Probability Theory and Statistics 3
Reader

Course 2024-2025

dr. Bram Wouters

Original manuscript (in Dutch): prof. dr. Cees Diks
Translation: Joeri den Heijer

Amsterdam, October 2024

Contents

Chapter 7

Limit Theorems

In the previous chapter, we learned how to derive the pdf and CDF of functions of random variables. However, in many cases, it is impossible to find a closed-form expression for the pdf or CDF. With a large number of observations, though, it is possible to approximate the CDF. Before proceeding with this, we first need to acquire some tools.

7.2 Sequences of Random Variables

Consider a sequence of random variables $Y_1, Y_2, \ldots$ with a corresponding sequence of CDF's $G_1(y)$, $G_2(y), \ldots$ so that

$$G_n(y) = \mathbb{P}\left[Y_n \leq y\right], \qquad n = 1, 2, ...$$

Definition 7.2.1.

If $Y_n \sim G_n(y)$ for each $n = 1, 2, \ldots$, and if for some CDF $G(y)$

$$\lim_{n\to\infty} G_n(y) = G(y)$$

for all values y at which $G(y)$ is continuous, then the sequence $Y_1, Y_2, \ldots$ is said to **converge in distribution** to the random variable Y, denoted by $Y_n \xrightarrow{d} Y$. The CDF $G(y)$ is called the **limiting distribution** of Y_n.

Example 7.2.1.

Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution, $X_i \sim$ UNIF$(0, 1)$, and let $Y_n := \max\limits_{1\leq i\leq n} X_i$. Then, the CDF of Y_n is

$$G_n(y) = \begin{cases} 0, & y \leq 0 \\ y^n, & 0 < y < 1 \\ 1, & y \geq 1 \end{cases} \Rightarrow \lim_{n\to\infty} G_n(y) = \begin{cases} 0, & y < 1 \\ 1, & y \geq 1. \end{cases}$$

This is the CDF of a degenerate random variable Y, with $\mathbb{P}\left[Y = 1\right] = 1$.

Definition 7.2.2

The function $G(y)$ is the CDF of a degenerate random variable with value c if

$$G(y) = \begin{cases} 0, & y < c \\ 1, & y \geq c, \end{cases} \qquad \text{in other words } \mathbb{P}\left[Y = c\right] = 1.$$

Definition 7.2.3

A sequence of random variables $Y_1, Y_2, \ldots$, is said to **converge stochastically** to a constant c if

$$\lim_{n\to\infty} G_n(y) = G(y) = \begin{cases} 0, & y < c \\ 1, & y \geq c. \end{cases}$$

Later on, we will consider an alternative and easier formulation of stochastic convergence. Not all limiting distributions are degenerate, as we will see in the next example. The following two limits can be useful in many problems:

$$\lim_{n\to\infty}\left(1 + \frac{c}{n}\right)^{nb} = e^{cb} \quad \text{and} \quad \lim_{n\to\infty}\left(1 + \frac{c}{n} + \frac{d(n)}{n}\right)^{nb} = e^{cb} \quad \text{if} \quad \lim_{n\to\infty} d(n) = 0.$$

Example 7.2.3.

Suppose that $X_1, \ldots, X_n$ is a random sample from a Pareto distribution, $X_i \sim \text{PAR}(1,1)$. The pdf and CDF of X_i are given by:

$$f_X(x) = \begin{cases} \dfrac{1}{(1+x)^2}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$F_X(x) = \int_0^x \frac{1}{(1+t)^2}\, dt = -\frac{1}{1+t}\bigg|_0^x = 1 - \frac{1}{1+x}, \quad x > 0; \quad 0 \text{ otherwise.}$$

Now, let $Y_n := n \min\limits_{1\leq i\leq n} X_i$. From this, it follows that:

$$\begin{aligned} \mathbb{P}\left[Y_n \leq y\right] &= \mathbb{P}\left[\min_{1\leq i\leq n} X_i \leq \frac{y}{n}\right] \\ &= 1 - \left(\mathbb{P}\left[X_1 > \frac{y}{n}\right]\right)^n \\ &= 1 - \left(1 + \frac{y}{n}\right)^{-n} = G_n(y) \underset{n\to\infty}{\longrightarrow} 1 - e^{-y}, \end{aligned}$$

which is the CDF of an exponential distribution, EXP(1).

Now, an example follows where no limiting distribution exists.

Example 7.2.4

Now, let

$$Y_n := \max_{1\leq i\leq n} X_i \Rightarrow G_n(y) = \begin{cases} \left(\frac{y}{1+y}\right)^n, y > 0 \\ 0, \quad \text{otherwise} \end{cases} \Rightarrow \lim_{n\to\infty} G_n(y) = 0.$$

In such cases, it is sometimes possible to achieve something through rescaling.

Example 7.2.5.

Now, let $Y_n := \frac{1}{n} \max\limits_{1\leq i\leq n} X_i$. Then:

$$G_n(y) = \begin{cases} \left(1 + \frac{1}{ny}\right)^{-n}, y > 0 \\ 0, \quad \text{otherwise} \end{cases} \Rightarrow G(y) = e^{-y^{-1}}, \quad y > 0.$$

7.3 The Central Limit Theorem

The following theorem is stated without proof.

Theorem 7.3.1

Let $Y_1, Y_2, \ldots$ be a sequence of random variables with respective CDFs $G_1(y), G_2(y), \ldots$ and MGFs $M_1(t), M_2(t), \ldots$. If $M(t)$ is the MGF of a random variable Y with CDF $G(y)$ and if $\lim\limits_{n\to\infty} M_n(t) = M(t)$ for all t in an open interval containing zero, $(-h, h)$ and $h > 0$, then

$$\lim_{n\to\infty} G_n(y) = G(y),$$

for all continuity points y of $G(y)$.

Example 7.3.2

Suppose, $Y_n \sim \text{BIN}(n, p)$, and $W_n := \frac{Y_n}{n} \Rightarrow$

$$\begin{aligned} M_{W_n}(t) &= M_{Y_n}\left(\tfrac{t}{n}\right) = \left(pe^{\frac{t}{n}} + q\right)^n \\ &= \left(p\left\{1 + \tfrac{t}{n} + \tfrac{t^2}{2n^2} + \tfrac{t^3}{3!n^3} + \cdots\right\} + q\right)^n \\ &= \left(1 + \tfrac{pt}{n} + \tfrac{d(n)}{n}\right)^n. \end{aligned}$$

Where $\frac{d(n)}{n}$ involves the disregarded terms of the series expansion, and also

$$\lim_{n\to\infty} d(n) = 0. \quad \text{Concluding:} \quad \lim_{n\to\infty} M_{W_n}(t) = e^{pt}.$$

Which is the MGF of Y with $\mathbb{P}[Y = p] = 1$, and thus W_n converges stochastically to p.

Theorem 7.3.2 Central Limit Theorem (CLT)

If $X_1, \dots, X_n$ is a random sample from a distribution with finite mean μ and variance σ^2, then the limiting distribution of

$$Z_n := \frac{\sum\limits_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma}$$

is the standard normal, $Z_n \xrightarrow{d} Z \sim \mathrm{N}(0,1)$. Or differently:

$$\lim_{n\to\infty} \mathbb{P}\left[Z_n \leq z\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}x^2}\,\mathrm{d}x = \Phi(z)$$

Proof. This limiting result holds for random samples from any distribution with finite mean and variance, but we prove it under the stronger assumption that the MGF of the distribution exists. Let $m(t)$ denote the MGF of $X - \mu$, $m(t) = M_{X-\mu}(t)$ and note that $m(0) = 1$, $m'(0) = \mathbb{E}[(X-\mu)] = 0$ and

$$\mathbb{E}[(X-\mu)^2] = \sigma^2.$$

Expanding $m(t)$ by the Taylor series formula about 0 gives

$$\begin{aligned} m(t) &= m(0) + m'(0)t + \frac{m''(\xi)t^2}{2} \quad \text{where} \quad 0 < |\xi| < |t| \\ &= 1 + \frac{m''(\xi)t^2}{2} = 1 + \frac{\sigma^2 t^2}{2} + \frac{\left(m''(\xi) - \sigma^2\right)t^2}{2}. \end{aligned}$$

Now we may write

$$Z_n = \frac{\sum\limits_{i=1}^{n} (X_i - \mu)}{\sqrt{n}\,\sigma}.$$

and

$$\begin{aligned} M_{Z_n}(t) &= M_{\sum\limits_{i=1}^{n}(X_i-\mu)}\left(\frac{t}{\sqrt{n}\,\sigma}\right) \\ &= \mathbb{E}\left[\exp\left\{\frac{t}{\sqrt{n}\,\sigma}\sum_{i=1}^{n}(X_i - \mu)\right\}\right] \\ &= \mathbb{E}\left[\exp\left\{\sum_{i=1}^{n}\frac{t}{\sqrt{n}\,\sigma}(X_i-\mu)\right\}\right] \underset{\text{ind.}}{=} \prod_{i=1}^{n} M_{X_i-\mu}\left(\frac{t}{\sqrt{n}\,\sigma}\right) \\ &= \left[M_{X_1-\mu}\left(\frac{t}{\sqrt{n}\,\sigma}\right)\right]^n = \left[m\left(\frac{t}{\sqrt{n}\,\sigma}\right)\right]^n \\ &= \left[1 + \frac{\sigma^2 t^2}{2n\sigma^2} + \frac{\left(m''(\xi) - \sigma^2\right)t^2}{2n\sigma^2}\right]^n, \end{aligned}$$

for $0 < |\xi| < \frac{|t|}{\sqrt{n}\,\sigma}$.

As $n \to \infty$, then $\frac{t}{\sqrt{n}} \to 0$, $\xi \to 0$, and $m''(\xi) - \sigma^2 \to 0$, so

$$M_{Z_n}(t) = \left[1 + \frac{t^2}{2n} + \frac{d(n)}{n}\right]^n,$$

where $d(n) \to 0$ as $n \to \infty$. It follows that

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{\frac{t^2}{2}}$$

which is the MGF of a $N(0,1)$ random variable. Often, the CLT is given in the following equivalent form:

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} Z \sim N(0,1).$$

Example 7.3.4.

Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution, $X_i \sim \text{UNIF}(0,1)$, and let

$$Y_n = \sum_{i=1}^{n} X_i, \quad \mathbb{E}[X_i] = \tfrac{1}{2}, \quad \text{Var}[X_i] = \tfrac{1}{12},$$

$$\sqrt{12}\tfrac{Y_n - \frac{n}{2}}{\sqrt{n}} \xrightarrow{d} Z \sim \text{N}(0,1).$$

For example, if $n = 12$, then approximately

$$Y_{12} - 6 \sim \text{N}(0,1).$$

7.4 See PTS2

7.5 Asymptotic Normality

We now know that $\bar{X}_n$ is approximately $\text{N}(\mu, \frac{\sigma^2}{n})$ distributed for large n. This is an example of a more general notion.

Definition 7.5.1.

If $Y_1, Y_2, \ldots$ is a sequence of random variables and m and c are constants such that

$$Z_n = \sqrt{n}\frac{Y_n - m}{c} \xrightarrow{d} Z \sim \text{N}(0,1)$$

as $n \to \infty$, then Y_n is said to have an **asymptotic normal distribution** with **asymptotic mean** m and **asymptotic variance** $\frac{c^2}{n}$.

Theorem 7.5.1.

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with a pdf $f(x)$ that is continuous and nonzero at the p-th percentile, x_p (meaning $\mathbb{P}[X < x_p] = p$). If $\frac{k}{n} \to p$ (with $k - np$ bounded), then the sequence of k-th order statistics, $X_{k:n}$, is asymptotic normal with asymptotic mean x_p en variance $\frac{c^2}{n}$, where

$$c^2 = \frac{p(1-p)}{[f_X(x_p)]^2}$$

Example 7.5.2.

Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution, $X_i \sim \text{EXP}(1)$, so that $F_X(x) = 1 - e^{-x}$, with $x_{0.5} = \log 2$ and $c^2 = \frac{0.5(1-0.5)}{\left[e^{-\log 2}\right]^2} = 1$. For odd n and $k = \frac{n+1}{2}$, then $X_{k:n}$ is asymptotically normal with asymptotic mean $x_{0.5} = \log 2$ and asymptotic variance $\frac{c^2}{n} = \frac{1}{n}$.

7.6 Properties of Stochastic Convergence

Theorem 7.6.1

The sequence $Y_1, Y_2, \ldots$ converges stochastically to c if and only if $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[|Y_n - c| < \varepsilon\right] = 1.$$

A sequence of random variables that satisfies Theorem (7.6.1), is also said to **converge in probability** to the constant c denoted by $\boxed{Y_n \xrightarrow{p} c.}$ (Another commonly used notation is $\operatorname*{plim}_{n\to\infty} Y_n = c$, where ‘plim’ stands for ‘probability limit’.)

Example 7.6.1 (continuation of 7.3.2)

Let Y be the number of successes in a BIN(n, p) distribution. In Example (7.3.2), we showed using the MGF that $\frac{Y}{n} \xrightarrow{P} p$. If we define $\hat{p} := \frac{Y}{n}$, then $\mathbb{E}[\hat{p}] = p, \text{and} \operatorname{Var}[\hat{p}] = \frac{pq}{n}$ which implies that:

$$\mathbb{P}\left[|\hat{p} - p| < \varepsilon\right] \underset{\text{Chebychev}}{\geq} 1 - \frac{pq}{\varepsilon^2 n}$$

for any $\varepsilon > 0$, so

$$\lim_{n\to\infty} \mathbb{P}\left[|\hat{p} - p| < \varepsilon\right] = 1.$$

Theorem 7.6.2 (Note, slightly different from book)

If $X_1, \ldots, X_n$ is a random sample with $\mathbb{E}[X_i] = \mu$ and $\operatorname{Var}[X_i] = \sigma^2$, for $i = 1, \ldots, n$, then:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$$

Proof. Since $\mathbb{E}[\bar{X}_n] = \mu$ and $\mathrm{Var}[\bar{X}_n] = \frac{\sigma^2}{n}$, it follows that:

$$\mathbb{P}\left[\left|\bar{X}_n - \mu\right| < \varepsilon\right] \underset{\text{Chebychev}}{\geq} 1 - \frac{\sigma^2}{\varepsilon^2 n} \xrightarrow[n\to\infty]{} 1.$$

This means that the sample mean provides a good estimate of the population mean, in the sense that the probability approaches 1 that X_n being arbitrarily close to μ as $n \to \infty$. We can also express this as follows: for any $\varepsilon > 0$ and $0 < \delta < 1$, if $n > \frac{\sigma^2}{\varepsilon^2\delta}$, then $\mathbb{P}\left[\mu - \varepsilon < \bar{X}_n < \mu + \varepsilon\right] \geq 1 - \delta$.

The following theorem , which is stated without proof, asserts that a sequence of asymptotically normal variables converges in probability to the asymptotic mean.

Theorem 7.6.3.

$$\text{If} \quad Z_n = \sqrt{n}\frac{Y_n - m}{c} \xrightarrow{d} Z \sim \mathrm{N}(0,1), \quad \text{then} \quad Y_n \xrightarrow{P} m.$$

7.7 Additional Limit Theorems

Definition 7.7.1.

The sequence of random variables $Y_1, Y_2, \ldots$ is said to **converge in probability** to the random variable Y, written $Y_n \xrightarrow{P} Y$ if

$$\lim_{n\to\infty} \mathbb{P}\left[|Y_n - Y| < \varepsilon\right] = 1,$$

for all $\varepsilon > 0$.

This form of convergence is stronger than convergence in distribution because it imposes requirements on the joint PDF of Y_n and Y, whereas convergence in distribution only concerns the convergence of their CDFs.

Theorem 7.7.1

If a sequence of random variables satisfies

$$Y_n \xrightarrow{P} Y$$

then it also holds that

$$Y_n \xrightarrow{d} Y$$

Note that the converse is not necessarily true, except in specific cases where the sequence of random variables converges in distribution to a constant.

Theorem 7.7.2.

If $Y_n \xrightarrow{P} c$, then for any function $g(y)$ that is continuous at $y = c$

$$g(Y_n) \xrightarrow{P} g(c).$$

Proof. Because $g(y)$ is continuous at c, it follows that for every $\varepsilon > 0$ a $\delta > 0$ exists such that $|y - c| < \delta$ implies $|g(y) - g(c)| < \varepsilon$. This, in turn, implies that

$$\mathbb{P}\left[|g(Y_n) - g(c)| < \varepsilon\right] \geq \mathbb{P}\left[|Y_n - c| < \delta\right]$$

because $\mathbb{P}[B] \geq \mathbb{P}[A]$ whenever $A \subset B$. But because $Y_n \xrightarrow{P} c$ it follows for every $\delta > 0$ that

$$\lim_{n\to\infty} \mathbb{P}\left[|g(Y_n) - g(c)| < \varepsilon\right] \geq \lim_{n\to\infty} \mathbb{P}\left[|Y_n - c| < \delta\right] = 1$$

Theorem 7.7.2 is also valid if Y_n and c are k-dimensional vectors.

Theorem 7.7.3.

If X_n and Y_n are two sequences of random variables such that $X_n \xrightarrow{P} c$ and $Y_n \xrightarrow{P} d$, then:

1. $aX_n + bY_n \xrightarrow{P} ac + bd$
2. $X_n Y_n \xrightarrow{P} cd$
3. $\frac{X_n}{c} \xrightarrow{P} 1, \text{ for } c \neq 0$
4. $\frac{1}{X_n} \xrightarrow{P} c^{-1}, \text{ if } \mathbb{P}[X_n \neq 0] = 1 \text{ for all } n, \text{ and } c \neq 0$
5. $\sqrt{X_n} \xrightarrow{P} \sqrt{c}, \text{ if } \mathbb{P}[X_n \geq 0] = 1 \text{ for all } n$

Example 7.7.1.

Suppose that $Y \sim$ BIN(n, p), we already know that $\hat{p} = \frac{Y}{n} \xrightarrow{P} p$. Thus it follows that $\hat{p}(1 - \hat{p}) \xrightarrow{P} p(1 - p)$.

Theorem 7.7.4. **Slutsky's Theorem**

If X_n and Y_n are two sequences of random variables such that $X_n \xrightarrow{P} c$ and $Y_n \xrightarrow{d} Y$, then:

1. $X_n + Y_n \xrightarrow{d} c + Y$
2. $X_n Y_n \xrightarrow{d} cY$
3. $\frac{Y_n}{X_n} \xrightarrow{d} \frac{Y}{c}, \qquad c \neq 0.$

Note that, as a special case X_n could be an ordinary numerical sequence such as $X_n = \frac{n}{n-1}$.

Example 7.7.2 Continuation of Example 7.7.1.

We know that

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} Z \sim \mathrm{N}(0,1)$$

We also know that $\hat{p}(1-\hat{p}) \xrightarrow{P} p(1-p)$. It follows that:

$$\frac{\hat{p}(1-\hat{p})}{p(1-p)} \xrightarrow{P} 1 \Rightarrow \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{p(1-p)}\frac{p(1-p)}{n}}} = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \xrightarrow{d} Z \sim \mathrm{N}(0,1).$$

Theorem 7.7.5.

If $Y_n \xrightarrow{d} Y$ and $g(y)$ is a continuous function that does not depend on n, then:

$$g(Y_n) \xrightarrow{d} g(Y).$$

The following theorem is very important; often one knows that a sequence of random variables is asymptotically normal, but one is more interested in a function of those random variables.

Theorem 7.7.6.

If $\sqrt{n}\frac{Y_n-m}{c} \xrightarrow{d} Z \sim \mathrm{N}(0,1)$ and $g(y)$ has a continuous derivative at $y=m$, $g'(m) \neq 0$, then:

$$\sqrt{n}\frac{g(Y_n)-g(m)}{cg'(m)} \xrightarrow{d} Z \sim \mathrm{N}(0,1).$$

Proof.

$$g(Y_n) \underset{\text{Taylor}}{=} g(m) + g'(\xi_n)(Y_n-m), \quad \text{where } \xi_n \text{ is between } Y_n \text{ and } m.$$

According to Theorem 7.6.3:

$$Y_n \xrightarrow{P} m \Rightarrow \xi_n \xrightarrow{P} m \underset{\text{thm.7.7.2}}{\Rightarrow} g'(\xi_n) \xrightarrow{P} g'(m),$$

$$\sqrt{n}\frac{g(Y_n)-g(m)}{c} = \sqrt{n}g'(\xi_n)\frac{Y_n-m}{c} \xrightarrow{d} g'(m)Z \sim N\left(0, \left[g'(m)\right]^2\right).$$

Example 7.7.3

$$\sqrt{n}\left(\bar{X}_n-\mu\right) \xrightarrow[\text{CLT}]{d} U \sim \mathrm{N}(0,\sigma^2) \Rightarrow \sqrt{n}\left(\bar{X}_n^2-\mu^2\right) \xrightarrow{d} V \sim N\left(0, \left[2\mu\right]^2\sigma^2\right).$$

Chapter 9

Point Estimation

Suppose we have some population from which we randomly (independently) draw an element that undergoes an experiment, resulting in a random variable X with pdf $f(x;\vartheta)$. Here, $f(x;\vartheta)$ is called the population pdf. If we repeat this process n times independently, we obtain $X_1,\ldots,X_n$, a sample of size n from $f(x;\vartheta)$. The joint pdf is then:

$$f(x_1,\ldots,x_n;\vartheta) = f(x_1;\vartheta)\times\cdots\times f(x_n;\vartheta) = \prod_{i=1}^{n} f(x_i;\vartheta). \tag{9.1.1}$$

This joint pdf establishes the connection between the observations and the mathematical model for the population. The key question is: how can we use this to estimate the unknown value of ϑ? Note that we assume $f(x;\vartheta)$ is fully specified except for the value of ϑ, for example,

$$f(x;\vartheta) = \begin{cases} \frac{1}{\vartheta}, & 0 < x < \vartheta \\ 0, & \text{otherwise} \end{cases} \quad \text{or} \quad f(x;\vartheta) = \begin{cases} \vartheta e^{-\vartheta x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Here, ϑ is called the **parameter**. If ϑ is a vector, we denote it by $\boldsymbol{\vartheta}$; $\vartheta \in \Omega$, where Ω is the set of all possible values that ϑ can take, also known as the **parameter space**. In what follows, we assume that $X_1,\ldots,X_n$ is a sample from $f(x;\vartheta)$ and $\tau(\vartheta)$ is some function of ϑ.

Definition 9.1.1

A statistic, $T = t(X_1,\ldots,X_n)$, that is used to estimate the value $\tau(\vartheta)$ is called an **estimator** of $\tau(\vartheta)$, and an observed value of the statistic, $t = t(x_1,\ldots,x_n)$, is called an **estimate** of $\tau(\vartheta)$.

A statistical quantity is a computable function of the sample elements, for example:
$\bar{X}, \sum_{i=1}^{n} X_i, S^2, X_{1:n}, X_{n:n}$ but not $X_{n:n}^{\vartheta}$ or $\bar{X} - \vartheta$. In other words, an estimator does not contain unknown parameters. Often, an estimator of ϑ is denoted by $\hat{\vartheta}$ or $\tilde{\vartheta}$. This breaks the usual convention of representing random variables with italic uppercase letters.

9.2 Some Methods of Estimation

Method of Moments

Method of moments estimator (MME): Consider a population pdf $f(x;\vartheta_1,\ldots,\vartheta_k)$ and define

$$\mu_j'(\vartheta_1,\ldots,\vartheta_k) := \mathbb{E}[X^j] \qquad j = 1,\ldots,k \tag{9.2.1}$$

Generally, $\mu_j'(\vartheta_1, \ldots, \vartheta_k)$ is a function of $\vartheta_1, \ldots, \vartheta_k$.

Definition 9.2.1

If $X_1, \ldots, X_n$ is a random sample from $f(x; \vartheta_1, \ldots, \vartheta_k)$, the first k **sample moments** are given by

$$M_j' := \frac{1}{n} \sum_{i=1}^{n} X_i^j, \qquad j = 1, \ldots, k \tag{9.2.2}$$

What we will do now is solve the system of k equations

$$M_j' = \mu_j'(\vartheta_1, \ldots, \vartheta_k), \qquad j = 1, \ldots, k \tag{9.2.3}$$

for $\vartheta_1, \ldots, \vartheta_k$. The obtained solutions, $\tilde{\vartheta}_1, \ldots, \tilde{\vartheta}_k$, are called the **moments estimators** of $\vartheta_1, \ldots, \vartheta_k$. Unlike in the book, we use $\tilde{\vartheta}$ for the MME of ϑ instead of $\hat{\vartheta}$.

Example 9.2.1

Consider a random sample from a distribution with two unknown parameters, the mean μ and the variance σ^2.

$$M_1' = \bar{X} = \tilde{\mu} \quad \text{and} \quad M_2' = \frac{1}{n} \sum_{i=1}^{n} X_i^2 = \mathbb{E}[X^2] = \tilde{\mu}^2 + \tilde{\sigma}^2$$

Gives:

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \tilde{\mu}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$$

Example 9.2.2

Consider a random sample from a two-parameter exponential distribution,
$X_i \sim \text{EXP}(1, \eta)$. With expectation $\mathbb{E}[X] = 1 + \eta$, then $\tilde{\eta} = \bar{X} - 1$.

Example 9.2.3

Consider now a random sample from an exponential distribution, $X_i \sim \text{EXP}(\vartheta)$. Notice that $\mathbb{E}[X] = \vartheta$, and therefore $\tilde{\vartheta} = \bar{X}$. Suppose we wish to estimate the probability $p(\vartheta) = \mathbb{P}[X \geq 1] = e^{-\frac{1}{\vartheta}}$. Expressed in terms of $p(\vartheta)$, the expectation is

$$\mathbb{E}[X] = \vartheta = -1/\log p(\vartheta).$$

We find the MME of $p(\vartheta)$ by setting this equal to $\bar{X}$ and solving for $p(\vartheta)$, yielding the solution

$$\widetilde{p(\vartheta)} = e^{-1/\bar{X}} = p(\tilde{\vartheta}).$$

If a class of estimators possesses this property, i.e., the estimator of $g(\vartheta) = g($ estimator of $\vartheta)$, it is said that this class of estimators has the invariance property. When we want to estimate $\tau(\vartheta)$ using the method of moments, there are two options. The first is to solve $\bar{X} = \mu(\tilde{\vartheta})$ and take $\widetilde{\tau} = \tau(\tilde{\vartheta})$ as the estimator. The second option is to express ϑ as a function of τ and then calculate $\tilde{\tau}$ by solving $\bar{X} = \mu(\tilde{\tau})$. In this case, $\tilde{\tau}$ does not necessarily equal $\tau(\tilde{\vartheta})$.

Example 9.2.4

Consider a random sample from a gamma distribution, $X_i \sim \text{GAM}(\vartheta, \kappa)$. The theoretical moments are

$$\begin{cases} \mu_1' &= \mathbb{E}[X] = \kappa\vartheta \\ \mu_2' &= \mathbb{E}[X^2] = \sigma^2 + \mu^2 = \kappa\vartheta^2 + \kappa^2\vartheta^2 \end{cases}$$

so that

$$\begin{cases} \kappa\vartheta &= \bar{X} \\ \kappa\,(1+\kappa)\,\vartheta^2 &= \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2 \end{cases} \quad \text{or} \quad \begin{cases} \kappa &= \bar{X}/\vartheta \\ \frac{\bar{X}}{\vartheta}\left(1+\frac{\bar{X}}{\vartheta}\right)\vartheta^2 &= \bar{X}\vartheta + \bar{X}^2 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2 \end{cases}$$

The resulting MMEs are $\tilde{\vartheta} = (n\bar{X})^{-1}\sum\limits_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = (n-1)(n\bar{X})^{-1}S^2$ and $\tilde{\kappa} = \bar{X}/\tilde{\vartheta}$.

Method of Maximum Likelihood

We will now consider a widely-used method that often yields estimators with desirable, especially large-sample, properties. This method involves selecting a parameter value that maximizes the joint probability density function (pdf) of the sample data. In other words, we seek the parameter value ϑ that makes the pdf as large as possible. For discrete data, this translates to finding the value of ϑ that maximizes the probability of the observed sample..

Example 9.2.5

Suppose that a coin is biased, and it is known that the average proportion of heads is one of the three values, $p = 0.2$, 0.3 or 0.8. An experiment consists of tossing the coin twice and observing the number of heads. This could be modeled mathematically as a random sample X_1, X_2 of size n = 2 from a Bernoulli distribution $X_i \sim \text{BIN}(1, p)$, where the parameter space is $p \in \Omega = \{\, 0.2, 0.3, 0.8\}$. Notice that the MME, $\mathbb{E}[X] = p \Rightarrow \tilde{p} = \bar{X}$, does not produce reasonable estimates in this example, because $\bar{x} = 0$, 0.5 or 1 are the only possibilities, and these are not values in Ω. Consider now the joint pdf of the random sample (X_1, X_2),

$$f(x_1, x_2) = p^{x_1+x_2}\,(1-p)^{2-x_1-x_2}\,, \qquad x_i = 0, 1$$

p	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
0.2	0.64	0.16	0.16	0.04
0.3	0.49	0.21	0.21	0.09
0.8	0.04	0.16	0.16	0.64

Table 9.1

Thus, the estimate that maximizes the "likelihood" for an observed pair (X_1, X_2) is

$$\hat{p} = \begin{cases} 0.2, & \text{if } (x_1, x_2) = (0,0) \\ 0.3, & \text{if } (x_1, x_2) = (0,1), (1,0) \\ 0.8, & \text{if } (x_1, x_2) = (1,1) \end{cases}$$

Definition 9.2.2

Likelihood Function The joint density function of n random variables $X_1, \dots, X_n$ evaluated in $x_1, \dots, x_n$, say $f(x_1, \dots, x_n; \vartheta)$, is referred to as a **likelihood function**. For fixed $x_1, \dots, x_n$ the likelihood function is a function of ϑ and often is denoted by $L(\vartheta)$. If $X_1, \dots, X_n$ represents a random sample from $f(x; \vartheta)$, then

$$L(\vartheta) = f(x_1; \vartheta) \times \ldots \times f(x_n; \vartheta) = \prod_{i=1}^{n} f(x_i; \vartheta).$$

For discrete random variables, given observations $x_1, \dots, x_n$, $L(\vartheta)$ represents the probability of these observations occurring. If $L(\vartheta_1) > L(\vartheta_2)$, then we prefer ϑ_1 as an estimator of ϑ over ϑ_2.

Definition 9.2.3

Maximum Likelihood Estimator (MLE) Let $L(\vartheta) = f(x_1, \dots, x_n; \vartheta)$, $\vartheta \in \Omega$ be the joint pdf of $X_1, \dots, X_n$. For a given set of observations $x_1, \dots, x_n$, a value $\hat{\vartheta} \in \Omega$ at which $L(\vartheta)$ is a maximum is called a **maximum likelihood estimate** (MLE). That is,

$$L(\hat{\vartheta}) = f(x_1, \dots, x_n; \hat{\vartheta}) = \max_{\vartheta \in \Omega} f(x_1, \dots, x_n; \vartheta). \tag{9.2.5}$$

Note that if each set of observations $x_1, \dots, x_n$ corresponds to a unique value $\hat{\vartheta}$, then this procedure defines a function, $\hat{\vartheta} = t(x_1, \dots, x_n)$. $\hat{\vartheta} = t(X_1, \dots, X_n)$ is called the **maximum likelihood estimator**, also denoted MLE. If Ω is an open interval, and if $L(\vartheta)$ is a differentiable function of ϑ and assumes a maximum on Ω, then the MLE will be a solution of the equation

$$\frac{d}{d\vartheta} L(\vartheta) = 0. \tag{9.2.6}$$

Note that any value of ϑ that maximizes $L(\vartheta)$ also will maximize the log-likelihood, $\log L(\vartheta)$, so for computational convenience the alternate form of the maximum likelihood equation

$$\frac{d}{d\vartheta} \log L(\vartheta) = 0. \tag{9.2.7}$$

often will be used. This last equation has the advantage that the logarithm converts a product into a sum, so the product rule does not need to be used.

Example 9.2.6

Consider a random sample of size n from a Poisson distribution, where $X_i \sim \mathrm{POI}(\vartheta)$. The (log-)likelihood function is given by

$$L(\vartheta) = \prod_{i=1}^{n} \frac{e^{-\vartheta} \vartheta^{x_i}}{x_i!} = \frac{e^{-n\vartheta} \vartheta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}, \quad \log L(\vartheta) = -n\vartheta + \sum_{i=1}^{n} x_i \log \vartheta.$$

Now, differentiate with respect to ϑ:

$$\frac{d}{d\vartheta}\log L(\vartheta) = \frac{d}{d\vartheta}\left(-n\vartheta + \sum_{i=1}^{n} x_i \log\vartheta\right) = -n + \frac{\sum_{i=1}^{n} x_i}{\vartheta}.$$

Setting this derivative to zero to find the maximum, we solve

$$-n + \frac{\sum_{i=1}^{n} x_i}{\vartheta} = 0 \Rightarrow \hat{\vartheta} = \bar{X},$$

It is possible to verify that this is a maximum by use of the second derivative, although this is often unnecessary. In this case $\lim\limits_{\vartheta\to 0} L(\vartheta) = \lim\limits_{\vartheta\to\infty} L(\vartheta) = 0$ and $L(\vartheta) > 0$ if $\vartheta \in (0,\infty)$. This confirms that we are dealing with a maximum. Now, suppose we want the MLE of $\tau = e^{-\vartheta} = \mathbb{P}\left[X = 0\right]$. We re-parameterize the model and then obtain a new likelihood function

$$L^*(\tau) = \frac{\tau^n\left(-\log\tau\right)^{\sum\limits_{i=1}^{n} x_i}}{\prod\limits_{i=1}^{n} x_i!}.$$

The new log-likelihood then becomes

$$\frac{d}{d\tau}\log L^*(\tau) = \frac{n}{\tau} + \sum_{i=1}^{n} x_i \frac{d}{d\tau}\log\left(-\log\tau\right) = \frac{n}{\tau} + \frac{\sum\limits_{i=1}^{n} x_i}{\tau\log\tau} = 0,$$

or equivalently,

$$\log\tau = -\bar{X}$$

with the solution $\hat{\tau} = e^{-\bar{X}}$. In general:

Theorem 9.2.1.

Invariance Property. If $\hat{\vartheta}$ is the MLE of ϑ and if $\tau(\vartheta)$ is a function of ϑ, then $\tau(\hat{\vartheta})$ is an MLE of $\tau(\vartheta)$

Example 9.2.7

Consider a random sample from an exponential distribution, $X_i \sim$ EXP(ϑ). For a sample of size n, the likelihood function is

$$L(\vartheta) = \vartheta^{-n} e^{-\vartheta^{-1}\sum_{i=1}^{n} x_i}.$$

Taking the derivative of the log-likelihood function with respect to ϑ gives

$$\frac{d}{d\vartheta}\log L(\vartheta) = -n\vartheta^{-1} + \vartheta^{-2}\sum_{i=1}^{n} x_i \quad \Rightarrow \quad \hat{\vartheta} = \bar{X}.$$

Therefore,

$$\widehat{p(\vartheta)} = \widehat{e^{-\vartheta^{-1}}} = e^{-\bar{X}^{-1}}.$$

There are cases where the MLE cannot be determined by taking a derivative.

Example 9.2.8

Consider a random sample from a two-parameter exponential distribution, $X_i \sim \text{EXP}(1, \eta)$. For a sample of size n, the likelihood function is

$$L(\eta) = \begin{cases} e^{-\sum\limits_{i=1}^{n}(x_i - \eta)}, & \text{if all } x_i \geq \eta \\ 0, & \text{otherwise} \end{cases}$$

or equivalently,

$$L(\eta) = \begin{cases} e^{-\sum\limits_{i=1}^{n}(x_i - \eta)}, & x_{1:n} \geq \eta \\ 0, & \text{otherwise.} \end{cases}$$

The derivative $\frac{d}{d\eta} \log L(\eta) = n$ does not depend on η. However, $L(\eta)$ is an increasing function of η, which implies that $\hat{\eta} = X_{1:n}$. In this example, the Method of Moments Estimator (MME) and Maximum Likelihood Estimator (MLE) do not coincide, with $\tilde{\eta} = \bar{X} - 1$ (Example 9.2.2). If the parameter is a vector, $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$ with $\vartheta \in \Omega \subset \mathbb{R}^k$, then Ω is usually the Cartesian product of k intervals. When this is the case, the maximum is achieved within this domain, and if the partial derivatives exist, one can solve for the maximum.

$$\frac{\partial}{\partial \vartheta_j} \log L(\vartheta_1, \ldots, \vartheta_k) = 0, \qquad j = 1, \ldots, k. \tag{9.2.8}$$

If the Hessian matrix H is negative definite, we have obtained a maximum.

$$H = \{h_{i,j}\}_{i,j=1}^{k}, \qquad h_{i,j} = \frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} \log L(\vartheta_1, \ldots, \vartheta_k).$$

Theorem 9.2.2

Invariance Property. If $\hat{\boldsymbol{\vartheta}} = (\hat{\vartheta}_1, \ldots, \hat{\vartheta}_k)$ denotes the MLE of $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_k)$, then the MLE of $\boldsymbol{\tau} = (\tau_1(\boldsymbol{\vartheta}), \ldots, \tau_r(\boldsymbol{\vartheta}))$ is $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_r) = \left(\tau_1(\hat{\boldsymbol{\vartheta}}), \ldots, \tau_r(\hat{\boldsymbol{\vartheta}})\right)$, for $1 \leq r \leq k$.

Example 9.2.10

For a set of random variables $X_i \sim \text{N}(\mu, \sigma)$, the MLEs of μ and $\vartheta = \sigma^2$ based on a random sample of size n are desired. We have

$$f(x; \mu, \vartheta) = (2\pi\vartheta)^{-\frac{1}{2}} e^{-\frac{(x-\mu)^2}{2\vartheta}} \quad \Rightarrow \quad L(\mu, \vartheta) = (2\pi\vartheta)^{-\frac{n}{2}} e^{-\frac{\sum\limits_{i=1}^{n}(x_i - \mu)^2}{2\vartheta}}$$

$$\frac{\partial}{\partial \mu} \log L(\mu, \vartheta) = \frac{\sum\limits_{i=1}^{n}(x_i - \mu)}{\vartheta}, \qquad \frac{\partial}{\partial \vartheta} \log L(\mu, \vartheta) = -\frac{n}{2\vartheta} + \frac{\sum\limits_{i=1}^{n}(x_i - \mu)^2}{2\vartheta^2}.$$

Setting these derivatives equal to zero and solving simultaneously for the solution values $\hat{\mu}$, $\hat{\vartheta}$ yields the MLEs

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\vartheta} = \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2.$$

To verify that this is indeed a maximum, we examine the Hessian matrix:

$$H = \begin{pmatrix} -\frac{n}{\vartheta} & -\frac{\sum_{i=1}^{n}(x_i-\mu)}{\vartheta^2} \\ -\frac{\sum_{i=1}^{n}(x_i-\mu)}{\vartheta^2} & \frac{n}{2\vartheta^2} - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\vartheta^3} \end{pmatrix},$$

$$H_{\hat{\mu},\hat{\vartheta}} = \begin{pmatrix} -\frac{n}{\hat{\vartheta}} & 0 \\ 0 & -\frac{n}{2\hat{\vartheta}^2} \end{pmatrix},$$

and this matrix is indeed negative definite. Now, let's look at an example where we need to combine techniques to find the MLEs.

Example 9.2.11

Consider a random sample from a two-parameter exponential distribution with both parameters unknown, $X_i \sim$ EXP(ϑ, η). The likelihood function is

$$L(\vartheta, \eta) = \vartheta^{-n} \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \eta)}{\vartheta}\right], \qquad \text{all } x_i \geq \eta,$$

$$\log L(\vartheta, \eta) = -n \log \vartheta - \frac{\sum_{i=1}^{n} x_i - n\eta}{\vartheta}, \qquad x_{1:n} \geq \eta.$$

We maximize this in two steps: first, holding ϑ constant, we maximize with respect to η, and then we maximize $\log L(\vartheta, \hat{\eta})$ with respect to ϑ. Since $\log L(\vartheta, \eta)$ is increasing in η, we have $\hat{\eta} = X_{1:n}$.

$$\frac{d}{d\vartheta} \log L(\vartheta, \hat{\eta}) = -\frac{n}{\vartheta} + \frac{\sum_{i=1}^{n}(x_i - x_{1:n})}{\vartheta^2} \Rightarrow \hat{\vartheta} = \frac{\sum_{i=1}^{n}(X_i - X_{1:n})}{n}.$$

The α percentile, x_α, such that $F(x_\alpha) = \alpha$ is given by $x_\alpha = -\vartheta \log(1-\alpha) + \eta$. This implies

$$\widehat{x}_\alpha = -\hat{\vartheta} \log(1-\alpha) + \hat{\eta}.$$

In some cases, closed-form expressions for the MLEs of parameters are not possible, and approximation formulas or iterative methods are needed.

Example 9.2.12

Let's consider Maximum Likelihood estimation for the parameters of a gamma distribution, $X_i \sim$ GAM(ϑ, κ), based on a random sample of size n. The likelihood function is

$$L(\vartheta, \kappa) = \vartheta^{-n\kappa} \left(\Gamma(\kappa)\right)^{-n} \exp\left[-\frac{\sum_{i=1}^{n} x_i}{\vartheta}\right] \left(\prod_{i=1}^{n} x_i\right)^{\kappa-1},$$

and

$$\log L(\vartheta, \kappa) = -n\kappa \log \vartheta - n \log \Gamma(\kappa) + (\kappa - 1) \log \prod_{i=1}^{n} x_i - \frac{\sum_{i=1}^{n} x_i}{\vartheta}.$$

The partial derivatives are

$$\frac{\partial}{\partial \vartheta} \log L(\vartheta, \kappa) = -\frac{n\kappa}{\vartheta} + \frac{\sum_{i=1}^{n} x_i}{\vartheta^2} = 0,$$

$$\frac{\partial}{\partial \kappa} \log L(\vartheta, \kappa) = -n \log \vartheta - n \frac{\Gamma'(\kappa)}{\Gamma(\kappa)} + \log \prod_{i=1}^{n} x_i = 0.$$

If we let $\tilde{x} := (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$ denote the geometric mean of the sample and let $\psi(\kappa) := \frac{\Gamma'(\kappa)}{\Gamma(\kappa)}$ denote the psi function, then setting the derivatives equal to zero gives the equations

$$\hat{\vartheta} = \frac{\bar{x}}{\hat{\kappa}} \quad \text{and} \quad -n \log \frac{\bar{x}}{\hat{\kappa}} - n\psi(\hat{\kappa}) + n \log \tilde{x} = 0.$$

In the book, approximation formulas for this case are provided on page 300. For the parameter $\mu = \mathbb{E}[X] = \kappa\vartheta$, the MLE has a closed form, given by $\hat{\mu} = \hat{\kappa}\hat{\vartheta} = \bar{x}$.

Example 9.2.13

Consider a random sample of size n from a Weibull distribution with both scale and shape parameters unknown, $X_i \sim \text{WEI}(\vartheta, \beta)$. The likelihood function and ML equations are

$$L(\vartheta, \beta) = \left(\frac{\beta}{\vartheta}\right)^n \prod_{i=1}^{n} \left(\frac{x_i}{\vartheta}\right)^{\beta-1} \exp\left[-\sum_{i=1}^{n} \left(\frac{x_i}{\vartheta}\right)^{\beta}\right],$$

$$\frac{\partial}{\partial \vartheta} \log L(\vartheta, \beta) = -\frac{n}{\vartheta} - \frac{n(\beta-1)}{\vartheta} + \frac{\beta}{\vartheta} \sum_{i=1}^{n} \left(\frac{x_i}{\vartheta}\right)^{\beta} = -\frac{n\beta}{\vartheta} + \frac{\beta}{\vartheta} \sum_{i=1}^{n} \left(\frac{x_i}{\vartheta}\right)^{\beta},$$

$$\frac{\partial}{\partial \beta} \log L(\vartheta, \beta) = \frac{n}{\beta} + \sum_{i=1}^{n} \log\left(\frac{x_i}{\vartheta}\right) - \sum_{i=1}^{n} \left(\frac{x_i}{\vartheta}\right)^{\beta} \log\left(\frac{x_i}{\vartheta}\right).$$

After some algebra, it follows that $\hat{\beta}$ and $\hat{\vartheta}$ are the solutions of

$$g(\beta) = \frac{\sum_{i=1}^{n} x_i^{\beta} \log x_i}{\sum_{i=1}^{n} x_i^{\beta}} - \frac{1}{\beta} - \frac{\sum_{i=1}^{n} \log x_i}{n} = 0 \quad \text{and} \quad \vartheta = \left(\frac{\sum_{i=1}^{n} x_i^{\beta}}{n}\right)^{\frac{1}{\beta}}.$$

A numerical approach to solve for $\hat{\beta}$ is the Newton-Raphson method, with the iterative sequence

$$\hat{\beta}_m = \hat{\beta}_{m-1} - \frac{g(\hat{\beta}_{m-1})}{g'(\hat{\beta}_{m-1})},$$

where $\beta_0 > 0$ is an initial guess, and the following holds $\lim\limits_{m \to \infty} \hat{\beta}_m = \hat{\beta}$.

9.3 Criteria for Evaluating Estimators

Definition 9.3.1

Unbiased Estimator An estimator T is said to be an **unbiased estimator** of $\tau(\vartheta)$ if

$$\mathbb{E}[T] = \tau(\vartheta) \tag{9.3.1}$$

for all $\vartheta \in \Omega$. Otherwise, we say that T is a **biased estimator** of $\tau(\vartheta)$.

Example 9.3.2

Consider a random sample of size n from an exponential distribution, $X_i \sim \text{EXP}(\vartheta)$. For the MLE $\hat{\vartheta}$, it holds that $\hat{\vartheta} = \bar{X}$ and $\mathbb{E}[\bar{X}] = \vartheta$. Thus, $\hat{\vartheta}$ is an unbiased estimator of ϑ.

If we define $\tau(\vartheta) = \frac{1}{\vartheta}$, then $\widehat{\tau(\vartheta)} = \tau(\hat{\vartheta}) = (\bar{X})^{-1}$. Let $S := \sum_{i=1}^n X_i$. The function of S is given by

$$f_S(s) = \begin{cases} \frac{\vartheta^{-n} s^{n-1} e^{-\vartheta^{-1} s}}{(n-1)!}, & s > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Using this, we can compute $\mathbb{E}[S^{-1}]$ as follows:

$$\mathbb{E}[S^{-1}] = \int_0^\infty \frac{\vartheta^{-n} s^{n-2} e^{-\vartheta^{-1} s}}{(n-1)!}\, ds = \frac{\vartheta^{-1}}{n-1} \int_0^\infty \frac{\vartheta^{-(n-1)} s^{n-2} e^{-\vartheta^{-1} s}}{(n-2)!}\, ds = \frac{\vartheta^{-1}}{n-1}.$$

Thus,

$$\mathbb{E}\left[\frac{n-1}{S}\right] = \mathbb{E}\left[\frac{n-1}{n\bar{X}}\right] = \frac{1}{\vartheta}.$$

Note that $\mathbb{E}[\frac{1}{S}] \neq \frac{1}{\mathbb{E}[S]}$. In the previous example, we obtained an unbiased estimator by multiplying an otherwise biased estimator by a constant. However, this is not always possible. For instance, we have

$$X_{1:n} \sim \text{EXP}\left(\frac{\vartheta}{n}\right) \Rightarrow \mathbb{E}[nX_{1:n}] = \vartheta,$$

however, $\mathbb{E}\left[(nX_{1:n})^{-1}\right]$ does not exist.

There are multiple estimators for a parameter since any statistic can be considered an estimator. This raises the obvious question of how to decide which estimators are "best" in some sense. Suppose T_1 and T_2 are estimators for $\tau(\vartheta)$. We say T_1 is **more concentrated** around $\tau(\vartheta)$ than T_2 if

$$\mathbb{P}\left[\tau(\vartheta) - \varepsilon < T_1 < \tau(\vartheta) + \varepsilon\right] \geq \mathbb{P}\left[\tau(\vartheta) - \varepsilon < T_2 < \tau(\vartheta) + \varepsilon\right] \tag{9.3.2}$$

for all $\varepsilon > 0$ and all $\vartheta \in \Omega$.

This criterion seems reasonable, but it usually does not hold universally. Often, (9.3.2) holds for certain values of ϑ, but the inequality reverses for other values of ϑ. If T is an unbiased estimator of $\tau(\vartheta)$, then by Chebyshev's inequality,

$$\mathbb{P}\left[\tau(\vartheta)-\varepsilon<T<\tau(\vartheta)+\varepsilon\right]\geq 1-\frac{\operatorname{Var}(T)}{\varepsilon^2}.$$

The smaller $\operatorname{Var}(T)$, the greater the right-hand side, and thus the greater the minimum probability that $T\in(\tau(\vartheta)-\varepsilon,\tau(\vartheta)+\varepsilon)$. For unbiased estimators, we prefer those with the smallest variance.

Example 9.3.3

Consider a sample from EXP(ϑ). From Example 9.3.2, we know that $T_1=\bar{X}$ and $T_2=nX_{1:n}$ are unbiased estimators of ϑ, with variances $\operatorname{Var}(T_1)=\frac{\vartheta^2}{n}$ and $\operatorname{Var}(T_2)=n^2\operatorname{Var}(X_{1:n})=\vartheta^2$. Thus, we prefer the first estimator.

Example 9.3.I (not in the book)

If $T_1\sim\mathrm{N}(\tau(\vartheta),\sigma_1^2)$ and $T_2\sim\mathrm{N}(\tau(\vartheta),\sigma_2^2)$, then T_1 is more concentrated around $\tau(\vartheta)$ than T_2 if $\sigma_1^2<\sigma_2^2$. Indeed,

$$\mathbb{P}\left[\tau(\vartheta)-\varepsilon<T_i<\tau(\vartheta)+\varepsilon\right]=\mathbb{P}\left[-\frac{\varepsilon}{\sigma_i}<\frac{T_i-\tau(\vartheta)}{\sigma_i}<\frac{\varepsilon}{\sigma_i}\right]=\Phi\left(\frac{\varepsilon}{\sigma_i}\right)-\Phi\left(-\frac{\varepsilon}{\sigma_i}\right),$$

but

$$\Phi\left(\frac{\varepsilon}{\sigma_1}\right)-\Phi\left(-\frac{\varepsilon}{\sigma_1}\right)>\Phi\left(\frac{\varepsilon}{\sigma_2}\right)-\Phi\left(-\frac{\varepsilon}{\sigma_2}\right).$$

Sometimes, one can demonstrate that a specific unbiased estimator has a smaller variance than all other unbiased estimators.

Definition 9.3.2

Let $X_1,\dots,X_n$ be a random sample of size n from $f(x;\vartheta)$. An estimator T^* of $\tau(\vartheta)$ is called a **uniformly minimum variance unbiased estimator** (UMVUE) of $\tau(\vartheta)$ if

1. $\mathbb{E}[T^*]=\tau(\vartheta)$.
2. for any other estimator T with $\mathbb{E}[T]=\tau(\vartheta)$ holds $\operatorname{var}(T^*)\leq\operatorname{var}(T)$ for all $\vartheta\in\Omega$.

In some cases, lower bounds can be derived for the variance of unbiased estimators. If an unbiased estimator can be found that attains such a lower bound, then it follows that the estimator is a UMVUE.

Theorem 9.3.I

Cramér-Rao Lower Bound (CRLB)
Suppose the joint pdf $f(x_1,\dots,x_n;\vartheta)$ satisfies the following conditions:

1. $\mathbb{E}\left[\left\{\frac{\partial}{\partial\vartheta}\log f(X_1,\dots,X_n;\vartheta)\right\}^2\right]$ exists in an open neighborhood of ϑ.
2. $\frac{d}{d\vartheta}\mathbb{E}[t(X_1,\dots,X_n)]=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}t(x_1,\dots,x_n)\frac{\partial}{\partial\vartheta}f(x_1,\dots,x_n;\vartheta)\,dx_1\cdots dx_n.$

3.

$$\begin{aligned}
&\textstyle\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}f(x_1,\ldots,x_n;\vartheta)\,dx_1\cdots dx_n\\
&\quad=\textstyle\frac{d}{d\vartheta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,\ldots,x_n;\vartheta)\,dx_1\cdots dx_n\\
&\quad=\textstyle\frac{d}{d\vartheta}1=0.
\end{aligned}$$

If $\mathbb{E}[T]=\mathbb{E}[t(X_1,\ldots,X_n)]=\tau(\vartheta)$, then it holds that:

$$\operatorname{Var}(T)\geq\frac{(\tau'(\vartheta))^2}{\mathbb{E}\left[\left\{\frac{\partial}{\partial\vartheta}\log f(X_1,\ldots,X_n;\vartheta)\right\}^2\right]}.$$

For a sample $X_1,\ldots,X_n$ from $f(x;\vartheta)$, we have

$$\operatorname{Var}(T)\geq\frac{(\tau'(\vartheta))^2}{n\,\mathbb{E}\left[\left\{\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right\}^2\right]}.$$

Proof. Consider the function

$$u(x_1,\ldots,x_n;\vartheta):=\frac{\partial}{\partial\vartheta}\log f(x_1,\ldots,x_n;\vartheta)=\frac{1}{f(x_1,\ldots,x_n;\vartheta)}\frac{\partial}{\partial\vartheta}f(x_1,\ldots,x_n;\vartheta) \tag{9.3.5}$$

$$\begin{aligned}
\mathbb{E}[U]&=\mathbb{E}[u(X_1,\ldots,X_n;\vartheta)]\\
&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}u(x_1,\ldots,x_n;\vartheta)f(x_1,\ldots,x_n;\vartheta)\,dx_1\cdots dx_n\\
&\underset{(9.3.5)}{=}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}f(x_1,\ldots,x_n;\vartheta)\,dx_1\cdots dx_n=0.
\end{aligned}$$

Define

$$\tau(\vartheta)=\mathbb{E}[T]=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}t(x_1,\ldots,x_n)f(x_1,\ldots,x_n;\vartheta)\,dx_1\cdots dx_n.$$

Then

$$\begin{aligned}
\tau'(\vartheta)&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}t(x_1,\ldots,x_n)\frac{\partial}{\partial\vartheta}f(x_1,\ldots,x_n;\vartheta)\,dx_1\cdots dx_n\\
&\underset{(9.3.5)}{=}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}t(x_1,\ldots,x_n)u(x_1,\ldots,x_n;\vartheta)f(x_1,\ldots,x_n;\vartheta)\,dx_1\cdots dx_n.
\end{aligned}$$

From the fact that

$$\mathbb{E}[U]=0, \quad \operatorname{Var}(U)=\mathbb{E}[U^2] \quad \text{and} \quad \tau'(\vartheta)=\mathbb{E}[TU]=\operatorname{cov}(T,U)$$

it follows that

$$(\operatorname{cov}(T,U))^2 \leq \operatorname{Var}(T)\cdot\operatorname{Var}(U), \quad \text{since} \quad (\rho(T,U))^2 \leq 1.$$

and consequently

$$\operatorname{Var}(T) \geq \frac{(\tau'(\vartheta))^2}{\operatorname{Var}(U)} = \frac{(\tau'(\vartheta))^2}{\mathbb{E}\left[\left\{\frac{\partial}{\partial\vartheta}\log f(X_1,\ldots,X_n;\vartheta)\right\}^2\right]}.$$

When $X_1,\ldots,X_n$ represent a random sample from $f(x;\vartheta)$,

$$f(x_1,\ldots,x_n;\vartheta)=f(x_1;\vartheta)\times\ldots\times f(x_n;\vartheta)$$

so that

$$u(x_1,\ldots,x_n;\vartheta)=\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\log f(x_i;\vartheta).$$

In this case,

$$\begin{aligned}\operatorname{Var}(U)&=\operatorname{Var}\left(\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\log f(X_i;\vartheta)\right)=n\operatorname{Var}\left(\frac{\partial}{\partial\vartheta}\log f(X_1;\vartheta)\right)\\&=n\,\mathbb{E}\left[\left\{\frac{\partial}{\partial\vartheta}\log f(X_1;\vartheta)\right\}^2\right].\end{aligned}$$

Since $X_1,\ldots,X_n$ are i.i.d., the terms

$$\left\{\frac{\partial}{\partial\vartheta}\log f(X_i;\vartheta)\right\}_{i=1}^{n}$$

are also independent and identically distributed. □

Remark 9.3.I

If certain differentiability conditions are satisfied, then

$$\mathbb{E}\left[\left\{\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right\}^2\right]=-\mathbb{E}\left[\frac{\partial^2}{\partial\vartheta^2}\log f(X;\vartheta)\right] \tag{9.3.I}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\tfrac{\partial^2}{\partial\vartheta^2}\log f(X;\vartheta)\right] &= \mathbb{E}\left[\tfrac{\partial}{\partial\vartheta}\tfrac{f'(X;\vartheta)}{f(X;\vartheta)}\right] \\
&= \mathbb{E}\left[\tfrac{f''(X;\vartheta)f(X;\vartheta)-\{f'(X;\vartheta)\}^2}{\{f(X;\vartheta)\}^2}\right] \\
&= \int_{-\infty}^{\infty} f''(x;\vartheta)\,dx - \int_{-\infty}^{\infty}\left(\tfrac{f'(x;\vartheta)}{f(x;\vartheta)}\right)^2 f(x;\vartheta)\,dx.
\end{aligned}$$

If $\frac{d^2}{d\vartheta^2}\int_{-\infty}^{\infty} f(x;\vartheta)\,dx = \int_{-\infty}^{\infty} f''(x;\vartheta)\,dx = \frac{d^2}{d\vartheta^2}1 = 0$, the proof is complete. □

Remark 9.3.II

Not every function $\tau(\vartheta)$ has an unbiased estimator whose variance actually attains the CRLB. Equality is achieved if $\rho(T,U) = \pm 1$. According to Theorem 5.3.1, this means there exist constants $a \neq 0$ and b such that $T = aU + b$. Since $\mathbb{E}[U] = 0$ and $\mathbb{E}[T] = \tau(\vartheta)$, it follows that $b = \tau(\vartheta)$ and $U = \frac{1}{a}(T - \tau(\vartheta))$. In the case of a sample, the variance of an unbiased estimator equals the CRLB if

$$U = \sum_{i=1}^{n} \frac{\partial}{\partial\vartheta}\log f(X_i;\vartheta) = \frac{1}{a}\left(T - \tau(\vartheta)\right). \tag{9.3.II}$$

In other words, if $\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\log f(X_i;\vartheta)$ can be written in the form $\frac{1}{a}(T-\tau(\vartheta))$, we may conclude that T is the UMVUE for $\tau(\vartheta)$.

Note: The constant a must not depend on the X-values, but it may still depend on the unknown parameter ϑ.

Remark 9.3.III

If the support of the pdf depends on the parameter, conditions 2 and 3 of Theorem 9.3.I are generally not satisfied.

Example 9.3.4

Consider a random sample from an exponential distribution, $X_i \sim \text{EXP}(\vartheta)$. Because

$$\begin{aligned}
\log f(x;\vartheta) &= -\frac{x}{\vartheta} - \log\vartheta \\
\frac{\partial}{\partial\vartheta}\log f(x;\vartheta) &= \frac{x}{\vartheta^2} - \frac{1}{\vartheta} = \frac{x-\vartheta}{\vartheta^2}.
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\left\{\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right\}^2\right] = \mathbb{E}\left[\left(\frac{X-\vartheta}{\vartheta^2}\right)^2\right] = \frac{\text{Var}(X)}{\vartheta^4} = \frac{1}{\vartheta^2}.$$

The CRLB for unbiased estimators of ϑ is therefore

$$\frac{\left(\frac{d}{d\vartheta}\vartheta\right)^2}{\frac{n}{\vartheta^2}} = \frac{\vartheta^2}{n}.$$

Because $\mathbb{E}[\bar{X}] = \vartheta$ and $\mathrm{Var}(\bar{X}) = \frac{\vartheta^2}{n}$, it follows that$\bar{X}$ is the UMVUE for ϑ.
We can also derive this using (9.3.II) without needing to compute the CRLB:

$$\sum_{i=1}^{n} \frac{\partial}{\partial\vartheta} \log f(X_i;\vartheta) = \sum_{i=1}^{n} \frac{X_i - \vartheta}{\vartheta^2} = \frac{n}{\vartheta^2}\left(\bar{X} - \vartheta\right) \quad \Rightarrow \quad \bar{X} \text{ is the UMVUE for } \vartheta.$$

In calculating the CRLB, we could also have used (9.3.1):

$$\frac{\partial^2}{\partial\vartheta^2} \log f(x;\vartheta) = -2\frac{x}{\vartheta^3} + \frac{1}{\vartheta^2} \Rightarrow \mathbb{E}\left[\frac{\partial^2}{\partial\vartheta^2} \log f(X;\vartheta)\right] = -\frac{1}{\vartheta^2}.$$

Example 9.3.5

We take a random sample from an geometric distribution, $X_i \sim \mathrm{GEO}(\vartheta)$, and we wish to find a UMVUE for $\tau(\vartheta) = \frac{1}{\vartheta}$.

$$\begin{aligned}
\log f(x;\vartheta) &= \log\vartheta + (x-1)\log(1-\vartheta) \\
\frac{\partial}{\partial\vartheta} \log f(x;\vartheta) &= \frac{1}{\vartheta} - \frac{x - \frac{1}{\vartheta}}{1-\vartheta} = \frac{\frac{1}{\vartheta} - x}{1-\vartheta} \\
\sum_{i=1}^{n} \frac{\partial}{\partial\vartheta} \log f(X_i;\vartheta) &= \frac{1}{\vartheta - 1}\left(\sum_{i=1}^{n} X_i - \frac{n}{\vartheta}\right) = \frac{n}{\vartheta - 1}\left(\bar{X} - \frac{1}{\vartheta}\right)
\end{aligned}$$

$\Rightarrow \bar{X}$ is UMVUE for $\tau(\vartheta) = \frac{1}{\vartheta}$.

Theorem 9.3.1

If an unbiased estimator for $\tau(\vartheta)$ exists, the variance of which achieves the CRLB, then only a linear function of $\tau(\vartheta)$ will admit an unbiased estimator, the variance of which achieves the corresponding CRLB.

Thus, in the previous example, there is no unbiased estimator whose variance attains the CRLB for unbiased estimators of ϑ because ϑ is not a linear function of $\frac{1}{\vartheta}$. In other words, there is a unique function of ϑ for which an unbiased estimator exists with variance equal to the CRLB. This also applies to linear functions of that unique function.

Definition 9.3.3

Efficiency. The **relative efficiency** of an unbiased estimator T of $\tau(\vartheta)$ to another unbiased estimator T^* of $\tau(\vartheta)$ is given by

$$\mathrm{re}(T, T^*) = \frac{\mathrm{Var}(T^*)}{\mathrm{Var}(T)} \tag{9.3.7}$$

An unbiased estimator T^* of $\tau(\vartheta)$ is said to be **efficient** if $\mathrm{re}(T, T^*) \leq 1$ for all unbiased estimators T of $\tau(\vartheta)$, and all $\vartheta \in \Omega$. The efficiency of an unbiased estimator T of $\tau(\vartheta)$ is given by

$$\mathrm{e}(T) = \mathrm{re}(T, T^*) \tag{9.3.8}$$

if $T^*$is an efficient estimator of $\tau(\vartheta)$.

Note that with this definition, an efficient estimator is just a UMVUE. We can now only find UMVUEs that attain the CRLB; in Chapter 10, we will learn to construct UMVUEs that do not attain the CRLB.

Example 9.3.7

Recall in Example 9.3.3 that we had unbiased estimators $T_1 = \bar{X}$ and $T_2 = nX_{1:n}$. In Example 9.3.4, we found that T_1 is a UMVUE. Thus, T_1 is an efficient estimator and the efficiency of T_2 is

$$\mathrm{e}(T_2) = \mathrm{re}(T_2, T_1) = \frac{\vartheta^2/n}{\vartheta^2} = \frac{1}{n};$$

and thus T_2 is a very poor estimator because its efficiency is small for large n. A slightly biased estimator that is highly concentrated about the parameter of interest may be preferable to an unbiased estimator that is less concentrated.

Definition 9.3.4

If T is an estimator of $\tau(\vartheta)$, then the **bias** is given by

$$b(T) = \mathbb{E}[T] - \tau(\vartheta) \tag{9.3.9}$$

and the **mean squared error (MSE)** of T is given by

$$\mathrm{MSE}(T) = \mathbb{E}[T - \tau(\vartheta)]^2 \tag{9.3.10}$$

Theorem 9.3.2

If T is an estimator of $\tau(\vartheta)$, then

$$\mathrm{MSE}(T) = \mathrm{Var}(T) + [b(T)]^2 \tag{9.3.11}$$

Proof:

$$\begin{aligned}\mathrm{MSE}(T) &= \mathbb{E}\left[T - \tau(\vartheta)\right]^2 \\ &= \mathbb{E}\left[T - \mathbb{E}[T] + \mathbb{E}[T] - \tau(\vartheta)\right]^2 \\ &= \mathbb{E}\left[T - \mathbb{E}[T]\right]^2 + 2b(T)\,\mathbb{E}\left[T - \mathbb{E}[T]\right] + [b(T)]^2 \\ &= \mathrm{Var}(T) + [b(T)]^2\end{aligned}$$

Example 9.3.9

Consider a sample from an EXP$(1, \eta)$ distribution. We want to compare the MME $\tilde{\eta} = \bar{X} - 1$ and the MLE $\hat{\eta} = X_{1:n}$:

$$\mathbb{E}[\tilde{\eta}] = \mathbb{E}[\bar{X}] - 1 = \eta + 1 - 1 = \eta \quad \text{and} \quad \mathbb{E}[\hat{\eta}] = \mathbb{E}[X_{1:n}] = \frac{1}{n} + \eta,$$

since

$$f_{X_{1:n}}(x) = \begin{cases} ne^{-(x-\eta)}, & x \geq \eta \\ 0, & \text{otherwise.} \end{cases}$$

The MSEs are

$$\mathrm{MSE}(\tilde{\eta}) = \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{1}{n} \quad \text{and} \quad \mathrm{MSE}(\hat{\eta}) = \mathrm{Var}(X_{1:n}) + \frac{1}{n^2} = \frac{2}{n^2}$$

For $n > 2$, we find that $\mathrm{MSE}(\hat{\eta}) < \mathrm{MSE}(\tilde{\eta})$.

9.4 Large-Sample Properties

An estimator may have undesirable properties for small n, but still be a reasonable estimator in certain applications if it has good asymptotic properties as the sample size increases.

Definition 9.4.1

Simple Consistency. Let $\{T_n\}$ be a sequence of estimators of $\tau(\vartheta)$. These estimators are said to be **consistent** estimators if

$$\lim_{n\to\infty} \mathbb{P}\left[|T_n - \tau(\vartheta)|\right] \leq \varepsilon\right] = 1 \tag{9.4.1}$$

for all $\varepsilon > 0$ and for every $\vartheta \in \Omega$.

In the terminology of Chapter 7, T_n converges in probability to $\tau(\vartheta)$, denoted as $T_n \xrightarrow{p} \tau(\vartheta) \quad \text{as} \quad n \to \infty$. Sometimes this also is referred to as **simple** consistency. One interpretation of consistency is that as n increases, the probability that T_n lies within an ε-interval around $\tau(\vartheta)$ becomes greater. Another slightly stronger type of consistency is based on the MSE.

Definition 9.4.2

MSE consistency. If $\{T_n\}$ is a sequence of estimators $\tau(\vartheta)$, then they are called **mean squared error (mse) consistent** if

$$\lim_{n\to\infty} \mathbb{E}\left[T_n - \tau(\vartheta)\right]^2 = 0 \tag{9.4.2}$$

When using the estimator T_n instead of $\tau(\vartheta)$, the resulting error is $T_n - \tau(\vartheta)$. The requirement is that the expectation of the squared error is small for large sample sizes.

Definition 9.4.3

Asymptotically Unbiased. A sequence $\{T_n\}$ is called asymptotically unbiased for $\tau(\vartheta)$ if

$$\lim_{n\to\infty} \mathbb{E}[T_n] = \tau(\vartheta) \tag{9.4.3}$$

Theorem 9.4.1

A sequence $\{T_n\}$ of estimators of $\tau(\vartheta)$ is mean squared error consistent if and only if it is asymptotically unbiased and

$$\lim_{n\to\infty} \operatorname{Var}(T_n) = 0$$

Proof. This follows immediately from Theorem 9.3.2, because

$$\text{MSE}(T_n) = \operatorname{Var}(T_n) + [b(T_n)]^2$$

Because both terms on the right are nonnegative, $\text{MSE}(T_n) \to 0$, implies both $\operatorname{Var}(T_n) \to 0$ and $[b(T_n)]^2 \to 0$. The converse is obvious. □

Theorem 9.4.2

If a sequence $\{T_n\}$ is MSE-consistent, it also is simply consistent.

Proof. The proof follows from Markov's inequality (2.4.12):

$$\mathbb{P}\left[|X| \geq c\right] \leq \frac{\mathbb{E}\left[|X|^r\right]}{c^r}, \qquad \text{for every } r > 0.$$

It follows that

$$\mathbb{P}\left[|T_n - \tau(\vartheta)| < \varepsilon\right] = 1 - \mathbb{P}\left[|T_n - \tau(\vartheta)| \geq \varepsilon\right] \geq 1 - \frac{\mathbb{E}\left[\{T_n - \tau(\vartheta)\}^2\right]}{\varepsilon^2} = 1 - \frac{\text{MSE}(T_n)}{\varepsilon^2} \xrightarrow[n\to\infty]{} 1$$

□

Example 9.4.2

Let $X_1, \ldots, X_n$ be a random sample from a distribution with finite mean μ and variance σ^2. It was shown in Chapter 7 that $\bar{X}_n \xrightarrow{p} \mu$ and $S_n^2 \xrightarrow{p} \sigma^2$. We also know that both estimators are unbiased, with $\operatorname{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$, which implies that $\bar{X}_n$ is MSE-consistent. If $\mu_4 = \mathbb{E}\left[(X-\mu)^4\right]$ exists, then (from 8.2.11), $\operatorname{Var}(S_n^2) = (\mu_4 - \frac{n-3}{n-1}\sigma^4)/n$, which also implies that S_n^2 is MSE-consistent.

If the distribution is exponential, $X_i \sim \text{EXP}(\vartheta)$, then it follows that $\bar{X}_n$ is MSE-consistent for ϑ, but the estimator $\hat{\vartheta} = nX_{1:n}$ is not even simply consistent because $nX_{1:n} \sim \text{EXP}(\vartheta)$.

Theorem 9.4.3

If $\{T_n\}$ is simply consistent for $\tau(\vartheta)$ and if $g(t)$ is continuous at each value of $\tau(\vartheta)$, then $\{g(T_n)\}$ is simply consistent for $g(\tau(\vartheta))$, see Theorem 7.7.2.

Definition 9.4.4

Asymptotic Efficiency. Let $\{T_n\}$ and $\{T_n^*\}$ be two asymptotically unbiased sequences of estimators for $\tau(\vartheta)$. The **asymptotic relative efficiency** of T_n relative to T_n^* is given by

$$\text{are}(T_n, T_n^*) = \lim_{n\to\infty} \frac{\text{Var}(T_n^*)}{\text{Var}(T_n)} \tag{9.4.4}$$

The sequence $\{T_n^*\}$ is said to be **asymptotically efficient** if $\text{are}(T_n, T_n^*) \leq 1$ for all other asymptotically unbiased sequences $\{T_n\}$ and all $\vartheta \in \Omega$. The **asymptotic efficiency** of an asymptotically unbiased sequence $\{T_n\}$ is given by

$$\text{ae}(T_n) = \text{are}(T_n, T_n^*) \tag{9.4.5}$$

if T_n^* is asymptotically efficient.

Asymptotic Properties of MLEs

If certain regularity conditions are satisfied, such as those in Theorem 9.3.I, then the solutions, $\hat{\vartheta}_n$, of the maximum likelihood equations have the following properties:

1. $\hat{\vartheta}_n$ exists and is unique.
2. $\hat{\vartheta}_n$ is a consistent estimator of ϑ.
3. $\sqrt{n}\left(\hat{\vartheta}_n - \vartheta\right) \xrightarrow{d} N\left(0, \frac{1}{\mathbb{E}\left[\left(\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right)^2\right]}\right)$.
4. $\hat{\vartheta}_n$ is asymptotically efficient.

Example 9.4.7

Consider a random sample from a Pareto distribution, $X_i \sim \text{PAR}(1, \kappa)$, where κ is unknown. Since

$$f(x;\kappa) = \kappa(1+x)^{-\kappa-1}, \qquad x > 0$$

it follows that

$$\log L(\kappa) = n\log\kappa - (\kappa+1)\sum_{i=1}^{n}\log(1+x_i)$$

and the ML equation is

$$\frac{d}{d\kappa}\log L(\kappa) = \frac{n}{\kappa} - \sum_{i=1}^{n}\log(1+x_i) = 0 \quad \Rightarrow \quad \hat{\kappa} = \frac{n}{\sum_{i=1}^{n}\log(1+x_i)}$$

To find the CRLB, note that

$$\begin{aligned} \log f(x;\kappa) &= \log\kappa - (\kappa+1)\log(1+x), \\ \frac{\partial}{\partial\kappa}\log f(x;\kappa) &= \frac{1}{\kappa} - \log(1+x), \\ \mathbb{E}\left[\left\{\frac{\partial}{\partial\kappa}\log f(X;\kappa)\right\}^2\right] &= -\mathbb{E}\left[\frac{\partial^2}{\partial\kappa^2}\log f(X;\kappa)\right] = \frac{1}{\kappa^2}. \end{aligned}$$

Thus, the CRLB is

$$\text{CRLB} = \frac{1}{n\cdot\kappa^{-2}} = \frac{\kappa^2}{n} \quad \Rightarrow \quad \sqrt{n}(\hat{\kappa}-\kappa) \xrightarrow{d} N(0,\kappa^2)$$

9.5 Bayes and MiniMax Estimators

We will not cover this topic in great detail; only a few key concepts will be discussed. An estimator will rarely be exactly equal to the quantity being estimated, so when using the estimator, some loss is inevitable.

Definition 9.5.1

Loss Function. If T is an estimator of $\tau(\vartheta)$, then a **loss function** is any real-valued function $L(t;\vartheta)$, such that

$$L(t;\vartheta) \geq 0 \quad \text{for every } t \tag{9.5.1}$$

and

$$L(t;\vartheta) = 0 \qquad \text{when } t = \tau(\vartheta) \tag{9.5.2}$$

The loss function itself is, of course, a random variable, and we aim to find an estimator for which the expected loss is small.

Definition 9.5.2

Risk. The **risk function** is defined to be the expected loss,

$$R_T(\vartheta) = \mathbb{E}[L(T;\vartheta)] \tag{9.5.3}$$

Examples of loss functions include $[T-\tau(\vartheta)]^2$, $|T-\tau(\vartheta)|$, and $\frac{[T-\tau(\vartheta)]^2}{[\tau(\vartheta)]^2}$. Similar to MSE, it is generally not possible to find estimators with uniformly smallest risk for all $\vartheta \in \Omega$, given a specific loss function.

Definition 9.5.3

Admissible Estimator. An estimator T_1 is a **better estimator** than T_2 if and only if

$$R_{T_1}(\vartheta) \leq R_{T_2}(\vartheta) \quad \text{for all } \vartheta \in \Omega$$

and

$$R_{T_1}(\vartheta) < R_{T_2}(\vartheta) \quad \text{for at least one } \vartheta$$

An estimator T is **admissible** if and only if there is no better estimator.

Definition 9.5.4

Minimax Estimator. An estimator T_1 is a **minimax estimator** if

$$\max_{\vartheta} R_{T_1}(\vartheta) \leq \max_{\vartheta} R_T(\vartheta) \tag{9.5.4}$$

for every estimator T.

The minimax principle is a conservative approach, because it attempts to protect against the worst risk that can occur. It may happen that an estimator has higher risk over a very small subset of Ω, but a much lower risk over the rest of Ω. Another approach is the Bayes approach, which assumes that the parameter itself is random and has a probability density function, called the prior pdf.

Definition 9.5.5

Bayes Risk. For a random sample from $f(x;\vartheta)$, the **Bayes risk** of an estimator T relative to a risk function $R_T(\vartheta)$ and a prior function $p(\vartheta)$ is given by

$$A_T = \mathbb{E}_{\vartheta}\left[R_T(\vartheta)\right] = \int_{\Omega} R_T(\vartheta)p(\vartheta)\,\mathrm{d}\vartheta. \tag{9.5.6}$$

If an estimator has the smallest Bayes risk, then it is referred to as a Bayes estimator. (see Definition 9.5.6). In the Bayesian framework, we consider a sample $X_1,\ldots,X_n$ to be drawn from the conditional pdf $f(x|\vartheta)$.

Definition 9.5.7

Posterior Distribution. The conditional distribution of ϑ, given the sample outcomes $\mathbf{x} = (x_1,\ldots,x_n)$, is called the **posterior pdf**, and is given by

$$f_{\vartheta|\mathbf{x}}(\vartheta) = \frac{f(x_1,\ldots,x_n;\vartheta)}{\int f(x_1,\ldots,x_n;\vartheta)p(\vartheta)\,\mathrm{d}\vartheta} \tag{9.5.8}$$

The Bayes estimator is the estimator that minimizes the average risk over ϑ, $\mathbb{E}_{\vartheta}\left[R_T(\vartheta)\right]$. However,

$$\mathbb{E}_{\vartheta}\left[R_T(\vartheta)\right] = \mathbb{E}_{\vartheta}\mathbb{E}_{\mathbf{X}|\vartheta}\left[L(T;\vartheta)\right] = \mathbb{E}_{\mathbf{X}}\mathbb{E}_{\vartheta|\mathbf{X}}\left[L(T;\vartheta)\right] \tag{9.5.9}$$

Theorem 9.5.1

The Bayes estimator is the estimator T that minimizes

$$\mathbb{E}_{\vartheta|\mathbf{X}}\left[L(T;\vartheta)\right] = \int_{\Omega} L(t;\vartheta) f_{\vartheta|\mathbf{x}}(\vartheta)\,\mathrm{d}\vartheta.$$

In the book, Bayes estimators are provided for various loss functions.

Chapter 10

Sufficiency and Completeness

10.1 Introduction

Chapter 9 presented methods for deriving point estimators based on a random sample to estimate unknown parameters of the population distribution. In some cases, it is possible to show, in a certain sense, that a particular statistic or set of statistics contains all of the ”information” in the sample about the parameters. It then would be reasonable to restrict attention to such statistics when estimating or otherwise making inferences about the parameters. For example, consider a random sample from an exponential distribution, where the sum of all sample values contains all the information about the parameter. Therefore, we only need to consider $\sum_{i=1}^n X_i$.

More generally, the idea of sufficiency involves the reduction of a data set to a more concise set of statistics with no loss of information about the unknown parameter. Roughly, a statistic S will be considered a ”sufficient” statistic for a parameter ϑ if the conditional distribution of any other statistic T given the value of S does not involve ϑ. In other words, once the value of a sufficient statistic is known, the observed value of any other statistic does not contain any further information about the parameter.

Example 10.1.I

Consider a random sample $X_1, \dots, X_n$ from an Poisson distribution, $X_i \sim$ POI(ϑ). Define $S := \sum_{i=1}^n X_i \sim$ POI$(n\vartheta)$. It follows that

$$f(x_1, \dots, x_n; \vartheta) = \prod_{i=1}^n \frac{e^{-\vartheta}\vartheta^{x_i}}{x_i!} = \frac{e^{-n\vartheta}\vartheta^{x_1+\cdots+x_n}}{\prod_{i=1}^n x_i!} \quad \text{and} \quad f_S(s;\vartheta) = \frac{e^{-n\vartheta}\vartheta^s n^s}{s!}$$

with the conditional probability function

$$f_{X_1,\dots,X_n|S}(x_1, \dots, x_n|S = s) = \frac{\mathbb{P}\left[X_1 = x_1, \dots, X_n = x_n,\, S = s\right]}{\mathbb{P}\left[S = s\right]}$$

which can be computed as:

$$= \begin{cases} 0, & \text{if } s \neq \sum_{i=1}^n x_i, \\ \frac{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)}{\mathbb{P}(S=s)} = \frac{e^{-n\vartheta}\vartheta^{x_1+\cdots+x_n} s!}{e^{-n\vartheta}\vartheta^s n^s \prod_{i=1}^n x_i!} = \frac{s!}{n^s \prod_{i=1}^n x_i!}, & \text{if } s = \sum_{i=1}^n x_i. \end{cases}$$

In the book, Example 10.1.1 also discusses S in the context of the Bernoulli distribution.

10.2 Sufficient Statistics

Definition 10.2.1

Jointly Sufficient Statistics. Let $\mathbf{X} = (X_1, \ldots, X_n)$ have a joint pdf $f(\mathbf{x}, \boldsymbol{\vartheta})$, and let $\mathbf{S} = (S_1, \ldots, S_k)$ be a k-dimensional statistic. Then $S_1, \ldots, S_k$ is a set of **jointly sufficient statistics for $\boldsymbol{\vartheta}$** if, for any other vector of statistics, $\mathbf{T}$, the conditional pdf of $\mathbf{T}$ given $\mathbf{S} = s$, denoted by $f_{\mathbf{T}|\mathbf{s}}(\mathbf{t})$, does not depend on $\boldsymbol{\vartheta}$. In the one-dimensional case, we say that S is a **sufficient statistic** for ϑ.

The sample itself is, of course, a sufficient statistic, as is the ordered sample, but the question is whether a smaller set of sufficient statistics exists that contains all the information.
Without proof, we state that if

$$\frac{f(x_1, \ldots, x_n; \boldsymbol{\vartheta})}{f_{\mathbf{S}}(\mathbf{s}; \boldsymbol{\vartheta})} \qquad \text{where} \quad \mathbf{s} = s(x_1, \ldots, x_n) \tag{10.2.1}$$

is independent of $\boldsymbol{\vartheta}$, then $\mathbf{S}$ is jointly sufficient for $\boldsymbol{\vartheta}$.

Definition 10.2.2

Minimal Sufficient. A set of statistics is called a **minimal sufficient** set if the members of the set are jointly sufficient for the parameters and if they are a function of every other set of jointly sufficient statistics.

This is a fairly challenging definition to work with, but there are constructive methods to find the minimal sufficient statistics, which we will not explore further here.

Example 10.2.1

Consider a sample from the EXP(ϑ) distribution, meaning:

$$f(x_1, \ldots, x_n; \vartheta) = \vartheta^{-n} \exp\left(-\frac{\sum_{i=1}^n x_i}{\vartheta}\right), \qquad x_i > 0.$$

It appears that $S = \sum_{i=1}^n X_i$ is a candidate, as we know:

$$f_S(s; \vartheta) = \frac{1}{\vartheta^n \Gamma(n)} s^{n-1} \exp\left(-\frac{s}{\vartheta}\right), \qquad s > 0.$$

If $s = \sum_{i=1}^n x_i$, then:

$$\frac{f(x_1, \ldots, x_n; \vartheta)}{f_S(s; \vartheta)} = \frac{\Gamma(n)}{s^{n-1}}.$$

This is independent of ϑ, making S sufficient for ϑ. Another simpler criterion exists for finding a sufficient statistic.

Theorem 10.2.1

Factorization Criterion. If $X_1, \ldots, X_n$ have a joint pdf $f(x_1, \ldots, x_n; \boldsymbol{\vartheta})$, and if S $= (S_1, \ldots, S_k)$, then $S_1, \ldots, S_k$ are jointly sufficient for $\boldsymbol{\vartheta}$ **if and only if**

$$f(x_1, \ldots, x_n; \boldsymbol{\vartheta}) = g(\mathbf{s}; \boldsymbol{\vartheta}) h(x_1, \ldots, x_n), \tag{10.2.3}$$

where $g(\mathbf{s}; \boldsymbol{\vartheta})$ depends on $x_1, \ldots, x_n$ only through $\mathbf{s}$, and $h(x_1, \ldots, x_n)$ does not depend on $\boldsymbol{\vartheta}$.

Example 10.2.2

Consider a sample from BIN$(1, \vartheta)$:

$$f(x_1, \ldots, x_n; \vartheta) = \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i} = \vartheta^s(1-\vartheta)^{n-s} = g(s;\vartheta)h(x_1, \ldots, x_n),$$

where

$$g(s;\vartheta) = \vartheta^s(1-\vartheta)^{n-s} \quad \text{and} \quad h(x_1, \ldots, x_n) = 1 \quad \text{for all } x_i = 0, 1.$$

Now, a slightly more complex example, where the support is partly determined by the parameter.

Example 10.2.II

For a sample $X_1, \ldots, X_n$, consider the following

$$f(x;\vartheta) = \frac{4\vartheta^4}{x^5}, \qquad 0 < \vartheta \le x \quad \text{and} \quad f(x_1, \ldots, x_n; \vartheta) = \frac{4^n \vartheta^{4n}}{\prod_{i=1}^n x_i^5}, \qquad 0 < \vartheta \le x_{1:n}.$$

We can rewrite this as:

$$f(x_1, \ldots, x_n; \vartheta) = g(x_{1:n}; \vartheta)\, h(x_1, \ldots, x_n),$$

where

$$g(x_{1:n};\vartheta) = \begin{cases} 0, & \text{if } \vartheta > x_{1:n}, \\ \vartheta^{4n}, & \text{if } \vartheta \le x_{1:n}, \end{cases} \quad \text{and} \quad h(x_1, \ldots, x_n) = \frac{4^n}{\prod_{i=1}^n x_i^5}.$$

This type of problem, where the support of the pdf depends on the parameter, can be approached as follows.

Definition 10.2.3

Indicator Function. If A is a set, the **indicator function** of A, denoted $I_A(\cdot)$, is defined as:

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A. \end{cases}$$

In the previous example, we set $A = [\vartheta, \infty)$, so

$$f(x; \vartheta) = \frac{4\vartheta^4}{x^5}\, I_{[\vartheta,\infty)}(x),$$

and therefore,

$$f(x_1, \dots, x_n; \vartheta) = \frac{4^n \vartheta^{4n}}{\prod_{i=1}^n x_i^5} \prod_{i=1}^n I_{[\vartheta,\infty)}(x_i) = \frac{4^n \vartheta^{4n}}{\prod_{i=1}^n x_i^5} I_{[\vartheta,\infty)}(x_{1:n}).$$

We observe that:

$$g(x_{1:n}; \vartheta) = \vartheta^{4n}\, I_{[\vartheta,\infty)}(x_{1:n}) \quad \text{and} \quad h(x_1, \dots, x_n) = \frac{4^n}{\prod_{i=1}^n x_i^5}.$$

Conventions: square brackets [ ] or round brackets ( ) indicate a continuous range, while curly braces { } are reserved for sets with countable elements.

Example 10.2.4

Consider a random sample from a normal distribution, $X_i \sim \text{N}(\mu, \sigma)$, where both μ and σ^2 are unknown. The joint pdf is

$$\begin{aligned} f(x_1, \dots, x_n; \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2\right)\right] \\ &= g(s_1, s_2; \mu, \sigma^2)\, h(x_1, \dots, x_n), \end{aligned}$$

where

$$g(s_1, s_2; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \left(s_2 - 2\mu s_1 + n\mu^2\right)\right]$$

and

$$s_1 = \sum_{i=1}^n x_i, \qquad s_2 = \sum_{i=1}^n x_i^2, \qquad h(x_1, \dots, x_n) = 1.$$

Example 10.2.5

Consider a random sample from an uniform distribution, $X_i \sim \text{UNIF}(\vartheta, \vartheta + 1)$. Notice that the length of the interval is one unit, but the endpoints are assumed to be unknown. The pdf of X is the indicator function of the interval, $f(x; \vartheta) = I_{[\vartheta,\vartheta+1]}(x)$, so the joint pdf of $X_1, \dots, X_n$ is

$$f(x_1, \dots, x_n; \vartheta) = \prod_{i=1}^n I_{[\vartheta,\vartheta+1]}(x_i) = I_{[\vartheta, x_{n:n}]}(x_{1:n})\, I_{[x_{1:n}, \vartheta+1]}(x_{n:n}).$$

The pair $(X_{1:n}, X_{n:n})$ is sufficient for ϑ. Note that:

$$I_{[\vartheta, x_{n:n}]}(x_{1:n})\, I_{[x_{1:n}, \vartheta+1]}(x_{n:n}) = I_{[\vartheta,\infty]}(x_{1:n})\, I_{(-\infty,\vartheta+1]}(x_{n:n}).$$

10.3 Other Properties of Sufficient Statistics

Theorem 10.3.1

If $S_1, \ldots, S_k$ are jointly sufficient for $\boldsymbol{\vartheta}$ and if $\hat{\boldsymbol{\vartheta}}$ is a unique MLE of $\boldsymbol{\vartheta}$, then $\hat{\boldsymbol{\vartheta}}$ is a function of $\mathbf{S} = (S_1, \ldots, S_k)$.

Proof. According to the factorization criterion,

$$L(\boldsymbol{\vartheta}) = f(x_1, \ldots, x_n; \boldsymbol{\vartheta}) = g(\mathbf{s}; \boldsymbol{\vartheta}) h(x_1, \ldots, x_n)$$

Maximizing the likelihood here involves maximizing $g(\mathbf{s}; \boldsymbol{\vartheta})$. If $\hat{\boldsymbol{\vartheta}}$ is unique, then it must be a function of $\mathbf{s}$. □

Theorem 10.3.2

If S is sufficient for ϑ, then any Bayes estimator will be a function of S.

Proof. Because the function $h(x_1, \ldots, x_n)$ in the factorization creterion does not depend on ϑ, it can be eliminated in equation (9.5.8), and the posterior density $f_{\vartheta|x}(\vartheta)$ can be replaced by

$$f_{\vartheta|x}(\vartheta) = \frac{g(s; \vartheta) p(\vartheta)}{\int g(s; \vartheta) p(\vartheta)\, d\vartheta}$$

As mentioned earlier, the order statistics are jointly sufficient. □

Theorem 10.3.3

If $X_1, \ldots, X_n$ is a random sample from a continuous distribution with pdf $f(x; \vartheta)$, then the order statistics form a jointly sufficient set for $\boldsymbol{\vartheta}$.

Proof. For fixed $x_{1:n}, \ldots, x_{n:n}$, and associated $x_1, \ldots, x_n$

$$\frac{f(x_1; \boldsymbol{\vartheta}) \cdots f(x_n; \boldsymbol{\vartheta})}{n! f(x_{1:n}; \boldsymbol{\vartheta}) \cdots f(x_{n:n}; \boldsymbol{\vartheta})} = \frac{1}{n!}, \quad x_{1:n} = \min(x_i), \ldots, x_{n:n} = \max(x_i)$$

and zero otherwise. □

Generally, sufficient statistics are involved in the construction of UMVUEs.

Now follows an important theorem, known as the Rao-Blackwell theorem, or the **improvement theorem**.

Theorem 10.3.4 Rao-Blackwell (Improvement Theorem)

Let $X_1, \dots, X_n$ have a joint pdf $f(x_1, \dots, x_n; \boldsymbol{\vartheta})$, and let $\mathsf{S} = (S_1, \dots, S_k)$ be a vector of jointly sufficient statistics for $\boldsymbol{\vartheta}$. If T is any unbiased estimator of $\tau(\boldsymbol{\vartheta})$, and if $T^* := \mathbb{E}\left[T \mid \mathsf{S}\right]$, then:

1. T^* is an unbiased estimator of $\tau(\boldsymbol{\vartheta})$,
2. T^* is a function of S, and
3. $\text{Var}(T^*) \leq \text{Var}(T)$ for every $\boldsymbol{\vartheta}$, and $\text{Var}(T^*) < \text{Var}(T)$ for some $\boldsymbol{\vartheta}$ unless $T^* = T$ with probability 1.

Proof. By the sufficiency of S, $f_{T|\mathsf{S}}(t)$ is independent of $\boldsymbol{\vartheta}$, and the function $t^*(\mathsf{S}) := \mathbb{E}\left[T \mid \mathsf{S} = \mathsf{s}\right]$ does not depend on $\boldsymbol{\vartheta}$. It follows that $T^* = t^*(\mathsf{S}) = \mathbb{E}\left[T \mid \mathsf{S}\right]$ is a function of S. Furthermore,

$$\begin{aligned}\mathbb{E}[T^*] &= \mathbb{E}_{\mathsf{S}}[T^*] \\ &= \mathbb{E}_{\mathsf{S}}\left[\mathbb{E}\left[T \mid \mathsf{S}\right]\right] \\ &= \mathbb{E}[T] \\ &= \tau(\boldsymbol{\vartheta})\end{aligned}$$

by Theorem 5.4.1. From Theorem 5.4.3,

$$\begin{aligned}\text{Var}(T) &= \text{Var}\left[\mathbb{E}\left[T \mid \mathsf{S}\right]\right] + \mathbb{E}\left[\text{Var}\left(T \mid \mathsf{S}\right)\right] \\ &\geq \text{Var}\left[\mathbb{E}\left[T \mid \mathsf{S}\right]\right] \\ &= \text{Var}(T^*).\end{aligned}$$

with equality if and only if $\mathbb{E}\left[\text{Var}\left(T \mid \mathsf{S}\right)\right] = 0$, which occurs if and only if $\text{Var}\left(T \mid \mathsf{S}\right) = 0$ with probability 1, Since

$$\text{Var}\left(T \mid \mathsf{S}\right) = \mathbb{E}\left[\left(T - \mathbb{E}\left[T \mid \mathsf{S}\right]\right)^2 \mid \mathsf{S}\right] = 0,$$

or equivalently $T = \mathbb{E}\left[T \mid \mathsf{S}\right] = T^*$. □

Example 10.3.I

Consider a random sample from an uniform distribution, $X_i \sim \text{BIN}(p)$. We seek a good unbiased estimator for $p(1-p) = pq$. We first construct a very simple unbiased estimator for pq, namely:

$$T = \begin{cases} 1, & \text{if } X_1 = 1 \text{ and } X_2 = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned}\mathbb{E}[T] &= 1 \times \mathbb{P}\left[X_1 = 1, X_2 = 0\right] + 0 \times \mathbb{P}\left[(X_1, X_2) \neq (1, 0)\right] \\ &= pq\end{aligned}$$

and

$$\begin{aligned} f(x_1,\dots,x_n;\vartheta) &= p^{\sum_{i=1}^n x_i} q^{n-\sum_{i=1}^n x_i} \\ &= g\left(\sum_{i=1}^n x_i; p\right) h(x_1,\dots,x_n) \end{aligned}$$

with

$$g\left(\sum_{i=1}^n x_i; p\right) = p^{\sum_{i=1}^n x_i} q^{n-\sum_{i=1}^n x_i} \quad \text{and} \quad h(x_1,\dots,x_n) = 1.$$

We observe that $S := \sum_{i=1}^n x_i$ is sufficient for p. Furthermore,

$$T^* := \mathbb{E}\left(T \mid S\right), \qquad t^* = \mathbb{E}\left(T \mid S = s\right)$$

and

$$\begin{aligned} t^* &= 1 \times \mathbb{P}\left[T = 1 \mid S = s\right] + 0 \times \mathbb{P}\left[T = 0 \mid S = s\right] \\ &= \frac{\mathbb{P}\left[X_1 = 1, X_2 = 0, S = s\right]}{\mathbb{P}\left[S = s\right]} \\ &= \frac{\mathbb{P}\left[X_1 = 1, X_2 = 0, \sum_{i=3}^n X_i = s-1\right]}{\mathbb{P}\left[S = s\right]} \\ &= \frac{pq\binom{n-2}{s-1}p^{s-1}q^{n-2-(s-1)}}{\binom{n}{s}p^s q^{n-s}} \\ &= \frac{\binom{n-2}{s-1}}{\binom{n}{s}} = \frac{s(n-s)}{n(n-1)}. \end{aligned}$$

We conclude that

$$T^* = \frac{S(n-S)}{n(n-1)}$$

is an unbiased estimator with smaller variance than T. In the next section, we will see that T^* is the UMVUE of pq.

10.4 Completeness and the Exponential Family

Definition 10.4.1

Completeness. A family of density functions $\{f_{\mathbf{T}}(\mathbf{t}; \boldsymbol{\vartheta} \in \boldsymbol{\Omega})\}$ is called **complete** if $\mathbb{E}[u(\mathbf{T})] = 0$ for all $\boldsymbol{\vartheta} \in \boldsymbol{\Omega})$ implies that $u(\mathbf{T}) = 0$ with probability 1 for all $\boldsymbol{\vartheta} \in \boldsymbol{\Omega})$.

Important: Completeness is a property of a family of pdfs, not an individual pdf. Sometimes it is said that ”there are no non-trivial unbiased estimators of 0.” If a statistic belongs to a complete family and is also sufficient, then any function of that statistic that is an unbiased estimator for $\tau(\boldsymbol{\vartheta})$ is the UMVUE of $\tau(\boldsymbol{\vartheta})$ and is unique. We prove this in the following theorem.

Theorem 10.4.1 (Lehmann-Scheffé)

> Let $X_1, \dots, X_n$ have a joint pdf $f(x_1, \dots, x_n; \boldsymbol{\vartheta})$, and let S be a vector of jointly complete sufficient statistics for $\boldsymbol{\vartheta}$. If $T^* = t^*(\mathrm{S})$ with $\mathbb{E}[T^*] = \tau(\boldsymbol{\vartheta})$, then T^* is the UMVUE of $\tau(\boldsymbol{\vartheta})$. Furthermore, T^* is the only function of S that is an unbiased estimator of $\tau(\boldsymbol{\vartheta})$.

Proof. It follows by completeness that any statistic that is a function of S and an unbiased estimator of $\tau(\boldsymbol{\vartheta})$ must be equal to T^* with probability 1. If T is any other statistic that is an unbiased estimator of $\tau(\boldsymbol{\vartheta})$, then by the Rao-Blackwell theorem, $\mathbb{E}(T \mid S)$ also is unbiased for $\tau(\boldsymbol{\vartheta})$ and a function of S. So by uniqueness, $T^* = \mathbb{E}(T \mid S)$ with probability 1. Furthermore, $\mathrm{Var}(T^*) \leq \mathrm{Var}(T)$ for all $\boldsymbol{\vartheta}$. Thus, T^* is a UMVUE of $\tau(\boldsymbol{\vartheta})$. □

Example 10.4.I

Let $X_1, \dots, X_n$ be a sample from the BIN$(1, p)$ distribution.

$$f(x_1, \dots, x_n; p) = p^s q^{n-s}, \qquad \text{where} \quad s = \sum_{i=1}^{n} x_i.$$

By the factorization criterion, $S = \sum_{i=1}^{n} X_i$ is sufficient. But is it also complete? We know that $\mathbb{P}[S = s] = \binom{n}{s} p^s q^{n-s}$, which implies:

$$\mathbb{E}[u(S)] = \sum_{s=0}^{n} u(s) \binom{n}{s} p^s q^{n-s} = q^n \sum_{s=0}^{n} u(s) \binom{n}{s} \left(\frac{p}{q}\right)^s = 0.$$

This results in a polynomial of degree n in $\frac{p}{q}$ that is 0 for all values of $\frac{p}{q}$, which can only occur if all coefficients are 0. Therefore, the estimator T^* from Example 10.3.I is the UMVUE for $p(1-p)$.

There exists a family of distributions that provides complete and sufficient statistics directly.

Definition 10.4.2

Exponential Class. A density function is said to be a member of the **regular exponential class (REC)** if it can be expressed in the form

$$f(x;\boldsymbol{\vartheta}) = c(\boldsymbol{\vartheta})h(x)\exp\left(\sum_{j=1}^{k} q_j(\boldsymbol{\vartheta})t_j(x)\right), \qquad x \in A \qquad (10.4.1)$$

and zero otherwise, where $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_k)$ is a vector of k unknown parameters, if the parameter space has the form

$$\Omega = \{\boldsymbol{\vartheta} \mid a_i \leq \vartheta_i \leq b_i, i = 1, \ldots, k\}$$

(note that $a_i = -\infty$ and $b_i = \infty$ are permissible values), and if it satisfies regularity conditions 1, 2, and 3a or 3b given by:

1 The set $A = \{x \mid f(x;\boldsymbol{\vartheta}) > 0\}$ does not depend on $\boldsymbol{\vartheta}$.

2 The functions $q_j(\boldsymbol{\vartheta})$ are non-trivial, functionally independent, and continuous functions of ϑ_j.

3a For a continuous random variable, the derivatives $t'_j(x)$ are linearly independent, continuous functions of x over A.

3b For a discrete random variable, the $t_j(x)$ are non-trivial, linearly independent functions of x on A.

Examples: Binomial, Poisson, Gamma, Normal, Weibull distributions.

Theorem 10.4.2

If $X_1, \ldots, X_n$ is a random sample from $f(x;\boldsymbol{\vartheta})$ that satisfies Definition 10.4.2, then the statistics

$$S_1 = \sum_{i=1}^{n} t_1(X_i), \ldots, S_k = \sum_{i=1}^{n} t_k(X_i)$$

are a minimal set of complete sufficient statistics for $\vartheta_1, \ldots, \vartheta_k$.

Example 10.4.II (continuation of Example 10.3.I)

Consider a random sample from an uniform distribution, $X_i \sim \text{BIN}(p)$. It follows that

$$f(x;p) = p^x q^{1-x} I_{\{0,1\}}(x) = q\left(\frac{p}{q}\right)^x I_{\{0,1\}}(x) = qI_{\{0,1\}}(x)e^{x\log\left(\frac{p}{q}\right)}.$$

Then $S = \sum_{i=1}^{n} X_i$ is complete and sufficient, and according to Example 10.3.I, $\frac{S(n-S)}{n(n-1)}$ is the UMVUE for pq.

In the book, definitions, theorems, and examples are given for the case where A depends on the parameter. We do not address this case here but instead will make use of the fact that, in such cases, the sufficient statistic is usually the smallest or largest observation, or both.

Example 10.4.III

Consider a sample from the pdf $f(x;\vartheta) = \frac{5\vartheta^5}{x^6} I_{[\vartheta,\infty)}(x)$, $\vartheta > 0$, then

$$L(\vartheta) = \frac{5^n \vartheta^{5n}}{\prod_{i=1}^n x_i^6} \prod_{i=1}^n I_{[\vartheta,\infty)}(x_i) = \frac{5^n \vartheta^{5n}}{\prod_{i=1}^n x_i^6} I_{[\vartheta,\infty)}(x_{1:n}).$$

We see that $X_{1:n}$ is sufficient, but is it complete? To verify this,

$$f_{X_{1:n}}(x) = \frac{5n\vartheta^5}{x^6} \left(\int_x^\infty \frac{5\vartheta^5}{t^6}\, dt \right)^{n-1} = \frac{5n\vartheta^{5n}}{x^{5n+1}} I_{[\vartheta,\infty)}(x),$$

and

$$\mathbb{E}[u(X_{1:n})] = \int_\vartheta^\infty u(x) \frac{5n\vartheta^{5n}}{x^{5n+1}}\, dx = 0 \quad \Leftrightarrow \quad \int_\vartheta^\infty \frac{u(x)}{x^{5n+1}}\, dx = 0.$$

Differentiating with respect to ϑ yields $-\frac{u(\vartheta)}{\vartheta^{5n+1}} = 0$ for all $\vartheta > 0$. Conclusion: $X_{1:n}$ is complete.

$$\mathbb{E}[X_{1:n}] = \int_\vartheta^\infty \frac{5n\vartheta^{5n}}{x^{5n}}\, dx = \frac{5n}{5n-1}\vartheta \quad \Rightarrow \quad \frac{5n-1}{5n} X_{1:n} \text{ , UMVUE for } \vartheta.$$

Theorem 10.4.7 (Basu)

Let $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots, x_n; \boldsymbol{\vartheta})$; $\vartheta \in \Omega$, suppose that $\mathbf{S} = (S_1, \ldots, S_k)$ where $(S_1, \ldots, S_k)$ are jointly complete sufficient statistics for $\boldsymbol{\vartheta}$, and suppose that $\mathbf{T}$ is any other statistic. If the distribution of $\mathbf{T}$ does not involve $\boldsymbol{\vartheta}$, then $\mathbf{T}$ and $\mathbf{S}$ are stochastically independent.

Proof. We will consider the discrete case. Denote by $f(\mathbf{t})$, $f(\mathbf{s}; \boldsymbol{\vartheta})$, and $f(\mathbf{t} \mid \mathbf{s})$ the pdf's of $\mathbf{T}$, $\mathbf{S}$, and the conditional pdf of $\mathbf{T}$ given $\mathbf{S} = \mathbf{s}$, respectively. Consider the following expected value relative to the distribution of $\mathbf{S}$:

$$\begin{aligned}
\mathbb{E}_{\mathbf{S}}\left[f(\mathbf{t}) - f(\mathbf{t} \mid \mathbf{S})\right] &= f(\mathbf{t}) - \sum_{\mathbf{s}} f(\mathbf{t} \mid \mathbf{s}) f(\mathbf{s}; \boldsymbol{\vartheta}) \\
&= f(\mathbf{t}) - \sum_{\mathbf{s}} f(\mathbf{s}, \mathbf{t}; \boldsymbol{\vartheta}) \\
&= f(\mathbf{t}) - f(\mathbf{t}) = 0
\end{aligned}$$

Because $\mathbf{S}$ is a complete sufficient statistic, $f(\mathbf{t} \mid \mathbf{s}) = f(\mathbf{t})$, which means that $\mathbf{S}$ and $\mathbf{T}$ are stochastically independent. The continuous case is similar. □

It can be difficult to calculate the pdf of $\mathbf{T}$ and demonstrate that it does not depend on the parameter. Often, however, a simpler approach can be used.

Example 10.4.IV

Consider a sample from the pdf $f(x;\vartheta) = \frac{5\vartheta^5}{x^6} I_{[\vartheta,\infty)}(x)$, $\vartheta > 0$. From Example 10.4.III, we know that $X_{1:n}$ is complete and sufficient for ϑ. We will now show that

$$\frac{X_{k:n}}{X_{j:n}}, \qquad k = 2,\dots,n; \quad j = 1,\dots,n; \quad k \neq j$$

and $X_{1:n}$ are independent. If $Z_i := \frac{X_i}{\vartheta}$, then $f_{Z_i}(z) = 5z^{-6} I_{[1,\infty)}(z)$, which is independent of ϑ. Therefore, $\frac{Z_{k:n}}{Z_{j:n}}$ is also independent of ϑ. Since $\frac{Z_{k:n}}{Z_{j:n}} = \frac{X_{k:n}}{X_{j:n}}$, we conclude that the latter is also independent of ϑ, thus proving the statement.

Example 10.4.9

Let $X_1,\dots,X_n$ be a sample from $\mathrm{N}(\mu,\sigma^2)$. The MLEs are given by

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}.$$

For known σ^2, $\bar{X}$ is complete and sufficient for μ, and since

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1),$$

the pdf does not depend on μ. Thus, $\bar{X}$ and $\hat{\sigma}^2$ are independent. In the general case, $\bar{X}$ and $\hat{\sigma}^2$ are complete and sufficient for μ and σ^2. It follows that

$$\frac{X_i - \bar{X}}{\sigma} \sim \mathrm{N}\left(0, \frac{n-1}{n}\right) \quad \text{and} \quad \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1).$$

According to Basu's theorem, $\frac{X_i - \bar{X}}{\hat{\sigma}}$ and $\frac{n\hat{\sigma}^2}{\sigma}$ are stochastically independent of $\bar{X}$ and $\hat{\sigma}^2$.

Chapter 11

Interval Estimation

11.1 Introduction

The problem of point estimation was discussed in Chapter 9. Along with a point estimate of the value of a parameter, we want to have some understanding of how close we can expect our estimate to be to the true value. Some information on this question is provided by knowing the variance or the MSE of the estimator. Another approach would be to consider interval estimates; one then could consider the probability that such an interval will contain the true parameter value. Indeed, one could adjust the interval to achieve some prescribed probability level, and thus a measure of its accuracy would be incorporated automatically into the interval estimate.

11.2 Confidence Intervals

Let $X_1, \dots, X_n$ have joint pdf $f(x_1, \dots, x_n; \vartheta)$; $\vartheta \in \Omega$, where ω is an interval. Suppose that L and U are statistics, say$L = l(X_1, \dots, X_n)$ and $U = u(X_1, \dots, X_n)$. If an experiment yields data $x_1, \dots, x_n$, then we have the observed values $l(x_1, \dots, x_n)$ and $u(x_1, \dots, x_n)$.

Definition 11.2.1

Confidence Interval An interval $(l(x_1, \dots, x_n), u(x_1, \dots, x_n))$ is called a $100\gamma\%$ **confidence interval** for ϑ if

$$\mathbb{P}\left[l(X_1, \dots, X_n) \leq \vartheta \leq u(X_1, \dots, X_n)\right] = \gamma \qquad (11.2.1)$$

with $0 < \gamma < 1$. The observed values $l(x_1, \dots, x_n)$ and $u(x_1, \dots, x_n)$ are called the **lower and upper bounds** of the confidence interval.

Now, ϑ is a constant: it is either within the confidence interval or not. You cannot claim that it lies within the interval with probability γ. What is true is that if you take many samples and compute many confidence intervals, approximately $100\gamma\%$ of these intervals will contain ϑ. Sometimes, one is interested only in either a lower or upper bound, but not both.

Definition 11.2.2

One-Sided Confidence Limits

1. If
$$\mathbb{P}\left[l(X_1,\dots,X_n) < \vartheta\right] = \gamma \tag{11.2.2}$$
then $l(x_1,\dots,x_n)$ is called a **one-sided lower 100γ% confidence limit** for ϑ.

2. If
$$\mathbb{P}\left[\vartheta < u(X_1,\dots,X_n)\right] = \gamma \tag{11.2.3}$$
then $u(x_1,\dots,x_n)$ is called a **one-sided upper 100γ% confidence limit** for ϑ.

It may not always be clear how to obtain confidence limits that satisfy Definitions 11.2.1 or 11.2.2. The concept of sufficiency often offers some aid in this problem. If a single sufficient statistic S exists, then one might consider finding confidence limits that are functions of S. Otherwise, another reasonable statistic, such as an MLE, might be considered.

Example 11.2.1

Let $X_1,\dots,X_n$ be a sample from an EXP(ϑ) distribution. And we wish to derive a one-sided lower 100γ% confidence limit for ϑ. We know that $\bar{X}$ is sufficient for ϑ and also that $S := \sum_{i=1}^n X_i \sim$ GAM$(\vartheta, n) \Rightarrow \frac{2n\bar{X}}{\vartheta} \sim \chi^2(2n)$. Thus,

$$\gamma = \mathbb{P}\left[\frac{2n\bar{X}}{\vartheta} < \chi^2_\gamma(2n)\right] = \mathbb{P}\left[\frac{2n\bar{X}}{\chi^2_\gamma(2n)} < \vartheta\right]$$

If $\bar{x}$ is observed, then a one-sided lower 100γ% confidence limit is given by

$$l(\boldsymbol{x}) = \frac{2n\bar{x}}{\chi^2_\gamma(2n)} \tag{11.2.4}$$

Similarly, a one-sided 100γ% upper bound is given by

$$u(\boldsymbol{x}) = \frac{2n\bar{x}}{\chi^2_{1-\gamma}(2n)} \tag{11.2.5}$$

Since $\mathbb{P}\left[\frac{2n\bar{X}}{\vartheta} > \chi^2_{1-\gamma}(2n)\right] = \gamma$. If a 100γ% confidence interval for ϑ is desired, we choose values $\alpha_1 > 0$ and $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = 1-\gamma$, giving

$$\mathbb{P}\left[\chi^2_{\alpha_1}(2n) < \frac{2n\bar{X}}{\vartheta} < \chi^2_{1-\alpha_2}(2n)\right] = 1 - \alpha_1 - \alpha_2 = \gamma$$

often, we set $\alpha_1 = \alpha_2 = \frac{1-\gamma}{2}$, which is known as the equal-tailed interval choice.

11.3 Pivotal Quantity Method

Suppose that $X_1,\dots,X_n$ has joint pdf $f(x_1,\dots,x_n;\vartheta)$, and we wish to obtain confidence limits for ϑ where other unknown nuisance parameters also may be present.

Definition 11.3.1

Pivotal Quantity If $Q = q(X_1, \dots, X_n; \vartheta)$ is a random variable that is a function only of $X_1, \dots, X_n$ and ϑ, then Q is called a **pivotal quantity** if its distribution does not depend on ϑ or any other unknown parameters.

Examples.

$$\begin{aligned}
&\sqrt{n}\frac{\bar{X} - \mu}{3} \sim \mathrm{N}(0,1) \quad \text{if } X_i \sim \mathrm{N}(\mu, 9),\\
&\frac{2n\bar{X}}{\vartheta} \sim \chi^2(2n) \quad \text{if } X_i \sim \mathrm{EXP}(\vartheta),\\
&Q = \frac{X_{n:n}}{\vartheta} \Rightarrow f_Q(q) = nq^{n-1} I_{(0,1)}(q) \quad \text{if } X_i \sim \mathrm{UNIF}(0, \vartheta).
\end{aligned}$$

Example 11.3.1

In Example 11.2.1, we encountered a chi-square distributed random variable, which will be denoted here as $Q = \frac{2n\bar{X}}{\vartheta}$ and which clearly satisfies the definition of a pivotal quantity In that example we were able to proceed from a probability statement about Q to obtain confidence limits for ϑ. More generally, if Q is a pivotal quantity for a parameter ϑ and if percentiles of Q say q_1 and q_2, are available such that

$$\mathbb{P}\left[q_1 < q(X_1, \dots, X_n; \vartheta) < q_2\right] = \gamma \tag{11.3.1}$$

then for an observed sample $x_1, \dots, x_n$, a $100\gamma\%$ confidence set for ϑ is given by the set

$$\{\vartheta \in \Omega \mid q_1 < q(x_1, \dots, x_n; \vartheta) < q_2\}. \tag{11.3.2}$$

Such a set need not be an interval, but we always have an interval if $q(x_1, \dots, x_n; \vartheta)$ is a monotone function of ϑ for fixed $x_1, \dots, x_n$. In some cases, the MLEs yield pivotal quantities.

Definition 11.3.I

Location and Scale Parameters A parameter ϑ associated with the pdf $f(x; \vartheta)$ is called:

1. a **location parameter** if $f(x; \vartheta) = f_0(x - \vartheta)$ where $f_0(z)$ is independent of any unknown parameters.
2. a **scale parameter** if $f(x; \vartheta) = \frac{1}{\vartheta} f_0\left(\frac{x}{\vartheta}\right)$ where $f_0(z)$ meets the requirement of (1).
3. a **location-scale parameter** if $f(x; \vartheta_1, \vartheta_2) = \frac{1}{\vartheta_2} f_0\left(\frac{x - \vartheta_1}{\vartheta_2}\right)$, with $f_0(z)$ meeting the requirement of (1).

Theorem 11.3.1

Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf $f(x; \vartheta)$, for $\vartheta \in \Omega$, and assume that an MLE $\hat{\vartheta}$ exists.

1. If ϑ is a location parameter, $Q = \hat{\vartheta} - \vartheta$ is a pivotal quantity.
2. If ϑ is a scale parameter, $Q = \frac{\hat{\vartheta}}{\vartheta}$ is a pivotal quantity.

We already have seen examples of pivotal quantities that are slight variations of the ones suggested in this theorem. Specifically, if $X_i \sim \mathrm{N}(\mu, \sigma^2)$. With σ^2 known, μ is a location parameter, and the MLE is $\bar{X}$; thus $\bar{X} - \mu$ is a pivotal quantity. In Example 11.2.1, $X_i \sim \mathrm{EXP}(\vartheta)$, so that ϑ is a scale parameter and the MLE is $\bar{X}$; thus $\bar{X}/\vartheta$ is a pivotal quantity. Notice that it sometimes is convenient to make a slight modification, such as multiplying by a known scale factor, so that the pivotal quantity has a known distribution. For example, we know that $2n\bar{X}/\vartheta \sim \chi^2(2n)$, which has tabulated percentiles, so it might be better to let this be our pivotal quantity rather than $\bar{X}/\vartheta$.

Theorem 11.3.2

Let $X_1, \ldots, X_n$ e a random sample from a distribution with location-scale parameters

$$f(x; \vartheta_1, \vartheta_2) = \frac{1}{\vartheta_2} f_0\left(\frac{x - \vartheta_1}{\vartheta_2}\right).$$

If MLEs $\hat{\vartheta}_1$ and $\hat{\vartheta}_2$ exist, then $\frac{\hat{\vartheta}_1 - \vartheta_1}{\hat{\vartheta}_2}$ and $\frac{\hat{\vartheta}_2}{\vartheta_2}$ are pivotal quantities for ϑ_1 and ϑ_2, respectively.

Notice also that $\frac{\hat{\vartheta}_1 - \vartheta_1}{\vartheta_2}$ has a distribution that is free of unknown parameters, but it is not a pivotal quantity unless ϑ_2 is known. If sufficient statistics exist, then MLEs can be found that are functions of them, and the method should provide good results.

Example 11.3.2

Consider a random sample from a normal distribution, $X_i \sim \mathrm{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We know that $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}} = S\sqrt{\frac{n-1}{n}}$. According to Theorem 11.3.2, $\frac{\hat{\mu} - \mu}{\hat{\sigma}}$ and $\frac{\hat{\sigma}}{\sigma}$ are pivotal quantities. Using results from Chapter 8, Theorems 8.4.3 and 8.3.6(3), we get:

$$\sqrt{n}\frac{\bar{X} - \mu}{S} \sim t(n-1) \tag{11.3.3}$$

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \tag{11.3.4}$$

Lt may not always be possible to find a pivotal quatity based on MLEs, but for a sample from a continuous distribution with a single unknown parameter, at least one pivotal quantity can always be derived by use of the probability integral transform. From Chapter 6, Theorem 6.3.3, if $X_i \sim f(x; \vartheta)$ and $F(x; \vartheta)$ is the CDF of X_i, then $F(X_i; \vartheta) \sim \mathrm{UNIF}(0, 1)$ and

$-\log F(X_i;\vartheta) \sim$ EXP(1). For a random sample $X_1, \ldots, X_n$, it follows that

$$-2\sum_{i=1}^{n} \log F(X_i;\vartheta) \sim \chi^2(2n) \tag{11.3.10}$$

leading to

$$\mathbb{P}\left[\chi^2_{\alpha/2}(2n) < -2\sum_{i=1}^{n} \log F(X_i;\vartheta) < \chi^2_{1-\alpha/2}\right] = 1-\alpha \tag{11.3.11}$$

If $F(X;\vartheta) \sim \text{UNIF}(0,1)$, then $1 - F(X;\vartheta) \sim \text{UNIF}(0,1)$ and

$$-2\sum_{i=1}^{n} \log\left(1 - F(X_i;\vartheta)\right) \sim \chi^2(2n) \tag{11.3.12}$$

Generally, (11.3.10) and (11.3.12) yield different results. The most computationally convenient expression is used, see Example 11.3.14.

Approximate Confidence Intervals

For some discrete random variables and problems with multiple parameters, a pivotal quantity may not exist. In such cases, the Central Limit Theorem (CLT) is used:

$$\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \xrightarrow{d} Z \sim \text{N}(0,1)$$

If σ is known, this yields an asymptotic pivotal quantity for μ. If σ is unknown but has a consistent estimator $\hat{\sigma}$, then:

$$\sqrt{n}\frac{\bar{X}-\mu}{\hat{\sigma}} \xrightarrow{d} Z \sim \text{N}(0,1)$$

Example 11.3.5

Consider a random sample from a Bernoulli distribution, $X_i \sim \text{BIN}(1,p)$. The MLE of p is $\hat{p} = \bar{X}$. We also know that $\hat{p}$ is sufficient and that $\sum_{i=1}^{n} X_i \sim \text{BIN}(n,p)$, but there is no pivotal quantity for p. However, by the CLT,

$$\sqrt{n}\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow{d} Z \sim \text{N}(0,1) \tag{11.3.17}$$

Thus, for large n, we also have the approximate result

$$\mathbb{P}\left[-z_{1-\alpha/2} < \sqrt{n}\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}} < z_{1-\alpha/2}\right] \approx 1-\alpha \tag{11.3.18}$$

This statement is much easier to invert, and approximate confidence limits for p are given by

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})}. \tag{11.3.19}$$

Other important distributions also admit approximate pivotal quantities, see Example 11.3.6 for the Poisson distribution.

11.4 The General Method

If a pivotal quantity is not available, then it is still possible to determine a confidence region for a parameter ϑ if a statistic exists with a distribution that depends on ϑ but not on any other unknown nuisance parameters. Specifically, let $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots, x_n; \vartheta)$, and $S = s(X_1, \ldots, X_n) \sim g(s; \vartheta)$. Preferably S will be sufficient for ϑ, or possibly some reasonable estimator such as an MLE, but this is not required. Now, for each possible value of ϑ , assume that we can find values $h_1(\vartheta)$ and $h_2(\vartheta)$ such that

$$\mathbb{P}\left[h_1(\vartheta) < S < h_2(\vartheta)\right] = 1 - \alpha. \tag{11.4.1}$$

If we observe $S = s$, then the set of values

$$\{\vartheta \in \Omega \mid h_1(\vartheta) < s < h_2(\vartheta)\}$$

is a $100(1-\alpha)\%$ confidence region for ϑ. In other words, if ϑ_0 is the true value of ϑ, then ϑ_0 will be in the confidence region if and only if $h_1(\vartheta_0) < s < h_2(\vartheta_0)$, which has $100(1-\alpha)\%$ confidence level because equation (11.4.1) holds with $\vartheta = \vartheta_0$ in this case. Quite often $h_1(\vartheta)$ and $h_2(\vartheta)$ will be monotonic increasing (or decreasing) functions of ϑ, and the resulting confidence region will be an interval.

Example 11.4.1

Consider a random sample of size n from the continuous distribution with pdf

$$f(x; \vartheta) = \begin{cases} \frac{1}{\vartheta^2} \exp\left[-\frac{(x-\vartheta)}{\vartheta^2}\right] & \text{if } x \geq \vartheta, \\ 0 & \text{if } x < \vartheta \end{cases}$$

with $\vartheta > 0$. There is no single sufficient statistic, but $X_{1:n}$ and $\sum X_i$ are jointly sufficient for ϑ. It is desired to derive a 90% confidence interval for ϑ based on the statistic $S = X_{1:n}$. The CDF of S is

$$G(s; \vartheta) = \begin{cases} 1 - \exp\left[-n\frac{(s-\vartheta)}{\vartheta^2}\right] & \text{if } s \geq \vartheta, \\ 0 & \text{if } s < \vartheta \end{cases}$$

One possible choice of functions $h_1(\vartheta)$ and $h_2(\vartheta)$ that satisfies (11.4.1) can be obtained by solving

$$G(h_1(\vartheta); \vartheta) = 0.05 \quad \text{and} \quad G(h_2(\vartheta); \vartheta) = 0.95.$$

This yields:

$$\begin{aligned} h_1(\vartheta) &= \vartheta - \log(0.95)\vartheta^2/n \approx \vartheta + 0.0513\vartheta^2/n, \\ h_2(\vartheta) &= \vartheta - \log(0.05)\vartheta^2/n \approx \vartheta + 2.996\vartheta^2/n. \end{aligned}$$

Suppose now that a sample of size $n = 10$ yields a minimum observation $s = x_{1:10} = 2.50$. The solutions of $2.50 = h_1(\vartheta)$ and $2.50 = h_2(\vartheta)$ are $\vartheta_1 = 2.469$ and $\vartheta_2 = 1.667$. Because $h_1(\vartheta)$ and $h_2(\vartheta)$ are increasing, the set of all $\vartheta > 0$ such that $h_1(\vartheta) < 2.50 < h_2(\vartheta)$ is the interval $(1.667, 2.469)$, which is a 90% confidence interval for ϑ.
Calling the left and right limits ϑ_L and ϑ_U, respectively. If $h_i(\vartheta)$ are increasing functions, then $h_2(\vartheta_L) = s$ and $h_1(\vartheta_U) = s$. If $h_i(\vartheta)$ are decreasing, then $h_1(\vartheta_L) = s = h_2(\vartheta_U)$.

Theorem 11.4.1

Let the statistic S be continuous with CDF $G(s;\vartheta)$, and suppose that $h_1(\vartheta)$ and $h_2(\vartheta)$ are functions that satisfy

$$G(h_1(\vartheta);\vartheta) = \alpha_1 \quad \text{and} \quad G(h_2(\vartheta);\vartheta) = 1-\alpha_2 \qquad (11.4.2)\ \&\ (11.4.3)$$

for each $\vartheta \in \Omega$, where $0 < \alpha_i < 1$. Let s be an observed value of S. If $h_1(\vartheta)$ and $h_2(\vartheta)$ are increasing functions of ϑ, then the following statements hold:

1. A one-sided lower $100(1-\alpha_2)\%$ confidence limit, ϑ_L, is the solution of
$$h_2(\vartheta_L) = s. \qquad (11.4.4)$$
2. A one-sided upper $100(1-\alpha_1)\%$ confidence limit ϑ_U, is the solution of
$$h_1(\vartheta_U) = s. \qquad (11.4.5)$$
3. If $\alpha = \alpha_1 + \alpha_2$ and $0 < \alpha < 1$, then $(\vartheta_L, \vartheta_U)$ is a $100(1-\alpha)\%$ confidence interval for ϑ.

If $h_i(\vartheta)$ are decreasing functions, the roles are reversed. If $G(s;\vartheta)$ is a decreasing function of ϑ for each fixed s, then $h_1(\vartheta)$ and $h_2(\vartheta)$ are increasing functions of ϑ. This leads to

Theorem 11.4.2

Suppose that the statistic S is continuous with CDF $G(s;\vartheta)$, and let s be an observed value of S. If $G(s;\vartheta)$ is a decreasing function of ϑ, then the following statements hold:

1. A one-sided lower $100(1-\alpha_2)\%$ confidence limit, ϑ_L, is the solution of
$$G(s;\vartheta_L) = 1-\alpha_2. \qquad (11.4.6)$$
2. A one-sided upper $100(1-\alpha_1)\%$ confidence limit ϑ_U, is the solution of
$$G(s;\vartheta_U) = \alpha_1. \qquad (11.4.7)$$
3. If $\alpha = \alpha_1 + \alpha_2$ and $0 < \alpha < 1$, then $(\vartheta_L, \vartheta_U)$ is a $100(1-\alpha)\%$ confidence interval for ϑ.

A similar theorem can be stated for the case where $G(s;\vartheta)$ is an increasing function of ϑ, giving $G(s;\vartheta_U) = 1-\alpha_2$ and $G(s;\vartheta_L) = \alpha_1$.

Example 11.4.3

Consider the statistic $S = X_{1:n}$ of Example 11.4.1.

$$G(s;\vartheta) = \left(1 - \exp\left[-\frac{n(s-\vartheta)}{\vartheta^2}\right]\right) I_{[\vartheta,\infty)}(s).$$

The function $G(s;\vartheta)$ is decreasing in ϑ. Setting $\alpha_2 = 0.05 = \alpha_1$, $n = 10$, $s = 2.5$, then, $G(2.5;\vartheta_L) = 0.95$ and $G(2.5;\vartheta_U) = 0.05$.

It also is possible to state a more general theorem that includes discrete cases, but it is not always possible to achieve a prescribed confidence level when the observed statistic is discrete. However, **conservative confidence intervals**, in general, can be obtained.

Definition 11.4.1

An observed confidence interval $(\vartheta_L, \vartheta_U)$ is called a **conservative** $100(1-\alpha)\%$ **confidence interval** for ϑ if the corresponding random interval contains the true value of ϑ with probability **at least** $(\geq)$ $1-\alpha$.

Theorem 11.4.3

Let S be a statistic with CDF $G(s;\vartheta)$, and let $h_1(\vartheta)$ and $h_2(\vartheta)$ be functions satisfying

$$G(h_1(\vartheta);\vartheta) = \alpha_1 \quad \text{and} \quad \mathbb{P}\left[S < h_2(\vartheta);\vartheta\right] = 1-\alpha_2. \tag{11.4.8}$$

with $0 < \alpha_i < 1$, $i = 1, 2$.

1. If $h_i(\vartheta)$ are increasing functions, then a conservative one-sided lower $100(1-\alpha_2)\%$ confidence limit ϑ_L, is the solution of

   $$\mathbb{P}\left[S < s;\vartheta_L\right] = 1-\alpha_2. \tag{11.4.9}$$

   A conservative one-sided upper $100(1-\alpha_1)\%$ confidence limit ϑ_U is a solution of $G(s;\vartheta_U) = \alpha_1$.

2. If $h_i(\vartheta)$ are decreasing functions, then ϑ_L and ϑ_U are found from

   $$G(s;\vartheta_L) = \alpha_1 \quad \text{and} \quad \mathbb{P}\left[S < s;\vartheta_U\right] = 1-\alpha_2. \tag{11.4.10}$$

3. In either case, if $\alpha = \alpha_1 + \alpha_2$ with $0 < \alpha < 1$, then $(\vartheta_L, \vartheta_U)$ is a conservative $100(1-\alpha)\%$ confidence interval for ϑ.

Example 11.4.4

We now desire to derive a conservative one-sided $(1-\alpha)100\%$ confidence limit for p. We know that $S = \sum X_i$ is a sufficient statistic, and $S \sim$ BIN(n,p). We will not find explicit expressions for $h_1(p)$ and $h_2(p)$ in this example, but note that $G(s;p)$ is a decreasing function of p:

$$G(s;p) = \sum_{y=0}^{s} \binom{n}{y} p^y (1-p)^{n-y}.$$

Taking the derivative with respect to p

$$\begin{aligned}\tfrac{d}{dp}G(s;p) &= \sum_{y=0}^{s}\tbinom{n}{y}yp^{y-1}(1-p)^{n-y}-\sum_{y=0}^{s}\tbinom{n}{y}(n-y)p^{y}(1-p)^{n-y-1}\\ &= n\sum_{y=1}^{s}\tbinom{n-1}{y-1}p^{y-1}(1-p)^{n-y}-n\sum_{y=0}^{s}\tbinom{n-1}{y}p^{y}(1-p)^{n-y-1}\\ &= n\sum_{t=0}^{s-1}\tbinom{n-1}{t}p^{t}(1-p)^{n-1-t}-n\sum_{y=0}^{s}\tbinom{n-1}{y}p^{y}(1-p)^{n-y-1}\\ &= -n\tbinom{n-1}{s}p^{s}(1-p)^{n-s-1}.\end{aligned}$$

Thus, for an observed value s, a solution of

$$\alpha_1=\sum_{y=0}^{s}\binom{n}{y}p_U^y(1-p_U)^{n-y}$$

is a conservative one-sided upper limit, and a solution of

$$1-\alpha_2=\sum_{y=0}^{s-1}\binom{n}{y}p_L^y(1-p_L)^{n-y}$$

is a conservative one-sided lower limit. If $\alpha=\alpha_1+\alpha_2$, then a conservative $100(1-\alpha)\%$ confidence interval is given by (p_L,p_U).

11.5 Two Sample Problems

Quite often random samples are taken for the purpose of comparing two or more populations. One may be interested in comparing the mean yields of two processes or the relative variation in yields of two processes. Confidence intervals are quite informative in making such comparisons.

Consider independent random samples of sizes n_1 and n_2 from two normally distributed populations $X_i\sim\mathrm{N}(\mu_1,\sigma_1^2)$ and $Y_j\sim\mathrm{N}(\mu_2,\sigma_2^2)$, respectively. Denote by $\bar{X},\bar{Y},S_1^2$, and S_2^2 the sample means and sample variances. Suppose we wish to know whether one population has a smaller variance (or expectation) than the other.

Procedure for Variances

A confidence interval for the ratio $\frac{\sigma_2^2}{\sigma_1^2}$ can be derived using Snedecor's F distribution as suggested in Example 8.4.1. Specifically, we know that

$$\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}\sim F(n_1-1,n_2-1) \tag{11.5.1}$$

which provides a pivotal quantity for $\frac{\sigma_2^2}{\sigma_1^2}$. Percentiles for the F distribution with $\nu_1=n_1-1$ and $\nu_2=n_2-1$ can be obtained from Table 7 (Appendix C), so

$$\mathbb{P}\left[f_{\alpha/2}(\nu_2,\nu_1)<\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}<f_{1-\alpha/2}(\nu_2,\nu_1)\right]=1-\alpha \tag{11.5.2}$$

Thus, if s_1^2 and s_2^2 are estimates, then a $(1-\alpha)100\%$ confidence interval for $\frac{\sigma_2^2}{\sigma_1^2}$ is given by:

$$\left(\frac{s_2^2}{s_1^2} f_{\alpha/2}(n_1-1, n_2-1), \frac{s_2^2}{s_1^2} f_{1-\alpha/2}(n_1-1, n_2-1)\right) \tag{11.5.3}$$

Procedure for Means

If the variances, σ_1^2 and σ_2^2, are known, then a pivotal quantity for the difference $\mu_2 - \mu_1$ is easily obtained. Specifically, because

$$\bar{Y} - \bar{X} - (\mu_2 - \mu_1) \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right), \tag{11.5.4}$$

it follows that

$$Z = \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1). \tag{11.5.5}$$

With this choice of Z, the statement

$$\mathbb{P}\left[-z_{1-\alpha/2} < Z < z_{1-\alpha/2}\right] = 1 - \alpha$$

can be solved to obtain a $100(1-\alpha)\%$ confidence interval for $\mu_2 - \mu_1$:

$$\bar{y} - \bar{x} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \tag{11.5.6}$$

In most cases, the variances will not be known, but in some cases it will be reasonable to assume that the variances are unknown but equal. For instance, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the common variance can be eliminated using a pooled estimate. A pooled estimator of the common variance is the weighted average

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \tag{11.5.7}$$

and if

$$V = \frac{(n_1+n_2-2)S_p^2}{\sigma^2} \tag{11.5.8}$$

then

$$V = \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_1+n_2-2). \tag{11.5.9}$$

Since $\bar{X}$ and $\bar{Y}$ are independent of S_1^2 and S_2^2, it follows that

$$T = \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{Z}{\sqrt{\frac{V}{n_1+n_2-2}}} \sim t(n_1+n_2-2). \tag{11.5.10}$$

Thus, the $(1-\alpha)100\%$ confidence interval for $\mu_2 - \mu_1$ is

$$\bar{y} - \bar{x} \pm t_{1-\alpha/2}(n_1+n_2-2)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{11.5.11}$$

Chapter 12

Tests of Hypotheses

12.1 Introduction

Someone who thinks they have Bechterew's disease visits their general practitioner for an examination. There are several possible outcomes:

1. The doctor concludes that the patient does not have the disease, while in fact, they do.
2. The doctor concludes that the patient does have the disease, while in fact, they do not.
3. The doctor correctly concludes that the patient does not have the disease.
4. The doctor correctly concludes that the patient does have the disease.

The doctor proceeds as follows: the patient undergoes k tests with outcomes $X_1, \ldots, X_k$. If $(X_1, \ldots, X_k)$ lies in $R \subset \mathbb{R}^k$, the conclusion is, "the patient does not have the disease." Otherwise, the conclusion is, "the patient does have the disease." In cases 3 and 4, the correct conclusions are drawn; in the other cases, the decisions are incorrect. This is a simple example of hypothesis testing. In hypothesis testing, we deal with a **hypothesis** about the parameter(s) of a random variable X, more precisely:

Definition 12.1.1

If $X \sim f(x; \boldsymbol{\vartheta})$ with $\boldsymbol{\vartheta} \in \Omega$, a **statistical hypothesis** is a statement about the distribution of X, such as $\boldsymbol{\vartheta} \in \Omega_0 \subset \Omega$. If the hypothesis completely specifies $f(x; \boldsymbol{\vartheta})$, then it is referred to as a **simple hypothesis**; otherwise it is called a **composite hypothesis**.

Quite often the distribution in question has a known parametric form, such as the exponential distribution, with a single unknown parameter ϑ, and the hypothesis consists of a statement about ϑ. We distinguish between the **null hypothesis** and the **alternative hypothesis**. If the parameter space is Ω, then:

$$H_0 : \boldsymbol{\vartheta} \in \Omega_0 \subset \Omega \qquad \text{versus} \qquad H_a : \boldsymbol{\vartheta} \in \Omega - \Omega_0$$

In some books, the index of the alternative hypothesis is 1 instead of a. The set of all possible outcomes of the sample is called the sample space S. If the sample falls in $C \subset S$, we reject the null hypothesis; if the sample falls in $S - C$, we do not reject the null hypothesis.

Definition 12.1.2

The **critical region** for a test of hypotheses is the subset of the sample space that corresponds to rejecting the null hypothesis.

The crucial question is, of course, how to determine the critical region. From Section 12.6 onward, we will construct critical regions based on certain principles; until then, we will proceed heuristically.

Example 12.1.1

A theory proposes that the yield of a certain chemical reaction is normally distributed, $X \sim \mathrm{N}(\mu, 16)$. Past experience indicates that $\mu = 10$ if a certain mineral is not present, and $\mu = 11$ if the mineral is present. Our experiment would be to take a random sample of size n. On the basis of that sample, we would try to decide which case is true. That is, we wish to test the "null hypothesis" $H_0 : \mu = \mu_0 = 10$ against the "alternative hypothesis" $H_a : \mu = \mu_a = 11$.

In our example, $\bar{X}$ is a sufficient statistic for ϑ, so we may conveniently express the critical region directly in terms of the univariate variable $\bar{X}$, and we will refer to $\bar{X}$ as the **test statistic**. If the alternative hypothesis is true, we expect $\bar{X}$ to be larger. Therefore, we will reject H_0 if $\bar{X}$ is sufficiently large. The critical regions are of the form $C = \{(x_1, \dots, x_n) \mid \bar{x} > c\}$. In general, two types of errors can be made with this procedure:

1. Type I error: Reject a true H_0.

2. Type II error: Fail to reject a false H_0.

In Example 12.1.1:

$$\begin{aligned}
\mathbb{P}\left[\text{Type I error}\right] &= \mathbb{P}\left[\bar{X} > c|\mu = 10\right] = \mathbb{P}\left[\sqrt{n}\tfrac{\bar{X}-10}{4} > \sqrt{n}\tfrac{c-10}{4}\right] = 1 - \Phi\left(\sqrt{n}\tfrac{c-10}{4}\right) \\
\mathbb{P}\left[\text{Type II error}\right] &= \mathbb{P}\left[\sqrt{n}\tfrac{\bar{X}-11}{4} \leq \sqrt{n}\tfrac{c-11}{4}\right] = \Phi\left(\sqrt{n}\tfrac{c-11}{4}\right)
\end{aligned}$$

We hope to choose a test statistic and a critical region so that we would have a small probability of making these two errors. We will adopt the following notations for these error probabilities:

1. $\mathbb{P}\left[\text{Type I error}\right] = \mathbb{P}\left[\text{TI}\right] = \alpha$

2. $\mathbb{P}\left[\text{Type II error}\right] = \mathbb{P}\left[\text{TII}\right] = \beta$

Definition 12.1.3

For a simple null hypothesis, H_0, the probability of a Type I error $\alpha = \mathbb{P}[\text{TI}]$ is referred to as the **significance level** of the test. For a composite null hypothesis $H_0 : \vartheta \in \Omega_0$, the **size** of the test (or size of the critical region) is $\max_{\vartheta \in \Omega_0} \mathbb{P}[\text{TI} \mid \vartheta]$.

In the patient and doctor example, the first case describes a Type II error, and the second case describes a Type I error, assuming the null hypothesis states the patient does not have the disease. We proceed by choosing a significance level for the Type I error, usually 0.05, 0.025, or 0.01, and then try to find a critical region that minimizes the Type II error. In Example 12.1.1, if $\alpha = 0.05$ and $n = 25$. Then:

$$0.05 = 1 - \Phi\left(5\frac{c-10}{4}\right) \quad \Rightarrow \quad \Phi\left(5\frac{c-10}{4}\right) = 0.95 \quad \Rightarrow \quad c = \frac{12.5 + 1.645}{1.25} = 11.316$$

The probability of Type II error for the critical region C is

$$\begin{aligned}\beta = \mathbb{P}\,[\text{TII}] &= \mathbb{P}\left[Z \leq 5\frac{11.316 - 11}{4}\right] \\ &= \mathbb{P}\,[Z \leq 0.395] = 0.654\end{aligned}$$

Notice that we have a relatively large Type II error, so we increase the sample size to 100. To maintain a critical region of size $\alpha = 005$, we would now use

$$c_2 = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 10 + 1.645(4)/10 = 10.658$$

The value β in this case is

$$\begin{aligned}\beta = \mathbb{P}\,[\text{TII}] &= \mathbb{P}\left[Z \leq 10\frac{10.658 - 11}{4}\right] \\ &= \mathbb{P}\,[Z \leq -0.855] = 0.196\end{aligned}$$

The larger the sample size, the smaller the Type II error for a fixed Type I error rate. This means that, in practice, we focus our interest on the alternative hypothesis. With a fixed probability of committing a Type I error, we then examine whether there is sufficient statistical evidence for the alternative hypothesis to reject the null hypothesis.

The large Type II error for $n = 25$ in the example above indicates that, even if the alternative hypothesis is true, there is a high probability that a sample of size $n = 25$ will not provide sufficient statistical evidence to reject the null hypothesis. Under the null hypothesis, we reject H_0 in 5% of cases, while under the alternative hypothesis, rejection occurs in 34.6% of cases. This latter probability increases as the sample size (and therefore the amount of statistical information available) increases.

Definition 12.1.4

The **power function**, $\pi(\vartheta)$, of a test of H_0 is the probability of rejecting H_0 when the true value of the parameter is ϑ.

For simple hypotheses $H_0 : \vartheta = \vartheta_0$ versus $H_a : \vartheta = \vartheta_a$, we have $\pi(\vartheta_0) = \mathbb{P}[\text{TI}] = \alpha$ and $\pi(\vartheta_a) = 1 - \mathbb{P}[\text{TII}] = 1 - \beta$. For composite hypotheses, say, $H_0 : \vartheta \in \Omega_0$ versus $H_a : \vartheta \in \Omega - \Omega_0$, the size of the test (or critical region) is

$$\alpha = \max_{\vartheta \in \Omega_0} \pi(\vartheta). \tag{12.1.5}$$

and if the true value $\vartheta \in \Omega - \Omega_0$, then $\pi(\vartheta) = 1 - \mathbb{P}\,[\text{TII}]$, where we note that $\mathbb{P}\,[\text{TII}]$ depends on ϑ.

12.2 Composite Hypotheses

Again, we assume again that $X \sim \mathrm{N}(\mu, \sigma^2)$ with σ^2 known, and we wish to test $H_0 : \mu = \mu_0$ against the composite alternative $H_a : \mu > \mu_0$. Ii was suggested in the previous example that the critical region should be located on the right hand tail for any alternative $\mu_1 > \mu_0$, but the value of the critical value c did not depend on the value of μ_1 Thus it is clear that the test for the simple alternative also is valid for this composite alternative A test at significance level still would reject H_0 if

$$z_0 = \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma} \geq z_{1-\alpha} \tag{12.2.1}$$

The power of this test at any value μ is

$$\begin{aligned} \pi(\mu) &= \mathbb{P}\left[\sqrt{n}\tfrac{\bar{X}-\mu_0}{\sigma} \geq z_{1-\alpha} \mid \mu\right] \\ &= \mathbb{P}\left[\sqrt{n}\tfrac{\bar{X}-\mu}{\sigma} \geq z_{1-\alpha} + \sqrt{n}\tfrac{\mu_0-\mu}{\sigma} \mid \mu\right] \\ &= 1 - \Phi\left(z_{1-\alpha} + \sqrt{n}\tfrac{\mu_0-\mu}{\sigma}\right). \end{aligned} \tag{12.2.2}$$

We also may consider a composite null hypothesis Suppose that we wish to test $H_0 : \mu \leq \mu_0$ against $H_a : \mu > \mu_0$, and we reject H_0 if inequality (12.2.1) is satisfied. which results in the same test, as it follows from (12.2.2) that

$$\max_{\mu \leq \mu_0} \pi(\mu) = \pi(\mu_0) = \alpha$$

p-Value

When a test is conducted, the experimenter only knows whether the null hypothesis was rejected or not, based on the chosen size of the test. However, the experimenter might prefer to use a different significance level. For this reason, we use the concept of the p-value.

The ***p*-value** is defined as the smallest size α at which H_0 can be rejected, based on the observed value of the test statistic.

Example 12.2.1

On the basis of a sample of size $n = 25$ from a normal distribution, $X_i \sim \mathrm{N}(\mu, 16)$, we wish to test $H_0 : \mu = 10$ versus $H_a : \mu = 11$. Suppose that we observe $\bar{x} = 11.4$. The p-value is

$$\begin{aligned} \mathbb{P}\left[\bar{X} \geq 11.4 \mid \mu = 10\right] &= \mathbb{P}\left[Z \geq 5\frac{11.4 - 10}{4}\right] \\ &= \mathbb{P}\left[Z \geq 1.75\right] = 0.0401 \end{aligned}$$

Because $0.01 < 0.0401 < 0.05$, the test would reject at the 0.05 level but not at the 0.01 level. If the p-value is reported, then interested readers can apply their own criteria.

The test with a critical region of form (12.2.1) is called an **upper one-sided test**, and the form (12.2.3) corresponds to a **lower one-sided test**. Another common type of test involves a **two-sided alternative**.

Consider a sample from a normal distribution with known variance, and test $H_0 : \mu = \mu_0$ against the alternative $H_a : \mu \neq \mu_0$, with the test statistic $z_0 = \sqrt{n}\frac{\bar{X}-\mu_0}{\sigma}$. A good compromise is to use a two-sided critical region and reject H_0 if

$$z_0 \leq -z_{1-\alpha/2} \qquad \text{or} \qquad z_0 \geq z_{1-\alpha/2}. \tag{12.2.4}$$

The power function for the two-sided test is

$$\begin{aligned} \pi(\mu) &= 1 - \mathbb{P}\left[-z_{1-\alpha/2} < Z_0 < z_{1-\alpha/2} \mid \mu\right] \\ &= 1 - \Phi\left(z_{1-\alpha/2} + \sqrt{n}\tfrac{\mu_0-\mu}{\sigma}\right) + \Phi\left(-z_{1-\alpha/2} + \sqrt{n}\tfrac{\mu_0-\mu}{\sigma}\right). \end{aligned} \tag{12.2.5}$$

12.3 Tests for the Normal Distribution

In this section, we will state theorems that summarize the common test procedures for the parameters of a normal distribution. All results are based on Corollary 8.3.1 and Theorem 8.3.6, which state that for a sample from $\mathrm{N}(\mu, \sigma^2)$

1. $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$
2. $\bar{X}$ and S^2 are independent
3. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

Tests for the Mean (σ^2 Known)

The results discussed in the previous section are summarized in the following theorem.

Theorem 12.3.1

Suppose that $x_1, \ldots, x_n$ is an observed random sample from $\mathrm{N}(\mu, \sigma^2)$, and let

$$z_0 = \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma} \tag{12.3.1}$$

1. A size α test of $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$ consists of rejecting H_0 if $z_0 \geq z_{1-\alpha}$. The power of this test is

$$\pi(\mu) = 1 - \Phi\left(z_{1-\alpha} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}\right). \tag{12.3.2}$$

2. A test of size α of $H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$ consists of rejecting H_0 if $z_0 \leq -z_{1-\alpha}$. The power of this test is

$$\pi(\mu) = \Phi\left(-z_{1-\alpha} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}\right). \tag{12.3.3}$$

Theorem 12.3.1, continued

3. A test of size α for $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ consists of rejecting H_0 if $z_0 \leq -z_{1-\alpha/2}$ or $z_0 \geq z_{1-\alpha/2}$.
4. The sample size needed to achieve power $1 - \beta$ at size α is given by

$$n = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \sigma^2}{\left(\mu_0 - \mu\right)^2} \tag{12.3.4}$$

for a one-sided test, and

$$n \approx \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \sigma^2}{\left(\mu_0 - \mu\right)^2} \tag{12.3.5}$$

for a two-sided test.

Note that all these tests are determined by the MLE; we reject when the MLE is too large, too small, or either, depending on the null hypothesis. If the variance is unknown, we replace σ with s in the test statistics.

Theorem 12.3.2

Let $x_1, \ldots, x_n$ be an observed random sample from $\mathrm{N}(\mu, \sigma^2)$, and let

$$t_0 = \sqrt{n}\frac{\bar{x} - \mu_0}{s} \tag{12.3.6}$$

The null hypotheses and alternatives are the same as in the previous theorem, but we replace z_0 with t_0 and $z_{1-\alpha}$ ($z_{1-\alpha/2}$) with $t_{1-\alpha}(n-1)$ ($t_{1-\alpha/2}(n-1)$). The power formulas become slightly more complex, but not significantly. Determining the necessary sample size for a given power is more challenging.

Test for Variance

It is possible to construct tests of hypotheses such as $H_0 : \sigma^2 = \sigma_0^2 \quad \text{versus} \quad H_a : \sigma^2 > \sigma_0^2$ based on the test statistic

$$V_0 = (n-1)S^2/\sigma_0^2 \tag{12.3.8}$$

because $V_0 \sim \chi^2(n-1)$ when H_0 is true. An observed value of the sample variance, S^2, that is large relative to σ_0^2 would support H_a. In the following theorem, $H(c; v)$ is the CDF of $\chi^2(v)$.

Theorem 12.3.3

Let $x_1, \ldots, x_n$ be an observed sample from $\mathrm{N}(\mu, \sigma^2)$, and consider

$$v_0 = \frac{(n-1)s^2}{\sigma_0^2} \tag{12.3.9}$$

1. A test of size α for $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_a : \sigma^2 > \sigma_0^2$ consists of rejecting H_0 if $v_0 \geq \chi^2_{1-\alpha}(n-1)$. The power of this test is

$$\pi(\sigma^2) = 1 - H\left[\frac{\sigma_0^2}{\sigma^2}\chi^2_{1-\alpha}(n-1); n-1\right] \tag{12.3.10}$$

2. A test of size α for $H_0 : \sigma^2 \geq \sigma_0^2$ versus $H_a : \sigma^2 < \sigma_0^2$ consists of rejecting H_0 if $v_0 \leq \chi^2_{\alpha}(n-1)$. The power of this test is

$$\pi(\sigma^2) = H\left[\frac{\sigma_0^2}{\sigma^2}\chi^2_{\alpha}(n-1); n-1\right]. \tag{12.2.11}$$

3. A test of size α for $H_0 : \sigma^2 = \sigma_0^2$ versus $H_a : \sigma^2 \neq \sigma_0^2$ consists of rejecting H_0 if $v_0 \leq \chi^2_{\alpha/2}(n-1)$ or $v_0 \geq \chi^2_{1-\alpha/2}(n-1)$.

Proof. We derive the power of part 1; the other cases follow similarly

$$\begin{aligned}
\pi(\sigma^2) &= \mathbb{P}\left[V_0 \geq \chi^2_{1-\alpha}(n-1) \mid \sigma^2\right] \\
&= \mathbb{P}\left[\tfrac{(n-1)S^2}{\sigma^2} \geq \tfrac{\sigma_0^2}{\sigma^2}\chi^2_{1-\alpha}(n-1) \mid \sigma^2\right] \\
&= 1 - H\left[\tfrac{\sigma_0^2}{\sigma^2}\chi^2_{1-\alpha}(n-1); n-1\right].
\end{aligned}$$

Notice that in particular, $\pi(\sigma^2) = 1 - H\left[\chi^2_{1-\alpha}(n-1); n-1\right] = 1-(1-\alpha) = \alpha$, and because $\pi(\sigma^2)$ is increasing, the size of the critical region is α. The MLE plays a significant role here as well.

Two-Sample Tests

It is possible to construct tests of hypotheses concerning the variances of two normal distributions, such as $H_0 : \frac{\sigma_2^2}{\sigma_1^2} = d_0$, based on an F statistic. In particular, consider the test statistic

$$F_0 = \frac{S_1^2}{S_2^2}d_0 \sim F(n_1-1, n_2-1). \tag{12.3.14}$$

where $F_0 \sim F(n_1 - 1, n_2 - 1$ if H_0 is true. This leads to the following theorem:

Theorem 12.3.4

Suppose that $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ are observed values of independent random samples from $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$, respectively, and let

$$f_0 = \frac{s_1^2}{s_2^2} d_0 \tag{12.3.15}$$

1. A test of size α for $H_0 : \frac{\sigma_2^2}{\sigma_1^2} \leq d_0$ versus $H_a : \frac{\sigma_2^2}{\sigma_1^2} > d_0$ consists of rejecting H_0 if $f_0 \leq \frac{1}{f_{1-\alpha}(n_2-1,n_1-1)}$.
2. A test of size α for $H_0 : \frac{\sigma_2^2}{\sigma_1^2} \geq d_0$ versus $H_a : \frac{\sigma_2^2}{\sigma_1^2} < d_0$ consists of rejecting H_0 if $f_0 \geq f_{1-\alpha}(n_1 - 1, n_2 - 1)$.
3. A test of size α for $H_0 : \frac{\sigma_2^2}{\sigma_1^2} = d_0$ versus $H_a : \frac{\sigma_2^2}{\sigma_1^2} \neq d_0$ consists of rejecting H_0 if $f_0 \leq \frac{1}{f_{1-\alpha/2}(n_2-1,n_1-1)}$ or $f_0 \geq f_{1-\alpha/2}(n_1 - 1, n_2 - 1)$.

if the variances are unknown but equal, then tests of hypotheses concerning the means such as $H_0 : \mu_1 - \mu_2 = d_0$ can be constructed based on the t distribution. In particular, let

$$t_0 = \frac{\bar{y} - \bar{x} - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{12.3.16}$$

where s_p^2 is the pooled estimate defined by equation (11.5.7)

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{11.5.7}$$

Theorem 12.3.5

Suppose that $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ are observed values of independent random samples from $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$, respectively, where $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

1. A test of size α for $H_0 : \mu_2 - \mu_1 \leq d_0$ versus $H_a : \mu_2 - \mu_1 > d_0$ consists of rejecting H_0 if $t_0 \geq t_{1-\alpha}(n_1 + n_2 - 2)$.
2. A test of size α for $H_0 : \mu_2 - \mu_1 \geq d_0$ versus $H_a : \mu_2 - \mu_1 < d_0$ consists of rejecting H_0 if $t_0 \leq -t_{1-\alpha}(n_1 + n_2 - 2)$.
3. A test of size α for $H_0 : \mu_2 - \mu_1 = d_0$ versus $H_a : \mu_2 - \mu_1 \neq d_0$ consists of rejecting H_0 if $t_0 \leq -t_{1-\alpha/2}(n_1 + n_2 - 2)$ or $t_0 \geq t_{1-\alpha/2}(n_1 + n_2 - 2)$.

Theorem 12.3.6 and Sections 12.4 and 12.5 address some special cases that are worth reading.

12.6 Most Powerful Tests

In the previous sections, the terminology of hypothesis testing has been developed, and some intuitively appealing tests have been described based on pivotal quantities or appropriate sufficient statistics. The tests presented earlier were based on reasonable test statistics, but no rationale was provided to suggest that they are best in any sense. From this point forward, we will consider a method for deriving critical regions corresponding to tests that are most powerful tests of a given size for testing simple hypotheses.
Let $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots, x_n; \vartheta)$, and consider a critical region C. The notation for the power function corresponding to C is

$$\pi_C(\vartheta) = \mathbb{P}\left[X_1, \ldots, X_n \in C \mid \vartheta\right] \quad \text{or} \quad \mathbb{P}_\vartheta\left[X_1, \ldots, X_n \in C\right]. \tag{12.6.1}$$

Definition 12.6.1

A test of $H_0 : \vartheta = \vartheta_0$ versus $H_a : \vartheta = \vartheta_1$ based on a critical region C^* is said to be a **most powerful test** of size α if

1. $\pi_{C^*}(\vartheta_0) = \alpha$, and
2. $\pi_{C^*}(\vartheta_1) \geq \pi_C(\vartheta_1)$ for any other critical region C of size α [that is, $\pi_C(\vartheta_0) = \alpha$]

Such a critical region, C^*, is called a **most powerful critical region** of size α. The following theorem shows how to derive a most powerful critical region for testing simple hypotheses.

Theorem 12.6.1 Neyman-Pearson Lemma

Suppose that $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots, x_n; \vartheta)$. Let

$$\lambda\left(x_1, \ldots, x_n; \vartheta_0, \vartheta_1\right) := \frac{f(x_1, \ldots, x_n; \vartheta_0)}{f(x_1, \ldots, x_n; \vartheta_1)} \tag{12.6.2}$$

and let C^* be the set

$$C^* := \{(x_1, \ldots, x_n) \mid \lambda\left(x_1, \ldots, x_n; \vartheta_0, \vartheta_1\right) \leq k\} \tag{12.6.3}$$

where k is a constant such that

$$\mathbb{P}_{\vartheta_0}\left[(X_1, \ldots, X_n) \in C^*\right] = \alpha \tag{12.6.4}$$

Then C^* is a most powerful critical region of size α for testing $H_0 : \vartheta = \vartheta_0$ versus $H_a : \vartheta = \vartheta_1$.

Proof. For convenience, we will adopt vector notation, $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{x} = (x_1, \ldots, x_n)$. Also, if A be a subset of the sample space, let

$$\mathbb{P}_\vartheta\left[\boldsymbol{X} \in A\right] = \int_A \cdots \int f(x_1, \ldots, x_n; \vartheta)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n \tag{12.6.5}$$

in the continuous case, where in the discrete case, the integrals are replaced by summations. If $A \subset C^*$, then

$$\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in A\right] \leq k\mathbb{P}_{\vartheta_1}\left[\boldsymbol{X} \in A\right] \tag{12.6.6}$$

since

$$\int \dots \int_A f(\boldsymbol{x}; \vartheta_0)\,\mathrm{d}x_1 \dots \mathrm{d}x_n \leq \int \dots \int_A kf(\boldsymbol{x}; \vartheta_1)\,\mathrm{d}x_1 \dots \mathrm{d}x_n.$$

Similarly, if $A \subset \bar{C}^*$ (notation: $\bar{H}$ denotes the complement of set H)

$$\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in A\right] \geq k\mathbb{P}_{\vartheta_1}\left[\boldsymbol{X} \in A\right]. \tag{12.6.7}$$

Note that for any critical region C

$$C^* = (C^* \cap C) \cup (C^* \cap \bar{C}) \quad \text{and} \quad C = (C \cap C^*) \cup (C \cap \bar{C}^*).$$

Then

$$\pi_{C^*}(\vartheta) = \mathbb{P}_{\vartheta}\left[\boldsymbol{X} \in C^* \cap C\right] + \mathbb{P}_{\vartheta}\left[\boldsymbol{X} \in C^* \cap \bar{C}\right]$$

and

$$\pi_{C}(\vartheta) = \mathbb{P}_{\vartheta}\left[\boldsymbol{X} \in C^* \cap C\right] + \mathbb{P}_{\vartheta}\left[\boldsymbol{X} \in C \cap \bar{C}^*\right]$$

and the difference is

$$\pi_{C^*}(\vartheta) - \pi_C(\vartheta) = \mathbb{P}_{\vartheta}\left[\boldsymbol{X} \in C^* \cap \bar{C}\right] - \mathbb{P}_{\vartheta}\left[\boldsymbol{X} \in C \cap \bar{C}^*\right] \tag{12.6.8}$$

Hence:

$$\begin{aligned}
\pi_{C^*}(\vartheta_1) - \pi_C(\vartheta_1) &= \mathbb{P}_{\vartheta_1}\left[\boldsymbol{X} \in C^* \cap \bar{C}\right] - \mathbb{P}_{\vartheta_1}\left[\boldsymbol{X} \in C \cap \bar{C}^*\right] \\
&\geq \frac{1}{k}\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in C^* \cap \bar{C}\right] - \frac{1}{k}\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in C \cap \bar{C}^*\right] \\
&= \frac{1}{k}\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in C^* \cap \bar{C}\right] + \frac{1}{k}\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in C^* \cap C\right] \\
&\quad - \frac{1}{k}\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in C^* \cap C\right] - \frac{1}{k}\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in C \cap \bar{C}^*\right] \\
&= \frac{1}{k}\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in C^*\right] - \frac{1}{k}\mathbb{P}_{\vartheta_0}\left[\boldsymbol{X} \in C\right] \\
&= \frac{1}{k}(\alpha - \alpha) = 0
\end{aligned}$$

Note that we could have also required $\pi_C(\vartheta_0) \leq \alpha$. Then, it still holds that:

$$\pi_{C^*}(\vartheta_1) - \pi_C(\vartheta_1) \geq 0.$$

In other words, the Neyman-Pearson approach maximizes the number of outcomes more likely under the alternative hypothesis within the critical region until the Type I error is precisely achieved, at least in the case of continuous random variables.

Example 12.6.I

Let $X_1, \ldots, X_n$ be a sample from a GAM(ϑ, κ) distribution with κ known. Consider $H_0 : \vartheta = \vartheta_0$ against $H_a : \vartheta = \vartheta_1$, where $\vartheta_1 > \vartheta_0$. According to Neyman-Pearson, H_0 is rejected if

$$\lambda\left(x_1, \ldots, x_n; \vartheta_0, \vartheta_1\right) = \frac{\vartheta_0^{-n} \exp\left[-\vartheta_0^{-1} \sum x_i\right]}{\vartheta_1^{-n} \exp\left[-\vartheta_1^{-1} \sum x_i\right]} \leq k$$

with

$$\mathbb{P}_{\vartheta_0}\left[\left\{\left(X_1, \ldots, X_n\right) \mid \lambda\left(X_1, \ldots, X_n; \vartheta_0, \vartheta_1\right) \leq k\right\}\right] = \alpha.$$

Now, we have

$$\mathbb{P}_{\vartheta_0}\left[\left\{\left(X_1, \ldots, X_n\right) \mid \lambda\left(X_1, \ldots, X_n; \vartheta_0, \vartheta_1\right) \leq k\right\}\right] = \mathbb{P}_{\vartheta_0}\left[\sum X_i \left(\vartheta_1^{-1} - \vartheta_0^{-1}\right) \leq k_1\right],$$

and

$$C^* = \left\{\left(x_1, \ldots, x_n\right) \mid \sum x_i \geq k_2\right\} \qquad \text{with} \qquad \mathbb{P}_{\vartheta_0}\left[\sum X_i \geq k_2\right] = \alpha.$$

Since $\frac{2\sum X_i}{\vartheta} \sim \chi^2(2n\kappa)$, we have $k_2 = \frac{1}{2}\vartheta_0 \chi^2_{1-\alpha}(2n\kappa)$. If $\vartheta_1 < \vartheta_0$, a similar method gives: reject if $\sum x_i \leq \frac{\vartheta_0}{2}\chi^2_{\alpha}(2n\kappa)$. Note that the critical region depends only on the null hypothesis, and for $\kappa = 1$, this example is identical to Example 12.6.1.

Example 12.6.2

Consider a random sample of size n from a normal distribution with mean zero, $X_i \sim \mathrm{N}(0, \sigma^2)$. We wish to test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_a : \sigma^2 = \sigma_1^2 > \sigma_0^2$. In this case

$$\lambda\left(x_1, \ldots, x_n; \sigma_0^2, \sigma_1^2\right) = \left(\frac{\sigma_1}{\sigma_0}\right)^n e^{-\frac{\sum x_i^2}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)}$$

reject if

$$-\frac{\sum x_i^2}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right) \leq k \quad \Rightarrow \quad \sum x_i^2 \geq k_1$$

with

$$\mathbb{P}_{\sigma_0^2}\left[\sum X_i^2 \geq k_1\right] = \alpha; \qquad \frac{\sum X_i^2}{\sigma^2} \sim \chi^2(n) \quad \Rightarrow \quad k_1 = \sigma_0^2 \chi^2_{1-\alpha}(n).$$

Similarly, if $\sigma_1^2 < \sigma_0^2$, reject if $\sum_{i=1}^n x_i^2 \leq \sigma_0^2 \chi^2_{\alpha}(n)$.

Example 12.6.4

We have a random sample of size n, and we wish to test $H_0 : X_i \sim \mathrm{UNIF}(0,1)$ against $H_a : X_i \sim \mathrm{EXP}(1)$. Note: there is an error in the derivation in the book! The test derived does not necessarily reject the null hypothesis if the largest observation, $X_{n:n}$, is greater than 1, a case where H_0 should evidently be rejected in favor of H_a. The issue is resolved by noting that λ is zero in this case. We find:

$$\lambda = \frac{f_0(x_1, \ldots, x_n)}{f_1(x_1, \ldots, x_n)} = \begin{cases} \frac{1}{e^{-\sum x_i}} = e^{\sum x_i} & \text{if } x_{n:n} \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The value $\lambda = 0$, which occurs if $X_{n:n} > 1$, is smaller than any possible value of λ under H_0 (since under H_0, $\lambda \geq 1$). A test of size $\alpha > 0$ rejects if $\lambda \leq k$, with $k > 1$, and rejection is certain if $\lambda = 0$. The derivation of the critical region when $\lambda \neq 0$ still holds (see the book for details).

12.7 Uniformly Most Powerful Tests

In the last section we saw that in some cases the same test is most powerful against several different alternative values. If a test is most powerful against every possible value in a composite alternative, then it will be called a uniformly most powerful test (UMP).

Definition 12.7.1

Let $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots, x_n; \vartheta)$ for $\vartheta \in \Omega$ and consider hypotheses of the form $H_0 : \vartheta \in \Omega_0$ versus $H_a : \vartheta \in \Omega - \Omega_0$, where $\Omega_0 \subset \Omega$. A critical region C^*, and the associated test are said to be **uniformly most powerful** (UMP) of size α if

$$\max_{\vartheta \in \Omega_0} \pi_{C^*}(\vartheta) = \alpha \tag{12.7.1}$$

and

$$\pi_{C^*}(\vartheta) \geq \pi_C(\vartheta) \tag{12.7.2}$$

for all $\vartheta \in \Omega - \Omega_0$ and all critical regions C of size α.

That is, C^* defines a UMP test of size α if it has size α, and if for all parameter values in the alternative, it has maximum power relative to all critical regions of size α.

Example 12.7.I (continuation of Example 12.6.2)

Let $X_1, \ldots, X_n$ be a sample from $\mathrm{N}(0, \sigma^2)$. We test $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_a : \sigma^2 > \sigma_0^2$. We begin by testing $H_0 : \sigma^2 = \sigma_*^2 \leq \sigma_0^2$ against $H_a : \sigma^2 = \sigma_1^2 > \sigma_0^2$. We reject if $\sum x_i^2 > \sigma_*^2 \chi^2_{1-\alpha^*}(n)$; this is independent of σ_1^2. The size is

$$\max_{\sigma_*^2 \leq \sigma_0^2} \mathbb{P}_{\sigma_*^2} \left[\sum X_i^2 > \sigma_*^2 \chi^2_{1-\alpha^*} \right]$$

and the maximum is clearly reached when $\sigma_*^2 = \sigma_0^2$. General results along these lines can be stated for any pdf that satisfies a property known as the ”monotone likelihood ratio.”

Definition 12.7.2

A joint pdf $f(x_1, \ldots, x_n; \vartheta)$ is said to have a **monotone likelihood ratio (MLR)** in the statistic $T = t(X_1, \ldots, X_n)$ if for any two values of the parameter $\vartheta_1 < \vartheta_2$, the ratio

$$\frac{f(x_1, \ldots, x_n; \vartheta_2)}{f(x_1, \ldots, x_n; \vartheta_1)}$$

depends on $(x_1, \ldots, x_n)$ only through the function $t(x_1, \ldots, x_n)$, and this ratio is a non-decreasing function of $t(x_1, \ldots, x_n)$.

If the ratio is a non-increasing function of $t(x_1, \ldots, x_n)$, it is a non-decreasing function of $-t(x_1, \ldots, x_n)$. In dealing with MLR, we use some special conventions

1. $\frac{c}{0} = \infty$ for $c > 0$,
2. $\frac{0}{0} = 1$.

Example 12.7.2

Consider a random sample of size n from an exponential distribution, $X_i \sim \text{EXP}(\vartheta)$. Because $f(x_1, \ldots, x_n; \vartheta) = \left(\frac{1}{\vartheta}\right)^n \exp\left(-\sum \frac{x_i}{\vartheta}\right)$, we have

$$\frac{f(x_1, \ldots, x_n; \vartheta_2)}{f(x_1, \ldots, x_n; \vartheta_1)} = \left(\frac{\vartheta_1}{\vartheta_2}\right)^n \exp\left[-\left(\frac{1}{\vartheta_2} - \frac{1}{\vartheta_1}\right) \sum x_i\right],$$

which is a non-decreasing function of $t(x_1, \ldots, x_n) = \sum x_i$ because $\frac{1}{\vartheta_2} - \frac{1}{\vartheta_1} < 0$ since $\vartheta_2 > \vartheta_1$. Thus, $f(x_1, \ldots, x_n; \vartheta$ has the MLR property in the statistic $T = \sum X_i$. Notice that the MLR property also holds for the statistic $\bar{X}$, because it is an increasing function of T.

Example 12.7.II

Let $x_1, \ldots, x_n$ be a sample from UNIF$(0, \vartheta)$

$$\begin{aligned}\frac{f(x_1, \ldots, x_n; \vartheta_2)}{f(x_1, \ldots, x_n; \vartheta_1)} &= \left(\frac{\vartheta_1}{\vartheta_2}\right)^n \frac{I_{(0,x_{n:n}]}(x_{1:n})}{I_{(0,x_{n:n}]}(x_{1:n})} \frac{I_{(0,\vartheta_2]}(x_{n:n})}{I_{(0,\vartheta_1]}(x_{n:n})} \\ &= \left(\frac{\vartheta_1}{\vartheta_2}\right)^n \begin{cases} 1, & x_{n:n} \leq \vartheta_1 \\ \infty, & \vartheta_1 < x_{n:n} \leq \vartheta_2. \end{cases}\end{aligned}$$

Here, we have a non-decreasing function of $x_{n:n}$. The case where $x_{n:n} > \vartheta_2$ is impossible and is thus excluded.

Theorem 12.7.1

If a joint pdf $f(x_1, \ldots, x_n; \vartheta)$ has the MLR property in the statistic $T = t(X_1, \ldots, X_n)$, then a UMP test of size α for $H_0 : \vartheta \leq \vartheta_0$ versus $H_a : \vartheta > \vartheta_0$ is to reject H_0 if $t(x_1, \ldots, x_n) \geq k$, where k is such that

$$\mathbb{P}_{\vartheta_0}\left[t(X_1, \ldots, X_n) \geq k\right] = \alpha.$$

If we have $H_0 : \vartheta \geq \vartheta_0$ against $H_a : \vartheta < \vartheta_0$, the UMP test is to reject H_0 if

$$t(x_1, \ldots, x_n) \leq k \quad \text{with} \quad \mathbb{P}_{\vartheta_0}\left[t(X_1, \ldots, X_n) \leq k\right] = \alpha.$$

Remark 12.7.I

If there is a minimal sufficient statistic S, then T is a function of S. Verify this yourself.

Theorem 12.7.2

Suppose that $X_1, \ldots, X_n$ have joint pdf of the form

$$f(x_1, \ldots, x_n; \vartheta) = c(\vartheta) h(x_1, \ldots, x_n) \exp\left[q(\vartheta) t(x_1, \ldots, x_n)\right] \tag{12.7.5}$$

where $q(\vartheta)$ is an increasing function of ϑ. Then Theorem 12.7.1 applies with $T = t(X_1, \ldots, X_n)$.

An obvious application of the theorem occurs when $X_1, \ldots, X_n$, is a random sample from a member of the regular exponential class (REC), say $c(\vartheta)h(x_1, \ldots, x_n) \exp\left[q(\vartheta)t(x_1, \ldots, x_n)\right]$ with $t(x_1, \ldots, x_n) = \sum u(x_i)$ and $q(\vartheta)$ is an increasing function of ϑ.

Unbiased Tests

it was mentioned earlier that in some cases where a UMP test may not exist, particularly for a two-sided alternative, there may exist a UMP test among the restricted class of "unbiased" tests.

Definition 12.7.3

A test of $H_0 : \vartheta \in \Omega_0$ versus $H_a : \vartheta \in \Omega - \Omega_0$ is **unbiased** if

$$\min_{\vartheta \in \Omega - \Omega_0} \pi(\vartheta) \geq \max_{\vartheta \in \Omega_0} \pi(\vartheta) \tag{12.7.6}$$

In other words, the probability of rejecting H_0 when it is false is at least as large as the probability of rejecting H_0 when it is true.

12.8 Generalized Likelihood Ratio Tests

The hypotheses we have examined so far consisted of simple null and alternative hypotheses, or they consisted of two intervals. In other words, the sets Ω_0 and $\Omega - \Omega_0$ were points or intervals. For these cases, we were able to derive most powerful and uniformly most powerful (UMP) tests. In other cases, the principle is to find a function of the observations that behaves differently under the null hypothesis compared to the alternative. We call such a function the test statistic. Given a suitable test statistic, then, as with the Neyman-Pearson lemma, sample points should be included in the critical region that are less likely to occur under H_0 and more likely to occur under H_a. The generalized likelihood ratio test is a generalization of the Neyman-Pearson test, and it provides a desirable test in many applications.

Definition 12.8.1

Let $(X_1, \ldots, X_n)$ have joint pdf $f(x_1, \ldots, x_n; \boldsymbol{\vartheta})$ for $\boldsymbol{\vartheta} \in \Omega$, and consider the hypothesis $H_0 : \boldsymbol{\vartheta} \in \Omega_0$ versus $H_a : \boldsymbol{\vartheta} \in \Omega - \Omega_0$. The **generalized likelihood ratio** (GLR) is defined by

$$\lambda(x_1, \ldots, x_n) = \frac{\max_{\boldsymbol{\vartheta} \in \Omega_0} f(x_1, \ldots, x_n; \boldsymbol{\vartheta})}{\max_{\boldsymbol{\vartheta} \in \Omega} f(x_1, \ldots, x_n; \boldsymbol{\vartheta})} = \frac{f(x_1, \ldots, x_n; \hat{\boldsymbol{\vartheta}}_{\mathbf{0}})}{f(x_1, \ldots, x_n; \hat{\boldsymbol{\vartheta}})} \tag{12.8.1}$$

where $\hat{\boldsymbol{\vartheta}}$ denotes the usual MLE of $\boldsymbol{\vartheta}$, and $\hat{\boldsymbol{\vartheta}}_{\mathbf{0}}$denotes the MLE under the restriction that H_0 is true

Note that the maximization in the denominator is over the entire parameter space and not just the alternative. Ideally, the restricted MLE and the unrestricted MLE are equal, and
$\lambda(x_1, \ldots, x_n) = 1$; generally, $\lambda(x_1, \ldots, x_n) \leq 1$. We will reject if $\lambda(x_1, \ldots, x_n) < k$, with

$$\max_{\vartheta \in \Omega_0} \mathbb{P}_{\vartheta}\left[\lambda(X_1, \ldots, X_n) < k\right] = \alpha \tag{12.8.I}$$

where α is the desired significance level. Equation (12.8.I) indicates how to determine k for a given significance level. Clearly, if $\lambda(x_1,\dots,x_n)=1$, the null hypothesis will not be rejected. Note that the test statistic is defined without any unknown parameters. It is often difficult to calculate the exact distribution of $\lambda(x_1,\dots,x_n)\le k$. However, if we have a large number of observations, then

$$(X_1,\dots,X_n)\sim f(x_1,\dots,x_n;\vartheta_1,\dots,\vartheta_k)$$

and

$$H_0:(\vartheta_1,\dots,\vartheta_r)=(\vartheta_{10},\dots,\vartheta_{r0}),\quad r<k$$

and if the regularity conditions are met, including the requirement that the MLEs are asymptotically normal, then

$$-2\log\lambda(X_1,\dots,X_n)\sim\chi^2(r). \tag{12.8.2}$$

We reject the null hypothesis if

$$-2\log\lambda(x_1,\dots,x_n)\ge\chi^2_{1-\alpha}(r). \tag{12.8.3}$$

The true size of the test is approximately α.

Example 12.8.1

Suppose that $X_i\sim\mathrm{N}(\mu,\sigma^2)$, where σ^2 is known, and we wish to test $H_0:\mu=\mu_0$ against $H_a:\mu\ne\mu_0$. The usual (unrestricted) MLE is $\hat\mu=\bar X$, and the GLR is

$$\lambda(x_1,\dots,x_n)=\frac{f(x_1,\dots,x_n;\mu_0)}{f(x_1,\dots,x_n;\bar x)}=e^{-\frac{1}{2\sigma^2}n(\bar x-\mu_0)^2}<k.$$

Rejecting H_0 if $\lambda(x)<k$ is equivalent to rejecting H_0 if

$$z^2=\left(\sqrt{n}\frac{\bar x-\mu_0}{\sigma}\right)^2>k_1\quad\text{where}\quad\left(\sqrt{n}\frac{\bar X-\mu_0}{\sigma}\right)^2\sim\chi^2(1)$$

$$\Rightarrow\ k_1=\chi^2_{1-\alpha}(1)$$

Note that in the above example, $Z^2=-2\log\lambda(X_1,\dots,X_n)$. Sometimes, when using the GLR, care must be taken as the parameter space can become unusual.

Example 12.8.2

Using the same data as in Example 12.8.1, but now with $H_0:\mu=\mu_0$ against $H_a:\mu>\mu_0$, $\Omega=[\mu_0,\infty)$ instead of $(-\infty,\infty)$, we have

$$\hat\mu=\begin{cases}\bar x, & \text{if }\bar x>\mu_0,\\ \mu_0, & \text{if }\bar x\le\mu_0.\end{cases}$$

We have

$$\lambda(x_1,\dots,x_n)=\begin{cases}e^{-\frac{n}{2}\left(\frac{\bar x-\mu_0}{\sigma}\right)^2}, & \text{if }\bar x>\mu_0,\\ 1, & \text{if }\bar x\le\mu_0.\end{cases}$$

Under H_0, $\mathbb{P}[\bar X>\mu_0]=0.5$. Thus, we reject H_0 if $\bar x>\mu_0$ and $z^2>k_1=z^2_{1-\alpha}$.
It sometimes is desired to test a hypothesis about one unknown parameter in the presence of another unknown nuisance parameter.

Example 12.8.3

Suppose now that $X_i \sim \mathrm{N}(\mu, \sigma^2)$ where σ^2 is assumed unknown, and we wish to test $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$.This does not represent a simple null hypothesis, because the distribution is is not specified completely under H_0. The parameter space is two-dimensional,

$$\Omega = \left\{(\mu,\, \sigma^2) \mid -\infty < \mu < \infty, \sigma^2 > 0\right\} \quad \text{and} \quad \Omega_0 \;\; = \left\{(\mu,\, \sigma^2) \mid \mu = \mu_0, \sigma^2 > 0\right\}$$

Maximizing $f(x_1, \dots, x_n; \mu, \sigma^2)$ over Ω yields $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, but over Ω_0, we obtain $\hat{\mu}_0 = \mu_0$ and $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$. Thus,

$$\lambda(x_1, \dots, x_n) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{n}{2}}$$

after some simplification. We reject if

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2}$$

and consequently,

$$\begin{aligned}
\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}\left(x_i - \bar{x} + \bar{x} - \mu_0\right)^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2} \\
&= \frac{1}{1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} < k,
\end{aligned}$$

which implies

$$\frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} > k_1 \quad \text{or equivalently} \quad \frac{n(\bar{x} - \mu_0)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} > k_2.$$

This gives $k_2 = F_{1-\alpha}(1, n-1)$. Since $F(1, n-1) \sim t^2(n-1)$, we also reject if

$$\left|\sqrt{n}\frac{\bar{x} - \mu_0}{s}\right| > t_{1-\frac{\alpha}{2}}(n-1).$$

The GLR method is very useful when we have k samples from k normal distributions and want to test if the means are equal while knowing the variances are the same. To keep it simple, we limit ourselves to two samples. For the general case, see Theorem 12.8.1. We have $X_1, \dots, X_n$ from $\mathrm{N}(\mu_1, \sigma^2)$ and $Y_1, \dots, Y_n$ from $\mathrm{N}(\mu_2, \sigma^2)$; testing $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$, we find

$$\begin{aligned}
L &= L\left(\mu_1, \mu_2, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - \mu_1)^2}{2\sigma^2}} \prod_{j=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_j - \mu_2)^2}{2\sigma^2}} \\
&= \left(\tfrac{1}{2\pi\sigma^2}\right)^{\frac{n+m}{2}} e^{-\frac{1}{2\sigma^2}\left[\sum\limits_{i=1}^{n}(x_i - \mu_1)^2 + \sum\limits_{j=1}^{m}(y_j - \mu_2)^2\right]}
\end{aligned}$$

and

$$\log L = -\frac{n+m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left(x_i-\mu_1\right)^2 + \sum_{j=1}^{m}\left(y_j-\mu_2\right)^2\right].$$

Maximizing w.r.t μ and σ^2

$$\frac{d}{d\mu_1}\log L = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu_1\right) = 0$$

$$\frac{d}{d\sigma^2}\log L = -\frac{n+m}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{n}\left(x_i-\mu_1\right)^2 + \sum_{j=1}^{m}\left(y_j-\mu_2\right)^2\right] = 0$$

leading to the solutions

$$\hat{\mu}_1 = \bar{x}, \quad \hat{\mu}_2 = \bar{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n+m}\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 + \sum_{j=1}^{m}\left(y_j-\bar{y}\right)^2\right]$$

Under H_0, $\mu_1 = \mu_2 = \mu$:

$$\frac{d}{d\mu}\log L = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(x_i-\mu\right) + \sum_{j=1}^{m}\left(y_j-\mu\right)\right] = 0$$

$$\frac{d}{d\sigma^2}\log L = -\frac{n+m}{2\sigma^2} + \frac{1}{\sigma^4}\left[\sum_{i=1}^{n}\left(x_i-\mu\right)^2 + \sum_{j=1}^{m}\left(y_j-\mu\right)^2\right] = 0$$

with solutions

$$\hat{\mu}_0 = \frac{n\bar{x}+m\bar{y}}{n+m} \quad \text{and} \quad \hat{\sigma}_0^2 = \frac{1}{n+m}\left[\sum_{i=1}^{n}\left(x_i-\hat{\mu}_0\right)^2 + \sum_{j=1}^{m}\left(y_j-\hat{\mu}_0\right)^2\right]$$

Then it follows that

$$\lambda\left(x_1,\ldots,x_n\right) = \frac{L\left(\hat{\mu}_0,\hat{\sigma}_0^2\right)}{L\left(\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma}^2\right)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{n+m}{2}} = \left[\frac{\sum\limits_{i=1}^{n}\left(x_i-\bar{x}\right)^2 + \sum\limits_{j=1}^{m}\left(y_j-\bar{y}\right)^2}{\sum\limits_{i=1}^{n}\left(x_i-\hat{\mu}_0\right)^2 + \sum\limits_{j=1}^{m}\left(y_j-\hat{\mu}_0\right)^2}\right]$$

Next,

$$\sum_{i=1}^{n}\left(x_i-\hat{\mu}_0\right)^2 = \sum_{i=1}^{n}\left(x_i-\bar{x}+\bar{x}-\hat{\mu}_0\right)^2 = \sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 + n\left(\bar{x}-\hat{\mu}_0\right)^2$$

and

$$n\left(\bar{x}-\hat{\mu}_0\right)^2 = n\left(\bar{x}-\frac{n\bar{x}+m\bar{y}}{n+m}\right)^2 = nm^2\left(\frac{\bar{x}-\bar{y}}{n+m}\right)^2,$$

it follows that

$$\begin{aligned}&\sum_{i=1}^{n}(x_i-\hat{\mu}_0)^2+\sum_{j=1}^{m}(y_i-\hat{\mu}_0)^2\\&=\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{j=1}^{m}(y_i-\bar{y})^2+nm^2\left(\frac{\bar{x}-\bar{y}}{n+m}\right)^2+mn^2\left(\frac{\bar{x}-\bar{y}}{n+m}\right)^2\\&=(n+m)\,\hat{\sigma}^2+\frac{nm}{n+m}(\bar{x}-\bar{y})^2\end{aligned}$$

This implies

$$\lambda(x_1,\dots,x_n)=\left[\frac{\hat{\sigma}^2}{(n+m)\hat{\sigma}^2+\frac{nm}{n+m}(\bar{x}-\bar{y})^2}\right]^{\frac{n+m}{2}}<k$$

which is equivalent to

$$\frac{(\bar{x}-\bar{y})^2}{\hat{\sigma}^2}>k_1\quad\text{and}\quad\bar{X}-\bar{Y}\sim\mathrm{N}\left(0,\frac{\sigma^2}{n}+\frac{\sigma^2}{m}\right)=\mathrm{N}\left(0,\sigma^2\frac{n+m}{nm}\right)$$

Under H_0,

$$\bar{X}-\bar{Y}\quad\text{and}\quad\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{j=1}^{m}(Y_j-\bar{Y})^2=:D_p^2$$

are independent, with

$$\frac{S_p^2}{\sigma^2}\sim\chi^2(n+m-2)$$

This leads to:

$$\frac{(n+m-2)nm}{n+m}\frac{(\bar{X}-\bar{Y})^2}{D_p^2}\sim F(1,n+m-2)$$