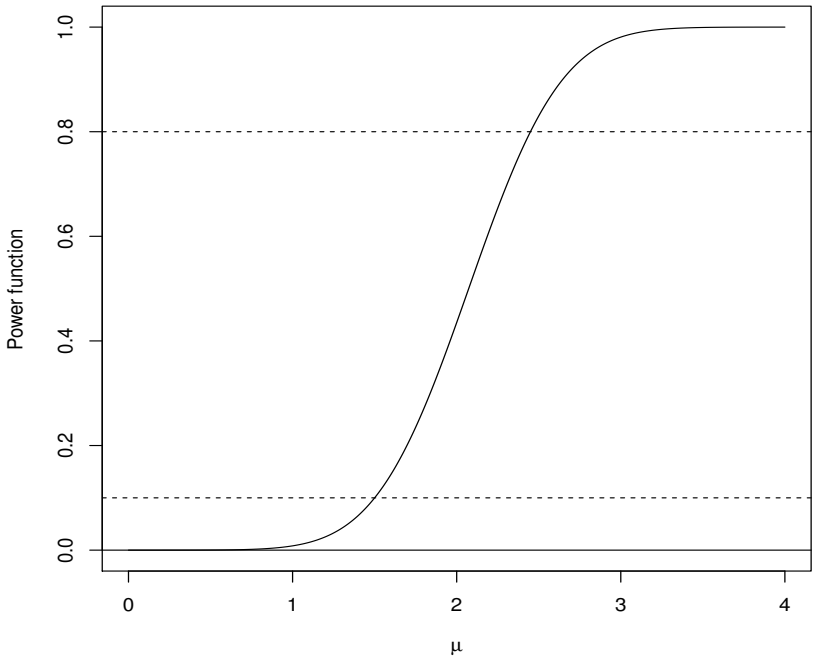

Figure 8.2: Power function $\beta(\mu)$ in Example 8.10 with $c = 1.28$, $n = 5$, $\mu_0 = 1.5$ and $\sigma_0 = 1$. Horizontal lines at 0.10 and 0.80 have been added.

Solution. Note that

$$\begin{aligned}\frac{\partial}{\partial\mu}\beta(\mu) &= \frac{\partial}{\partial\mu}\left[1 - F_Z\left(c + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right)\right] \\ &= \frac{\sqrt{n}}{\sigma_0} f_Z\left(c + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right) > 0;\end{aligned}$$

i.e., $\beta(\mu)$ is an **increasing** function of μ. Therefore,

$$\sup_{\mu \leq \mu_0} \beta(\mu) = \beta(\mu_0) = 1 - F_Z(c) \stackrel{\text{set}}{=} 0.10 \implies c = 1.28,$$

the 0.90 quantile of the $\mathcal{N}(0,1)$ distribution. Also, because $\beta(\mu)$ is increasing,

$$\begin{aligned}\inf_{\mu \geq \mu_0 + \sigma_0} \beta(\mu) = \beta(\mu_0 + \sigma_0) &= 1 - F_Z(1.28 - \sqrt{n}) \stackrel{\text{set}}{=} 0.80 \\ &\implies 1.28 - \sqrt{n} = -0.84 \\ &\implies n = 4.49,\end{aligned}$$

which would be rounded up to $n = 5$. The resulting power function with $c = 1.28$, $n = 5$, $\mu_0 = 1.5$ and $\sigma_0 = 1$ is shown in Figure 8.2.

Definition: A test $\phi(\mathbf{x})$ with power function $\beta(\theta)$ is a **size** α test if

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

The test $\phi(\mathbf{x})$ is a **level** α test if

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

Note that if $\phi(\mathbf{x})$ is a size α test, then it is also level α. The converse is not true. In other words,

$$\{\text{class of size } \alpha \text{ tests}\} \subset \{\text{class of level } \alpha \text{ tests}\}.$$

Remark: Often, it is unnecessary to differentiate between the two classes of tests. However, in testing problems involving discrete distributions (e.g., binomial, Poisson, etc.), it is generally not possible to construct a size α test for a specified value of α; e.g., $\alpha = 0.05$. Thus (unless one randomizes), we may have to settle for a level α test.

Important: As the definition above indicates, the **size** of any test $\phi(\mathbf{x})$ is calculated by maximizing the power function over the null parameter space Θ_0 identified in H_0.

Example 8.11. Suppose X_1, X_2 are iid Poisson(θ), where $\theta > 0$, and consider testing

$$H_0 : \theta \geq 3$$
$$\text{versus}$$
$$H_1 : \theta < 3.$$

We consider the two tests

$$\begin{aligned} \phi_1 = \phi_1(x_1, x_2) &= I(x_1 = 0) \\ \phi_2 = \phi_2(x_1, x_2) &= I(x_1 + x_2 \leq 1). \end{aligned}$$

The power function for the first test is

$$\beta_1(\theta) = E_\theta[I(X_1 = 0)] = P_\theta(X_1 = 0) = e^{-\theta}.$$

Recall that $T = T(X_1, X_2) = X_1 + X_2 \sim$ Poisson(2θ). The power function for the second test is

$$\beta_2(\theta) = E_\theta[I(X_1 + X_2 \leq 1)] = P_\theta(X_1 + X_2 \leq 1) = e^{-2\theta} + 2\theta e^{-2\theta}.$$

I have plotted both power functions in Figure 8.3 (next page).

Size calculations: The size of each test is calculated as follows. For the first test,

$$\alpha = \sup_{\theta \geq 3} \beta_1(\theta) = \beta_1(3) = e^{-3} \approx 0.049787.$$

For the second test,

$$\alpha = \sup_{\theta \geq 3} \beta_2(\theta) = \beta_2(3) = e^{-6} + 6e^{-6} \approx 0.017351.$$

Both ϕ_1 and ϕ_2 are level $\alpha = 0.05$ tests.

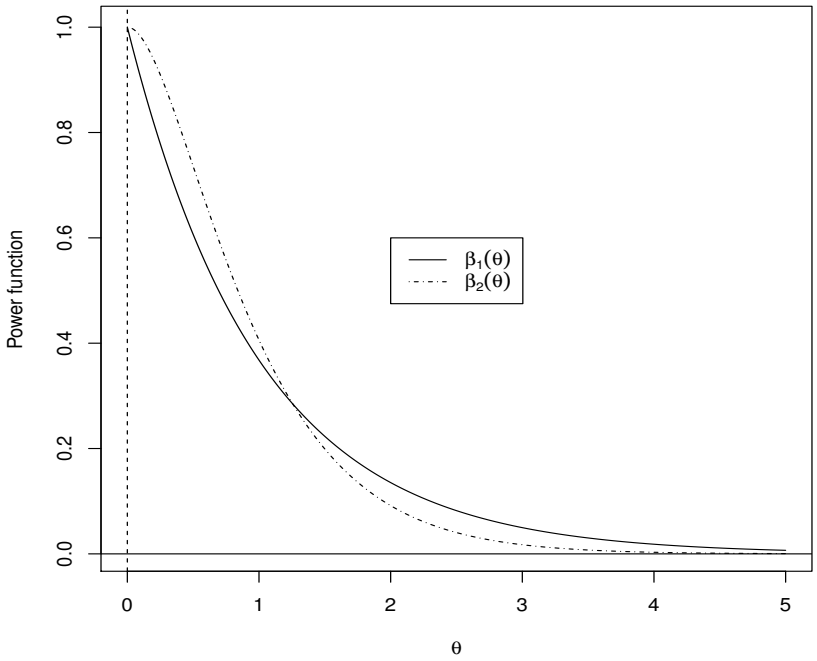

Figure 8.3: Power functions $\beta_1(\theta)$ and $\beta_2(\theta)$ in Example 8.11.

Example 8.12. Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta) = e^{-(x-\theta)}I(x \geq \theta)$, where $-\infty < \theta < \infty$. In Example 8.6 (notes, pp 72-74), we considered testing

$$H_0 : \theta \leq \theta_0$$
versus
$$H_1 : \theta > \theta_0$$

and derived the LRT to take the form $\phi(\mathbf{x}) = I(x_{(1)} \geq c')$. Find the value of c' that makes $\phi(\mathbf{x})$ a size α test.

Solution. The pdf of $X_{(1)}$ is $f_{X_{(1)}}(x|\theta) = ne^{-n(x-\theta)}I(x \geq \theta)$. We set

$$\begin{aligned}
\alpha = \sup_{\theta \leq \theta_0} E_\theta[\phi(\mathbf{X})] &= \sup_{\theta \leq \theta_0} P_\theta(X_{(1)} \geq c') \\
&= \sup_{\theta \leq \theta_0} \int_{c'}^{\infty} ne^{-n(x-\theta)}dx \\
&= \sup_{\theta \leq \theta_0} e^{-n(c'-\theta)} = e^{-n(c'-\theta_0)}.
\end{aligned}$$

Therefore, $c' = \theta_0 - n^{-1}\ln\alpha$. A size α LRT uses $\phi(\mathbf{x}) = I(x_{(1)} \geq \theta_0 - n^{-1}\ln\alpha)$.

8.3.2 Most powerful tests

Definition: Let $\mathcal{C}$ be a class of tests for testing

$$H_0 : \theta \in \Theta_0$$
versus
$$H_1 : \theta \in \Theta_0^c,$$

where $\Theta_0^c = \Theta \setminus \Theta_0$. A test in $\mathcal{C}$ with power function $\beta(\theta)$ is a **uniformly most powerful (UMP) class $\mathcal{C}$ test** if

$$\beta(\theta) \geq \beta^*(\theta) \text{ for all } \theta \in \Theta_0^c,$$

where $\beta^*(\theta)$ is the power function of any other test in $\mathcal{C}$. The "uniformly" part in this definition refers to the fact that the power function $\beta(\theta)$ is larger than (i.e., at least as large as) the power function of any other class $\mathcal{C}$ test **for all** $\theta \in \Theta_0^c$.

Important: In this course, we will restrict attention to tests $\phi(\mathbf{x})$ that are level α tests. That is, we will take

$$\mathcal{C} = \{\text{all level } \alpha \text{ tests}\}.$$

This restriction is analogous to the restriction we made in the "optimal estimation problem" in Chapter 7. Recall that we restricted attention to unbiased estimators first; we then wanted to find the one with the smallest variance (uniformly, for all $\theta \in \Theta$). In the same spirit, we make the same type of restriction here by considering only those tests that are level α tests. This is done so that we can avoid having to consider "silly tests," e.g.,

$$\phi(\mathbf{x}) = 1 \text{ for all } \mathbf{x} \in \mathcal{X}.$$

The power function for this test is $\beta(\theta) = 1$, for all $\theta \in \Theta$. This test cannot be beaten in terms of power when H_1 is true! Unfortunately, it is not a very good test when H_0 is true.

Recall: A test $\phi(\mathbf{x})$ with power function $\beta(\theta)$ is a **level** α test if

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

That is, $P(\text{Type I Error}|\theta)$ can be **no larger** than α for all $\theta \in \Theta_0$.

Starting point: We start by considering the **simple-versus-simple** test:

$$H_0 : \theta = \theta_0$$
versus
$$H_1 : \theta = \theta_1.$$

Both H_0 and H_1 specify exactly one probability distribution.

Remark: This type of test is rarely of interest in practice. However, it is the "building block" situation for more interesting problems.

Theorem 8.3.12 (**Neyman-Pearson Lemma**). Consider testing

$$H_0 : \theta = \theta_0$$
versus
$$H_1 : \theta = \theta_1$$

and denote by $f_{\mathbf{X}}(\mathbf{x}|\theta_0)$ and $f_{\mathbf{X}}(\mathbf{x}|\theta_1)$ the pdfs (pmfs) of $\mathbf{X} = (X_1, X_2, ..., X_n)$ corresponding to θ_0 and θ_1, respectively. Consider the test function

$$\phi(\mathbf{x}) = \begin{cases} 1, & \dfrac{f_{\mathbf{X}}(\mathbf{x}|\theta_1)}{f_{\mathbf{X}}(\mathbf{x}|\theta_0)} > k \\ \\ 0, & \dfrac{f_{\mathbf{X}}(\mathbf{x}|\theta_1)}{f_{\mathbf{X}}(\mathbf{x}|\theta_0)} < k, \end{cases}$$

for $k \geq 0$, where

$$\alpha = P_{\theta_0}(\mathbf{X} \in R) = E_{\theta_0}[\phi(\mathbf{X})]. \tag{8.1}$$

Sufficiency: Any test satisfying the definition of $\phi(\mathbf{x})$ above and Equation (8.1) is a **most powerful** (**MP**) **level** α **test**.

Remarks:

- The necessity part of the Neyman-Pearson (NP) Lemma is less important for our immediate purposes (see CB, pp 388).
- In a simple-versus-simple test, any MP level α test is obviously also UMP level α. Recall that the "uniformly" part in UMP refers to all $\theta \in \Theta_0^c$. However, in a simple H_1, there is only one value of $\theta \in \Theta_0^c$. I choose to distinguish MP from UMP in this situation (whereas the authors of CB do not).

Example 8.13. Suppose that $X_1, X_2, ..., X_n$ are iid beta$(\theta, 1)$, where $\theta > 0$; i.e., the population pdf is

$$f_X(x|\theta) = \theta x^{\theta-1} I(0 < x < 1).$$

Derive the MP level α test for

$$H_0 : \theta = 1$$
versus
$$H_1 : \theta = 2.$$

Solution. The pdf of $\mathbf{X} = (X_1, X_2, ..., X_n)$ is, for $0 < x_i < 1$,

$$f_{\mathbf{X}}(\mathbf{x}|\theta) \overset{\text{iid}}{=} \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1}.$$

Form the ratio

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\theta_1)}{f_{\mathbf{X}}(\mathbf{x}|\theta_0)} = \frac{f_{\mathbf{X}}(\mathbf{x}|2)}{f_{\mathbf{X}}(\mathbf{x}|1)} = \frac{2^n \left(\prod_{i=1}^n x_i\right)^{2-1}}{1^n \left(\prod_{i=1}^n x_i\right)^{1-1}} = 2^n \prod_{i=1}^n x_i.$$

The NP Lemma says that the MP level α test uses the rejection rejection

$$R = \left\{ \mathbf{x} \in \mathcal{X} : \ 2^n \prod_{i=1}^n x_i > k \right\},$$

where the constant k satisfies

$$\alpha = P_{\theta=1}(\mathbf{X} \in R) = P\left(2^n \prod_{i=1}^n X_i > k \ \Big| \ \theta = 1 \right).$$

Instead of finding the constant k that satisfies this equation, we rewrite the rejection rule $\{2^n \prod_{i=1}^n x_i > k\}$ in a way that makes our life easier. Note that

$$\begin{aligned} 2^n \prod_{i=1}^n x_i > k \ &\iff \ \prod_{i=1}^n x_i > 2^{-n}k \\ &\iff \ \sum_{i=1}^n -\ln x_i < -\ln(2^{-n}k) = k', \text{ say.} \end{aligned}$$

We have rewritten the rejection rule $\{2^n \prod_{i=1}^n x_i > k\}$ as $\{\sum_{i=1}^n -\ln x_i < k'\}$. Therefore,

$$\alpha = P\left(2^n \prod_{i=1}^n X_i > k \ \Big| \ \theta = 1 \right) = P\left(\sum_{i=1}^n -\ln X_i < k' \ \Big| \ \theta = 1 \right).$$

We have now changed the problem to choosing k' to solve this equation above.

Q: Why did we do this?
A: Because it is easier to find the distribution of $\sum_{i=1}^n -\ln X_i$ when $H_0 : \theta = 1$ is true.

Recall that

$$\begin{aligned} X_i \overset{H_0}{\sim} \mathcal{U}(0,1) \ &\implies \ -\ln X_i \overset{H_0}{\sim} \text{ exponential}(1) \\ &\implies \ \sum_{i=1}^n -\ln X_i \overset{H_0}{\sim} \text{ gamma}(n,1). \end{aligned}$$

Therefore, to satisfy the equation above, we take $k' = g_{n,1,1-\alpha}$, the (lower) α quantile of a gamma$(n,1)$ distribution. This notation for quantiles is consistent with how CB have defined them on pp 386. Thus, the MP level α test of $H_0 : \theta = 1$ versus $H_1 : \theta = 2$ has rejection region

$$R = \left\{ \mathbf{x} \in \mathcal{X} : \sum_{i=1}^n -\ln x_i < g_{n,1,1-\alpha} \right\}.$$

Special case: If $n = 10$ and $\alpha = 0.05$, then $g_{10,1,0.95} \approx 5.425$.

Q: What is $\beta(2)$, the **power** of this MP test (when $n = 10$ and $\alpha = 0.05$)?
A: We calculate

$$\beta(2) = P\left(\sum_{i=1}^{10} -\ln X_i < 5.425 \;\middle|\; \theta = 2\right).$$

Recall that

$$\begin{aligned} X_i \overset{H_1}{\sim} \text{ beta}(2,1) &\implies -\ln X_i \overset{H_1}{\sim} \text{ exponential}(1/2) \\ &\implies \sum_{i=1}^{10} -\ln X_i \overset{H_1}{\sim} \text{ gamma}(10, 1/2). \end{aligned}$$

Therefore,

$$\beta(2) = \int_0^{5.425} \underbrace{\frac{1}{\Gamma(10)\left(\frac{1}{2}\right)^{10}} u^9 e^{-2u}}_{\text{gamma}(10,1/2) \text{ pdf}} du \approx 0.643.$$

Proof of NP Lemma. We prove the sufficiency part only. Define the test function

$$\phi(\mathbf{x}) = \begin{cases} 1, & \dfrac{f_{\mathbf{X}}(\mathbf{x}|\theta_1)}{f_{\mathbf{X}}(\mathbf{x}|\theta_0)} > k \\ 0, & \dfrac{f_{\mathbf{X}}(\mathbf{x}|\theta_1)}{f_{\mathbf{X}}(\mathbf{x}|\theta_0)} < k, \end{cases}$$

where $k \geq 0$ and

$$\alpha = P_{\theta_0}(\mathbf{X} \in R) = E_{\theta_0}[\phi(\mathbf{X})];$$

i.e., $\phi(\mathbf{x})$ is a size α test. We want to show that $\phi(\mathbf{x})$ is MP level α. Therefore, let $\phi^*(\mathbf{x})$ be the test function for any other level α test of H_0 versus H_1. Note that

$$\begin{aligned} E_{\theta_0}[\phi(\mathbf{X})] &= \alpha \\ E_{\theta_0}[\phi^*(\mathbf{X})] &\leq \alpha. \end{aligned}$$

Thus,

$$E_{\theta_0}[\phi(\mathbf{X}) - \phi^*(\mathbf{X})] = \underbrace{E_{\theta_0}[\phi(\mathbf{X})]}_{=\alpha} - \underbrace{E_{\theta_0}[\phi^*(\mathbf{X})]}_{\leq \alpha} \geq 0.$$

Define

$$b(\mathbf{x}) = [\phi(\mathbf{x}) - \phi^*(\mathbf{x})][f_{\mathbf{X}}(\mathbf{x}|\theta_1) - k f_{\mathbf{X}}(\mathbf{x}|\theta_0)].$$

We want to show that $b(\mathbf{x}) \geq 0$, for all $\mathbf{x} \in \mathcal{X}$.

- **Case 1:** Suppose $f_{\mathbf{X}}(\mathbf{x}|\theta_1) - k f_{\mathbf{X}}(\mathbf{x}|\theta_0) > 0$. Then, by definition, $\phi(\mathbf{x}) = 1$. Because $0 \leq \phi^*(\mathbf{x}) \leq 1$, we have

$$b(\mathbf{x}) = \underbrace{[\phi(\mathbf{x}) - \phi^*(\mathbf{x})]}_{\geq 0} \underbrace{[f_{\mathbf{X}}(\mathbf{x}|\theta_1) - k f_{\mathbf{X}}(\mathbf{x}|\theta_0)]}_{> 0} \geq 0.$$

- **Case 2:** Suppose $f_{\mathbf{X}}(\mathbf{x}|\theta_1) - kf_{\mathbf{X}}(\mathbf{x}|\theta_0) < 0$. Then, by definition, $\phi(\mathbf{x}) = 0$. Because $0 \leq \phi^*(\mathbf{x}) \leq 1$, we have

$$b(\mathbf{x}) = \underbrace{[\phi(\mathbf{x}) - \phi^*(\mathbf{x})]}_{\leq\ 0}\underbrace{[f_{\mathbf{X}}(\mathbf{x}|\theta_1) - kf_{\mathbf{X}}(\mathbf{x}|\theta_0)]}_{<\ 0} \geq 0.$$

- **Case 3:** Suppose $f_{\mathbf{X}}(\mathbf{x}|\theta_1) - kf_{\mathbf{X}}(\mathbf{x}|\theta_0) = 0$. It is then obvious that $b(\mathbf{x}) = 0$.

We have shown that $b(\mathbf{x}) = [\phi(\mathbf{x}) - \phi^*(\mathbf{x})][f_{\mathbf{X}}(\mathbf{x}|\theta_1) - kf_{\mathbf{X}}(\mathbf{x}|\theta_0)] \geq 0$. Therefore,

$$\begin{aligned}&[\phi(\mathbf{x}) - \phi^*(\mathbf{x})]f_{\mathbf{X}}(\mathbf{x}|\theta_1) - k[\phi(\mathbf{x}) - \phi^*(\mathbf{x})]f_{\mathbf{X}}(\mathbf{x}|\theta_0) \geq 0 \\ &\qquad\qquad \iff [\phi(\mathbf{x}) - \phi^*(\mathbf{x})]f_{\mathbf{X}}(\mathbf{x}|\theta_1) \geq k[\phi(\mathbf{x}) - \phi^*(\mathbf{x})]f_{\mathbf{X}}(\mathbf{x}|\theta_0).\end{aligned}$$

Integrating both sides, we get

$$\int_{\mathbb{R}^n} [\phi(\mathbf{x}) - \phi^*(\mathbf{x})]f_{\mathbf{X}}(\mathbf{x}|\theta_1)d\mathbf{x} \geq k\int_{\mathbb{R}^n} [\phi(\mathbf{x}) - \phi^*(\mathbf{x})]f_{\mathbf{X}}(\mathbf{x}|\theta_0)d\mathbf{x},$$

that is,

$$E_{\theta_1}[\phi(\mathbf{X}) - \phi^*(\mathbf{X})] \geq k\underbrace{E_{\theta_0}[\phi(\mathbf{X}) - \phi^*(\mathbf{X})]}_{\geq\ 0,\ \text{shown above}} \geq 0.$$

Therefore, $E_{\theta_1}[\phi(\mathbf{X}) - \phi^*(\mathbf{X})] \geq 0$ and hence $E_{\theta_1}[\phi(\mathbf{X})] \geq E_{\theta_1}[\phi^*(\mathbf{X})]$. This shows that $\phi(\mathbf{x})$ is more powerful than $\phi^*(\mathbf{x})$. Because $\phi^*(\mathbf{x})$ is an arbitrary level α test, we are done. □

Corollary 8.3.13 (NP Lemma with a sufficient statistic T). Consider testing

$$\begin{gathered}H_0 : \theta = \theta_0 \\ \text{versus} \\ H_1 : \theta = \theta_1,\end{gathered}$$

and suppose that $T = T(\mathbf{X})$ is a sufficient statistic. Denote by $g_T(t|\theta_0)$ and $g_T(t|\theta_1)$ the pdfs (pmfs) of T corresponding to θ_0 and θ_1, respectively. Consider the test function

$$\phi(t) = \begin{cases} 1, & \dfrac{g_T(t|\theta_1)}{g_T(t|\theta_0)} > k \\ 0, & \dfrac{g_T(t|\theta_1)}{g_T(t|\theta_0)} < k, \end{cases}$$

for $k \geq 0$, where, with rejection region $S \subset \mathcal{T}$,

$$\alpha = P_{\theta_0}(T \in S) = E_{\theta_0}[\phi(T)].$$

The test that satisfies these specifications is a MP level α test.
Proof. See CB (pp 390).

Implication: In search of a MP test, we can immediately restrict attention to those tests based on a sufficient statistic.

Example 8.14. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and σ_0^2 is known. Find the MP level α test for

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu = \mu_1,$$

where $\mu_1 < \mu_0$.
Solution. The sample mean $T = T(\mathbf{X}) = \overline{X}$ is a sufficient statistic for the $\mathcal{N}(\mu, \sigma_0^2)$ family. Furthermore,

$$T \sim \mathcal{N}\left(\mu, \frac{\sigma_0^2}{n}\right) \quad \Longrightarrow \quad g_T(t|\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2/n}}\, e^{-\frac{n}{2\sigma_0^2}(t-\mu)^2},$$

for $t \in \mathbb{R}$. Form the ratio

$$\frac{g_T(t|\mu_1)}{g_T(t|\mu_0)} = \frac{\frac{1}{\sqrt{2\pi\sigma_0^2/n}}\, e^{-\frac{n}{2\sigma_0^2}(t-\mu_1)^2}}{\frac{1}{\sqrt{2\pi\sigma_0^2/n}}\, e^{-\frac{n}{2\sigma_0^2}(t-\mu_0)^2}} = e^{-\frac{n}{2\sigma_0^2}[(t-\mu_1)^2-(t-\mu_0)^2]}.$$

Corollary 8.3.13 says that the MP level α test rejects H_0 when

$$e^{-\frac{n}{2\sigma_0^2}[(t-\mu_1)^2-(t-\mu_0)^2]} > k \quad \Longleftrightarrow \quad t < \frac{2\sigma_0^2 n^{-1}\ln k - (\mu_1^2 - \mu_0^2)}{2(\mu_0 - \mu_1)} = k', \text{ say.}$$

Therefore, the MP level α test uses the rejection region

$$S = \left\{t \in \mathcal{T} : \frac{g_T(t|\theta_1)}{g_T(t|\theta_0)} > k\right\} = \{t \in \mathcal{T} : t < k'\},$$

where k' satisfies

$$\begin{aligned} \alpha = P_{\mu_0}(T < k') &= P\left(Z < \frac{k' - \mu_0}{\sigma_0/\sqrt{n}}\right) \\ &\Longrightarrow \frac{k' - \mu_0}{\sigma_0/\sqrt{n}} = -z_\alpha \\ &\Longrightarrow k' = \mu_0 - z_\alpha\sigma_0/\sqrt{n}. \end{aligned}$$

Therefore, the MP level α test rejects H_0 when $\overline{X} < \mu_0 - z_\alpha\sigma_0/\sqrt{n}$. This is the same test we would have gotten using $f_{\mathbf{X}}(\mathbf{x}|\mu_0)$ and $f_{\mathbf{X}}(\mathbf{x}|\mu_1)$ with the original version of the NP Lemma (Theorem 8.3.12).

8.3.3 Uniformly most powerful tests

Remark: So far, we have discussed "test related optimality" in the context of simple-versus-simple hypotheses. We now extend the idea of "most powerful" to more realistic situations involving composite hypotheses; e.g., $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

Definition: A family of pdfs (pmfs) $\{g_T(t|\theta); \theta \in \Theta\}$ for a univariate random variable T has **monotone likelihood ratio** (**MLR**) if for all $\theta_2 > \theta_1$, the ratio

$$\frac{g_T(t|\theta_2)}{g_T(t|\theta_1)}$$

is a nondecreasing function of t over the set $\{t : g_T(t|\theta_1) > 0 \text{ or } g_T(t|\theta_2) > 0\}$.

Example 8.15. Suppose $T \sim b(n, \theta)$, where $0 < \theta < 1$. The pmf of T is

$$g_T(t|\theta) = \binom{n}{t}\theta^t(1-\theta)^{n-t},$$

for $t = 0, 1, 2, ..., n$. Suppose $\theta_2 > \theta_1$. Consider

$$\frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} = \frac{\binom{n}{t}\theta_2^t(1-\theta_2)^{n-t}}{\binom{n}{t}\theta_1^t(1-\theta_1)^{n-t}} = \left(\frac{1-\theta_2}{1-\theta_1}\right)^n \left[\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}\right]^t.$$

Note that $\left(\frac{1-\theta_2}{1-\theta_1}\right)^n > 0$ and is free of t. Also, because $\theta_2 > \theta_1$, both

$$\frac{\theta_2}{\theta_1} > 1 \quad \text{and} \quad \frac{1-\theta_1}{1-\theta_2} > 1.$$

Therefore,

$$\frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} = \underbrace{c(\theta_1, \theta_2)}_{>0} a^t,$$

where $a > 1$. This is an increasing function of t over $\{t : t = 0, 1, 2, ..., n\}$. Therefore, the family $\{g_T(t|\theta) : 0 < \theta < 1\}$ has MLR.

Remark: Many common families of pdfs (pmfs) have MLR. For example, if

$$T \sim g_T(t|\theta) = h(t)c(\theta)e^{w(\theta)t},$$

i.e., T has pdf (pmf) in the one-parameter exponential family, then $\{g_T(t|\theta); \theta \in \Theta\}$ has MLR if $w(\theta)$ is a nondecreasing function of θ.
Proof. Exercise.

Q: Why is MLR useful?
A: It makes getting UMP tests easy.

Theorem 8.3.17 (**Karlin-Rubin**). Consider testing

$$H_0 : \theta \leq \theta_0$$
versus
$$H_1 : \theta > \theta_0.$$

Suppose that T is sufficient. Suppose that $\{g_T(t|\theta); \theta \in \Theta\}$ has MLR. The test that rejects H_0 iff $T > t_0$ is a UMP level α test, where

$$\alpha = P_{\theta_0}(T > t_0).$$

Similarly, when testing

$$H_0 : \theta \geq \theta_0$$
versus
$$H_1 : \theta < \theta_0,$$

the test that rejects H_0 iff $T < t_0$ is UMP level α, where $\alpha = P_{\theta_0}(T < t_0)$.

Example 8.16. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(θ), where $0 < \theta < 1$, and consider testing

$$H_0 : \theta \leq \theta_0$$
versus
$$H_1 : \theta > \theta_0.$$

We know that

$$T = \sum_{i=1}^{n} X_i$$

is a sufficient statistic and $T \sim b(n, \theta)$. In Example 8.15, we showed that the family $\{g_T(t|\theta) : 0 < \theta < 1\}$ has MLR. Therefore, the Karlin-Rubin Theorem says that the UMP level α test is

$$\phi(t) = I(t > t_0),$$

where t_0 solves

$$\alpha = P_{\theta_0}(T > t_0) = \sum_{t=\lfloor t_0 \rfloor + 1}^{n} \binom{n}{t} \theta_0^t (1 - \theta_0)^{n-t}.$$

Special case: I took $n = 30$ and $\theta_0 = 0.2$. I used R to calculate the following:

t_0	$P_{\theta_0}(T \geq \lfloor t_0 \rfloor + 1)$
$7 \leq t_0 < 8$	$P(T \geq 8 \mid \theta = 0.2) = 0.2392$
$8 \leq t_0 < 9$	$P(T \geq 9 \mid \theta = 0.2) = 0.1287$
$9 \leq t_0 < 10$	$P(T \geq 10 \mid \theta = 0.2) = 0.0611$
$10 \leq t_0 < 11$	$P(T \geq 11 \mid \theta = 0.2) = 0.0256$
$11 \leq t_0 < 12$	$P(T \geq 12 \mid \theta = 0.2) = 0.0095$

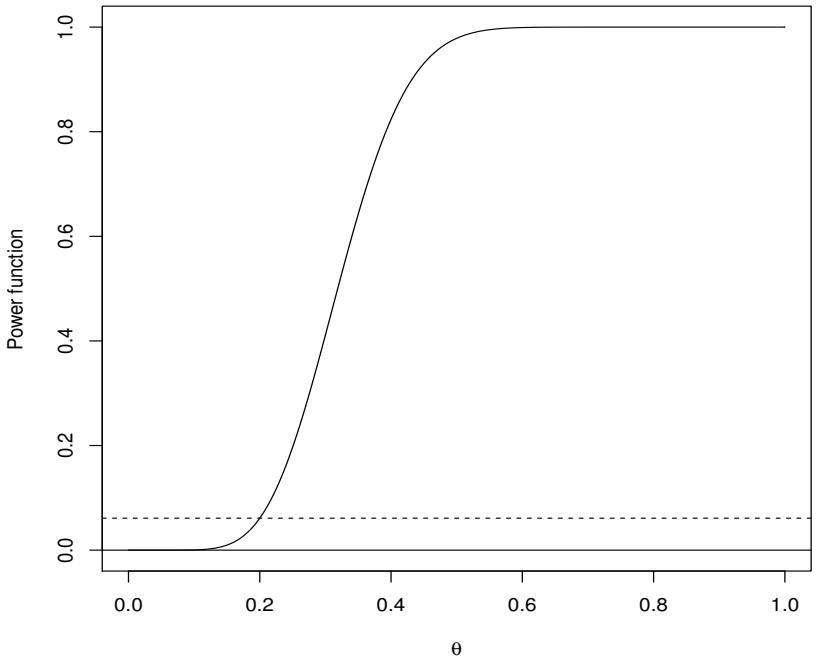

Figure 8.4: Power function $\beta(\theta)$ for the UMP level $\alpha = 0.0611$ test in Example 8.16 with $n = 30$ and $\theta_0 = 0.2$. A horizontal line at $\alpha = 0.0611$ has been added.

Therefore, the UMP level $\alpha = 0.0611$ test of $H_0 : \theta \leq 0.2$ versus $H_1 : \theta > 0.2$ uses $I(t \geq 10)$. The UMP level $\alpha = 0.0256$ test uses $I(t \geq 11)$. Note that (without randomizing) it is not possible to write a UMP level $\alpha = 0.05$ test in this problem. For the level $\alpha = 0.0611$ test, the power function is

$$\beta(\theta) = P_\theta(T \geq 10) = \sum_{t=10}^{30} \binom{30}{t} \theta^t (1-\theta)^{30-t},$$

which is depicted in Figure 8.4 (above).

Example 8.17. Suppose that $X_1, X_2, ..., X_n$ are iid with population distribution

$$f_X(x|\theta) = \theta e^{-\theta x} I(x > 0),$$

where $\theta > 0$. Note that this population distribution is an exponential distribution with mean $1/\theta$. Derive the UMP level α test for

$$H_0 : \theta \geq \theta_0$$
versus
$$H_1 : \theta < \theta_0.$$

Solution. It is easy to show that

$$T = \sum_{i=1}^{n} X_i$$

is a sufficient statistic and $T \sim \text{gamma}(n, 1/\theta)$. Suppose $\theta_2 > \theta_1$ and form the ratio

$$\frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} = \frac{\frac{1}{\Gamma(n)\left(\frac{1}{\theta_2}\right)^n}\, t^{n-1}e^{-\theta_2 t}}{\frac{1}{\Gamma(n)\left(\frac{1}{\theta_1}\right)^n}\, t^{n-1}e^{-\theta_1 t}} = \left(\frac{\theta_2}{\theta_1}\right)^n e^{-t(\theta_2-\theta_1)}.$$

Because $\theta_2 - \theta_1 > 0$, we see that the ratio

$$\frac{g_T(t|\theta_2)}{g_T(t|\theta_1)}$$

is a decreasing function of t over $\{t : t > 0\}$. However, the ratio is an increasing function of $t^* = -t$, and $T^* = T^*(\mathbf{X}) = -\sum_{i=1}^{n} X_i$ is still a sufficient statistic (it is a one-to-one function of T). Therefore, we can apply the Karlin-Rubin Theorem using $T^* = -T$ instead. Specifically, the UMP level α test is

$$\phi(t^*) = I(t^* < t_0),$$

where t_0 satisfies

$$\begin{aligned} \alpha = E_{\theta_0}[\phi(T^*)] &= P_{\theta_0}(T^* < t_0) \\ &= P_{\theta_0}(T > -t_0). \end{aligned}$$

Because $T \sim \text{gamma}(n, 1/\theta)$, we take $-t_0 = g_{n,1/\theta_0,\alpha}$, the (upper) α quantile of a gamma$(n, 1/\theta_0)$ distribution. Therefore, the UMP level α test is $I(t > g_{n,1/\theta_0,\alpha})$; i.e., the UMP level α rejection region is

$$R = \left\{ \mathbf{x} \in \mathcal{X} : \sum_{i=1}^{n} x_i > g_{n,1/\theta_0,\alpha} \right\}.$$

Using χ^2 critical values: We can also write this rejection region in terms of a χ^2 quantile. To see why, note that when $\theta = \theta_0$, the quantity $2\theta_0 T \sim \chi^2_{2n}$ so that

$$\begin{aligned} \alpha = P_{\theta_0}(T > -t_0) &= P_{\theta_0}(2\theta_0 T > -2\theta_0 t_0) \\ &\Longrightarrow -2\theta_0 t_0 \stackrel{\text{set}}{=} \chi^2_{2n,\alpha}. \end{aligned}$$

Therefore, the UMP level α rejection region can be written as

$$R = \left\{ \mathbf{x} \in \mathcal{X} : 2\theta_0 \sum_{i=1}^{n} x_i > \chi^2_{2n,\alpha} \right\} = \left\{ \mathbf{x} \in \mathcal{X} : \sum_{i=1}^{n} x_i > \frac{\chi^2_{2n,\alpha}}{2\theta_0} \right\}.$$

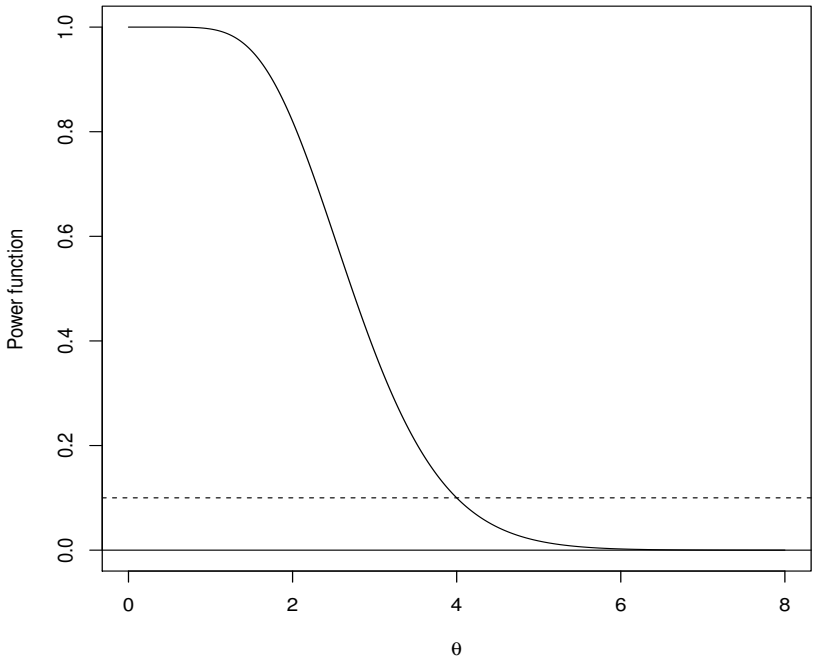

Figure 8.5: Power function $\beta(\theta)$ for the UMP level $\alpha = 0.10$ test in Example 8.17 with $n = 10$ and $\theta_0 = 4$. A horizontal line at $\alpha = 0.10$ has been added.

Remark: One advantage of writing the rejection region in this way is that it depends on a χ^2 quantile, which, historically, may have been available in probability tables (i.e., in times before computers and R). Another small advantage is that we can express the power function $\beta(\theta)$ in terms of a χ^2 cdf instead of a more general gamma cdf.

Power function: The power function of the UMP level α test is given by

$$\begin{aligned}\beta(\theta) = P_\theta(\mathbf{X} \in R) = P_\theta\left(T > \frac{\chi^2_{2n,\alpha}}{2\theta_0}\right) &= P_\theta\left(2\theta T > \frac{\theta\chi^2_{2n,\alpha}}{\theta_0}\right) \\ &= 1 - F_{\chi^2_{2n}}\left(\frac{\theta\chi^2_{2n,\alpha}}{\theta_0}\right),\end{aligned}$$

where $F_{\chi^2_{2n}}(\cdot)$ is the χ^2_{2n} cdf. A graph of this power function, when $n = 10$, $\alpha = 0.10$, and $\theta_0 = 4$, is shown in Figure 8.5 (above).

Proof of Karlin-Rubin Theorem. We will prove this theorem in parts. The first part is a lemma.

Lemma 1: If $g(x) \uparrow_{\text{nd}} x$ and $h(x) \uparrow_{\text{nd}} x$, then

$$\text{cov}[g(X), h(X)] \geq 0.$$

Proof. Take X_1, X_2 to be iid with the same distribution as X. Then

$$\begin{aligned} &E\{[h(X_1) - h(X_2)][g(X_1) - g(X_2)]\} \\ &\quad = E[h(X_1)g(X_1)] - E[h(X_2)g(X_1)] - E[h(X_1)g(X_2)] + E[h(X_2)g(X_2)] \\ &\quad \overset{X_1 \perp\!\!\!\perp X_2}{=} \underbrace{E[h(X_1)g(X_1)] - E[h(X_2)]E[g(X_1)]}_{= \text{cov}[g(X),h(X)]} \underbrace{-E[h(X_1)]E[g(X_2)] + E[h(X_2)g(X_2)]}_{= \text{cov}[g(X),h(X)]} \end{aligned}$$

which equals $2\text{cov}[g(X), h(X)]$. Therefore,

$$\text{cov}[g(X), h(X)] = \frac{1}{2}E\{[h(X_1) - h(X_2)][g(X_1) - g(X_2)]\}.$$

However, note that

$$[h(x_1) - h(x_2)][g(x_1) - g(x_2)] = \begin{cases} (\geq 0)(\geq 0), & x_1 > x_2 \\ 0, & x_1 = x_2 \\ (\leq 0)(\leq 0), & x_1 < x_2, \end{cases}$$

showing that $[h(x_1) - h(x_2)][g(x_1) - g(x_2)] \geq 0$, for all $x_1, x_2 \in \mathbb{R}$. By Theorem 2.2.5 (CB, pp 57), $E\{[h(X_1) - h(X_2)][g(X_1) - g(X_2)]\} \geq 0$. □

Remark: Our frame of reference going forward is testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Proving the other case stated in the Karlin-Rubin Theorem is analogous.

Lemma 2. Suppose the family $\{g_T(t|\theta) : \theta \in \Theta\}$ has MLR. If $\psi(t) \uparrow_{\text{nd}} t$, then $E_\theta[\psi(T)] \uparrow_{\text{nd}} \theta$.
Proof. Suppose that $\theta_2 > \theta_1$. Because $\{g_T(t|\theta) : \theta \in \Theta\}$ has MLR, we know that

$$\frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} \uparrow_{\text{nd}} t$$

over the set $\{t : g_T(t|\theta_1) > 0 \text{ or } g_T(t|\theta_2) > 0\}$. Therefore, by Lemma 1, we know

$$\begin{aligned} \text{cov}_{\theta_1}\left[\psi(T), \frac{g_T(T|\theta_2)}{g_T(T|\theta_1)}\right] \geq 0 \quad &\Longrightarrow \quad \underbrace{E_{\theta_1}\left[\psi(T)\frac{g_T(T|\theta_2)}{g_T(T|\theta_1)}\right]}_{= E_{\theta_2}[\psi(T)]} \geq E_{\theta_1}[\psi(T)] \underbrace{E_{\theta_1}\left[\frac{g_T(T|\theta_2)}{g_T(T|\theta_1)}\right]}_{= 1} \\ &\Longrightarrow \quad E_{\theta_2}[\psi(T)] \geq E_{\theta_1}[\psi(T)]. \end{aligned}$$

Because θ_1 and θ_2 are arbitrary, the result follows. □

Lemma 3. Under the same assumptions stated in Lemma 2,

$$P_\theta(T > t_0) \uparrow_{\text{nd}} \theta$$

for all $t_0 \in \mathbb{R}$. In other words, the family $\{g_T(t|\theta) : \theta \in \Theta\}$ is stochastically increasing in θ.

Proof. This is a special case of Lemma 2. Fix t_0. Take $\psi(t) = I(t > t_0)$. Clearly,

$$\psi(t) = \begin{cases} 1, & t > t_0 \\ 0, & t \leq t_0 \end{cases}$$

is a nondecreasing function of t (with t_0 fixed). From Lemma 2, we know that

$$E_\theta[\psi(T)] = E_\theta[I(T > t_0)] = P_\theta(T > t_0) \uparrow_{\text{nd}} \theta.$$

Because $t_0 \in \mathbb{R}$ was chosen arbitrarily, this result is true for all $t_0 \in \mathbb{R}$. □

Implication: In the statement of the Karlin-Rubin Theorem (for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$), we have now shown that the power function

$$\beta(\theta) = P_\theta(T > t_0)$$

is a nondecreasing function of θ. This explains why α satisfies

$$\alpha = P_{\theta_0}(T > t_0).$$

Why? Because $P_\theta(T > t_0)$ is a nondecreasing function of θ,

$$\alpha = \sup_{H_0} \beta(\theta) = \sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = P_{\theta_0}(T > t_0).$$

This shows that $\phi(t) = I(t > t_0)$ is a size α (and hence level α) test function. Thus, all that remains is to show that this test is *uniformly* most powerful (i.e., most powerful $\forall \theta > \theta_0$). Remember that we are considering the test

$$H_0 : \theta \leq \theta_0$$
versus
$$H_1 : \theta > \theta_0.$$

Let $\phi^*(\mathbf{x})$ be any other level α test of H_0 versus H_1. Fix $\theta_1 > \theta_0$ and consider the test of

$$H_0^* : \theta = \theta_0$$
versus
$$H_1^* : \theta = \theta_1$$

instead. Note that

$$E_{\theta_0}[\phi^*(\mathbf{X})] \leq \sup_{\theta \leq \theta_0} E_\theta[\phi^*(\mathbf{X})] \leq \alpha$$

because $\phi^*(\mathbf{x})$ is a level α test of H_0 versus H_1. This also means that $\phi^*(\mathbf{x})$ is a level α test of H_0^* versus H_1^*. However, Corollary 8.3.13 (Neyman Pearson with a sufficient statistic T) says that $\phi(t)$ is the most powerful (MP) level α test of H_0^* versus H_1^*. This means that

$$E_{\theta_1}[\phi(T)] \geq E_{\theta_1}[\phi^*(\mathbf{X})].$$

Because $\theta_1 > \theta_0$ was chosen arbitrarily and because $\phi^*(\mathbf{x})$ was too, we have

$$E_\theta[\phi(T)] \geq E_\theta[\phi^*(\mathbf{X})]$$

for all $\theta > \theta_0$ and for any level α test $\phi^*(\mathbf{x})$ of H_0 versus H_1. Because $\phi(t)$ is a level α test of H_0 versus H_1 (shown above), we are done. □

Note: In single parameter exponential families, we can find UMP tests for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ (or for $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$). Unfortunately,

- once we get outside this setting (even with a one-sided H_1), UMP tests do become scarce.
- with a two-sided H_1, that is $H_1 : \theta \neq \theta_0$, UMP tests do not exist.

In other words, the collection of problems for which a UMP test exists is somewhat small. In many ways, this should not be surprising. Requiring a test to outperform **all other** level α tests **for all** θ in the alternative space Θ_0^c is asking a lot. The "larger" Θ_0^c is, the harder it is to find a UMP test.

Example 8.18. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and σ_0^2 is known. Consider testing

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu \neq \mu_0.$$

There is no UMP test for this problem. A UMP test would exist if we could find a test whose power function "beats" the power function for all other level α tests. For one-sided alternatives, it is possible to find one. However, a two-sided alternative space is too large. To illustrate, suppose we considered testing

$$H_0' : \mu \leq \mu_0$$
versus
$$H_1' : \mu > \mu_0.$$

The UMP level α test for H_0' versus H_1' uses

$$\phi'(\mathbf{x}) = I\left(\overline{x} > \frac{z_\alpha \sigma_0}{\sqrt{n}} + \mu_0\right)$$

and has power function

$$\beta'(\mu) = 1 - F_Z\left(z_\alpha + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right),$$

where $F_Z(\cdot)$ is the standard normal cdf. This is a size (and level) α test for H_0' versus H_1'. It is also a size (and level) test for H_0 versus H_1 because

$$\sup_{\mu \in \Theta_0} \beta'(\mu) = \sup_{\mu = \mu_0} \beta'(\mu) = \beta'(\mu_0) = 1 - F_Z(z_\alpha) = \alpha.$$

Now consider testing

$$H_0'' : \mu \geq \mu_0$$
versus
$$H_1'' : \mu < \mu_0.$$

The UMP level α test for H_0'' versus H_1'' uses

$$\phi''(\mathbf{x}) = I\left(\overline{x} < -\frac{z_\alpha \sigma_0}{\sqrt{n}} + \mu_0\right)$$

and has power function

$$\beta''(\mu) = F_Z\left(-z_\alpha + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right).$$

This is a size (and level) α test for H_0 versus H_1 because

$$\sup_{\mu \in \Theta_0} \beta''(\mu) = \sup_{\mu = \mu_0} \beta''(\mu) = \beta''(\mu_0) = F_Z(-z_\alpha) = \alpha.$$

Therefore, we have concluded that

- $\phi'(\mathbf{x})$ is UMP level α when $\mu > \mu_0$
- $\phi''(\mathbf{x})$ is UMP level α when $\mu < \mu_0$.

However, $\phi'(\mathbf{x}) \neq \phi''(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Therefore, no UMP test can exist for H_0 versus H_1.

Q: How do we find an "optimal" test in situations like this (e.g., a two-sided H_1)?
A: We change what we mean by "optimal."

Definition: Consider the test of

$$H_0 : \theta \in \Theta_0$$
versus
$$H_1 : \theta \in \Theta_0^c.$$

A test with power function $\beta(\theta)$ is **unbiased** if $\beta(\theta') \geq \beta(\theta'')$ for all $\theta' \in \Theta_0^c$ and for all $\theta'' \in \Theta_0$. That is, the power is always larger in the alternative parameter space than it is in the null parameter space.

- Therefore, when no UMP test exists, we could further restrict attention to those tests that are level α **and** are unbiased. Conceptually, define

$$\mathcal{C}^U = \{\text{all level } \alpha \text{ tests that are unbiased}\}.$$

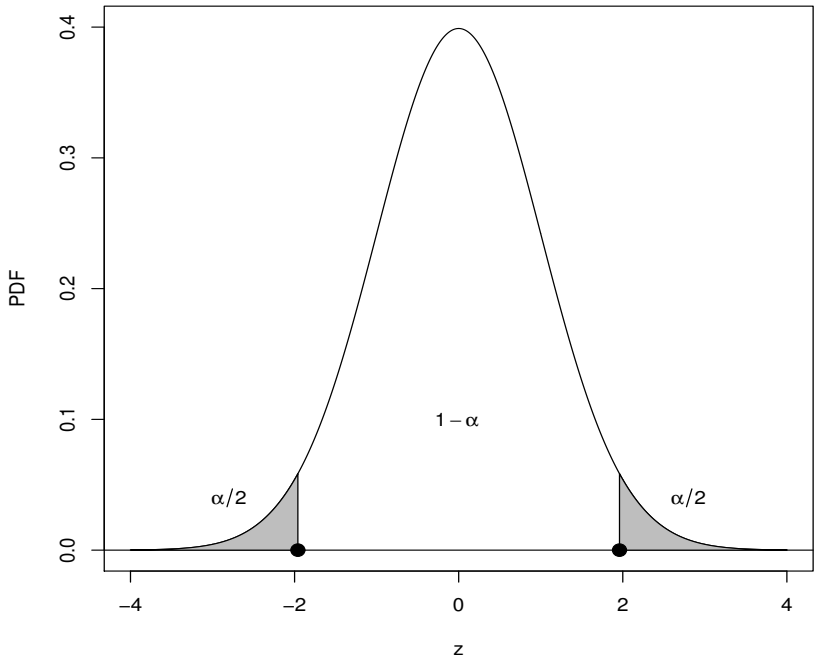

Figure 8.6: Pdf of $Z \sim \mathcal{N}(0,1)$. The UMPU level α rejection region in Example 8.18 is shown shaded.

- The test in $\mathcal{C}^U$ that is UMP is called the **uniformly most powerful unbiased** (**UMPU**) test. The UMPU test has power function $\beta(\theta)$ that satisfies

$$\beta(\theta) \geq \beta^*(\theta) \quad \text{for all } \theta \in \Theta_0^c,$$

where $\beta^*(\theta)$ is the power function of any other (unbiased) test in $\mathcal{C}^U$.

Example 8.18 (continued). Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and σ_0^2 is known. Consider testing

$$H_0 : \mu = \mu_0$$
$$\text{versus}$$
$$H_1 : \mu \neq \mu_0.$$

The UMPU level α test uses

$$\phi(\mathbf{x}) = \begin{cases} 1, & \dfrac{\overline{x} - \mu_0}{\sigma_0/\sqrt{n}} < -z_{\alpha/2} \quad \text{or} \quad \dfrac{\overline{x} - \mu_0}{\sigma_0/\sqrt{n}} > z_{\alpha/2} \\ 0, & \text{otherwise.} \end{cases}$$

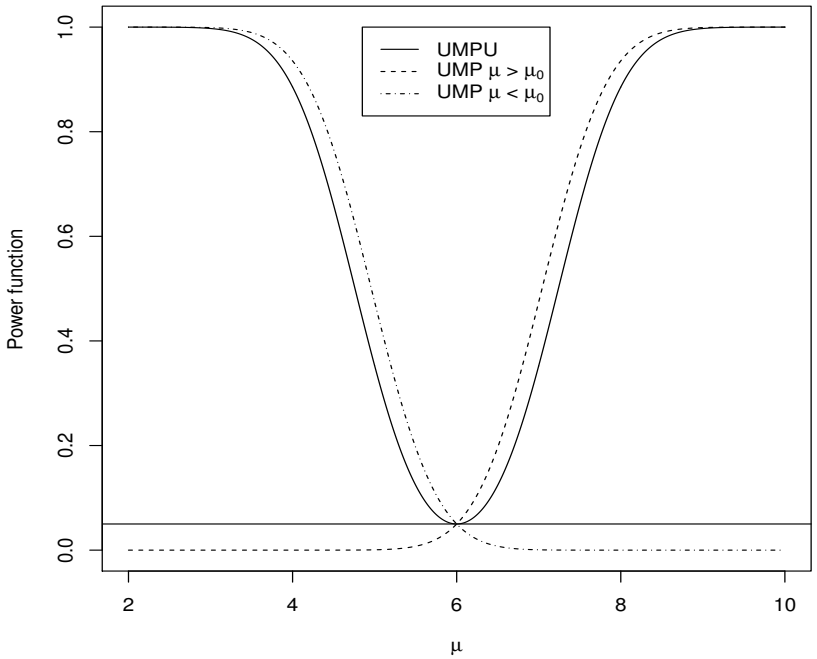

Figure 8.7: Power function $\beta(\mu)$ of the UMPU level $\alpha = 0.05$ test in Example 8.18 with $n = 10$, $\mu_0 = 6$, and $\sigma_0^2 = 4$. Also shown are the power functions corresponding to the two UMP level $\alpha = 0.05$ tests with $H_1 : \mu > \mu_0$ and $H_1 : \mu < \mu_0$.

In other words, the UMPU level α rejection region is

$$R = \{\mathbf{x} \in \mathcal{X} : \phi(\mathbf{x}) = 1\} = \{\mathbf{x} \in \mathcal{X} : \ |z| > z_{\alpha/2}\},$$

where

$$z = \frac{\overline{x} - \mu_0}{\sigma_0/\sqrt{n}}.$$

Note that, because $Z \overset{H_0}{\sim} \mathcal{N}(0,1)$,

$$\begin{aligned} P_{\mu_0}(\mathbf{X} \in R) = P_{\mu_0}(|Z| > z_{\alpha/2}) &= 1 - P_{\mu_0}(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= 1 - [F_Z(z_{\alpha/2}) - F_Z(-z_{\alpha/2})] \\ &= 1 - (1 - \alpha/2) + \alpha/2 = \alpha, \end{aligned}$$

which shows that R is a size (and hence level) α rejection region. The power function of the UMPU test $\phi(\mathbf{x})$ is

$$\beta(\mu) = P_\mu(\mathbf{X} \in R) = P_\mu(|Z| \geq z) \ = \ 1 - F_Z\left(z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right) + F_Z\left(-z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right).$$

Special case: I took $n = 10$, $\alpha = 0.05$, $\mu_0 = 6$, and $\sigma_0^2 = 4$. The UMPU level $\alpha = 0.05$ power function $\beta(\mu)$ is shown in Figure 8.7 (above). For reference, I have also plotted in

Figure 8.7 the UMP level $\alpha = 0.05$ power functions for the two one-sided tests (i.e., the tests with $H_1 : \mu > \mu_0$ and $H_1 : \mu < \mu_0$, respectively).

- It is easy to see that the UMPU test is an unbiased test. Note that $\beta(\mu)$ is always larger in the alternative parameter space $\{\mu \in \mathbb{R} : \mu \neq \mu_0\}$ than it is when $\mu = \mu_0$.
- The UMPU test's power function "loses" to each UMP test's power function in the region where that UMP test is most powerful. This is the price one must pay for restricting attention to unbiased tests. The best unbiased test for a two-sided H_1 will not beat a one-sided UMP test. However, the UMPU test is clearly better than the UMP tests in each UMP test's null parameter space.

8.3.4 Probability values

Definition: A **p-value** $p(\mathbf{X})$ is a test statistic, satisfying $0 \leq p(\mathbf{x}) \leq 1$, for all $\mathbf{x} \in \mathcal{X}$. Small values of $p(\mathbf{x})$ are evidence against H_0. A p-value is said to be **valid** if

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha,$$

for all $\theta \in \Theta_0$ and for all $0 \leq \alpha \leq 1$.

Remark: Quoting your authors (CB, pp 397),

> *"If $p(\mathbf{X})$ is a valid p-value, it is easy to construct a level α test based on $p(\mathbf{X})$. The test that rejects H_0 if and only if $p(\mathbf{X}) \leq \alpha$ is a level α test."*

It is easy to see why this is true. The validity requirement above guarantees that

$$\phi(\mathbf{x}) = I(p(\mathbf{x}) \leq \alpha)$$

is a level α test function. Why? Note that

$$\sup_{\theta \in \Theta_0} E_\theta[\phi(\mathbf{X})] = \sup_{\theta \in \Theta_0} P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha.$$

Therefore, rejecting H_0 when $p(\mathbf{x}) \leq \alpha$ is a level α decision rule.

Theorem 8.3.27. Let $W = W(\mathbf{X})$ be a test statistic such that large values of W give evidence against H_0. For each $\mathbf{x} \in \mathcal{X}$, define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq w),$$

where $w = W(\mathbf{x})$. Then $p(\mathbf{X})$ is a valid p-value. Note that the definition of $p(\mathbf{x})$ for when small values of W give evidence against H_0 would be analogous.

Proof. Fix $\theta \in \Theta_0$. Let $F_{-W}(w|\theta)$ denote the cdf of $-W = -W(\mathbf{X})$. When the test rejects for large values of W,

$$p_\theta(\mathbf{x}) \equiv P_\theta(W(\mathbf{X}) \geq w) = P_\theta(-W(\mathbf{X}) \leq -w) = F_{-W}(-w|\theta),$$

where $w = W(\mathbf{x})$. If $-W(\mathbf{X})$ is a continuous random variable, then

$$p_\theta(\mathbf{X}) \stackrel{d}{=} F_{-W}(-W|\theta) \stackrel{d}{=} \mathcal{U}(0,1),$$

by the Probability Integral Transformation (Chapter 2). If $-W(\mathbf{X})$ is discrete, then

$$p_\theta(\mathbf{X}) \stackrel{d}{=} F_{-W}(-W|\theta) \geq_{\text{ST}} \mathcal{U}(0,1),$$

where the notation $X \geq_{\text{ST}} Y$ means "the distribution of X is stochastically larger than the distribution of Y" (see Exercise 2.10, CB, pp 77). Combining both cases, we have

$$P_\theta(p_\theta(\mathbf{X}) \leq \alpha) \leq \alpha,$$

for all $0 \leq \alpha \leq 1$. Now, note that

$$p(\mathbf{x}) \equiv \sup_{\theta' \in \Theta_0} P_{\theta'}(W(\mathbf{X}) \geq w) \geq P_\theta(W(\mathbf{X}) \geq w) \equiv p_\theta(\mathbf{x}),$$

for all $\mathbf{x} \in \mathcal{X}$. Therefore,

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq P_\theta(p_\theta(\mathbf{X}) \leq \alpha) \leq \alpha.$$

Because we fixed $\theta \in \Theta_0$ arbitrarily, this result must hold for all $\theta \in \Theta_0$. We have shown that $p(\mathbf{X})$ is a valid p-value. $\square$

Example 8.19. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Set $\boldsymbol{\theta} = (\mu, \sigma^2)$. Consider testing

$$H_0 : \mu \leq \mu_0$$
versus
$$H_1 : \mu > \mu_0.$$

We have previously shown (see pp 75-76, notes) that large values of

$$W = W(\mathbf{X}) = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

are evidence against H_0 (i.e., this is a "one-sample t test," which is a LRT). The null parameter space is

$$\Theta_0 = \{\boldsymbol{\theta} = (\mu, \sigma^2) : \ \mu \leq \mu_0, \ \sigma^2 > 0\}.$$

Therefore, with observed value $w = W(\mathbf{x})$, the p-value for the test is

$$\begin{aligned} p(\mathbf{x}) = \sup_{\boldsymbol{\theta} \in \Theta_0} P_{\boldsymbol{\theta}}(W(\mathbf{X}) \geq w) &= \sup_{\boldsymbol{\theta} \in \Theta_0} P_{\boldsymbol{\theta}}\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq w\right) \\ &= \sup_{\boldsymbol{\theta} \in \Theta_0} P_{\boldsymbol{\theta}}\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} \geq w + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) \\ &= \sup_{\mu \leq \mu_0} P_{\boldsymbol{\theta}}\left(T_{n-1} \geq w + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) = P\left(T_{n-1} \geq w\right), \end{aligned}$$

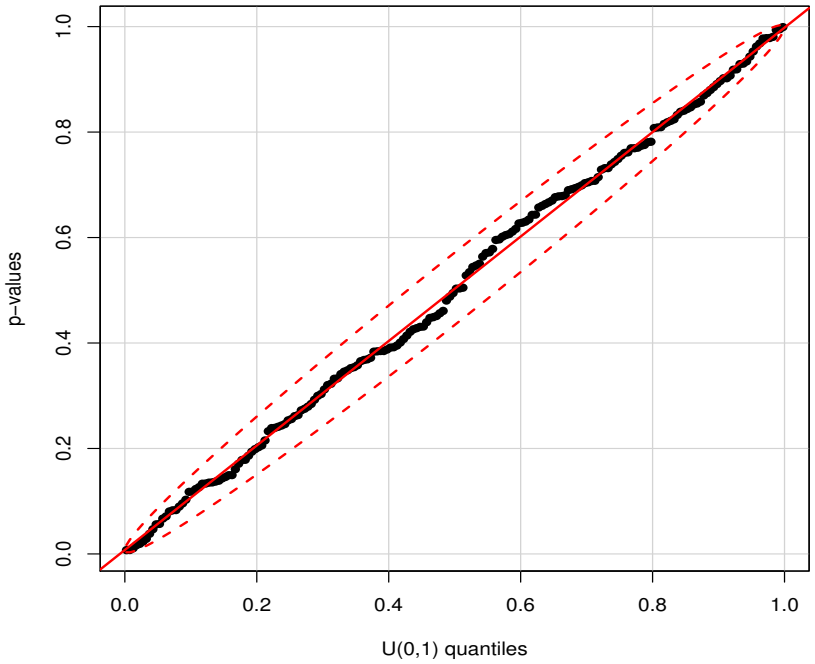

Figure 8.8: Uniform qq plot of $B = 200$ simulated p-values in Example 8.20.

where T_{n-1} is a t random variable with $n-1$ degrees of freedom. The penultimate equality holds because the distribution of $(\overline{X} - \mu)/(S/\sqrt{n})$ does not depend on σ^2. The last equality holds because $(\mu_0 - \mu)/(S/\sqrt{n})$ is a nonnegative random variable.

Remark: In Example 8.19, calculating the supremum over Θ_0 is relatively easy. In other problems, it might not be, especially when there are nuisance parameters. A very good discussion on this is given in Berger and Boos (1994). These authors propose another type of p-value by "suping" over subsets of Θ_0 formed from calculating confidence intervals first (which can make the computation easier).

Important: If H_0 is simple, say $H_0 : \theta = \theta_0$, and if a p-value $p(\mathbf{x})$ satisfies

$$P_{\theta_0}(p(\mathbf{X}) \leq \alpha) = \alpha,$$

for all $0 \leq \alpha \leq 1$, then $\phi(\mathbf{x}) = I(p(\mathbf{x}) \leq \alpha)$ is a size α test and $p(\mathbf{X}) \overset{H_0}{\sim} \mathcal{U}(0,1)$.

Example 8.20. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(0,1)$. I used R to simulate $B = 200$ independent samples of this type, each with $n = 30$. With each sample, I performed a t test for $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ and calculated the p-value for each test (note that H_0 is true). A uniform qq plot of the 200 p-values in Figure 8.8 shows agreement with the $\mathcal{U}(0,1)$ distribution. Using $\alpha = 0.05$, there were 9 tests (out of 200) that incorrectly rejected H_0.

9 Interval Estimation

Complementary reading: Chapter 9 (CB).

9.1 Introduction

Setting: We observe $\mathbf{X} = (X_1, X_2, ..., X_n) \sim f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. Usually, $X_1, X_2, ..., X_n$ will constitute a random sample (iid sample) from a population $f_X(x|\boldsymbol{\theta})$. We regard $\boldsymbol{\theta}$ as fixed and unknown.

Definition: An **interval estimate** of a real-valued parameter θ is any pair of functions

$$\begin{aligned} L(\mathbf{x}) &= L(x_1, x_2, ..., x_n) \\ U(\mathbf{x}) &= U(x_1, x_2, ..., x_n), \end{aligned}$$

satisfying $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. When $\mathbf{X} = \mathbf{x}$ is observed, the inference

$$L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$$

is made. The random version $[L(\mathbf{X}), U(\mathbf{X})]$ is called an **interval estimator**.

Remark: In the definition above, a **one-sided** interval estimate is formed when one of the endpoints is $\pm\infty$. For example, if $L(\mathbf{x}) = -\infty$, then the estimate is $(-\infty, U(\mathbf{x})]$. If $U(\mathbf{x}) = \infty$, the estimate is $[L(\mathbf{x}), \infty)$.

Definition: Suppose $[L(\mathbf{X}), U(\mathbf{X})]$ is an interval estimator for θ. The **coverage probability** of the interval is

$$P_\theta(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})).$$

It is important to note the following:

- In the probability above, it is the endpoints $L(\mathbf{X})$ and $U(\mathbf{X})$ that are random; not θ (it is fixed).
- The coverage probability is regarded as a function of θ. That is, the probability that $[L(\mathbf{X}), U(\mathbf{X})]$ contains θ may be different for different values of $\theta \in \Theta$. This is usually true when $\mathbf{X}$ is discrete.

Definition: The **confidence coefficient** of the interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ is

$$\inf_{\theta \in \Theta} P_\theta(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})).$$

An interval estimator with confidence coefficient equal to $1-\alpha$ is called a $1-\alpha$ **confidence interval**.

Remark: In some problems, it is possible that the estimator itself is not an interval. More generally, we use the term $1-\alpha$ **confidence set** to allow for these types of estimators. The notation $C(\mathbf{X})$ is used more generally to denote a confidence set.

Example 9.1. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0, \theta)$, where $\theta > 0$. We consider two interval estimators:

1. $(aX_{(n)}, bX_{(n)})$, where $1 \leq a < b$
2. $(X_{(n)} + c, X_{(n)} + d)$, where $0 \leq c < d$.

The pdf of $X_{(n)}$ is

$$\begin{aligned} f_{X_{(n)}}(x) &= nf_X(x)[F_X(x)]^{n-1} = n\left(\frac{1}{\theta}\right)\left(\frac{x}{\theta}\right)^{n-1} I(0 < x < \theta) \\ &= \frac{nx^{n-1}}{\theta^n} I(0 < x < \theta). \end{aligned}$$

By transformation, the pdf of

$$T = \frac{X_{(n)}}{\theta}$$

is

$$f_T(t) = nt^{n-1} I(0 < t < 1);$$

i.e., $T \sim \text{beta}(n, 1)$. The coverage probability for the first interval is

$$\begin{aligned} P_\theta(aX_{(n)} \leq \theta \leq bX_{(n)}) &= P_\theta\left(\frac{1}{bX_{(n)}} \leq \frac{1}{\theta} \leq \frac{1}{aX_{(n)}}\right) \\ &= P_\theta\left(\frac{1}{b} \leq \frac{X_{(n)}}{\theta} \leq \frac{1}{a}\right) \\ &= \int_{1/b}^{1/a} nt^{n-1} dt = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n, \end{aligned}$$

that is, the coverage probability is the same for all $\theta \in \Theta = \{\theta : \theta > 0\}$. The confidence coefficient of the interval $(aX_{(n)}, bX_{(n)})$ is therefore

$$\inf_{\theta > 0}\left[\left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n\right] = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n.$$

On the other hand, the coverage probability for the second interval is

$$\begin{aligned} P_\theta(X_{(n)} + c \leq \theta \leq X_{(n)} + d) &= P_\theta(c \leq \theta - X_{(n)} \leq d) \\ &= P_\theta\left(\frac{c}{\theta} \leq 1 - \frac{X_{(n)}}{\theta} \leq \frac{d}{\theta}\right) \\ &= P_\theta\left(1 - \frac{d}{\theta} \leq \frac{X_{(n)}}{\theta} \leq 1 - \frac{c}{\theta}\right) \\ &= \int_{1-\frac{d}{\theta}}^{1-\frac{c}{\theta}} nt^{n-1} dt = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n, \end{aligned}$$

which does depend on θ. Interestingly, the confidence coefficient of $(X_{(n)}+c, X_{(n)}+d)$ is

$$\inf_{\theta>0}\left[\left(1-\frac{c}{\theta}\right)^n-\left(1-\frac{d}{\theta}\right)^n\right]=0.$$

Example 9.2. Suppose that $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0<p<1$. A "$1-\alpha$ confidence interval" commonly taught in undergraduate courses is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

where $\widehat{p}$ is the sample proportion, that is,

$$\widehat{p}=\frac{Y}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

where $Y=\sum_{i=1}^{n}X_i \sim b(n,p)$, and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution. In Chapter 10, we will learn that this is a large-sample "Wald-type" confidence interval. An expression for the coverage probability of this interval is

$$\begin{aligned}
&P_p\left(\widehat{p}-z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \le p \le \widehat{p}+z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)\\
&\quad=E_p\left[I\left(\frac{Y}{n}-z_{\alpha/2}\sqrt{\frac{\frac{Y}{n}(1-\frac{Y}{n})}{n}} \le p \le \frac{Y}{n}+z_{\alpha/2}\sqrt{\frac{\frac{Y}{n}(1-\frac{Y}{n})}{n}}\right)\right]\\
&\quad=\sum_{y=0}^{n} I\left(\frac{y}{n}-z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}} \le p \le \frac{y}{n}+z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}\right)\underbrace{\binom{n}{y}p^y(1-p)^y}_{b(n,p)\text{ pmf}}.
\end{aligned}$$

Special case: I used R to graph this coverage probability function across values of $0<p<1$ when $n=40$ and $\alpha=0.05$; see Figure 9.1 (next page).

- The coverage probability rarely attains the nominal 0.95 level across $0<p<1$.
- The jagged nature of the coverage probability function (of p) arises from the discreteness of $Y \sim b(40,p)$.
- The confidence coefficient of the Wald interval (i.e., the infimum coverage probability across all $0<p<1$) is clearly 0.
- An excellent account of the performance of this confidence interval (and competing intervals) is given in Brown et al. (2001, *Statistical Science*).
- When $1-\alpha=0.95$, one competing interval mentioned in Brown et al. (2001) replaces y with $y^*=y+2$ and n with $n^*=n+4$. This "add two successes-add two failures" interval was proposed by Agresti and Coull (1998, *American Statistician*). Because this interval's coverage probability is much closer to the nominal level across $0<p<1$ (and because it is so easy to compute), it has begun to usurp the Wald confidence interval in introductory level courses.

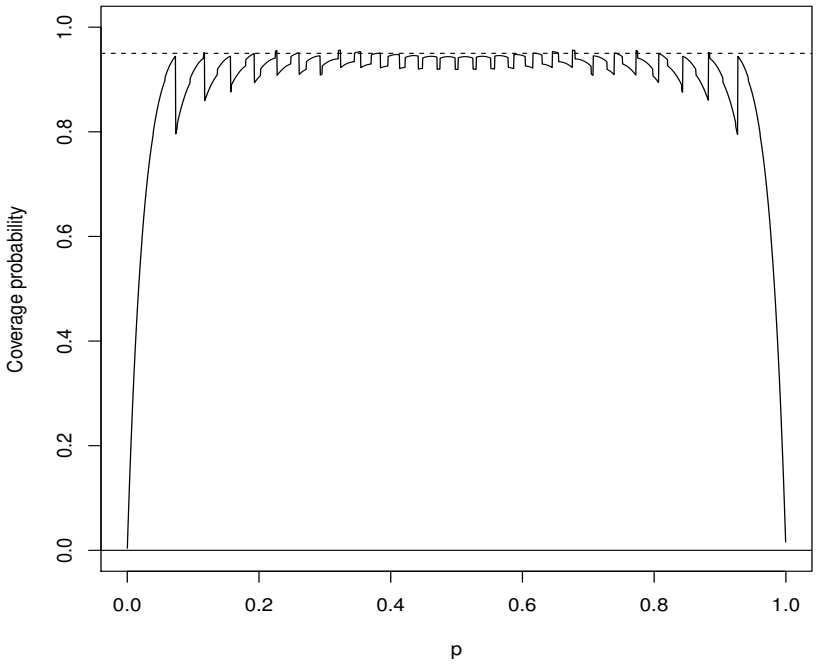

Figure 9.1: Coverage probability of the Wald confidence interval for a binomial proportion p when $n = 40$ and $\alpha = 0.05$. A dotted horizontal line at $1 - \alpha = 0.95$ has been added.

9.2 Methods of Finding Interval Estimators

Preview: The authors present four methods to find interval estimators:

1. Test inversion (i.e., inverting a test statistic)
2. Using pivotal quantities
3. "Guaranteeing an interval" by pivoting a cdf
4. Bayesian credible intervals.

Note: Large-sample interval estimators will be discussed in Chapter 10.

9.2.1 Inverting a test statistic

Remark: This method of interval construction is motivated by the strong duality between hypothesis testing and confidence intervals.

Motivation: Consider testing $H_0 : \theta = \theta_0$ using the (non-randomized) test function

$$\phi(\mathbf{x}) = I(\mathbf{x} \in R_{\theta_0}) = \begin{cases} 1, & \mathbf{x} \in R_{\theta_0} \\ 0, & \mathbf{x} \in R_{\theta_0}^c, \end{cases}$$

where

$$P_{\theta_0}(\mathbf{X} \in R_{\theta_0}) = E_{\theta_0}[\phi(\mathbf{X})] = \alpha;$$

i.e., $\phi(\mathbf{x})$ is a size α test. Note that we have used the notation R_{θ_0} to emphasize that the rejection region R depends on the value of θ_0. Let $A_{\theta_0} = R_{\theta_0}^c$ denote the "acceptance region" for the test, that is, A_{θ_0} is the set of all $\mathbf{x} \in \mathcal{X}$ that do **not** lead to H_0 being rejected. For each $\mathbf{x} \in \mathcal{X}$, define

$$C(\mathbf{x}) = \{\theta_0 : \mathbf{x} \in A_{\theta_0}\}.$$

From this definition, clearly $\theta_0 \in C(\mathbf{x}) \iff \mathbf{x} \in A_{\theta_0}$. Therefore,

$$P_{\theta_0}(\theta_0 \in C(\mathbf{X})) = P_{\theta_0}(\mathbf{X} \in A_{\theta_0}) = 1 - P_{\theta_0}(\mathbf{X} \in R_{\theta_0}) = 1 - \alpha.$$

However, this same argument holds for all $\theta_0 \in \Theta$; i.e., it holds regardless of the value of θ under H_0. Therefore,

$$C(\mathbf{X}) = \{\theta \in \Theta : \mathbf{X} \in A_\theta\}$$

is a $1 - \alpha$ confidence set.

Example 9.3. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. A size α likelihood ratio test (LRT) of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ uses the test function

$$\phi(\mathbf{x}) = \begin{cases} 1, & \dfrac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \geq t_{n-1,\alpha/2} \\ 0, & \text{otherwise.} \end{cases}$$

The "acceptance region" for this test is

$$A_{\mu_0} = \left\{ \mathbf{x} \in \mathcal{X} : \frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} < t_{n-1,\alpha/2} \right\},$$

where, note that

$$\begin{aligned} P_{\mu_0}(\mathbf{X} \in A_{\mu_0}) &= P_{\mu_0}\left(\frac{|\overline{X} - \mu_0|}{S/\sqrt{n}} < t_{n-1,\alpha/2} \right) \\ &= P_{\mu_0}\left(-t_{n-1,\alpha/2} < \frac{\overline{X} - \mu_0}{S/\sqrt{n}} < t_{n-1,\alpha/2} \right) = 1 - \alpha. \end{aligned}$$

Therefore, a $1 - \alpha$ confidence set for μ is

$$\begin{aligned} C(\mathbf{x}) = \{\mu \in \mathbb{R} : \mathbf{x} \in A_\mu\} &= \left\{ \mu : -t_{n-1,\alpha/2} < \frac{\overline{x} - \mu}{s/\sqrt{n}} < t_{n-1,\alpha/2} \right\} \\ &= \left\{ \mu : -t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \overline{x} - \mu < t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \right\} \\ &= \left\{ \mu : \overline{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \right\}. \end{aligned}$$

The random version of this confidence set (interval) is written as

$$\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right).$$

Remark: As Example 9.3 suggests, when we invert a two-sided hypothesis test, we get a two-sided confidence interval. This will be true in most problems. Analogously, inverting one-sided tests generally leads to one-sided intervals.

Example 9.4. Suppose $X_1, X_2, ..., X_n$ are iid exponential(θ), where $\theta > 0$. A uniformly most powerful (UMP) level α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ uses the test function

$$\phi(t) = \begin{cases} 1, & t \geq \dfrac{\theta_0}{2}\chi^2_{2n,\alpha} \\ 0, & \text{otherwise,} \end{cases}$$

where the sufficient statistic $t = \sum_{i=1}^n x_i$. The "acceptance region" for this test is

$$A_{\theta_0} = \left\{\mathbf{x} \in \mathcal{X} : t < \frac{\theta_0}{2}\chi^2_{2n,\alpha}\right\},$$

where, note that

$$P_{\theta_0}(\mathbf{X} \in A_{\theta_0}) = P_{\theta_0}\left(T < \frac{\theta_0}{2}\chi^2_{2n,\alpha}\right) = P_{\theta_0}\left(\frac{2T}{\theta_0} < \chi^2_{2n,\alpha}\right) = 1 - \alpha,$$

because $2T/\theta_0 \overset{H_0}{\sim} \text{gamma}(n, 2) \overset{d}{=} \chi^2_{2n}$. Therefore, a $1 - \alpha$ confidence set for θ is

$$\begin{aligned} C(\mathbf{x}) = \{\theta > 0 : \mathbf{x} \in A_\theta\} &= \left\{\theta : t < \frac{\theta}{2}\chi^2_{2n,\alpha}\right\} \\ &= \left\{\theta : \frac{2t}{\chi^2_{2n,\alpha}} < \theta\right\}. \end{aligned}$$

The random version of this confidence set is written as

$$\left(\frac{2T}{\chi^2_{2n,\alpha}},\ \infty\right),$$

where $T = \sum_{i=1}^n X_i$. This is a "one-sided" interval, as expected, because we have inverted a one-sided test.

Remark: The test inversion method makes direct use of the relationship between hypothesis tests and confidence intervals (sets). On pp 421, the authors of CB write,

> *"Both procedures look for consistency between sample statistics and population parameters. The hypothesis test fixes the parameter and asks what sample values (the acceptance region) are consistent with that fixed value. The confidence set fixes the sample value and asks what parameter values (the confidence interval) make this sample value most plausible."*

An illustrative figure (Figure 9.2.1, pp 421) displays this relationship in the $\mathcal{N}(\mu, \sigma_0^2)$ case; i.e., writing a confidence interval for a normal mean μ when σ_0^2 is known.

9.2.2 Pivotal quantities

Definition: A random variable $Q = Q(\mathbf{X}, \theta)$ is a **pivotal quantity** (or **pivot**) if the distribution of Q does not depend on θ. That is, Q has the same distribution for all $\theta \in \Theta$.

Remark: Finding pivots makes getting confidence intervals easy. If $Q = Q(\mathbf{X}, \theta)$ is a pivot, then we can set

$$1 - \alpha = P_\theta(a \leq Q(\mathbf{X}, \theta) \leq b),$$

where a and b are quantiles of the distribution of Q that satisfy the equation. Because Q is a pivot, the probability on the RHS will be the same for all $\theta \in \Theta$. Therefore, a $1 - \alpha$ confidence interval can be determined from this equation.

Example 9.5. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0, \theta)$, where $\theta > 0$. In Example 9.1, we showed that

$$Q = Q(\mathbf{X}, \theta) = \frac{X_{(n)}}{\theta} \sim \text{beta}(n, 1).$$

Because the distribution of Q is free of θ, we know that Q is a pivot. Let $b_{n,1,1-\alpha/2}$ and $b_{n,1,\alpha/2}$ denote the lower and upper $\alpha/2$ quantiles of a beta$(n, 1)$ distribution, respectively. We can then write

$$\begin{aligned} 1 - \alpha = P_\theta\left(b_{n,1,1-\alpha/2} \leq \frac{X_{(n)}}{\theta} \leq b_{n,1,\alpha/2}\right) &= P_\theta\left(\frac{1}{b_{n,1,1-\alpha/2}} \geq \frac{\theta}{X_{(n)}} \geq \frac{1}{b_{n,1,\alpha/2}}\right) \\ &= P_\theta\left(\frac{X_{(n)}}{b_{n,1,\alpha/2}} \leq \theta \leq \frac{X_{(n)}}{b_{n,1,1-\alpha/2}}\right). \end{aligned}$$

This shows that

$$\left(\frac{X_{(n)}}{b_{n,1,\alpha/2}}, \ \frac{X_{(n)}}{b_{n,1,1-\alpha/2}}\right)$$

is a $1 - \alpha$ confidence interval for θ.

Example 9.6. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$ and the x_i's are fixed constants (measured without error). Consider writing a confidence interval for

$$\theta = E(Y|x_0) = \beta_0 + \beta_1 x_0,$$

where x_0 is a specified value of x. In a linear models course, you have shown that

$$\widehat{\theta} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \sim \mathcal{N}\left(\theta, \ \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]\right),$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least-squares estimators of β_0 and β_1, respectively.

- If σ^2 is known (completely unrealistic), we can use

$$Q(\mathbf{Y}, \theta) = \frac{\widehat{\theta} - \theta}{\sqrt{\sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} \sim \mathcal{N}(0, 1)$$

  as a pivot to write a confidence interval for θ.

- More realistically, σ^2 is unknown and

$$Q(\mathbf{Y}, \theta) = \frac{\widehat{\theta} - \theta}{\sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} \sim t_{n-2},$$

  where MSE is the mean-squared error from the regression, is used as a pivot.

In the latter case, we can write

$$1 - \alpha = P_{\boldsymbol{\beta}, \sigma^2}\left(-t_{n-2,\alpha/2} \leq \frac{\widehat{\theta} - \theta}{\sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} \leq t_{n-2,\alpha/2}\right),$$

for all $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and σ^2. It follows that

$$\widehat{\theta} \pm t_{n-2,\alpha/2} \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}$$

is a $1 - \alpha$ confidence interval for θ.

Remark: As Examples 9.5 and 9.6 illustrate, interval estimates are easily obtained after writing $1 - \alpha = P_\theta(a \leq Q(\mathbf{X}, \theta) \leq b)$, for constants a and b (quantiles of Q). More generally, $\{\theta \in \Theta : Q(\mathbf{x}, \theta) \in \mathcal{A}\}$ is a set estimate for θ, where $\mathcal{A}$ satisfies $1 - \alpha = P_\theta(Q(\mathbf{X}, \theta) \in \mathcal{A})$. For example, in Example 9.5, we could have written

$$1 - \alpha = P_\theta\left(b_{n,1,1-\alpha} \leq \frac{X_{(n)}}{\theta} \leq 1\right) = P_\theta\left(X_{(n)} \leq \theta \leq \frac{X_{(n)}}{b_{n,1,1-\alpha}}\right)$$

and concluded that

$$\left(X_{(n)}, \frac{X_{(n)}}{b_{n,1,1-\alpha}}\right)$$

is a $1 - \alpha$ confidence interval for θ. How does this interval compare with

$$\left(\frac{X_{(n)}}{b_{n,1,\alpha/2}}, \frac{X_{(n)}}{b_{n,1,1-\alpha/2}}\right)?$$

Which one is "better?" For that matter, how should we define what "better" means?

Remark: The more general statement

$$1-\alpha = P_{\boldsymbol{\theta}}(Q(\mathbf{X},\boldsymbol{\theta}) \in \mathcal{A})$$

is especially useful when $\boldsymbol{\theta}$ is a vector and the goal is to find a confidence set (i.e., **confidence region**) for $\boldsymbol{\theta}$. In such cases, $\mathcal{A}$ will generally be a subset of $\mathbb{R}^k$ where $k = \dim(\boldsymbol{\theta})$.

Example 9.7. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Set $\boldsymbol{\theta} = (\mu, \sigma^2)$. We know that

$$Q_1 = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

that is, Q_1 is a pivot. Therefore,

$$\begin{aligned} 1-\alpha &= P_{\boldsymbol{\theta}}\left(-t_{n-1,\alpha/2} \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) \\ &= P_{\boldsymbol{\theta}}\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right), \end{aligned}$$

showing that

$$C_1(\mathbf{X}) = \left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

is a $1-\alpha$ confidence set for μ. Similarly, we know that

$$Q_2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

that is, Q_2 is also a pivot. Therefore,

$$\begin{aligned} 1-\alpha &= P_{\boldsymbol{\theta}}\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\right) \\ &= P_{\boldsymbol{\theta}}\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right), \end{aligned}$$

showing that

$$C_2(\mathbf{X}) = \left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}},\ \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

is a $1-\alpha$ confidence set for σ^2.

Extension: Suppose we wanted to write a confidence set (region) for $\boldsymbol{\theta} = (\mu, \sigma^2)$ in $\mathbb{R}^2$. From the individual pivots, we know that $C_1(\mathbf{X})$ and $C_2(\mathbf{X})$ are each $1-\alpha$ confidence sets.

Q: Is $C_1(\mathbf{X}) \times C_2(\mathbf{X})$, the Cartesian product of $C_1(\mathbf{X})$ and $C_2(\mathbf{X})$, a $1-\alpha$ confidence region for $\boldsymbol{\theta}$?
A: No. By Bonferroni's Inequality,

$$\begin{aligned} P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in C_1(\mathbf{X}) \times C_2(\mathbf{X})) &\geq P_{\boldsymbol{\theta}}(\mu \in C_1(\mathbf{X})) + P_{\boldsymbol{\theta}}(\sigma^2 \in C_2(\mathbf{X})) - 1 \\ &= (1-\alpha) + (1-\alpha) - 1 \\ &= 1-2\alpha. \end{aligned}$$

Therefore, $C_1(\mathbf{X}) \times C_2(\mathbf{X})$ is a $1-2\alpha$ confidence region for $\boldsymbol{\theta}$.

Bonferroni adjustment: Adjust $C_1(\mathbf{X})$ and $C_2(\mathbf{X})$ individually so that the confidence coefficient of each is $1-\alpha/2$. The adjusted set $C_1^*(\mathbf{X}) \times C_2^*(\mathbf{X})$ is a $1-\alpha$ confidence region for $\boldsymbol{\theta}$. This region has coverage probability larger than or equal to $1-\alpha$ for all $\boldsymbol{\theta}$ (so it is "conservative").

More interesting approach: Consider the quantity

$$Q = Q(\mathbf{X}, \boldsymbol{\theta}) = Q(\mathbf{X}, \mu, \sigma^2) = \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 + \frac{(n-1)S^2}{\sigma^2}.$$

It is easy to show that $Q \sim \chi^2_n$, establishing that Q is a pivot. Therefore, we can write

$$1-\alpha = P_{\boldsymbol{\theta}}(Q \leq \chi^2_{n,\alpha}) = P_{\boldsymbol{\theta}}\left(\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 + \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n,\alpha}\right),$$

which shows that

$$C(\mathbf{X}) = \{\boldsymbol{\theta} = (\mu, \sigma^2) : Q(\mathbf{X}, \mu, \sigma^2) \leq \chi^2_{n,\alpha}\}$$

is a $1-\alpha$ confidence region (in $\mathbb{R}^2$) for $\boldsymbol{\theta}$. To see that this set looks like, note that the boundary is

$$\begin{aligned} Q(\mathbf{x}, \mu, \sigma^2) = \chi^2_{n,\alpha} &\iff \left(\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}\right)^2 + \frac{(n-1)s^2}{\sigma^2} = \chi^2_{n,\alpha} \\ &\iff (\mu-\overline{x})^2 = \frac{\chi^2_{n,\alpha}}{n}\left[\sigma^2 - \frac{(n-1)s^2}{\chi^2_{n,\alpha}}\right], \end{aligned}$$

which is a parabola in $\Theta = \{\boldsymbol{\theta} = (\mu, \sigma^2) : -\infty < \mu < \infty,\ \sigma^2 > 0\}$. The parabola has vertex at

$$\left(\overline{x},\ \frac{(n-1)s^2}{\chi^2_{n,\alpha}}\right)$$

and it opens upward (because $\chi^2_{n,\alpha}/n > 0$). The confidence set is the interior of the parabola.

Discussion: Example 9.2.7 (CB, pp 427-428) provides tips on how to find pivots in location and scale (and location-scale) families.

Family	Parameter	Pivot examples
Location	μ	$\overline{X}-\mu$, $X_{(n)}-\mu$, $X_{(1)}-\mu$
Scale	σ	$\frac{\overline{X}}{\sigma}$, $\frac{X_{(n)}}{\sigma}$, $\frac{X_{(1)}}{\sigma}$

In general, **differences** are pivotal in location family problems; **ratios** are pivotal for scale parameters.

Exercise: Suppose $X_1, X_2, ..., X_n$ are iid from

$$f_X(x|\mu) = f_Z(x-\mu),$$

where $-\infty < \mu < \infty$ and $f_Z(\cdot)$ is a standard pdf. Show that $Q(\mathbf{X}, \mu) = \overline{X} - \mu$ is a pivotal quantity.

9.2.3 Pivoting the CDF

Example 9.8. Suppose $X_1, X_2, ..., X_n$ are iid with population pdf

$$f_X(x|\theta) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta \\ 0, & x < \theta, \end{cases}$$

where $-\infty < \theta < \infty$. How can we obtain a confidence set for θ? Note that, because $\{f_X(x|\theta) : -\infty < \theta < \infty\}$ is a location family, we could try working with $Q = Q(\mathbf{X}, \theta) = \overline{X} - \theta$. From our recent discussion, we know that Q is pivotal. In fact, it is easy to show that

$$X_i - \theta \sim \text{exponential}(1) \stackrel{d}{=} \text{gamma}(1, 1)$$

and hence

$$Q = \overline{X} - \theta = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta) \sim \text{gamma}(n, 1/n).$$

As expected, the distribution of Q is free of θ. Furthermore, $2nQ \sim \text{gamma}(n, 2) \stackrel{d}{=} \chi^2_{2n}$. Using $2nQ = 2n(\overline{X} - \theta)$ as a pivot, we can write

$$1 - \alpha = P_\theta(\chi^2_{2n,1-\alpha/2} \leq 2n(\overline{X} - \theta) \leq \chi^2_{2n,\alpha/2}) = P_\theta\left(\overline{X} - \frac{\chi^2_{2n,\alpha/2}}{2n} \leq \theta \leq \overline{X} - \frac{\chi^2_{2n,1-\alpha/2}}{2n}\right).$$

Therefore,

$$\left(\overline{X} - \frac{\chi^2_{2n,\alpha/2}}{2n},\ \overline{X} - \frac{\chi^2_{2n,1-\alpha/2}}{2n}\right)$$

is a $1 - \alpha$ confidence set for θ.

Criticism: Although this is a bonafide $1 - \alpha$ confidence set, it is not based on $T = T(\mathbf{X}) = X_{(1)}$, a sufficient statistic for θ. Let's find a pivot based on T instead. One example of such a pivot is $Q(T, \theta) = T - \theta \sim \text{exponential}(1/n)$. Another example is $Q(T, \theta) = F_T(T|\theta)$, the cdf of T, which is $\mathcal{U}(0, 1)$ by the Probability Integral Transformation.

CDF: It is easy to show that

$$F_T(t|\theta) = \begin{cases} 0, & t \leq \theta \\ 1 - e^{-n(t-\theta)}, & t > \theta. \end{cases}$$

Therefore, because $F_T(T|\theta) \sim \mathcal{U}(0, 1)$, we can write

$$\begin{aligned} 1 - \alpha &= P_\theta(\alpha/2 \leq F_T(T|\theta) \leq 1 - \alpha/2) \\ &= P_\theta(\alpha/2 \leq 1 - e^{-n(T-\theta)} \leq 1 - \alpha/2) \\ &= P_\theta\left(T + \frac{1}{n}\ln\left(\frac{\alpha}{2}\right) \leq \theta \leq T + \frac{1}{n}\ln\left(1 - \frac{\alpha}{2}\right)\right). \end{aligned}$$

Therefore,

$$\left(T + \frac{1}{n}\ln\left(\frac{\alpha}{2}\right),\ T + \frac{1}{n}\ln\left(1 - \frac{\alpha}{2}\right)\right)$$

is a $1 - \alpha$ confidence set for θ.

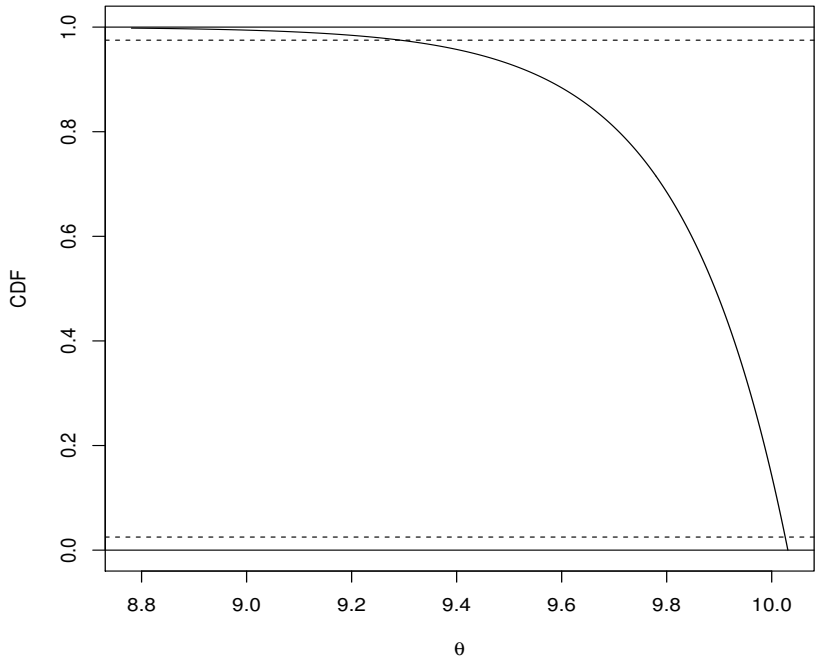

Figure 9.2: CDF of $T = X_{(1)}$ in Example 9.8, $F_T(t|\theta)$, plotted as a function of θ with t fixed. The value of t is 10.032, calculated based on an iid sample from $f_X(x|\theta)$ with $n = 5$. Dotted horizontal lines at $\alpha/2 = 0.025$ and $1 - \alpha/2 = 0.975$ have been added.

Special case: I used R to simulate an iid sample of size $n = 5$ from $f_X(x|\theta)$. The cdf of $T = X_{(1)}$ is plotted in Figure 9.2 as a function of θ with the observed value of $t = x_{(1)} = 10.032$ held fixed. A 0.95 confidence set is $(9.293, 10.026)$. The true value of θ is 10.

Theorem 9.2.12. Suppose T is a statistic with a continuous cdf $F_T(t|\theta)$. Suppose $\alpha_1 + \alpha_2 = \alpha$. Suppose for all $t \in \mathcal{T}$, the functions $\theta_L(t)$ and $\theta_U(t)$ are defined as follows:

- When $F_T(t|\theta)$ is a **decreasing** function of θ,
  - $F_T(t|\theta_U(t)) = \alpha_1$
  - $F_T(t|\theta_L(t)) = 1 - \alpha_2$.
- When $F_T(t|\theta)$ is an **increasing** function of θ,
  - $F_T(t|\theta_U(t)) = 1 - \alpha_2$
  - $F_T(t|\theta_L(t)) = \alpha_1$.

Then the random interval $(\theta_L(T), \theta_U(T))$ is a $1 - \alpha_1 - \alpha_2$ confidence set for θ.

Remark: Theorem 9.2.12 remains valid for any statistic T with continuous cdf. In practice, we would likely want T to be a sufficient statistic.

Remark: In practice, one often sets $\alpha_1 = \alpha_2 = \alpha/2$ so that $(\theta_L(T), \theta_U(T))$ is a $1-\alpha$ confidence set. This is not necessarily the "optimal" approach, but it is reasonable in most situations. One sided confidence sets are obtained by letting either α_1 or α_2 equal 0.

Remark: Pivoting the cdf always "works" because (if T is continuous), the cdf itself, when viewed as random, is a pivot. From the Probability Integral Transformation, we know that $F_T(T|\theta) \sim \mathcal{U}(0,1)$. Therefore, when $F_T(t|\theta)$ is a **decreasing** function of θ, we have

$$\begin{aligned} P_\theta(\theta_L(T) \leq \theta \leq \theta_U(T)) &= P_\theta(\alpha_1 \leq F_T(T|\theta) \leq 1-\alpha_2) \\ &= 1-\alpha_1-\alpha_2. \end{aligned}$$

The case wherein $F_T(t|\theta)$ is an increasing function of θ is analogous.

Implementation: To pivot the cdf, it is not necessary that $F_T(t|\theta)$ be available in closed form (as in Example 9.8). All we really have to do is solve

$$\int_{-\infty}^{t_0} f_T(t|\theta_1^*(t_0))dt \stackrel{\text{set}}{=} \alpha/2 \quad \text{and} \quad \int_{t_0}^{\infty} f_T(t|\theta_2^*(t_0))dt \stackrel{\text{set}}{=} \alpha/2$$

(in the equal $\alpha_1 = \alpha_2 = \alpha/2$ case, say), based on the observed value $T = t_0$. We solve these equations for $\theta_1^*(t_0)$ and $\theta_2^*(t_0)$. One of these will be the lower limit $\theta_L(t_0)$ and the other will be the upper limit $\theta_U(t_0)$, depending on whether $F_T(t|\theta)$ is an increasing or decreasing function of θ.

Remark: The discrete case (i.e., the statistic T has a discrete distribution) is handled in the same way except that the integrals above are replaced by sums.

Example 9.9. Suppose $X_1, X_2, ..., X_n$ are iid Poisson(θ), where $\theta > 0$. We now pivot the cdf of $T = \sum_{i=1}^n X_i$, a sufficient statistic, to write a $1-\alpha$ confidence set for θ. Recall that $T = \sum_{i=1}^n X_i \sim$ Poisson($n\theta$). If $T = t_0$ is observed, we set

$$\begin{aligned} P_\theta(T \leq t_0) = \sum_{k=0}^{t_0} \frac{(n\theta)^k e^{-n\theta}}{k!} &\stackrel{\text{set}}{=} \alpha/2 \\ P_\theta(T \geq t_0) = \sum_{k=t_0}^{\infty} \frac{(n\theta)^k e^{-n\theta}}{k!} &\stackrel{\text{set}}{=} \alpha/2 \end{aligned}$$

and solve each equation for θ. In practice, the solutions could be found by setting up a grid search over possible values of θ and then selecting the values that solve these equations (one solution will be the lower endpoint; the other solution will be the upper endpoint). In this example, however, it is possible to get closed-form expressions for the confidence set endpoints. To see why, we need to recall the following result which "links" the Poisson and gamma distributions.

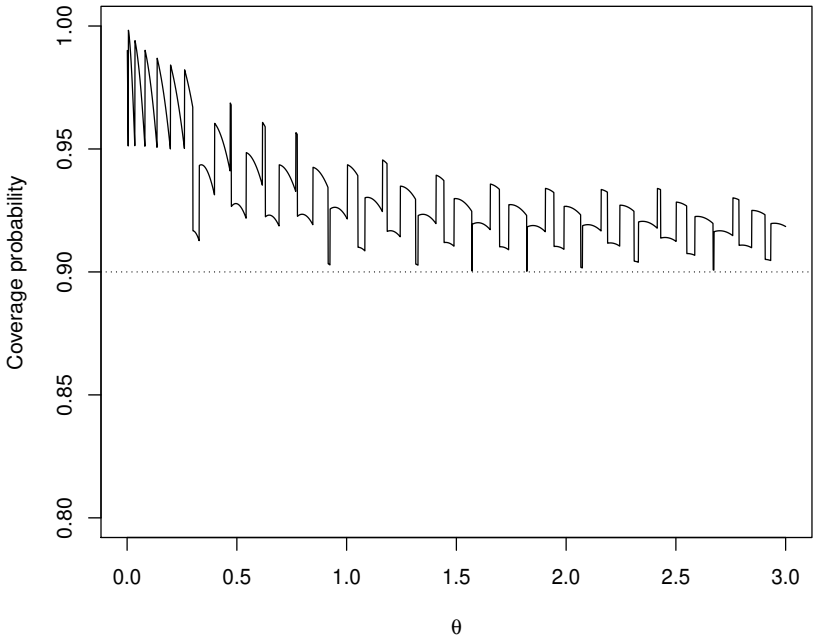

Figure 9.3: Coverage probability of the confidence interval in Example 9.9 when $n = 10$ and $\alpha = 0.10$. A dotted horizontal line at $1 - \alpha = 0.90$ has been added.

Result: If $X \sim \text{gamma}(a, b)$, $a \in \mathbb{N}$ (a positive integer), then

$$P(X \leq x) = P(Y \geq a),$$

where $Y \sim \text{Poisson}(x/b)$. This identity was stated in Example 3.3.1 (CB, pp 100-101).

Application: If we apply this result in Example 9.9 for the second equation to be solved, we have $a = t_0$, $x/b = n\theta$, and

$$\frac{\alpha}{2} \stackrel{\text{set}}{=} P_\theta(T \geq t_0) = P_\theta(X \leq bn\theta) = P_\theta\left(\frac{2X}{b} \leq 2n\theta\right) = P_\theta(\chi^2_{2t_0} \leq 2n\theta).$$

Therefore, we set

$$2n\theta = \chi^2_{2t_0,1-\alpha/2}$$

and solve for θ (this will give the lower endpoint). A similar argument shows that the upper endpoint solves

$$2n\theta = \chi^2_{2(t_0+1),\alpha/2}.$$

Therefore, a $1 - \alpha$ confidence set for θ is

$$\left(\frac{1}{2n}\chi^2_{2t_0,1-\alpha/2},\ \frac{1}{2n}\chi^2_{2(t_0+1),\alpha/2}\right).$$

Remark: When T is discrete, the coverage probability of $(\theta_L(T), \theta_U(T))$ found through pivoting the cdf will generally be a function of θ and the interval itself will be **conservative**, that is,

$$P_\theta(\theta_L(T) \leq \theta \leq \theta_U(T)) \geq 1 - \alpha,$$

for all $\theta \in \Theta$. This is true because when T is discrete, the cdf $F_T(T|\theta)$ is stochastically larger than a $\mathcal{U}(0,1)$ distribution. For example, consider Example 9.9 with $n = 10$ and $1 - \alpha = 0.90$. Figure 9.3 shows that the coverage probability of a nominal 0.90 confidence interval is always at least 0.90 and can, in fact, be much larger than 0.90.

Remark: Pivoting a discrete cdf can be used to write confidence sets for parameters in other discrete distributions. For example, a $1 - \alpha$ confidence interval for a binomial probability p when using this technique is given by

$$\left(\frac{1}{1 + \frac{n-x+1}{x} F_{2(n-x+1),2x,\alpha/2}}, \frac{\frac{x+1}{n-x} F_{2(x+1),2(n-x),\alpha/2}}{1 + \frac{x+1}{n-x} F_{2(x+1),2(n-x),\alpha/2}} \right),$$

where x is the realized value of $X \sim b(n,p)$ and $F_{a,b,\alpha/2}$ is the upper $\alpha/2$ quantile of an F distribution with degrees of freedom a and b. This is known as the **Clopper-Pearson** confidence interval for p and it can (not surprisingly) be very conservative; see Brown et al. (2001, *Statistical Science*). The interval arises by first exploiting the relationship between the binomial and beta distributions (see CB, Exercise 2.40, pp 82) and then the relationship which "links" the beta and F distributions (see CB, Theorem 5.3.8, pp 225).

9.2.4 Bayesian intervals

Recall: In the Bayesian paradigm, all inference is carried out using the posterior distribution $\pi(\theta|\mathbf{x})$. However, because the posterior $\pi(\theta|\mathbf{x})$ is itself a legitimate probability distribution (for θ, updated after seeing $\mathbf{x}$), we can calculate probabilities involving θ directly by using this distribution.

Definition: For any set $\mathcal{A} \subset \mathbb{R}$, the **credible probability** associated with $\mathcal{A}$ is

$$P(\theta \in \mathcal{A}|\mathbf{X} = \mathbf{x}) = \int_{\mathcal{A}} \pi(\theta|\mathbf{x}) d\theta.$$

If the credible probability is $1 - \alpha$, we call $\mathcal{A}$ a $1 - \alpha$ **credible set**. If $\pi(\theta|\mathbf{x})$ is discrete, we simply replace integrals with sums.

Note: Bayesian credible intervals are interpreted differently than confidence intervals.

- **Confidence interval interpretation:** "If we were to perform the experiment over and over again, each time under identical conditions, and if we calculated a $1 - \alpha$ confidence interval each time the experiment was performed, then $100(1 - \alpha)$ percent of the intervals we calculated would contain the true value of θ. Any specific interval we calculate represents one of these possible intervals."
- **Credible interval interpretation:** "The probability our interval contains θ is $1 - \alpha$."

Example 9.10. Suppose that $X_1, X_2, ..., X_n$ are iid Poisson(θ), where the prior distribution for $\theta \sim$ gamma(a, b), a, b known. In Example 7.10 (notes, pp 38-39), we showed that the posterior distribution

$$\theta|\mathbf{X} = \mathbf{x} \sim \text{gamma}\left(\sum_{i=1}^{n} x_i + a,\ \frac{1}{n + \frac{1}{b}}\right).$$

In Example 8.9 (notes, pp 77-78), we used this Bayesian model setup with the number of goals per game in the 2013-2014 English Premier League season and calculated the posterior distribution for the mean number of goals θ to be

$$\theta|\mathbf{X} = \mathbf{x} \sim \text{gamma}\left(1060 + 1.5,\ \frac{1}{380 + \frac{1}{2}}\right) \stackrel{d}{=} \text{gamma}(1061.5, 0.002628).$$

A 0.95 credible set for θ is $(2.62, 2.96)$.

```
> qgamma(0.025,1061.5,1/0.002628)
[1] 2.624309
> qgamma(0.975,1061.5,1/0.002628)
[1] 2.959913
```

Q: Why did we select the "equal-tail" quantiles (0.025 and 0.975) in this example?
A: It's easy!

Note: There are two types of Bayesian credible intervals commonly used: Equal-tail (ET) intervals and highest posterior density (HPD) intervals.

Definition: The set $\mathcal{A}$ is a **highest posterior density** (**HPD**) $1 - \alpha$ credible set if

$$\mathcal{A} = \{\theta : \pi(\theta|\mathbf{x}) \geq c\}$$

and the credible probability of $\mathcal{A}$ is $1 - \alpha$. ET and HPD intervals will coincide only when $\pi(\theta|\mathbf{x})$ is symmetric.

Remark: In practice, because Monte Carlo methods are often used to approximate posterior distributions, simple ET intervals are usually the preferred choice. HPD intervals can be far more difficult to construct and are rarely much better than ET intervals.

9.3 Methods of Evaluating Interval Estimators

Note: We will not cover all of the material in this subsection. We will have only a brief discussion of the relevant topics.

Evaluating estimators: When evaluating any interval estimator, there are two important criteria to consider:

1. **Coverage probability.** When the coverage probability is not equal to $1-\alpha$ for all $\theta \in \Theta$ (as is usually the case in discrete distributions), we would like it to be as close as possible to the nominal $1-\alpha$ level.
   - Some intervals maintain a coverage probability $\geq 1-\alpha$ for all $\theta \in \Theta$ but can be very conservative (as in Example 9.9).
   - Confidence intervals based on large-sample theory might confer a coverage probability $\leq 1-\alpha$ for some/all $\theta \in \Theta$, even though they are designed to be nominal as $n \to \infty$. Large-sample intervals are discussed in Chapter 10.
2. **Interval length.** Shorter intervals are more informative. Interval length (or expected interval length) depends on the interval's underlying confidence coefficient.
   - It only makes sense to compare two interval estimators (on the basis of interval length) when the intervals have the same coverage probability (or confidence coefficient).

Example 9.11. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Set $\boldsymbol{\theta} = (\mu, \sigma^2)$. A $1-\alpha$ confidence interval for μ is

$$C(\mathbf{X}) = \left(\overline{X} + a\frac{S}{\sqrt{n}},\ \overline{X} + b\frac{S}{\sqrt{n}} \right),$$

where the constants a and b are quantiles from the t_{n-1} distribution satisfying

$$1-\alpha = P_{\boldsymbol{\theta}}\left(\overline{X} + a\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + b\frac{S}{\sqrt{n}} \right).$$

Which choice of a and b is "best?" More precisely, which choice minimizes the **expected length**? The length of this interval is

$$L = (b-a)\frac{S}{\sqrt{n}},$$

which, of course, is random. The expected length is

$$E_{\boldsymbol{\theta}}(L) = (b-a)\frac{E_{\boldsymbol{\theta}}(S)}{\sqrt{n}} = (b-a)c(n)\sigma/\sqrt{n},$$

where the constant

$$c(n) = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\sqrt{n-1}\Gamma(\frac{n-1}{2})}.$$

Note that the expected length is **proportional** to $b-a$.

Theorem 9.3.2. Suppose $Q = Q(\mathbf{X}, \theta)$ is a pivotal quantity and

$$P_\theta(a \leq Q \leq b) = 1 - \alpha,$$

where a and b are constants. Let $f_Q(q)$ denote the pdf of Q. If

1. $\int_a^b f_Q(q)dq = 1 - \alpha$
2. $f_Q(a) = f_Q(b) > 0$
3. $f_Q'(a) > f_Q'(b)$,

then $b - a$ is minimized relative to Q.

Remark: The version of Theorem 9.3.2 stated in CB (pp 441-442) is slightly different than the one I present above; the authors' version requires that the pdf of Q be unimodal (mine requires that it be differentiable).

Application: Consider Example 9.11 with

$$Q = Q(\mathbf{X}, \boldsymbol{\theta}) = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

and

$$C(\mathbf{x}) = \left(\overline{x} + a\frac{s}{\sqrt{n}},\ \overline{x} + b\frac{s}{\sqrt{n}} \right).$$

If we choose $a = -t_{n-1,\alpha/2}$ and $b = t_{n-1,\alpha/2}$, then the conditions in Theorem 9.3.2 are satisfied. Therefore,

$$\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \right)$$

has the shortest expected length among all $1 - \alpha$ confidence intervals based on Q.

Proof of Theorem 9.3.2. Suppose $Q \sim f_Q(q)$, where

$$1 - \alpha = P_\theta(a \leq Q \leq b) = F_Q(b) - F_Q(a)$$

so that

$$F_Q(b) = 1 - \alpha + F_Q(a)$$

and

$$b = F_Q^{-1}[1 - \alpha + F_Q(a)] \equiv b(a), \text{ say.}$$

The goal is to minimize $b - a = b(a) - a$. Taking derivatives, we have (by the Chain Rule)

$$\begin{aligned} \frac{d}{da}[b(a) - a] &= \frac{d}{da}[F_Q^{-1}[1 - \alpha + F_Q(a)] - 1 \\ &= \frac{d}{da}[1 - \alpha + F_Q(a)]\ \frac{d}{d\eta}F_Q^{-1}(\eta) - 1, \end{aligned}$$

where $\eta = 1-\alpha+F_Q(a)$. However, note that by the inverse function theorem (from calculus),

$$\frac{d}{d\eta}F_Q^{-1}(\eta) = \frac{1}{F_Q'[F_Q^{-1}(\eta)]} = \frac{1}{F_Q'(b)} = \frac{1}{f_Q(b)}.$$

Therefore,

$$\frac{d}{da}[b(a)-a] = \frac{f_Q(a)}{f_Q(b)} - 1 \stackrel{\text{set}}{=} 0 \implies f_Q(a) = f_Q(b).$$

To finish the proof, all we need to show is that

$$\frac{d^2}{da^2}[b(a)-a] > 0$$

whenever $f_Q(a) = f_Q(b)$ and $f_Q'(a) > f_Q'(b)$. This will guarantee that the conditions stated in Theorem 9.3.2 lead to $b-a$ being minimized. □

Remark: The theorem we have just proven is applicable when an interval's length (or expected length) is proportional to $b-a$. This is often true when θ is a location parameter and $f_X(x|\theta)$ is a location family. When an interval's length is not proportional to $b-a$, then Theorem 9.3.2 is not directly applicable. However, we might be able to formulate a modified version of the theorem that is applicable.

Example 9.12. Suppose $X_1, X_2, ..., X_n$ are iid exponential(β), where $\beta > 0$. A pivotal quantity based on $T = T(\mathbf{X}) = \sum_{i=1}^n X_i$, a sufficient statistic, is

$$Q = Q(T,\beta) = \frac{2T}{\beta} \sim \chi^2_{2n}.$$

Therefore, we can write

$$1-\alpha = P_\beta(a \leq Q \leq b) = P_\beta\left(a \leq \frac{2T}{\beta} \leq b\right) = P_\beta\left(\frac{2T}{b} \leq \beta \leq \frac{2T}{a}\right),$$

where a and b are quantiles from the χ^2_{2n} distribution. In this example, the expected interval length is not proportional to $b-a$. Instead, the expected length is

$$E_\beta(L) = E_\beta\left(\frac{2T}{a} - \frac{2T}{b}\right) = \left(\frac{1}{a} - \frac{1}{b}\right)E_\beta(2T) = \left(\frac{1}{a} - \frac{1}{b}\right)2n\beta,$$

which is proportional to

$$\frac{1}{a} - \frac{1}{b}.$$

Theorem 9.3.2 is therefore not applicable here. To modify the theorem (towards finding a shortest expected length confidence interval based on Q), we would have to minimize

$$\frac{1}{a} - \frac{1}{b} = \frac{1}{a} - \frac{1}{b(a)}$$

with respect to a subject to the constraint that

$$\int_a^{b(a)} f_Q(q)dq = 1-\alpha,$$

where $f_Q(q)$ is the pdf of $Q \sim \chi^2_{2n}$. See CB (pp 444).

10 Asymptotic Evaluations

Complementary reading: Chapter 10 (CB).

10.1 Introduction

Preview: In this chapter, we revisit "large sample theory" and discuss three important topics in statistical inference:

1. **Point estimation** (Section 10.1)
   - Efficiency, consistency
   - Large sample properties of maximum likelihood estimators
2. **Hypothesis testing** (Section 10.3)
   - Wald, score, LRT
   - asymptotic distributions
3. **Confidence intervals** (Section 10.4)
   - Wald, score, LRT

Our previous inference discussions (i.e., in Chapters 7-9 CB) dealt with finite sample topics (i.e., unbiasedness, MSE, optimal estimators/tests, confidence intervals based on finite sample pivots, etc.). We now investigate **large sample inference**, a topic of utmost importance in statistical research.

10.2 Point Estimation

Setting: We observe $\mathbf{X} = (X_1, X_2, ..., X_n) \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Usually, $X_1, X_2, ..., X_n$ will constitute a random sample (an iid sample) from a population $f_X(x|\theta)$. We regard the scalar parameter θ as fixed and unknown. Define

$$W_n = W_n(\mathbf{X}) = W_n(X_1, X_2, ..., X_n)$$

to be a **sequence of estimators**. For example,

$$\begin{aligned} W_1 &= X_1 \\ W_2 &= \frac{X_1 + X_2}{2} \\ W_3 &= \frac{X_1 + X_2 + X_3}{3}, \end{aligned}$$

and so on, so that in general,

$$W_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note that we emphasize the dependence of this sequence on the sample size n.

Definition: A sequence of estimators W_n is **consistent** for a parameter θ if

$$W_n \xrightarrow{p} \theta \text{ for all } \theta \in \Theta.$$

That is, for all $\epsilon > 0$ and for all $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta(|W_n - \theta| \geq \epsilon) = 0.$$

An equivalent definition is

$$\lim_{n\to\infty} P_\theta(|W_n - \theta| < \epsilon) = 1.$$

We call W_n a **consistent estimator** of θ. What makes consistency "different" from our usual definition of convergence in probability is that we require $W_n \xrightarrow{p} \theta$ **for all** $\theta \in \Theta$. In other words, convergence of W_n must result for all members of the family $\{f_X(x|\theta) : \theta \in \Theta\}$.

Remark: From Markov's Inequality, we know that for all $\epsilon > 0$,

$$P_\theta(|W_n - \theta| \geq \epsilon) \leq \frac{E_\theta[(W_n - \theta)^2]}{\epsilon^2}.$$

Therefore, a sufficient condition for W_n to be consistent is

$$\frac{E_\theta[(W_n - \theta)^2]}{\epsilon^2} \to 0$$

for all $\theta \in \Theta$. However, note that

$$E_\theta[(W_n - \theta)^2] = \text{var}_\theta(W_n) + [E_\theta(W_n) - \theta]^2 = \text{var}_\theta(W_n) + [\text{Bias}_\theta(W_n)]^2.$$

This leads to the following theorem.

Theorem 10.1.3. If W_n is a sequence of estimators of a parameter θ satisfying

1. $\text{var}_\theta(W_n) \to 0$, as $n \to \infty$, for all $\theta \in \Theta$
2. $\text{Bias}_\theta(W_n) \to 0$, as $n \to \infty$, for all $\theta \in \Theta$,

then W_n is a consistent estimator of θ.

Weak Law of Large Numbers: Suppose that $X_1, X_2, ..., X_n$ are iid with $E_\theta(X_1) = \mu$ and $\text{var}_\theta(X_1) = \sigma^2 < \infty$. Let

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

denote the sample mean. As an estimator of μ, it is easy to see that the conditions of Theorem 10.1.3 are satisfied. Therefore, $\overline{X}_n$ is a consistent estimator of $E_\theta(X_1) = \mu$.

Continuity: Suppose W_n is a consistent estimator of θ. Suppose $g : \mathbb{R} \to \mathbb{R}$ is a continuous function. Then

$$g(W_n) \xrightarrow{p} g(\theta) \text{ for all } \theta \in \Theta.$$

That is, $g(W_n)$ is a consistent estimator of $g(\theta)$.

Consistency of MLEs: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta$. Let

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|\mathbf{x})$$

denote the maximum likelihood estimator (MLE) of θ. Under "certain regularity conditions," it follows that

$$\widehat{\theta} \xrightarrow{p} \theta \text{ for all } \theta \in \Theta,$$

as $n \to \infty$. That is, MLEs are consistent estimators.

Remark: Consistency also results for vector valued MLEs, say $\widehat{\boldsymbol{\theta}}$, but we herein restrict attention to the scalar case.

Sufficient conditions to prove consistency of MLEs:

1. $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$.
2. The parameter θ is **identifiable**, that is, for $\theta_1, \theta_2 \in \Theta$,

   $$f_X(x|\theta_1) = f_X(x|\theta_2) \implies \theta_1 = \theta_2.$$

   In other words, different values of θ cannot produce the same probability distribution.
3. The family of pdfs $\{f_X(x|\theta) : \theta \in \Theta\}$ has common support $\mathcal{X}$. This means that the support does not depend on θ. In addition, the pdf $f_X(x|\theta)$ is differentiable with respect to θ.
4. The parameter space Θ contains an open set where the true value of θ, say θ_0, resides as an interior point.

Remark: Conditions 1-4 generally hold for exponential families that are of full rank.

Example 10.1. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(0, \theta)$, where $\theta > 0$. The MLE of θ is

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

As an MLE, $\widehat{\theta} \xrightarrow{p} \theta$, for all $\theta > 0$; i.e., $\widehat{\theta}$ is a consistent estimator of θ.

Asymptotic normality of MLEs: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta$. Let $\widehat{\theta}$ denote the MLE of θ. Under "certain regularity conditions," it follows that

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

as $n \to \infty$, where the asymptotic variance

$$v(\theta) = \frac{1}{I_1(\theta)}.$$

Recall that $I_1(\theta)$, the **Fisher Information** based on one observation, is given by

$$I_1(\theta) = E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\} = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right].$$

Remark: The four regularity conditions on the last page were sufficient conditions for **consistency**. For **asymptotic normality**, there are two additional sufficient conditions:

5. The pdf/pmf $f_X(x|\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ, and $\int_{\mathbb{R}} f_X(x|\theta)dx$ can be differentiated three times under the integral sign.
6. There exists a function $M(x)$ such that
$$\left|\frac{\partial^3}{\partial\theta^3}\ln f_X(x|\theta)\right| \leq M(x)$$
for all $x \in \mathcal{X}$ for all $\theta \in N_c(\theta_0)$ $\exists c > 0$ and $E_{\theta_0}[M(X)] < \infty$.

Note: We now sketch a **casual proof** of the asymptotic normality result for MLEs. Let θ_0 denote the true value of θ. Let $S(\theta) = S(\theta|\mathbf{x})$ denote the score function; i.e.,

$$S(\theta) = \frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{x}|\theta).$$

Note that because $\widehat{\theta}$ is an MLE, it solves the score equation; i.e., $S(\widehat{\theta}) = 0$. Therefore, we can write (via Taylor series expansion about θ_0),

$$\begin{aligned} 0 &= S(\widehat{\theta}) \\ &= S(\theta_0) + \frac{\partial S(\theta_0)}{\partial\theta}(\widehat{\theta}-\theta_0) + \frac{1}{2}\frac{\partial^2 S(\widehat{\theta}_*)}{\partial\theta^2}(\widehat{\theta}-\theta_0)^2 \end{aligned}$$

where $\widehat{\theta}_*$ is between θ_0 and $\widehat{\theta}$. Therefore, we have

$$0 = S(\theta_0) + (\widehat{\theta}-\theta_0)\left[\frac{\partial S(\theta_0)}{\partial\theta} + \frac{1}{2}\frac{\partial^2 S(\widehat{\theta}_*)}{\partial\theta^2}(\widehat{\theta}-\theta_0)\right].$$

After simple algebra, we have

$$\begin{aligned} \sqrt{n}(\widehat{\theta}-\theta_0) &= \frac{-\sqrt{n}S(\theta_0)}{\frac{\partial S(\theta_0)}{\partial\theta} + \frac{1}{2}\frac{\partial^2 S(\widehat{\theta}_*)}{\partial\theta^2}(\widehat{\theta}-\theta_0)} \\ &= \frac{-\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta_0)}{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\ln f_X(X_i|\theta_0) + \frac{1}{2n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\ln f_X(X_i|\widehat{\theta}_*)(\widehat{\theta}-\theta_0)} = \frac{-\mathrm{A}}{\mathrm{B}+\mathrm{C}}, \end{aligned}$$

where

$$\begin{aligned} \text{A} &= \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta_0) \\ \text{B} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\ln f_X(X_i|\theta_0) \\ \text{C} &= \frac{1}{2n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\ln f_X(X_i|\widehat{\theta}_*)(\widehat{\theta}-\theta_0). \end{aligned}$$

The **first** term

$$\text{A} = \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta_0) \xrightarrow{d} \mathcal{N}(0, I_1(\theta_0)).$$

Proof. For general θ, define

$$Y_i = \frac{\partial}{\partial\theta}\ln f_X(X_i|\theta),$$

for $i = 1, 2, ..., n$. The Y_i's are iid with mean

$$\begin{aligned} E_\theta(Y) = E_\theta\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right] &= \int_{\mathbb{R}}\frac{\partial}{\partial\theta}\ln f_X(x|\theta) f_X(x|\theta)dx \\ &= \int_{\mathbb{R}}\frac{\partial}{\partial\theta} f_X(x|\theta)dx \\ &= \frac{d}{d\theta}\int_{\mathbb{R}} f_X(x|\theta)dx = 0 \end{aligned}$$

and variance

$$\text{var}_\theta(Y) = \text{var}_\theta\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right] = E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\} = I_1(\theta).$$

Therefore, the CLT says that

$$\text{A} = \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(X_i|\theta) = \sqrt{n}(\overline{Y}-0) \xrightarrow{d} \mathcal{N}(0, I_1(\theta)),$$

as $n \to \infty$. Note that when $\theta = \theta_0$, we have

$$-\text{A} = -\sqrt{n}(\overline{Y}-0) \xrightarrow{d} \mathcal{N}(0, I_1(\theta_0)),$$

because the $\mathcal{N}(0, I_1(\theta))$ limiting distribution above is symmetric about 0. □

The **second** term, by WLLN,

$$\text{B} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\ln f_X(X_i|\theta_0) \xrightarrow{p} E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta_0)\right] = -I_1(\theta_0).$$

The **third** term

$$\text{C} = \frac{1}{2n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\ln f_X(X_i|\widehat{\theta}_*)(\widehat{\theta}-\theta_0) \xrightarrow{p} 0.$$

Proof (very casual). We have

$$\text{C} = \frac{1}{2}(\widehat{\theta}-\theta_0)\,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\ln f_X(X_i|\widehat{\theta}_*).$$

Note that $\widehat{\theta}-\theta_0 \xrightarrow{p} 0$, because $\widehat{\theta}$ is consistent (i.e., $\widehat{\theta}$ converges in probability to θ_0). Therefore, it suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\ln f_X(X_i|\widehat{\theta}_*)$$

converges to something finite (in probability). Note that for n "large enough;" i.e., as soon as $\widehat{\theta}_* \in N_c(\theta_0)$ in Regularity Condition 6,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\ln f_X(X_i|\widehat{\theta}_*) \le \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3}{\partial\theta^3}\ln f_X(X_i|\widehat{\theta}_*)\right| \le \frac{1}{n}\sum_{i=1}^{n}M(X_i) \xrightarrow{p} E_{\theta_0}[M(X)] < \infty. \;\square$$

We have shown that $\text{C} \xrightarrow{p} 0$ and hence $\text{B}+\text{C} \xrightarrow{p} -I_1(\theta_0)$. Finally, note that

$$\frac{-\text{A}}{\text{B}+\text{C}} = \underbrace{-\text{A}}_{\xrightarrow{d}\mathcal{N}(0,I_1(\theta_0))}\;\underbrace{\frac{1}{\text{B}+\text{C}}}_{\xrightarrow{p}-\frac{1}{I_1(\theta_0)}} \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_1(\theta_0)}\right),$$

by Slutsky's Theorem. $\square$

Remark: We have shown that, under regularity conditions, an MLE $\widehat{\theta}$ satisfies

$$\sqrt{n}(\widehat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

where

$$v(\theta) = \frac{1}{I_1(\theta)}.$$

Now recall the **Delta Method** from Chapter 5; i.e., if $g:\mathbb{R}\to\mathbb{R}$ is differentiable at θ and $g'(\theta)\neq 0$, then

$$\sqrt{n}[g(\widehat{\theta})-g(\theta)] \xrightarrow{d} \mathcal{N}\left(0, [g'(\theta)]^2 v(\theta)\right).$$

Therefore, not only are MLEs asymptotically normal, but functions of MLEs are too.

Example 10.1 (continued). Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(0,\theta)$, where $\theta > 0$. The MLE of θ is

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}X_i^2.$$

We know that $\widehat{\theta} \xrightarrow{p} \theta$, as $n \to \infty$. We now derive the asymptotic distribution of $\widehat{\theta}$ (suitably centered and scaled). We know

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

where

$$v(\theta) = \frac{1}{I_1(\theta)}.$$

Therefore, all we need to do is calculate $I_1(\theta)$. The pdf of X is, for all $x \in \mathbb{R}$,

$$f_X(x|\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-x^2/2\theta}.$$

Therefore,

$$\ln f_X(x|\theta) = -\frac{1}{2}\ln(2\pi\theta) - \frac{x^2}{2\theta}.$$

The derivatives of $\ln f_X(x|\theta)$ are

$$\begin{aligned} \frac{\partial}{\partial\theta}\ln f_X(x|\theta) &= -\frac{1}{2\theta} + \frac{x^2}{2\theta^2} \\ \frac{\partial^2}{\partial\theta^2}\ln f_X(x|\theta) &= \frac{1}{2\theta^2} - \frac{x^2}{\theta^3}. \end{aligned}$$

Therefore,

$$I_1(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right] = E_\theta\left(\frac{X^2}{\theta^3} - \frac{1}{2\theta^2}\right) = \frac{\theta}{\theta^3} - \frac{1}{2\theta^2} = \frac{1}{2\theta^2}$$

and

$$v(\theta) = \frac{1}{I_1(\theta)} = 2\theta^2.$$

We have

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 2\theta^2).$$

Exercise: Use the Delta Method to derive the large sample distributions of $g_1(\widehat{\theta}) = \widehat{\theta}^2$, $g_2(\widehat{\theta}) = e^{\widehat{\theta}}$, and $g_3(\widehat{\theta}) = \ln\widehat{\theta}$, suitably centered and scaled.

Important: Suppose that an MLE $\widehat{\theta}$ (or any sequence of estimators) satisfies

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)).$$

Suppose that $v(\widehat{\theta})$ is a consistent estimator of $v(\theta)$, that is,

$$v(\widehat{\theta}) \xrightarrow{p} v(\theta),$$

for all $\theta \in \Theta$ as $n \to \infty$. We know that

$$Z_n = \frac{\widehat{\theta} - \theta}{\sqrt{\dfrac{v(\theta)}{n}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

In addition,

$$Z_n^* = \frac{\widehat{\theta}-\theta}{\sqrt{\frac{v(\widehat{\theta})}{n}}} = \underbrace{\frac{\widehat{\theta}-\theta}{\sqrt{\frac{v(\theta)}{n}}}}_{\xrightarrow{d}\mathcal{N}(0,1)} \underbrace{\sqrt{\frac{v(\theta)}{v(\widehat{\theta})}}}_{\xrightarrow{p}1} \xrightarrow{d} \mathcal{N}(0,1),$$

by Slutsky's Theorem. Note that

$$\sqrt{\frac{v(\theta)}{v(\widehat{\theta})}} \xrightarrow{p} 1$$

because of continuity. This technique is widely used in large sample arguments.

Summary:

1. We start with a sequence of estimators (e.g., an MLE sequence, etc.) satisfying
$$\sqrt{n}(\widehat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)).$$
2. We find a consistent estimator of the asymptotic variance, say $v(\widehat{\theta})$.
3. Slutsky's Theorem and continuity of convergence are used to show that
$$Z_n^* = \frac{\widehat{\theta}-\theta}{\sqrt{\frac{v(\widehat{\theta})}{n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

One can then use Z_n^* to formulate large sample (Wald) hypothesis tests and confidence intervals; see Sections 10.3 and 10.4, respectively.

Example 10.1 (continued). Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(0, \theta)$, where $\theta > 0$. The MLE of θ is

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

We have shown that

$$\sqrt{n}(\widehat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0, 2\theta^2) \iff Z_n = \frac{\widehat{\theta}-\theta}{\sqrt{\frac{2\theta^2}{n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

A consistent estimator of $v(\theta) = 2\theta^2$ is $v(\widehat{\theta}) = 2\widehat{\theta}^2$, by continuity. Therefore,

$$Z_n^* = \frac{\widehat{\theta}-\theta}{\sqrt{\frac{2\widehat{\theta}^2}{n}}} = \underbrace{\frac{\widehat{\theta}-\theta}{\sqrt{\frac{2\theta^2}{n}}}}_{\xrightarrow{d}\mathcal{N}(0,1)} \underbrace{\sqrt{\frac{2\theta^2}{2\widehat{\theta}^2}}}_{\xrightarrow{p}1} \xrightarrow{d} \mathcal{N}(0,1),$$

by Slutsky's Theorem.

Definition: Suppose we have two competing sequences of estimators (neither of which is necessarily an MLE sequence) denoted by W_n and V_n that satisfy

$$\begin{aligned}\sqrt{n}(W_n - \theta) &\xrightarrow{d} \mathcal{N}(0, \sigma_W^2) \\ \sqrt{n}(V_n - \theta) &\xrightarrow{d} \mathcal{N}(0, \sigma_V^2).\end{aligned}$$

Both estimators are consistent estimators of θ. Define the **asymptotic relative efficiency** (**ARE**) as

$$\text{ARE}(W_n \text{ to } V_n) = \frac{\sigma_W^2}{\sigma_V^2}.$$

With this definition, the following interpretations are used:

1. If ARE < 1, then W_n is more efficient than V_n.
2. If ARE $= 1$, then W_n is as efficient as V_n.
3. If ARE > 1, then W_n is less efficient than V_n.

The ARE is commonly used to compare the variances of two competing consistent estimators; the comparison is of course on the basis of each estimator's large sample distribution.

Remark: Before we do an example illustrating ARE, let's have a brief discussion about sample quantile estimators.

Sample quantiles: Suppose $X_1, X_2, ..., X_n$ are iid with continuous cdf F. Define

$$\phi_p = F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

We call ϕ_p the p**th quantile** of the distribution of X. Note that if F is strictly increasing, then $F^{-1}(p)$ is well defined by

$$\phi_p = F^{-1}(p) \iff F(\phi_p) = p.$$

The simplest definition of the sample pth quantile is $\widehat{F}_n^{-1}(p)$, where

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

is the **empirical distribution function** (**edf**). The edf is a non-decreasing step function that takes steps of size $1/n$ at each observed X_i. Therefore,

$$\widehat{\phi}_p \equiv \widehat{F}_n^{-1}(p) = \begin{cases} X_{(np)}, & np \in \mathbb{Z}^+ \\ X_{(\lfloor np \rfloor + 1)}, & \text{otherwise.} \end{cases}$$

This is just a fancy way of saying that the sample pth quantile is one of the order statistics (note that other books may define this differently; e.g., by averaging order statistics, etc.). Whenever I teach STAT 823, I prove that

$$\sqrt{n}(\widehat{\phi}_p - \phi_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(\phi_p)}\right),$$

where f is the population pdf of X. For example, if $p = 0.5$, then $\phi_p = \phi_{0.5}$ is the median of X and the sample median $\widehat{\phi}_{0.5}$ satisfies

$$\sqrt{n}(\widehat{\phi}_{0.5} - \phi_{0.5}) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4f^2(\phi_{0.5})}\right).$$

Example 10.2. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Consider the following two estimators $W_n = \overline{X}_n$ and $V_n = \widehat{\phi}_{0.5}$ as estimators of μ. Note that because the $\mathcal{N}(\mu, \sigma^2)$ population distribution is symmetric, the population median $\phi_{0.5} = \mu$ as well.

We know that

$$\sqrt{n}(\overline{X}_n - \mu) \stackrel{d}{=} \mathcal{N}(0, \sigma^2),$$

that is, this "limiting distribution" is the exact distribution of $\sqrt{n}(\overline{X}_n - \mu)$ for each n. From our previous discussion on sample quantiles, we know that

$$\sqrt{n}(\widehat{\phi}_{0.5} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4f^2(\phi_{0.5})}\right),$$

where (under the normal assumption),

$$\frac{1}{4f^2(\phi_{0.5})} = \frac{1}{4f^2(\mu)} = \frac{1}{4\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^2} = \frac{\pi}{2}\sigma^2.$$

Therefore, the asymptotic relative efficiency of the sample median $\widehat{\phi}_{0.5}$ when compared to the sample mean $\overline{X}_n$ is

$$\text{ARE}(\widehat{\phi}_{0.5} \text{ to } \overline{X}_n) = \frac{\frac{\pi}{2}\sigma^2}{\sigma^2} = \frac{\pi}{2} \approx 1.57.$$

Interpretation: The sample median $\widehat{\phi}_{0.5}$ would require 57 percent more observations to achieve the same level of (asymptotic) precision as $\overline{X}_n$.

Example 10.3. Suppose $X_1, X_2, ..., X_n$ are iid beta$(\theta, 1)$, where $\theta > 0$.

- Show that the MOM estimator of θ is

$$\widehat{\theta}_{\text{MOM}} = \frac{\overline{X}}{1 - \overline{X}}$$

and that $\widehat{\theta}_{\text{MOM}}$ satisfies

$$\sqrt{n}(\widehat{\theta}_{\text{MOM}} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\theta(\theta+1)^2}{\theta+2}\right).$$

Hint: Use CLT and Delta Method.

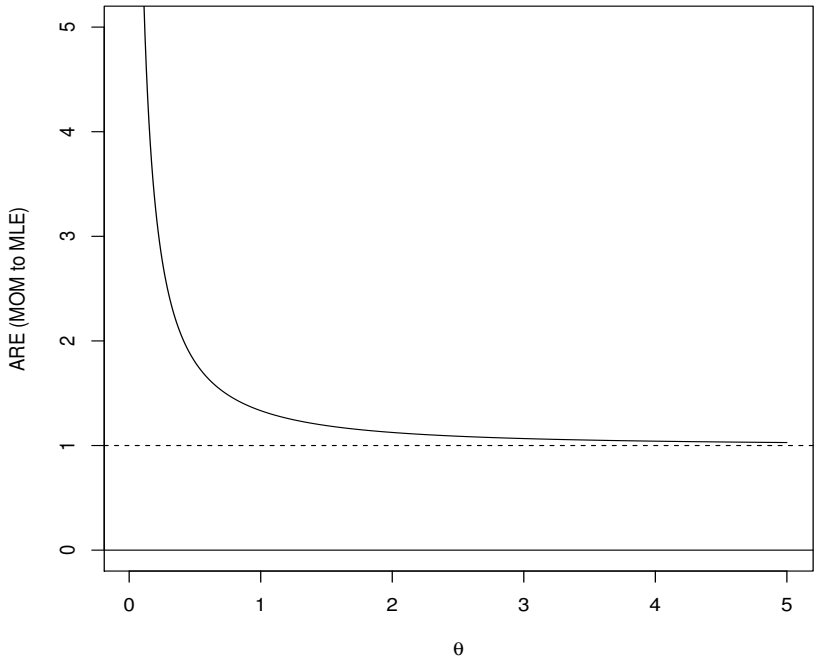

Figure 10.1: Plot of ARE($\widehat{\theta}_{\text{MOM}}$ to $\widehat{\theta}_{\text{MLE}}$) versus θ in Example 10.3.

- Show that the MLE of θ is
$$\widehat{\theta}_{\text{MLE}} = -\frac{n}{\sum_{i=1}^{n} \ln X_i}$$
and that $\widehat{\theta}_{\text{MLE}}$ satisfies
$$\sqrt{n}(\widehat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2).$$
Hint: Use large sample results for MLEs.

- Show that
$$\text{ARE}(\widehat{\theta}_{\text{MOM}} \text{ to } \widehat{\theta}_{\text{MLE}}) = \frac{(\theta+1)^2}{\theta(\theta+2)}.$$

- I graphed ARE($\widehat{\theta}_{\text{MOM}}$ to $\widehat{\theta}_{\text{MLE}}$) as a function of θ in Figure 10.1. Note that ARE is always greater than unity; i.e., the MOM estimator is not as efficient as the MLE.

10.3 Hypothesis Testing

Remark: In Chapter 8 (CB), we discussed methods to derive hypothesis tests and also optimality issues based on finite sample criteria. These discussions revealed that optimal tests (e.g., UMP tests) were available for just a small collection of problems (some of which were not realistic).

Preview: In this section, we present three large sample approaches to formulate hypothesis tests:

1. Wald (1943)
2. Score (1947, Rao); also known as "Lagrange multiplier tests"
3. Likelihood ratio (1928, Neyman-Pearson).

These are known as the "large sample likelihood based tests."

10.3.1 Wald tests

Recall: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. As long as suitable regularity conditions hold, we know that an MLE $\widehat{\theta}$ satisfies

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

where

$$v(\theta) = \frac{1}{I_1(\theta)}.$$

If $v(\theta)$ is a continuous function of θ, then

$$v(\widehat{\theta}) \xrightarrow{p} v(\theta),$$

for all θ; i.e., $v(\widehat{\theta})$ is a consistent estimator of $v(\theta)$, and

$$Z_n^* = \frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\widehat{\theta})}{n}}} = \underbrace{\frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\theta)}{n}}}}_{\xrightarrow{d}\mathcal{N}(0,1)} \underbrace{\sqrt{\frac{v(\theta)}{v(\widehat{\theta})}}}_{\xrightarrow{p}1} \xrightarrow{d} \mathcal{N}(0,1),$$

by Slutsky's Theorem. This forms the basis for the Wald test.

Wald statistic: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Consider testing

$$H_0 : \theta = \theta_0$$
versus
$$H_1 : \theta \neq \theta_0.$$

When H_0 is true, then

$$Z_n^W = \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{v(\widehat{\theta})}{n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Therefore,

$$R = \{\mathbf{x} \in \mathcal{X} : |z_n^W| \geq z_{\alpha/2}\},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution, is an approximate size α rejection region for testing H_0 versus H_1. One sided tests also use Z_n^W. The only thing that changes is the form of R.

Example 10.4. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$. Derive the Wald test of

$$H_0 : p = p_0$$
versus
$$H_1 : p \neq p_0.$$

Solution. We already know that the MLE of p is given by

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

the so-called "sample proportion." Because $\widehat{p}$ is an MLE, we know that

$$\sqrt{n}(\widehat{p} - p) \xrightarrow{d} \mathcal{N}(0, v(p)),$$

where

$$v(p) = \frac{1}{I_1(p)}.$$

We now calculate $I_1(p)$. The pmf of X is, for $x = 0, 1$,

$$f_X(x|p) = p^x(1-p)^{1-x}.$$

Therefore,

$$\ln f_X(x|p) = x \ln p + (1-x)\ln(1-p).$$

The derivatives of $\ln f_X(x|p)$ are

$$\begin{aligned} \frac{\partial}{\partial p} \ln f_X(x|p) &= \frac{x}{p} - \frac{1-x}{1-p} \\ \frac{\partial^2}{\partial p^2} \ln f_X(x|p) &= -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}. \end{aligned}$$

Therefore,

$$I_1(p) = -E_p\left[\frac{\partial^2}{\partial p^2} \ln f_X(X|p)\right] = E_p\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}$$

and

$$v(p) = \frac{1}{I_1(p)} = p(1-p).$$

We have

$$\sqrt{n}(\widehat{p} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

Because the asymptotic variance $v(p) = p(1-p)$ is a continuous function of p, it can be consistently estimated by $v(\widehat{p}) = \widehat{p}(1-\widehat{p})$. The Wald statistic to test $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ is given by

$$Z_n^W = \frac{\widehat{p} - p_0}{\sqrt{\frac{v(\widehat{p})}{n}}} = \frac{\widehat{p} - p_0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}.$$

An approximate size α rejection region is

$$R = \{\mathbf{x} \in \mathcal{X} : |z_n^W| \geq z_{\alpha/2}\}.$$

10.3.2 Score tests

Motivation: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Recall that the score function, when viewed as random, is

$$\begin{aligned} S(\theta|\mathbf{X}) &= \frac{\partial}{\partial\theta} \ln L(\theta|\mathbf{X}) \\ &\stackrel{\text{iid}}{=} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f_X(X_i|\theta), \end{aligned}$$

the sum of iid random variables. Recall that

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta} \ln f_X(X|\theta)\right] &= 0 \\ \text{var}_\theta\left[\frac{\partial}{\partial\theta} \ln f_X(X|\theta)\right] &= E_\theta\left\{\left[\frac{\partial}{\partial\theta} \ln f_X(X|\theta)\right]^2\right\} = I_1(\theta). \end{aligned}$$

Therefore, applying the CLT to the sum above, we have

$$\sqrt{n}\left(\frac{1}{n}S(\theta|\mathbf{X}) - 0\right) \xrightarrow{d} \mathcal{N}(0, I_1(\theta)),$$

which means

$$\frac{\frac{1}{n}S(\theta|\mathbf{X})}{\sqrt{\frac{I_1(\theta)}{n}}} = \frac{S(\theta|\mathbf{X})}{\sqrt{nI_1(\theta)}} \stackrel{\text{iid}}{=} \frac{S(\theta|\mathbf{X})}{\sqrt{I_n(\theta)}} \xrightarrow{d} \mathcal{N}(0,1),$$

where recall $I_n(\theta) = nI_1(\theta)$ is the Fisher information based on all n iid observations. Therefore, the score function divided by the square root of the Fisher information (based on all n observations) behaves asymptotically like a $\mathcal{N}(0,1)$ random variable. This fact forms the basis for the score test.

Score statistic: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Consider testing

$$H_0 : \theta = \theta_0$$
versus
$$H_1 : \theta \neq \theta_0.$$

When H_0 is true, then

$$Z_n^S = \frac{S(\theta_0|\mathbf{X})}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Therefore,

$$R = \{\mathbf{x} \in \mathcal{X} : |z_n^S| \geq z_{\alpha/2}\},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution, is an approximate size α rejection region for testing H_0 versus H_1. One sided tests also use Z_n^S. The only thing that changes is the form of R.

Example 10.5. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$. Derive the score test of

$$H_0 : p = p_0$$
versus
$$H_1 : p \neq p_0.$$

Solution. The likelihood function is given by

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

The log-likelihood function is

$$\ln L(p|\mathbf{x}) = \sum_{i=1}^{n} x_i \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p).$$

The score function is

$$S(p|\mathbf{x}) = \frac{\partial}{\partial p} \ln L(p|\mathbf{x}) = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p}.$$

Recall in Example 10.4, we calculated

$$I_1(p) = \frac{1}{p(1-p)}.$$

Therefore, the score statistic is

$$Z_n^S = \frac{S(p_0|\mathbf{X})}{\sqrt{I_n(p_0)}} = \frac{\frac{\sum_{i=1}^n X_i}{p_0} - \frac{n - \sum_{i=1}^n X_i}{1-p_0}}{\sqrt{\frac{n}{p_0(1-p_0)}}} = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

An approximate size α rejection region is

$$R = \{\mathbf{x} \in \mathcal{X} : |z_n^S| \geq z_{\alpha/2}\}.$$

Remark: It is insightful to compare

$$Z_n^W = \frac{\widehat{p} - p_0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \quad \text{with} \quad Z_n^S = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

The two statistics differ only in how the **standard error** of $\widehat{p}$ (as a point estimator of p) is calculated. The Wald statistic uses the estimated standard error. The score statistic uses the standard error calculated under the assumption that $H_0 : p = p_0$ is true (i.e., nothing is being estimated). This is an argument in favor of the score statistic.

10.3.3 Likelihood ratio tests

Setting: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Consider testing

$$H_0 : \theta = \theta_0$$
versus
$$H_1 : \theta \neq \theta_0.$$

The likelihood ratio test (LRT) statistic is defined as

$$\lambda(\mathbf{x}) = \frac{\sup\limits_{\theta \in \Theta_0} L(\theta|\mathbf{x})}{\sup\limits_{\theta \in \Theta} L(\theta|\mathbf{x})} = \frac{L(\theta_0|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})} = \frac{L(\theta_0)}{L(\widehat{\theta})}.$$

Suppose the regularity conditions needed for MLEs to be consistent and asymptotically normal hold. When H_0 is true,

$$-2\ln\lambda(\mathbf{X}) \xrightarrow{d} \chi_1^2.$$

Because small values of $\lambda(\mathbf{x})$ are evidence against H_0, large values of $-2\ln\lambda(\mathbf{x})$ are too. Therefore,

$$R = \{\mathbf{x} \in \mathcal{X} : -2\ln\lambda(\mathbf{x}) \geq \chi_{1,\alpha}^2\},$$

where $\chi_{1,\alpha}^2$ is the upper α quantile of the χ_1^2 distribution, is an approximate size α rejection region for testing H_0 versus H_1.

Proof. Our proof is casual. Suppose $H_0 : \theta = \theta_0$ is true. First, write $\ln L(\widehat{\theta})$ in a Taylor series expansion about θ_0, that is,

$$\begin{aligned}
\ln L(\widehat{\theta}) &= \ln L(\theta_0) + (\widehat{\theta} - \theta_0)\frac{\partial}{\partial\theta}\ln L(\theta_0) + \frac{1}{2}(\widehat{\theta} - \theta_0)^2\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_*) \\
&= \ln L(\theta_0) + \sqrt{n}(\widehat{\theta} - \theta_0)\underbrace{\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ln L(\theta_0)}_{\text{see Equation (10.2)}} + \frac{n}{2}(\widehat{\theta} - \theta_0)^2\,\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_*), \quad (10.1)
\end{aligned}$$

where $\widehat{\theta}_*$ is between $\widehat{\theta}$ and θ_0. Now write $\frac{\partial}{\partial\theta}\ln L(\theta_0)$ in a Taylor series expansion about $\widehat{\theta}$, that is,

$$\frac{\partial}{\partial\theta}\ln L(\theta_0) = \underbrace{\frac{\partial}{\partial\theta}\ln L(\widehat{\theta})}_{=\,0} + (\theta_0-\widehat{\theta})\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_{**}),$$

where $\widehat{\theta}_{**}$ is between θ_0 and $\widehat{\theta}$. Note that $\frac{\partial}{\partial\theta}\ln L(\widehat{\theta}) = 0$ because $\widehat{\theta}$ solves the score equation. From the last equation, we have

$$\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ln L(\theta_0) = \sqrt{n}(\widehat{\theta}-\theta_0)\left\{-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_{**})\right\}. \tag{10.2}$$

Combining Equations (10.1) and (10.2), we have

$$\ln L(\widehat{\theta}) = \ln L(\theta_0) + \sqrt{n}(\widehat{\theta}-\theta_0)\sqrt{n}(\widehat{\theta}-\theta_0)\left\{-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_{**})\right\} + \frac{n}{2}(\widehat{\theta}-\theta_0)^2\,\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_*)$$

so that

$$\ln L(\widehat{\theta}) - \ln L(\theta_0) = n(\widehat{\theta}-\theta_0)^2\left\{-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_{**})\right\} + \frac{n}{2}(\widehat{\theta}-\theta_0)^2\left\{\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_*)\right\}. \tag{10.3}$$

Because $\widehat{\theta}$ is consistent (and because H_0 is true), we know that $\widehat{\theta} \xrightarrow{p} \theta_0$, as $n\to\infty$. Therefore, because $\widehat{\theta}_*$ and $\widehat{\theta}_{**}$ are both trapped between $\widehat{\theta}$ and θ_0, both terms in the brackets, i.e.,

$$\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_{**}) \quad \text{and} \quad \frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L(\widehat{\theta}_*)$$

converge in probability to

$$E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right] = -I_1(\theta_0),$$

by the WLLN. Therefore, the RHS of Equation (10.3) will behave in the limit the same as

$$\begin{aligned}\frac{n}{2}(\widehat{\theta}-\theta_0)^2 I_1(\theta_0) &= \frac{1}{2}\,\sqrt{n}(\widehat{\theta}-\theta_0)\sqrt{n}(\widehat{\theta}-\theta_0)I_1(\theta_0)\\ &= \frac{1}{2}\,\underbrace{\frac{\sqrt{n}(\widehat{\theta}-\theta_0)}{\sqrt{\frac{1}{I_1(\theta_0)}}}}_{\xrightarrow{d}\mathcal{N}(0,1)}\,\underbrace{\frac{\sqrt{n}(\widehat{\theta}-\theta_0)}{\sqrt{\frac{1}{I_1(\theta_0)}}}}_{\xrightarrow{d}\mathcal{N}(0,1)} \xrightarrow{d} \frac{1}{2}\,\chi^2_1,\end{aligned}$$

by continuity. Therefore, when $H_0:\theta=\theta_0$ is true,

$$-2\ln\lambda(\mathbf{X}) = -2[\ln L(\theta_0) - \ln L(\widehat{\theta})] \xrightarrow{d} \chi^2_1. \;\square$$

Example 10.6. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$. Derive the large sample LRT test of

$$H_0 : p = p_0$$
versus
$$H_1 : p \neq p_0.$$

Solution. The likelihood ratio statistic is

$$\begin{aligned}\lambda(\mathbf{x}) = \frac{L(p_0|\mathbf{x})}{L(\widehat{p}|\mathbf{x})} &= \frac{p_0^{\sum_{i=1}^n x_i}(1-p_0)^{n-\sum_{i=1}^n x_i}}{\widehat{p}^{\sum_{i=1}^n x_i}(1-\widehat{p})^{n-\sum_{i=1}^n x_i}} \\ &= \left(\frac{p_0}{\widehat{p}}\right)^{\sum_{i=1}^n x_i}\left(\frac{1-p_0}{1-\widehat{p}}\right)^{n-\sum_{i=1}^n x_i}.\end{aligned}$$

Therefore,

$$\begin{aligned}-2\ln\lambda(\mathbf{X}) &= -2\left[\sum_{i=1}^n X_i \ln\left(\frac{p_0}{\widehat{p}}\right) + \left(n - \sum_{i=1}^n X_i\right)\ln\left(\frac{1-p_0}{1-\widehat{p}}\right)\right] \\ &= -2\left[n\widehat{p}\ \ln\left(\frac{p_0}{\widehat{p}}\right) + n(1-\widehat{p})\ \ln\left(\frac{1-p_0}{1-\widehat{p}}\right)\right].\end{aligned}$$

An approximate size α rejection region is

$$R = \{\mathbf{x} \in \mathcal{X} : -2\ln\lambda(\mathbf{x}) \geq \chi^2_{1,\alpha}\}.$$

Monte Carlo Simulation: When $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$, we have derived the Wald, score, and large sample LRT for testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. Each test is a large sample test, so the size of each one is approximately equal to α when n is large. We now perform a simulation to assess finite sample characteristics.

- Take $n = 20$, $n = 50$, $n = 100$
- Let $p_0 = 0.1$ and $p_0 = 0.3$
- At each configuration of n and p_0, we will
  - simulate $B = 10000$ Bernoulli(p_0) samples (i.e., H_0 is true)
  - calculate z_n^W, z_n^S, and $-2\ln\lambda(\mathbf{x})$ with each sample
  - record the percentage of times that H_0 is (incorrectly) rejected when $\alpha = 0.05$
  - this percentage is an estimate of the **true size** of the test (for a given configuration of n and p_0).

The results from this simulation study are shown in Table 10.1.

		Wald	Score	LRT
$p_0 = 0.1$	$n = 20$	0.1204	0.0441	0.1287
	$n = 50$	0.1189	0.0316	0.0627
	$n = 100$	0.0716	0.0682	**0.0456**
$p_0 = 0.3$	$n = 20$	**0.0538**	0.0243	**0.0538**
	$n = 50$	0.0646	**0.0447**	**0.0447**
	$n = 100$	**0.0506**	0.0637	**0.0506**

Table 10.1: Monte Carlo simulation. Size estimates of nominal $\alpha = 0.05$ Wald, score, and LRTs for a binomial proportion p when $n = 20, 50, 100$ and $p_0 = 0.1, 0.3$.

Important: Note that these sizes are really estimates of the true sizes (at each setting of n and p_0). Therefore, we should acknowledge that these are estimates and report the margin of error associated with them.

- Because these are nominal size 0.05 tests, the margin of error associated with each "estimate," assuming a 99 percent confidence level, is equal to

$$B = 2.58\sqrt{\frac{0.05(1-0.05)}{10000}} \approx 0.0056.$$

- Size estimates between 0.0444 and 0.0556 indicate that the test is operating at the nominal level. I have bolded the estimates in Table 10.1 that are within these bounds.
- Values <0.0444 suggest conservatism (the test rejects too often). Values >0.0556 suggest anti-conservatism (the test is not rejecting often enough).

Summary: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Assume that the regularity conditions needed for MLEs to be consistent and asymptotically normal (**CAN**) hold. We have presented three large sample procedures to test

$$H_0 : \theta = \theta_0$$
$$\text{versus}$$
$$H_1 : \theta \neq \theta_0.$$

- **Wald:**

$$Z_n^W = \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{v(\widehat{\theta})}{n}}} = \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{1}{nI_1(\widehat{\theta})}}} \xrightarrow{d} \mathcal{N}(0,1)$$

- **Score:**

$$Z_n^S = \frac{S(\theta_0|\mathbf{X})}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0,1)$$

- **LRT:**
$$-2\ln\lambda(\mathbf{X}) = -2[\ln L(\theta_0|\mathbf{X}) - \ln L(\widehat{\theta}|\mathbf{X})] \xrightarrow{d} \chi^2_1.$$

All convergence results are under $H_0 : \theta = \theta_0$.

- Note that $(Z_n^W)^2$, $(Z_n^S)^2$, and $-2\ln\lambda(\mathbf{X})$ each converge in distribution to a χ^2_1 distribution as $n \to \infty$.
- In terms of power (i.e., rejecting H_0 when H_1 is true), all three testing procedures are **asymptotically equivalent** when examining certain types of alternative sequences (i.e., Pitman sequences of alternatives). For these alternative sequences, $(Z_n^W)^2$, $(Z_n^S)^2$, and $-2\ln\lambda(\mathbf{X})$ each converge to the same (noncentral) $\chi^2_1(\lambda)$ distribution. However, the powers may be quite different in finite samples.

Remark: The large sample LRT procedure can be easily generalized to multi-parameter hypotheses.

Theorem 10.3.3. Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. Assume that the regularity conditions needed for MLEs to be CAN hold. Consider testing

$$H_0 : \boldsymbol{\theta} \in \Theta_0$$
$$\text{versus}$$
$$H_1 : \boldsymbol{\theta} \in \Theta_0^c$$

and define
$$\lambda(\mathbf{x}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta}|\mathbf{x})}{\sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\mathbf{x})} = \frac{L(\widehat{\boldsymbol{\theta}}_0|\mathbf{x})}{L(\widehat{\boldsymbol{\theta}}|\mathbf{x})}.$$

If $\boldsymbol{\theta} \in \Theta_0$, then
$$-2\ln\lambda(\mathbf{X}) = -2[\ln L(\widehat{\boldsymbol{\theta}}_0|\mathbf{X}) - \ln L(\widehat{\boldsymbol{\theta}}|\mathbf{X})] \xrightarrow{d} \chi^2_\nu,$$
where $\nu = \dim(\Theta) - \dim(\Theta_0)$, the number of "free parameters" between Θ and Θ_0.

Implication: Rejecting $H_0 : \boldsymbol{\theta} \in \Theta_0$ when $\lambda(\mathbf{x})$ is small is equivalent to rejecting H_0 when $-2\ln\lambda(\mathbf{x})$ is large. Therefore,

$$R = \{\mathbf{x} \in \mathcal{X} : -2\ln\lambda(\mathbf{x}) \geq \chi^2_{\nu,\alpha}\}$$

is an approximate size α rejection region. This means

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}}(\text{Reject } H_0) = \alpha \text{ for all } \boldsymbol{\theta} \in \Theta_0.$$

Example 10.7. McCann and Tebbs (2009) summarize a study examining perceived unmet need for dental health care for people with HIV infection. Baseline in-person interviews were

conducted with 2,864 HIV infected individuals (aged 18 years and older) as part of the HIV Cost and Services Utilization Study. Define

$$\begin{aligned} X_1 &= \text{number of patients with private insurance} \\ X_2 &= \text{number of patients with medicare and private insurance} \\ X_3 &= \text{number of patients without insurance} \\ X_4 &= \text{number of patients with medicare but no private insurance.} \end{aligned}$$

Set $\mathbf{X} = (X_1, X_2, X_3, X_4)$ and model $\mathbf{X} \sim \text{mult}(2864, p_1, p_2, p_3, p_4; \sum_{i=1}^4 p_i = 1)$. Under this assumption, consider testing

$$H_0 : p_1 = p_2 = p_3 = p_4 = \tfrac{1}{4}$$
versus
$$H_1 : H_0 \text{ not true.}$$

The null parameter space is

$$\Theta_0 = \{\boldsymbol{\theta} = (p_1, p_2, p_3, p_4) : \ p_1 = p_2 = p_3 = p_4 = 1/4\},$$

the singleton $(1/4, 1/4, 1/4, 1/4)$. The entire parameter space is

$$\Theta = \left\{ \boldsymbol{\theta} = (p_1, p_2, p_3, p_4) : 0 < p_1 < 1, 0 < p_2 < 1, 0 < p_3 < 1, 0 < p_4 < 1; \sum_{i=1}^4 p_i = 1 \right\},$$

a simplex in $\mathbb{R}^4$. The number of "free parameters" is $\nu = \dim(\Theta) - \dim(\Theta_0) = 3 - 0 = 3$. The observed data from the study are summarized by

$$\mathbf{x} = (658, 839, 811, 556).$$

The likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{x}) = L(p_1, p_2, p_3, p_4|\mathbf{x}) = \frac{2864!}{x_1!\ x_2!\ x_3!\ x_4!}\ p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4}.$$

Maximizing $L(p_1, p_2, p_3, p_4|\mathbf{x})$ over Θ, noting that $p_4 = 1 - p_1 - p_2 - p_3$, gives the (unrestricted) maximum likelihood estimates

$$\widehat{p}_1 = \frac{x_1}{2864}, \qquad \widehat{p}_2 = \frac{x_2}{2864}, \qquad \widehat{p}_3 = \frac{x_3}{2864}, \qquad \widehat{p}_4 = \frac{x_4}{2864}.$$

Therefore,

$$\begin{aligned} \lambda(\mathbf{x}) = \lambda(x_1, x_2, x_3, x_4) &= \frac{L(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})}{L(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \widehat{p}_4)} \\ &= \frac{\frac{2864!}{x_1!\ x_2!\ x_3!\ x_4!}\ (\frac{1}{4})^{x_1}(\frac{1}{4})^{x_2}(\frac{1}{4})^{x_3}(\frac{1}{4})^{x_4}}{\frac{2864!}{x_1!\ x_2!\ x_3!\ x_4!}\ (\frac{x_1}{2864})^{x_1}(\frac{x_2}{2864})^{x_2}(\frac{x_3}{2864})^{x_3}(\frac{x_4}{2864})^{x_4}} = \prod_{i=1}^4 \left(\frac{2864}{4x_i}\right)^{x_i}. \end{aligned}$$

The large sample LRT statistic is

$$-2\ln\lambda(\mathbf{x}) = -2\sum_{i=1}^{4} x_i \ln\left(\frac{2864}{4x_i}\right) \approx 75.69.$$

An approximate size $\alpha = 0.05$ rejection region is

$$R = \{\mathbf{x} \in \mathcal{X} : -2\ln\lambda(\mathbf{x}) \geq 7.81\}.$$

Therefore, we have very strong evidence against H_0.

10.4 Confidence Intervals

Remark: In Chapter 9 (CB), we discussed methods to derive confidence intervals based on exact (i.e., finite sample) distributions. We now present three large sample approaches:

1. Wald
2. Score
3. Likelihood ratio.

These are known as the "large sample likelihood based confidence intervals."

Definition: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. The random variable

$$Q_n = Q_n(\mathbf{X}, \theta)$$

is called a **large sample pivot** if its asymptotic distribution is free of all unknown parameters. If Q_n is a large sample pivot and if

$$P_\theta(Q_n(\mathbf{X}, \theta) \in \mathcal{A}) \approx 1 - \alpha,$$

then $C(\mathbf{X}) = \{\theta : Q_n(\mathbf{X}, \theta) \in \mathcal{A}\}$ is called an **approximate** $1-\alpha$ confidence set for θ.

10.4.1 Wald intervals

Recall: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. As long as suitable regularity conditions hold, we know that an MLE $\widehat{\theta}$ satisfies

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

where

$$v(\theta) = \frac{1}{I_1(\theta)}.$$

If $v(\theta)$ is a continuous function of θ, then $v(\widehat{\theta}) \xrightarrow{p} v(\theta)$, for all θ; i.e., $v(\widehat{\theta})$ is a consistent estimator of $v(\theta)$, and

$$Q_n(\mathbf{X}, \theta) = \frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\widehat{\theta})}{n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

by Slutsky's Theorem. Therefore, $Q_n(\mathbf{X}, \theta)$ is a large sample pivot and

$$\begin{aligned} 1-\alpha &\approx P_\theta(-z_{\alpha/2} \leq Q_n(\mathbf{X}, \theta) \leq z_{\alpha/2}) \\ &= P_\theta\left(-z_{\alpha/2} \leq \frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\widehat{\theta})}{n}}} \leq z_{\alpha/2}\right) = P_\theta\left(\widehat{\theta} - z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}} \leq \theta \leq \widehat{\theta} + z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}}\right). \end{aligned}$$

Therefore,

$$\widehat{\theta} \pm z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}}$$

is an approximate $1-\alpha$ confidence interval for θ.

Remark: We could have arrived at this same interval by inverting the large sample test of

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_1 : \theta \neq \theta_0 \end{gathered}$$

that uses the (Wald) test statistic

$$Z_n^W = \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{v(\widehat{\theta})}{n}}}$$

and rejection region

$$R = \{\mathbf{x} \in \mathcal{X} : |z_n^W| \geq z_{\alpha/2}\}.$$

This is why this type of large sample interval is called a **Wald confidence interval** (it is the interval that arises from inverting a large sample Wald test).

Extension: We can also write large sample Wald confidence intervals for functions of θ using the Delta Method. Recall that if $g : \mathbb{R} \to \mathbb{R}$ is differentiable at θ and $g'(\theta) \neq 0$, then

$$\sqrt{n}[g(\widehat{\theta}) - g(\theta)] \xrightarrow{d} \mathcal{N}\left(0, [g'(\theta)]^2 v(\theta)\right).$$

If $[g'(\theta)]^2 v(\theta)$ is a continuous function of θ, then we can find a consistent estimator for it, namely $[g'(\widehat{\theta})]^2 v(\widehat{\theta})$, because MLEs are consistent themselves and consistency is preserved under continuous mappings. Therefore,

$$Q_n(\mathbf{X}, \theta) = \frac{g(\widehat{\theta}) - g(\theta)}{\sqrt{\frac{[g'(\widehat{\theta})]^2 v(\widehat{\theta})}{n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

by Slutsky's Theorem and

$$g(\widehat{\theta}) \pm z_{\alpha/2}\sqrt{\frac{[g'(\widehat{\theta})]^2 v(\widehat{\theta})}{n}}$$

is an approximate $1-\alpha$ confidence interval for $g(\theta)$.

Example 10.8. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$.

(a) Derive a $1-\alpha$ (large sample) Wald confidence interval for p.

(b) Derive a $1-\alpha$ (large sample) Wald confidence interval for

$$g(p) = \ln\left(\frac{p}{1-p}\right),$$

the log odds of p.

Solution. (a) We already know that the MLE of p is given by

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

In Example 10.4, we showed that

$$v(p) = \frac{1}{I_1(p)} = p(1-p).$$

Therefore,

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is an approximate $1-\alpha$ Wald confidence interval for p. The problems with this interval (i.e., in conferring the nominal coverage probability) are well known; see Brown et al. (2001).

(b) Note that $g(p) = \ln[p/(1-p)]$ is a differentiable function and

$$g'(p) = \frac{1}{p(1-p)} \neq 0.$$

The Delta Method gives

$$\begin{aligned}\sqrt{n}\left[\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) - \ln\left(\frac{p}{1-p}\right)\right] &\xrightarrow{d} \mathcal{N}\left(0, \left[\frac{1}{p(1-p)}\right]^2 p(1-p)\right) \\ &\stackrel{d}{=} \mathcal{N}\left(0, \frac{1}{p(1-p)}\right).\end{aligned}$$

Because the asymptotic variance $1/p(1-p)$ can be consistently estimated by $1/\widehat{p}(1-\widehat{p})$, we have

$$\frac{\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) - \ln\left(\frac{p}{1-p}\right)}{\sqrt{\frac{1}{n\widehat{p}(1-\widehat{p})}}} \xrightarrow{d} \mathcal{N}(0,1)$$

by Slutsky's Theorem, and

$$\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm z_{\alpha/2}\sqrt{\frac{1}{n\widehat{p}(1-\widehat{p})}}$$

is an approximate $1-\alpha$ Wald confidence interval for $g(p) = \ln[p/(1-p)]$.

Remarks: As you can see, constructing (large sample) Wald confidence intervals is straightforward. We rely on the MLE being consistent and asymptotically normal (CAN) and also on being able to find a consistent estimator of the asymptotic variance of the MLE.

- More generally, if you have an estimator $\widehat{\theta}$ (not necessarily an MLE) that is asymptotically normal and if you can estimate its (large sample) variance consistently, you can do Wald inference. This general strategy for large sample inference is ubiquitous in statistical research.
- The problem, of course, is that because large sample standard errors must be estimated, the performance of Wald confidence intervals (and tests) can be poor in small samples. Brown et al. (2001) highlights this for the binomial proportion; however, this behavior is seen in other settings.
- I view Wald inference as a "fall back." It is what to do when no other large sample inference procedures are available; i.e., "having something is better than nothing."
- Of course, in very large sample settings (e.g., large scale Phase III clinical trials, public health studies with thousands of individuals, etc.), Wald inference is usually the default approach (probably because of its simplicity) and is generally satisfactory.

10.4.2 Score intervals

Recall: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. We have shown previously that

$$Q_n(\mathbf{X}, \theta) = \frac{S(\theta|\mathbf{X})}{\sqrt{I_n(\theta)}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $I_n(\theta) = nI_1(\theta)$ is the Fisher information based on the sample.

Motivation: Score confidence intervals arise from inverting (large sample) score tests. Recall that in testing $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$, the score statistic

$$Q_n(\mathbf{X}, \theta_0) = \frac{S(\theta_0|\mathbf{X})}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0,1)$$

when H_0 is true. Therefore,

$$R = \{\mathbf{x} \in \mathcal{X} : |Q_n(\mathbf{x}, \theta_0)| \geq z_{\alpha/2}\}$$

is an approximate size α rejection region for testing H_0 versus H_1. The acceptance region is

$$A = R^c = \{\mathbf{x} \in \mathcal{X} : |Q_n(\mathbf{x}, \theta_0)| < z_{\alpha/2}\}.$$

From inverting this acceptance region, we can conclude that

$$C(\mathbf{x}) = \{\theta : |Q_n(\mathbf{x}, \theta)| < z_{\alpha/2}\}$$

is an approximate $1-\alpha$ confidence set for θ. If $C(\mathbf{x})$ is an interval, then we call it a **score confidence interval**.

Example 10.9. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$. Derive a $1-\alpha$ (large sample) score confidence interval for p.
Solution. From Example 10.5, we have

$$Q_n(\mathbf{X}, p) = \frac{S(p|\mathbf{X})}{\sqrt{I_n(p)}} = \frac{\frac{\sum_{i=1}^n X_i}{p} - \frac{n - \sum_{i=1}^n X_i}{1-p}}{\sqrt{\frac{n}{p(1-p)}}} = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}.$$

From our discussion above, the (random) set

$$C(\mathbf{X}) = \{p : |Q_n(\mathbf{X}, p)| < z_{\alpha/2}\} = \left\{ p : \left| \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \right| < z_{\alpha/2} \right\}$$

forms the score interval for p. After observing $\mathbf{X} = \mathbf{x}$, this interval could be calculated numerically (e.g., using a grid search over values of p that satisfy this inequality). However, in the binomial case, we can get closed-form expressions for the endpoints. To see why, note that the boundary

$$|Q_n(\mathbf{x}, p)| = z_{\alpha/2} \iff (\widehat{p} - p)^2 = z_{\alpha/2}^2 \frac{p(1-p)}{n}.$$

After algebra, this equation becomes

$$\left(1 + \frac{z_{\alpha/2}^2}{n}\right) p^2 - \left(2\widehat{p} + \frac{z_{\alpha/2}^2}{n}\right) p + \widehat{p}^2 = 0.$$

The LHS of the last equation is a quadratic function of p. The roots of this equation, if they are real, delimit the score interval for p. Using the quadratic formula, the lower and upper limits are

$$p_L = \frac{(2\widehat{p} + z^2_{\alpha/2}/n) - \sqrt{(2\widehat{p} + z^2_{\alpha/2}/n)^2 - 4(1 + z^2_{\alpha/2}/n)\widehat{p}^2}}{2(1 + z^2_{\alpha/2}/n)}$$

$$p_U = \frac{(2\widehat{p} + z^2_{\alpha/2}/n) + \sqrt{(2\widehat{p} + z^2_{\alpha/2}/n)^2 - 4(1 + z^2_{\alpha/2}/n)\widehat{p}^2}}{2(1 + z^2_{\alpha/2}/n)},$$

respectively. Note that the score interval is much more complex than the Wald interval. However, the score interval (in this setting and elsewhere) typically confers very good coverage probability, that is, close to the nominal $1 - \alpha$ level, even for small samples. Therefore, although we have added complexity, the score interval is typically much better.

10.4.3 Likelihood ratio intervals

Recall: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. The LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\theta_0|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})}$$

and

$$R = \{\mathbf{x} \in \mathcal{X} : -2\ln\lambda(\mathbf{x}) \geq \chi^2_{1,\alpha}\}$$

is an approximate size α rejection region for testing H_0 versus H_1. Inverting the acceptance region,

$$C(\mathbf{x}) = \left\{\theta : -2\ln\left[\frac{L(\theta|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})}\right] < \chi^2_{1,\alpha}\right\}$$

is an approximate $1 - \alpha$ confidence set for θ. If $C(\mathbf{x})$ is an interval, then we call it a **likelihood ratio confidence interval**.

Example 10.10. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$. Derive a $1 - \alpha$ (large sample) likelihood ratio confidence interval for p.
Solution. From Example 10.6, we have

$$-2\ln\left[\frac{L(p|\mathbf{x})}{L(\widehat{p}|\mathbf{x})}\right] = -2\left[n\widehat{p}\ \ln\left(\frac{p}{\widehat{p}}\right) + n(1-\widehat{p})\ \ln\left(\frac{1-p}{1-\widehat{p}}\right)\right].$$

Therefore, the confidence interval is

$$C(\mathbf{x}) = \left\{p : -2\left[n\widehat{p}\ \ln\left(\frac{p}{\widehat{p}}\right) + n(1-\widehat{p})\ \ln\left(\frac{1-p}{1-\widehat{p}}\right)\right] < \chi^2_{1,\alpha}\right\}.$$

This interval must be calculated using numerical search methods.