STAT 512
MATHEMATICAL STATISTICS

Spring 2024

Lecture Notes

Joshua M. Tebbs
Department of Statistics
University of South Carolina

Contents

6 Functions of Random Variables

6.1 Introduction

Preview: We are now ready to address the following important questions in probability and distribution theory, namely,

1. "If we know the distribution of a random variable Y, what is the distribution of $U = h(Y)$, a function of Y?"
2. "If we know the joint distribution of the random variables $Y_1, Y_2, ..., Y_n$, what is the distribution of $U = h(Y_1, Y_2, ..., Y_n)$, a function of $Y_1, Y_2, ..., Y_n$?"

This chapter deals with finding distributions of **functions** of random variables. In the first question above, the function $h : \mathbb{R} \to \mathbb{R}$. In the second, $h : \mathbb{R}^n \to \mathbb{R}$. Therefore, $U = h(Y)$ or $U = h(Y_1, Y_2, ..., Y_n)$ are (univariate) random variables that have their own distributions. We will also consider functions $h : \mathbb{R}^n \to \mathbb{R}^n$ in Section 6.6 (WMS), paying particular attention to the bivariate ($n = 2$) case. This situation arises if we want to answer this question:

3. "If we know the bivariate distribution of $\mathbf{Y} = (Y_1, Y_2)$ and

$$\begin{aligned} U_1 &= h_1(Y_1, Y_2) \\ U_2 &= h_2(Y_1, Y_2), \end{aligned}$$

what is the bivariate distribution of $\mathbf{U} = (U_1, U_2)$?"

In the third question, $h : \mathbb{R}^2 \to \mathbb{R}^2$; i.e., h is a vector-valued function.

Examples: Here are some applications that motivate why these questions are important:

- Physicians measure the systolic blood pressure Y for pregnant women at elevated risk for preeclampsia. What is the distribution of $U = h(Y) = \ln Y$?
- Actuaries record the losses due to liability Y_1 and collision Y_2 for drivers in South Carolina. What is the distribution of the total loss

$$U = h(Y_1, Y_2) = Y_1 + Y_2?$$

- Researchers observe test scores $Y_1, Y_2, ..., Y_n$ for a sample of n students. What is the distribution of

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

the average test score, or perhaps

$$Y_{(n)} = \max\{Y_1, Y_2, ..., Y_n\},$$

the maximum test score? Note that both $\overline{Y}$ and $Y_{(n)}$ are functions of $Y_1, Y_2, ..., Y_n$.

Preview: This chapter deals with finding distributions of functions of random variables. We will investigate three methods for doing this:

1. Method of distribution functions (or "cdf technique")
2. Method of transformations
3. Method of moment generating functions (or "mgf technique").

6.2 Method of distribution functions

Setting: Suppose Y is a continuous random variable with cumulative distribution function (cdf) $F_Y(y)$ and probability density function (pdf) $f_Y(y)$. Define

$$U = h(Y).$$

The cdf technique is implemented by deriving the cdf of U, that is,

$$F_U(u) = P(U \leq u) = P(h(Y) \leq u),$$

for all $u \in \mathbb{R}$. The last probability suggests we can find $F_U(u)$ if we know the distribution of Y, because we can always integrate $f_Y(y)$ over the set $B = \{y : h(y) \leq u\}$. If we can derive $F_U(u)$ in this way, then the pdf of U is simply

$$f_U(u) = \frac{d}{du} F_U(u).$$

Note: The method of distribution functions (or "cdf technique") is especially useful when the cdf of Y exists in closed form; i.e., we know a formula for $F_Y(y)$. When this is true, we can usually write $F_U(u)$ in terms of this formula. This is illustrated in the next two examples.

Example 6.1. Suppose $Y \sim \mathcal{U}(0, 1)$; i.e., Y has a uniform distribution from 0 to 1. The pdf of Y is

$$f_Y(y) = \begin{cases} 1, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the distribution of $U = h(Y) = -\ln Y$.
(b) Calculate $E(U)$.

Solutions. (a) The pdf of Y is shown in Figure 6.1 (left). Note that the support of Y is

$$R_Y = \{y : 0 < y < 1\}.$$

A graph of the function $h(y) = -\ln y$ over $(0, 1)$ is shown in Figure 6.1 (right). Note that

$$0 < y < 1 \iff u = h(y) = -\ln y > 0.$$

Therefore, the support of $U = h(Y) = -\ln Y$ is

$$R_U = \{u : u > 0\}.$$

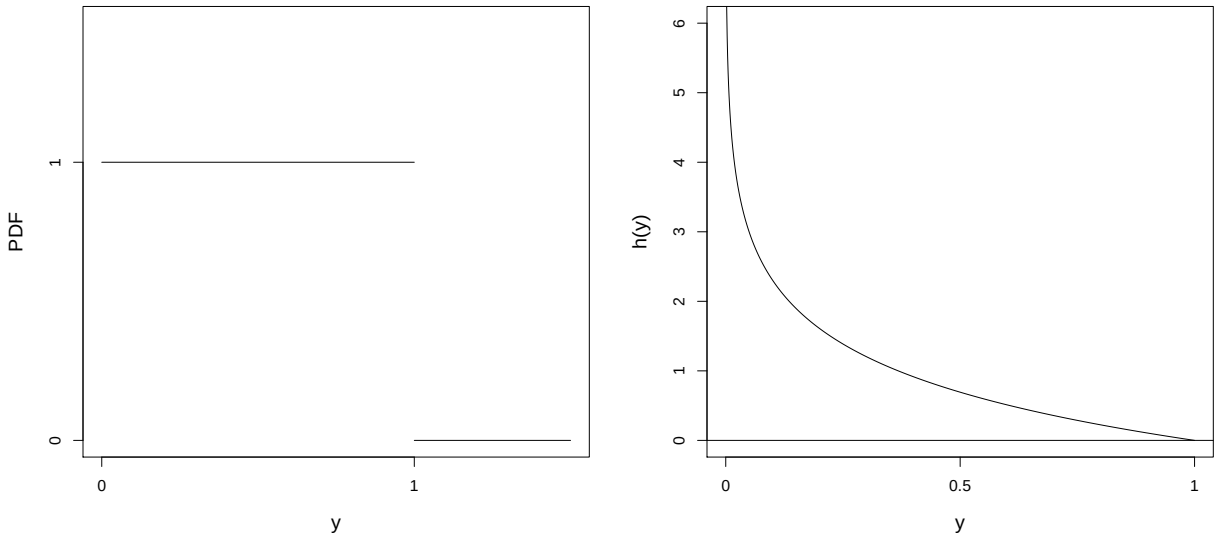

Figure 6.1: Left: Pdf of $Y \sim \mathcal{U}(0,1)$ in Example 6.1. Right: A graph of the function $h(y) = -\ln y$ over $(0,1)$; i.e., over the support $R_Y = \{y : 0 < y < 1\}$.

For $u > 0$, the cdf of U is

$$\begin{aligned} F_U(u) = P(U \leq u) &= P(-\ln Y \leq u) \\ &= P(\ln Y > -u) \\ &= P(Y > e^{-u}) = 1 - P(Y \leq e^{-u}) = 1 - F_Y(e^{-u}). \end{aligned}$$

Notice how we have written $F_U(u)$ in terms of $F_Y(y)$. We now recall the cdf of $Y \sim \mathcal{U}(0,1)$ is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ y, & 0 < y < 1 \\ 1, & y \geq 1. \end{cases}$$

Therefore, for $0 < y < 1 \Longleftrightarrow u > 0$, the cdf of $U = h(Y) = -\ln Y$ is

$$F_U(u) = 1 - F_Y(e^{-u}) = 1 - e^{-u}.$$

For $u > 0$, the pdf of U is

$$f_U(u) = \frac{d}{du} F_U(u) = \frac{d}{du}(1 - e^{-u}) = e^{-u}.$$

Summarizing,

$$f_U(u) = \begin{cases} e^{-u}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this as an exponential pdf with mean $\beta = 1$; i.e., $U \sim$ exponential(1).

(b) Because $U \sim$ exponential(1), we know $E(U) = 1$. However, would we get the same answer if we calculated

$$E(-\ln Y) = \int_{\mathbb{R}} -\ln y \; f_Y(y)dy = \int_0^1 -\ln y \; dy;$$

i.e., by using the distribution of Y? Let

$$\begin{array}{ll} u = -\ln y & du = -\frac{1}{y}dy \\ dv = dy & v = y. \end{array}$$

Indeed, integration by parts shows we get the same answer:

$$\int_0^1 -\ln y \; dy = -y \ln y \Big|_0^1 - \int_0^1 (-1)dy = (0-0) + 1 = 1.$$

This is not a coincidence. In fact, these calculations illustrate the following result.

Law of the Unconscious Statistician: Suppose Y is a continuous random variable with pdf $f_Y(y)$ and let $U = h(Y)$. We can calculate $E(U)$ in two ways:

$$\begin{aligned} E(U) \; = \; E[h(Y)] \; &= \; \int_{\mathbb{R}} h(y) f_Y(y) dy \quad \longleftarrow \text{STAT 511 way} \\ E(U) \; &= \; \int_{\mathbb{R}} u f_U(u) du, \end{aligned}$$

where $f_U(u)$ is the pdf of U. The Law of the Unconscious Statistician says $E(U) = E[h(Y)]$ in the sense that if one expectation exists, so does the other and they are equal. This result is also true in the discrete case.

Example 6.2. Suppose $Y \sim$ exponential(α); i.e., Y has an exponential distribution with mean $\alpha > 0$. The pdf of Y is

$$f_Y(y) = \begin{cases} \frac{1}{\alpha} e^{-y/\alpha}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

For $m > 0$, find the distribution of $U = h(Y) = Y^{1/m}$.

Solution. The pdf of Y is shown in Figure 6.2 (left). Note that the support of Y is

$$R_Y = \{y : y > 0\}.$$

For $m > 0$, a graph of $h(y) = y^{1/m}$ over $(0, \infty)$ is shown in Figure 6.2 (right). Note that

$$y > 0 \iff u = h(y) = y^{1/m} > 0.$$

Therefore, the support of $U = h(Y) = Y^{1/m}$ is

$$R_U = \{u : u > 0\}.$$

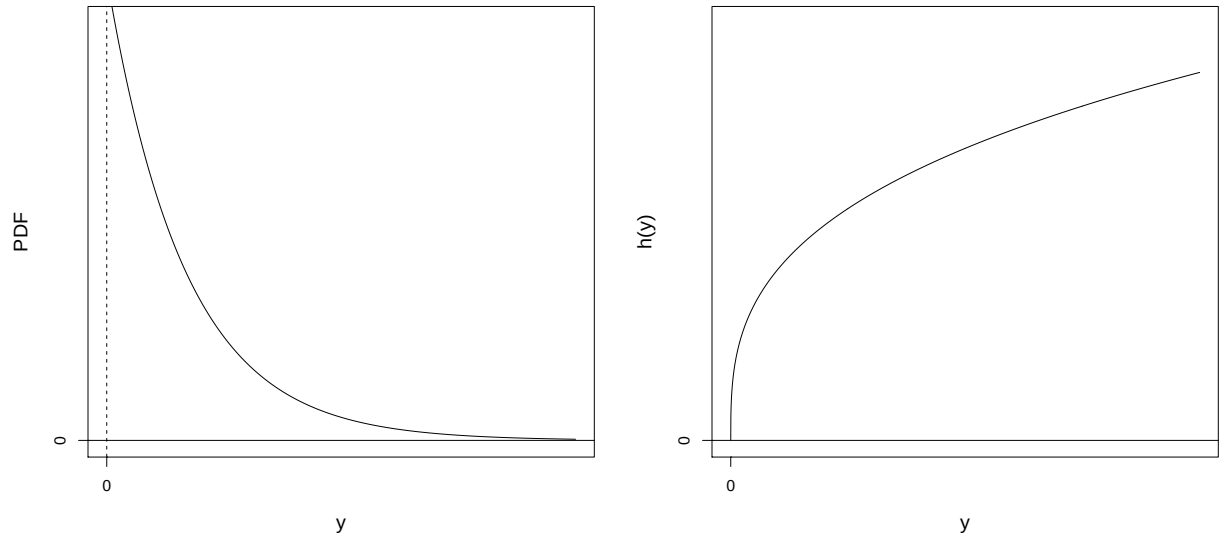

Figure 6.2: Left: Pdf of $Y \sim$ exponential(α) in Example 6.2. Right: A graph of the function $h(y) = y^{1/m}$ over $(0, \infty)$; i.e., over the support $R_Y = \{y : y > 0\}$.

For $u > 0$, the cdf of U is

$$F_U(u) = P(U \leq u) = P(Y^{1/m} \leq u) = P(Y \leq u^m) = F_Y(u^m).$$

Notice how we have written $F_U(u)$ in terms of $F_Y(y)$. We now recall the cdf of $Y \sim$ exponential(α) is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y/\alpha}, & y > 0. \end{cases}$$

Therefore, for $y > 0 \Longleftrightarrow u > 0$, the cdf of $U = h(Y) = Y^{1/m}$ is

$$F_U(u) = F_Y(u^m) = 1 - e^{-u^m/\alpha}.$$

For $u > 0$, the pdf of U is

$$f_U(u) = \frac{d}{du} F_U(u) = \frac{d}{du}(1 - e^{-u^m/\alpha}) = \frac{m}{\alpha} u^{m-1} e^{-u^m/\alpha}.$$

Summarizing,

$$f_U(u) = \begin{cases} \frac{m}{\alpha} u^{m-1} e^{-u^m/\alpha}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this as a Weibull pdf with parameters $m > 0$ and $\alpha > 0$; see Exercise 6.26 (WMS, pp 317). We write $U \sim$ Weibull(m, α). $\square$

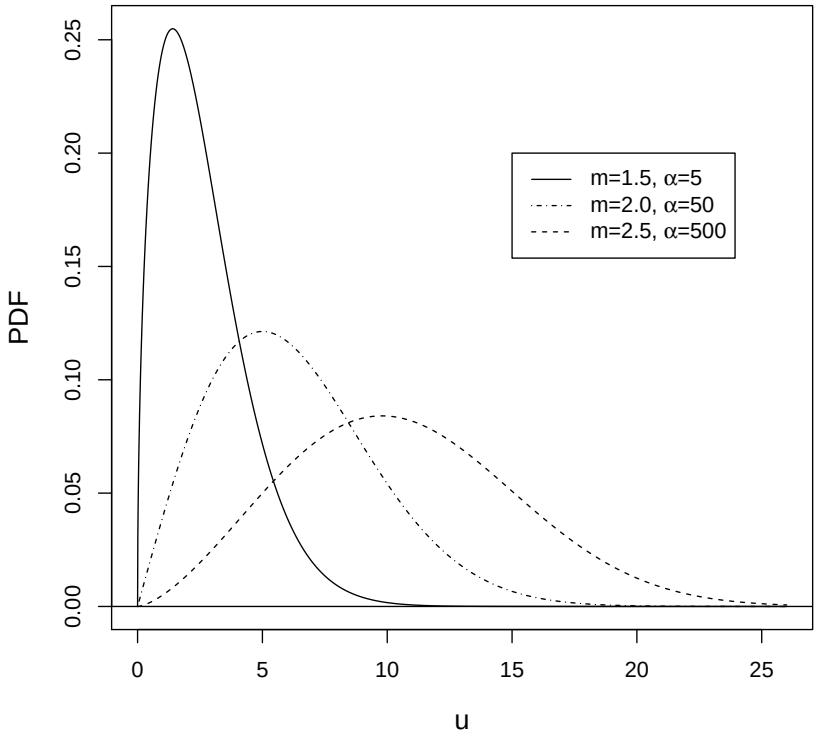

Figure 6.3: Weibull(m, α) pdfs for different combinations of m and α.

Remark: The Weibull(m, α) family of pdfs is a flexible family; see Figure 6.3 above. The Weibull distribution is most often used in engineering and the natural sciences to model positive quantities; e.g., time to part failure, breaking strength, wind speeds, etc.

Example 6.3. Suppose Y is a continuous random variable with cdf $F_Y(y)$ and pdf $f_Y(y)$. Derive a general expression for the pdf of $U = h(Y) = Y^2$.

Solution. The cdf of $U = h(Y) = Y^2$ is

$$F_U(u) = P(U \leq u) = P(Y^2 \leq u) = P(-\sqrt{u} \leq Y \leq \sqrt{u}) = F_Y(\sqrt{u}) - F_Y(-\sqrt{u}).$$

Notice how we have written $F_U(u)$ in terms of $F_Y(y)$. The pdf of U is

$$\begin{aligned} f_U(u) = \frac{d}{du} F_U(u) &= \frac{d}{du}\left[F_Y(\sqrt{u}) - F_Y(-\sqrt{u})\right] \\ &= f_Y(\sqrt{u})\frac{1}{2\sqrt{u}} - f_Y(-\sqrt{u})\left(-\frac{1}{2\sqrt{u}}\right) \\ &= \frac{1}{2\sqrt{u}}\left[f_Y(\sqrt{u}) + f_Y(-\sqrt{u})\right]. \end{aligned}$$

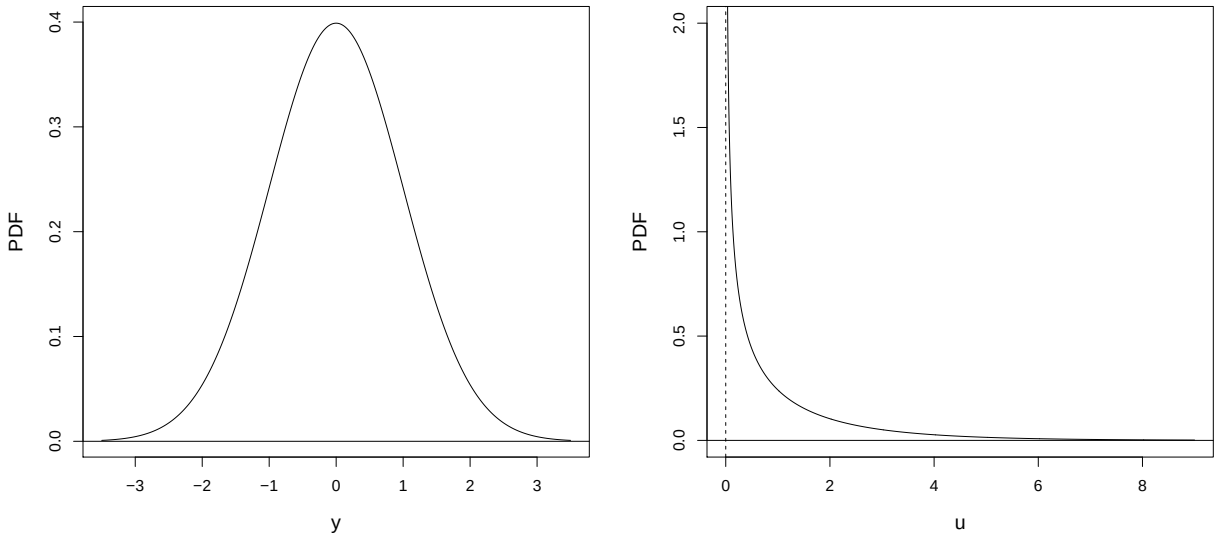

Figure 6.4: Left: Pdf of $Y \sim \mathcal{N}(0,1)$. Right: Pdf of $U = Y^2 \sim \chi^2(1)$.

Summarizing, a general formula for the pdf of $U = h(Y) = Y^2$ is

$$f_U(u) = \frac{1}{2\sqrt{u}}\left[f_Y(\sqrt{u}) + f_Y(-\sqrt{u})\right].$$

Example 6.4. Suppose $Y \sim \mathcal{N}(0,1)$; i.e., Y has a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. This is also called the standard normal distribution. The pdf of Y is

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}}e^{-y^2/2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Derive the pdf of $U = h(Y) = Y^2$.

Solution. We apply the result from Example 6.3. For $u > 0$, note that

$$\begin{aligned} f_Y(\sqrt{u}) &= \frac{1}{\sqrt{2\pi}}e^{-(\sqrt{u})^2/2} = \frac{1}{\sqrt{2\pi}}e^{-u/2} \\ f_Y(-\sqrt{u}) &= \frac{1}{\sqrt{2\pi}}e^{-(-\sqrt{u})^2/2} = \frac{1}{\sqrt{2\pi}}e^{-u/2}. \end{aligned}$$

Therefore, for $u > 0$,

$$\begin{aligned} f_U(u) = \frac{1}{2\sqrt{u}}\left(\frac{1}{\sqrt{2\pi}}e^{-u/2} + \frac{1}{\sqrt{2\pi}}e^{-u/2}\right) &= \frac{1}{\sqrt{u}}\frac{1}{\sqrt{2\pi}}e^{-u/2} \\ &= \frac{1}{\Gamma(\frac{1}{2})2^{1/2}}u^{\frac{1}{2}-1}e^{-u/2}. \end{aligned}$$

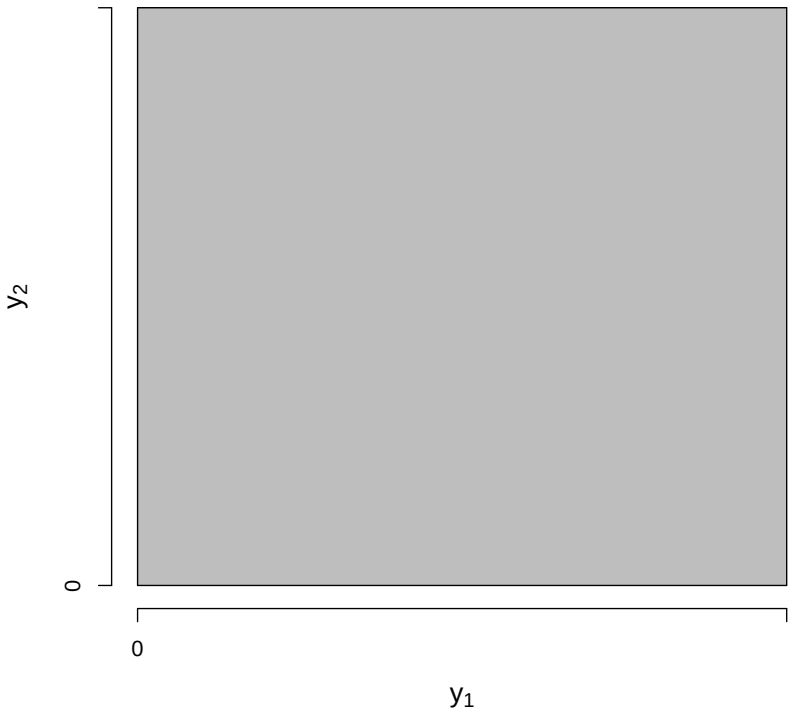

Figure 6.5: The support $R = \{(y_1, y_2) : y_1 > 0,\ y_2 > 0\}$ in Example 6.5; i.e., the entire first quadrant.

We recognize this as a gamma pdf with shape parameter $\alpha = 1/2$ and scale parameter $\beta = 2$, which is the same as the χ^2 pdf with 1 degree of freedom; i.e., $U \sim \chi^2(1)$. Therefore,

$$Y \sim \mathcal{N}(0,1) \implies U = Y^2 \sim \chi^2(1).$$

Both of these pdfs are shown in Figure 6.4 (see last page). □

Remark: The cdf technique is also useful in bivariate settings where Y_1 and Y_2 are continuous and we want to derive the distribution of $U = h(Y_1, Y_2)$. This is illustrated next.

Example 6.5. Suppose Y_1 and Y_2 are continuous random variables with joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} e^{-(y_1+y_2)}, & y_1 > 0,\ y_2 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the pdf of $U = h(Y_1, Y_2) = Y_1 + Y_2$ and calculate $E(U)$.

Solution. The bivariate support of (Y_1, Y_2) is $R_{Y_1,Y_2} = \{(y_1, y_2) : y_1 > 0,\ y_2 > 0\}$, the entire first quadrant; see Figure 6.5 (above). The joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ is a three-dimensional function which takes the value $e^{-(y_1+y_2)}$ over this region (and equals zero, otherwise).

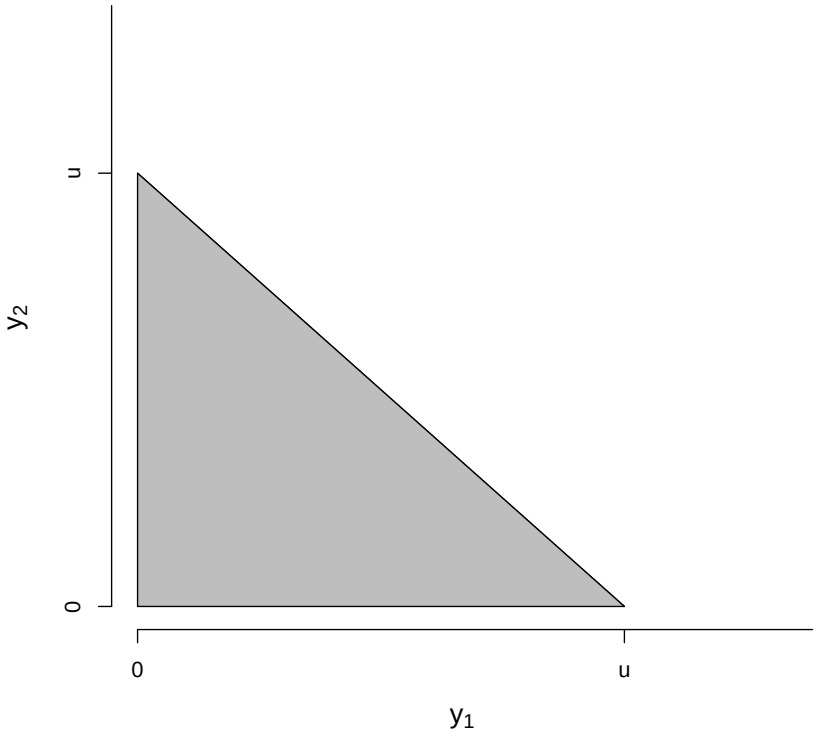

Figure 6.6: The set $B = \{(y_1, y_2) : y_1 > 0, y_2 > 0,\ y_1 + y_2 \leq u\}$ in Example 6.5. The upper boundary line is $y_2 = u - y_1$.

To find the distribution of $U = Y_1 + Y_2$, we use the cdf technique. First, observe that

$$y_1 > 0,\ y_2 > 0 \implies u = h(y_1, y_2) = y_1 + y_2 > 0.$$

Therefore, the support of $U = h(Y_1, Y_2) = Y_1 + Y_2$ is

$$R_U = \{u : u > 0\}.$$

For $u > 0$, the cdf of U is

$$\begin{aligned} F_U(u) = P(U \leq u) &= P(Y_1 + Y_2 \leq u) \\ &= \int\int_{(y_1,y_2)\in B} f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2 = \int\int_{(y_1,y_2)\in B} e^{-(y_1+y_2)}dy_1dy_2, \end{aligned}$$

where the set $B = \{(y_1, y_2) : y_1 > 0, y_2 > 0,\ y_1 + y_2 \leq u\}$ is shown in Figure 6.6 (see above). Note that the boundary of B is

$$y_1 + y_2 = u \implies y_2 = u - y_1,$$

a linear function of y_1 with slope -1 and intercept $u > 0$. Therefore,

$$\begin{aligned} F_U(u) = P(U \leq u) &= P(Y_1 + Y_2 \leq u) \\ &= \int_{y_1=0}^{u} \int_{y_2=0}^{u-y_1} e^{-(y_1+y_2)} dy_2 dy_1 \\ &= \int_{y_1=0}^{u} e^{-y_1} \left(-e^{-y_2} \Big|_{y_2=0}^{u-y_1} \right) dy_1 \\ &= \int_{y_1=0}^{u} e^{-y_1}[1 - e^{-(u-y_1)}] dy_1 \\ &= \int_{y_1=0}^{u} (e^{-y_1} - e^{-u}) dy_1 \;=\; \left(-e^{-y_1} - e^{-u} y_1\right) \Big|_{y_1=0}^{u} = 1 - e^{-u} - ue^{-u}. \end{aligned}$$

Therefore, the pdf of U, for $u > 0$, is

$$f_U(u) = \frac{d}{du} F_U(u) = \frac{d}{du}(1 - e^{-u} - ue^{-u}) = e^{-u} - (e^{-u} - ue^{-u}) = ue^{-u}.$$

Summarizing,

$$f_U(u) = \begin{cases} ue^{-u}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this as the gamma pdf with shape parameter $\alpha = 2$ and scale parameter $\beta = 1$; i.e., $U \sim$ gamma$(2, 1)$. Therefore, $E(U) = 2$. If you did not recognize $U \sim$ gamma$(2, 1)$, then you could calculate

$$E(U) = \int_{\mathbb{R}} u f_U(u) du = \int_0^{\infty} u^2 e^{-u} du = \Gamma(3) 1^3 = 2! = 2.$$

Remark: In the joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} e^{-(y_1+y_2)}, & y_1 > 0,\ y_2 > 0 \\ 0, & \text{otherwise,} \end{cases}$$

note that

$$f_{Y_1,Y_2}(y_1, y_2) = e^{-(y_1+y_2)} = e^{-y_1} e^{-y_2} = f_{Y_1}(y_1) f_{Y_2}(y_2),$$

for all $(y_1, y_2) \in \mathbb{R}^2$; i.e., Y_1 and Y_2 are independent exponential(1) random variables. Therefore, if all we wanted to do was find $E(U) = E(Y_1 + Y_2)$, we wouldn't have to derive the pdf of U. We could appeal to the Law of the Unconscious Statistician and simply write

$$E(U) = E(Y_1 + Y_2) = E(Y_1) + E(Y_2) = 1 + 1 = 2.$$

Note: If you didn't realize that Y_1 and Y_2 were exponential(1), then you could calculate

$$E(U) = E(Y_1 + Y_2) = \int_{y_1=0}^{\infty} \int_{y_2=0}^{\infty} (y_1 + y_2) e^{-(y_1+y_2)} dy_2 dy_1 = 2$$

by using the joint pdf of Y_1 and Y_2. $\square$

6.3 Method of transformations

Remark: We have just learned the method of distribution functions (i.e., the "cdf technique") to derive the distribution of $U = h(Y)$, a function of Y. The method of transformations is a special case of the cdf technique when $h : \mathbb{R} \to \mathbb{R}$ is a **one-to-one function** over R_Y, the support of Y. In this situation, we obtain a formula for $f_U(u)$, the pdf of U, in terms of $f_Y(u)$, the pdf of Y. Therefore, we can avoid having to work with cdfs.

Recall: By "one-to-one function," we mean either (a) h is strictly increasing over R_Y or (b) h is strictly decreasing over R_Y. Recall

- strictly increasing: $y_1 < y_2 \Longrightarrow h(y_1) < h(y_2)$; if h is differentiable, $h'(y) > 0$.
- strictly decreasing: $y_1 < y_2 \Longrightarrow h(y_1) > h(y_2)$; if h is differentiable, $h'(y) < 0$.

Setting: Suppose Y is a continuous random variable with cdf $F_Y(y)$ and pdf $f_Y(y)$ which is nonzero over the support R_Y. Let $U = h(Y)$, where h is a one-to-one function over R_Y.

Case 1: If h is strictly increasing, then

$$F_U(u) = P(U \leq u) = P(h(Y) \leq u) = P(Y \leq h^{-1}(u)) = F_Y(h^{-1}(u)).$$

Notice how we have written $F_U(u)$ in terms of $F_Y(y)$. The penultimate equality results from noting that $\{y : h(y) \leq u\} = \{y : y \leq h^{-1}(u)\}$. Taking derivatives, the pdf of U (where nonzero) is

$$f_U(u) = \frac{d}{du}F_U(u) = \frac{d}{du}F_Y(h^{-1}(u)) = f_Y(h^{-1}(u)) \underbrace{\frac{d}{du}h^{-1}(u)}_{>0}.$$

Recall: From calculus, recall that if h is strictly increasing (decreasing), then h^{-1} is also strictly increasing (decreasing).

Case 2: If h is strictly decreasing, then

$$F_U(u) = P(U \leq u) = P(h(Y) \leq u) = P(Y \geq h^{-1}(u)) = 1 - F_Y(h^{-1}(u)).$$

Notice how we have again written $F_U(u)$ in terms of $F_Y(y)$. The penultimate equality in this case results from noting that $\{y : h(y) \leq u\} = \{y : y \geq h^{-1}(u)\}$. Taking derivatives, the pdf of U (where nonzero) is

$$f_U(u) = \frac{d}{du}F_U(u) = \frac{d}{du}\left[1 - F_Y(h^{-1}(u))\right] = -f_Y(h^{-1}(u)) \underbrace{\frac{d}{du}h^{-1}(u)}_{<0}.$$

Combining both cases, we arrive at the following result.

Result: Suppose Y is a continuous random variable with pdf $f_Y(y)$ which is nonzero over the support R_Y. Let $U = h(Y)$, where h is a one-to-one function over R_Y. The pdf of U, where nonzero, is

$$f_U(u) = f_Y(h^{-1}(u)) \left|\frac{d}{du}h^{-1}(u)\right|.$$

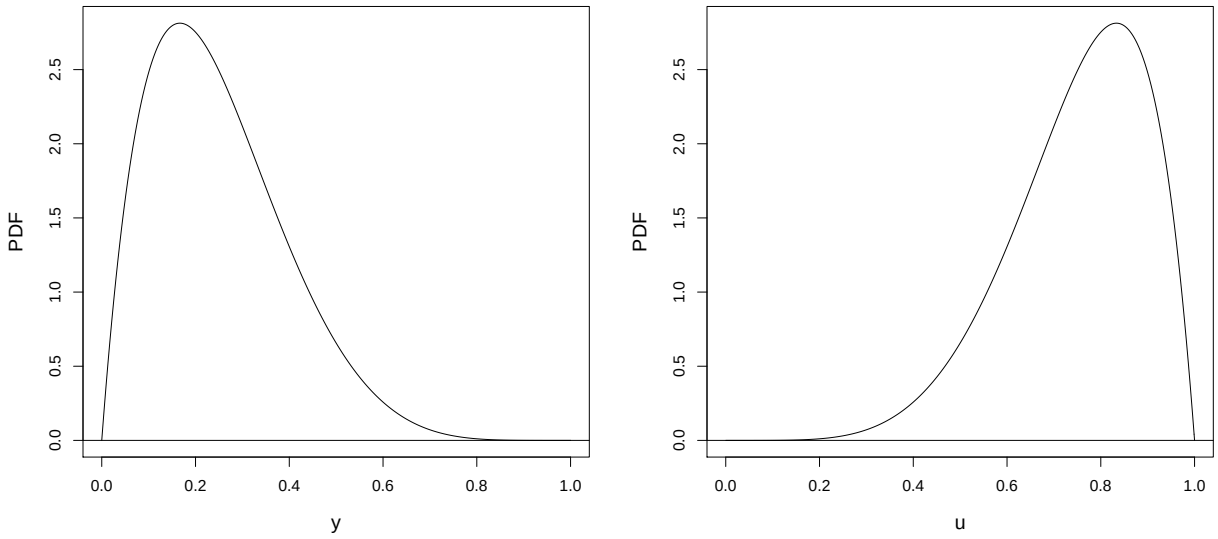

Figure 6.7: Left: Pdf of $Y \sim \text{beta}(2, 6)$. Right: Pdf of $U = 1 - Y \sim \text{beta}(6, 2)$.

Example 6.6. Suppose $Y \sim \text{beta}(\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$; i.e., the pdf of Y is

$$f_Y(y) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, y^{\alpha-1}(1-y)^{\beta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find the pdf of $U = h(Y) = 1 - Y$.

Solution. We use the transformation method. Note that $h(y) = 1 - y$ is a linear function of y with slope -1. Therefore, $h(y)$ is strictly decreasing and hence one-to-one over $R_Y = \{y : 0 < y < 1\}$. To find the support of U, note that

$$0 < y < 1 \iff 0 < 1 - y < 1.$$

Therefore, $R_U = \{u : 0 < u < 1\}$. We now find the inverse transformation:

$$u = h(y) = 1 - y \implies y = h^{-1}(u) = 1 - u.$$

The derivative of the inverse transformation is

$$\frac{d}{du}h^{-1}(u) = \frac{d}{du}(1-u) = -1.$$

Therefore, for $0 < u < 1$, the pdf of U is

$$\begin{aligned} f_U(u) &= f_Y(h^{-1}(u))\left|\frac{d}{du}h^{-1}(u)\right| \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,(1-u)^{\alpha-1}[1-(1-u)]^{\beta-1} \times |-1| = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,u^{\beta-1}(1-u)^{\alpha-1}. \end{aligned}$$

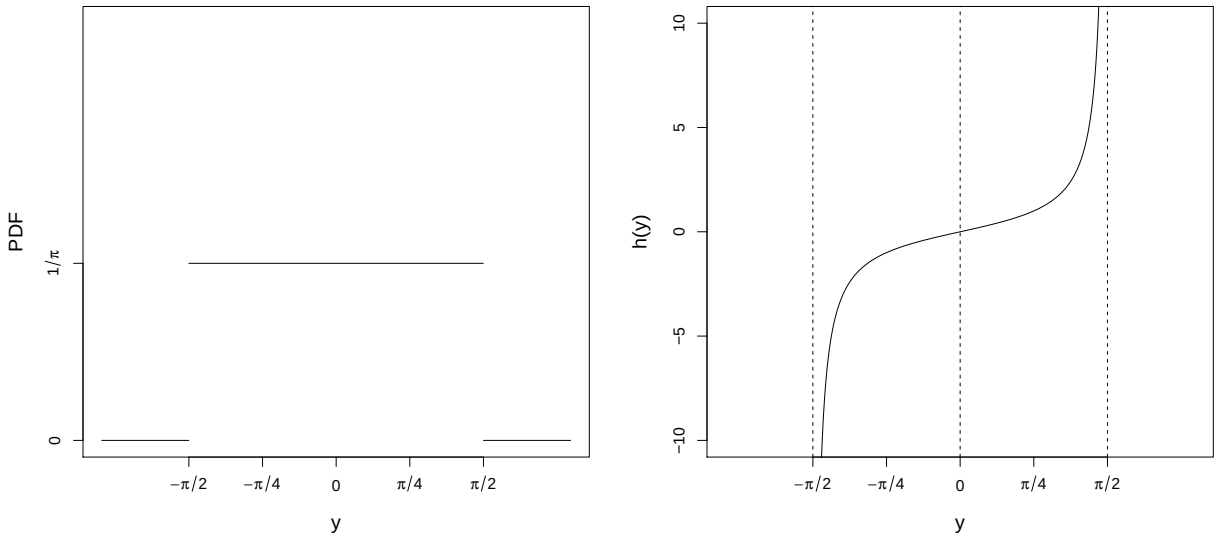

Figure 6.8: Left: Pdf of $Y \sim \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$. Right: A graph of the function $h(y) = \tan y$ over $(-\frac{\pi}{2}, \frac{\pi}{2})$; i.e., over the support $R_Y = \{y : -\frac{\pi}{2} < y < \frac{\pi}{2}\}$.

Summarizing, the pdf of $U = h(Y) = 1 - Y$ is

$$f_U(u) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, u^{\beta-1}(1-u)^{\alpha-1}, & 0 < u < 1 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this as a beta pdf with the roles of α and β reversed; i.e., $U \sim \text{beta}(\beta, \alpha)$. □

Example 6.7. Suppose $Y \sim \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$; i.e., Y has a uniform distribution from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$. The pdf of Y is

$$f_Y(y) = \begin{cases} \dfrac{1}{\pi}, & -\dfrac{\pi}{2} < y < \dfrac{\pi}{2} \\ 0, & \text{otherwise.} \end{cases}$$

Find the pdf of $U = h(Y) = \tan Y$.

Solution. We use the transformation method. The pdf of Y is shown in Figure 6.8 (left). Note that $h(y) = \tan y$ is not a one-to-one function over $\mathbb{R}$; however, it is one-to-one over $R_Y = \{y : -\frac{\pi}{2} < y < \frac{\pi}{2}\}$; see Figure 6.8 (right). To find the support of U, note that

$$-\frac{\pi}{2} < y < \frac{\pi}{2} \iff -\infty < \tan y < \infty.$$

Therefore, $R_U = \{u : -\infty < u < \infty\}$. We now find the inverse transformation. Note that

$$u = h(y) = \tan y \implies y = h^{-1}(u) = \tan^{-1} u.$$

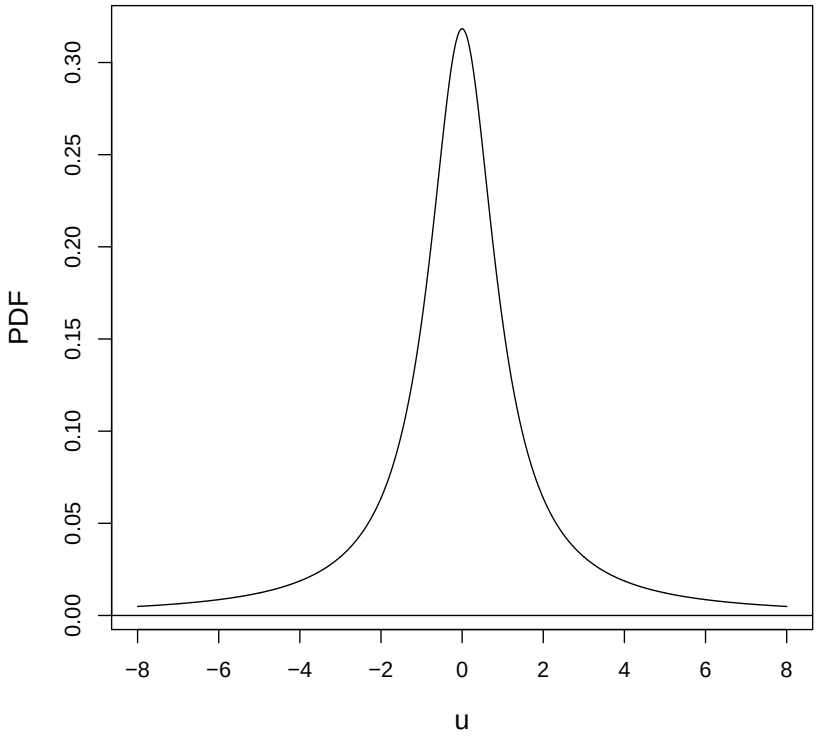

Figure 6.9: The standard Cauchy pdf.

The derivative of the inverse transformation is

$$\frac{d}{du}h^{-1}(u) = \frac{d}{du}\tan^{-1} u = \frac{1}{1+u^2}.$$

Therefore, for $-\infty < u < \infty$, the pdf of U is

$$f_U(u) = f_Y(h^{-1}(u))\left|\frac{d}{du}h^{-1}(u)\right| = \frac{1}{\pi}\left|\frac{1}{1+u^2}\right| = \frac{1}{\pi(1+u^2)}.$$

Summarizing, the pdf of $U = h(Y) = \tan Y$ is

$$f_U(u) = \begin{cases} \dfrac{1}{\pi(1+u^2)}, & -\infty < u < \infty \\ 0, & \text{otherwise.} \end{cases}$$

A random variable U with this pdf is said to have a standard **Cauchy distribution**. One interesting fact about the Cauchy distribution is that $E(U)$ does not exist (nor do any of the higher order moments). This is true because the integral

$$\int_{\mathbb{R}} u f_U(u)du = \int_{-\infty}^{\infty} \frac{u}{\pi(1+u^2)}du$$

does not converge absolutely. $\square$

Example 6.8. Suppose $Y \sim \mathcal{N}(\mu, \sigma^2)$; i.e., the pdf of Y is

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the pdf of $U = h(Y) = e^Y$.
(b) Find $E(U)$ and $V(U)$.

Solutions. (a) We use the transformation method. Note that $h(y) = e^y$ is a one-to-one function over $R_Y = \{y : -\infty < y < \infty\}$, the support of Y. To find the support of U, note

$$-\infty < y < \infty \iff u = e^y > 0.$$

Therefore, $R_U = \{u : u > 0\}$. We now find the inverse transformation. Note that

$$u = h(y) = e^y \implies y = h^{-1}(u) = \ln u.$$

The derivative of the inverse transformation is

$$\frac{d}{du} h^{-1}(u) = \frac{d}{du} \ln u = \frac{1}{u}.$$

Therefore, for $u > 0$, the pdf of U is

$$f_U(u) = f_Y(h^{-1}(u)) \left| \frac{d}{du} h^{-1}(u) \right| = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\ln u - \mu}{\sigma}\right)^2} \left| \frac{1}{u} \right| = \frac{1}{\sqrt{2\pi}\sigma u} e^{-\frac{1}{2}\left(\frac{\ln u - \mu}{\sigma}\right)^2}.$$

Summarizing, the pdf of $U = e^Y$ is

$$f_U(u) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma u} e^{-\frac{1}{2}\left(\frac{\ln u - \mu}{\sigma}\right)^2}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

A random variable U with this pdf is said to have a **lognormal distribution** with parameters μ and σ^2. We write $U \sim \text{lognormal}(\mu, \sigma^2)$. Figure 6.10 (next page) displays lognormal pdfs for different combinations of μ and σ^2.

(b) The mean of $U \sim \text{lognormal}(\mu, \sigma^2)$ is

$$E(U) = \int_{\mathbb{R}} u f_U(u) du = \int_0^\infty \frac{u}{\sqrt{2\pi}\sigma u} e^{-\frac{1}{2}\left(\frac{\ln u - \mu}{\sigma}\right)^2} du = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\ln u - \mu}{\sigma}\right)^2} du.$$

This integral is not easy. It is easier to use the Law of the Unconscious Statistician, write

$$E(U) = E(e^Y),$$

and then recognize that $E(e^Y)$ is the moment generating function (mgf) of $Y \sim \mathcal{N}(\mu, \sigma^2)$ when $t = 1$. We know the mgf of Y; recall that

$$m_Y(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Therefore,

$$E(U) = E(e^Y) = E(e^{tY})\Big|_{t=1} = m_Y(1) = \exp\left(\mu + \frac{\sigma^2}{2}\right).$$

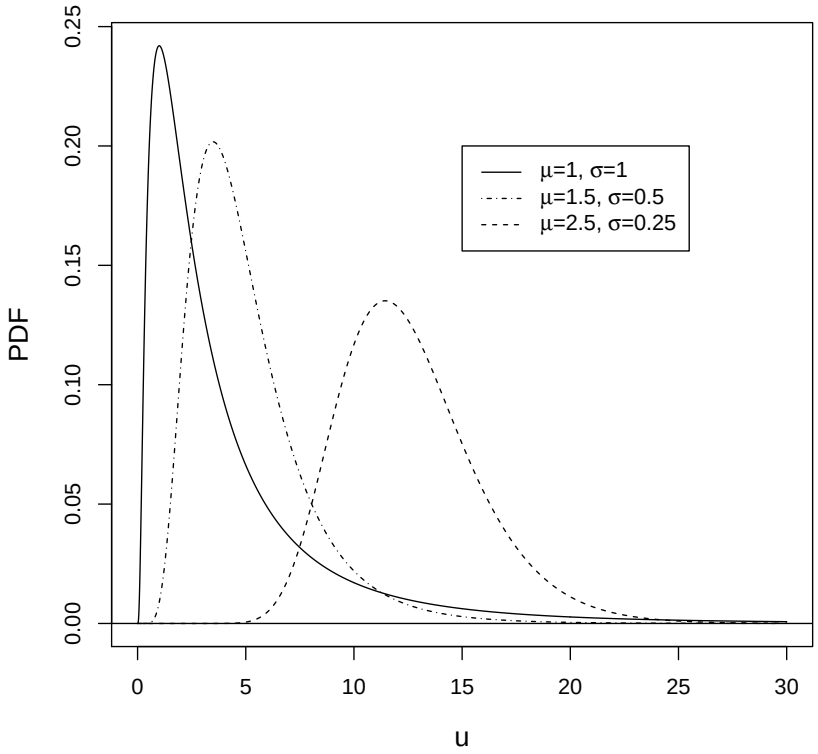

Figure 6.10: Lognormal(μ, σ^2) pdfs for different combinations of μ and σ^2.

To find $V(U)$, we find the second moment $E(U^2)$ and then use the variance computing formula. Note that

$$E(U^2) = E[(e^Y)^2] = E(e^{2Y}) = E(e^{tY})\Big|_{t=2} = m_Y(2) = e^{2(\mu+\sigma^2)}.$$

Therefore,

$$V(U) = E(U^2) - [E(U)]^2 = e^{2(\mu+\sigma^2)} - \left[\exp\left(\mu + \frac{\sigma^2}{2}\right)\right]^2 = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}.$$

Remark: The lognormal distribution is commonly used to model positive quantities (like the gamma and Weibull distributions). Note that

$$Y \sim \mathcal{N}(\mu, \sigma^2) \iff e^Y \sim \text{lognormal}(\mu, \sigma^2)$$

is equivalent to

$$U \sim \text{lognormal}(\mu, \sigma^2) \iff \ln U \sim \mathcal{N}(\mu, \sigma^2).$$

This is another reason the lognormal distribution is useful. In many applications, measurements are normal (or at least approximately normal) after taking logarithms. □

6.4 Method of moment generating functions

Recall: Suppose Y is a random variable. When Y is **discrete** with support R, the moment generating function (mgf) of Y is

$$m_Y(t) = E(e^{tY}) = \sum_{y \in R} e^{ty} p_Y(y),$$

where $p_Y(y)$ is the probability mass function (pmf) of Y. When Y is **continuous**, the mgf of Y is

$$m_Y(t) = E(e^{tY}) = \int_{\mathbb{R}} e^{ty} f_Y(y) dy,$$

where $f_Y(y)$ is the probability density function (pdf) of Y. In both cases, we require the expectation $E(e^{tY}) < \infty$ for all t in an open neighborhood about $t = 0$; i.e., $\exists b > 0$ such that $E(e^{tY}) < \infty \; \forall t \in (-b, b)$. If no such $b > 0$ exists, then the mgf of Y does not exist.

Moments: A random variable's mgf is a powerful tool. For one, we learned

$$E(Y^k) = m_Y^{(k)}(0),$$

where

$$m_Y^{(k)}(0) = \frac{d^k}{dt^k} m_Y(t) \bigg|_{t=0}.$$

In other words, the moments of Y can be found by differentiating the mgf.

Uniqueness: Another reason the mgf is important is that it uniquely identifies the distribution of Y. For example, suppose I have a random variable Y whose mgf is given by

$$m_Y(t) = e^{2.5(e^t - 1)}.$$

Then I know $Y \sim$ Poisson($\lambda = 2.5$) because this is the mgf of a Poisson random variable with mean $\lambda = 2.5$. Or, perhaps Y has the following mgf:

$$m_Y(t) = \left(\frac{1}{1 - 3.6t} \right)^2, \quad \text{for } t < 1/3.6.$$

In this case, I know $Y \sim$ gamma($\alpha = 2, \beta = 3.6$) because this is the mgf of a gamma random variable with shape parameter $\alpha = 2$ and scale parameter $\beta = 3.6$. The uniqueness property of mgfs stems from the uniqueness of LaPlace transforms in mathematical analysis.

Usefulness: Because a random variable's mgf uniquely identifies its distribution, we can exploit this to answer the question posed at the beginning of this chapter, namely,

> "If we know the distribution of Y, what is the distribution of $U = h(Y)$?"

Because mgfs are unique, we now have another approach to try when answering this question. We can derive the mgf of U and then match it to one that we know (e.g., Poisson, gamma, etc.). If we can do this, then we know U must have the distribution identified by that mgf. This is called the **method of moment generating functions** (i.e., the "mgf technique").

Example 6.9. Suppose that $Y \sim \text{gamma}(\alpha, \beta)$; i.e., Y has a gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. For the constant $c > 0$, find the distribution of $U = h(Y) = cY$.

Solution. We could apply the method of transformations here. Note that $u = h(y) = cy$ is a strictly increasing (and hence one-to-one) function over $R_Y = \{y : y > 0\}$, the support of Y. Let's use the mgf technique instead. Recall that the mgf of Y is

$$m_Y(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}, \quad t < \frac{1}{\beta}.$$

The mgf of $U = h(Y) = cY$ is therefore

$$\begin{aligned} m_U(t) = E(e^{tU}) &= E(e^{ctY}) \\ &= m_Y(ct) \\ &= \left(\frac{1}{1-\beta ct}\right)^{\alpha}, \end{aligned}$$

which exists for $ct < 1/\beta \Longleftrightarrow t < 1/\beta c$. We recognize $m_U(t)$ as the mgf of a gamma random variable with shape parameter α and scale parameter βc. Because mgfs are unique (i.e., they uniquely identify a distribution), we have $U = cY \sim \text{gamma}(\alpha, \beta c)$. $\square$

Special case: Suppose $Y \sim \text{gamma}(\alpha, \beta)$ and take $c = 2/\beta$ so that

$$U = h(Y) = cY = \frac{2Y}{\beta}.$$

The result in Example 6.9 says

$$Y \sim \text{gamma}(\alpha, \beta) \implies U = \frac{2Y}{\beta} \sim \text{gamma}(\alpha, 2) \stackrel{d}{=} \chi^2(2\alpha).$$

In other words, we can always convert a gamma random variable Y into a χ^2 random variable by using this specific transformation. This fact is important and will be used repeatedly. The symbol "$\stackrel{d}{=}$" is read "is equal in distribution" or "has the same distribution as."

Example 6.10. Suppose $Y \sim \mathcal{N}(\mu, \sigma^2)$. For constants $a, b \in \mathbb{R}$, derive the distribution of $U = h(Y) = aY + b$.

Solution. Again, we could apply the method of transformations here. Note that $u = h(y) = a + by$ is a linear (and hence one-to-one) function over $R_Y = \{y : -\infty < y < \infty\}$, the support of Y. Let's use the mgf technique instead. Recall that the mgf of Y is

$$m_Y(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

The mgf of $U = h(Y) = aY + b$ is

$$m_U(t) = E(e^{tU}) = E[e^{t(aY+b)}] = E(e^{atY+bt}) = E(e^{atY}e^{bt}) = e^{bt}E(e^{atY}) = e^{bt}m_Y(at).$$

Therefore,

$$m_U(t) = e^{bt}\exp\left[\mu(at) + \frac{\sigma^2(at)^2}{2}\right] = \exp\left[(a\mu+b)t + \frac{(a^2\sigma^2)t^2}{2}\right].$$

We recognize $m_U(t)$ as the mgf of a normal random variable with mean $a\mu+b$ and variance $a^2\sigma^2$. Because mgfs are unique, we have shown

$$Y \sim \mathcal{N}(\mu,\sigma^2) \implies U = h(Y) = aY + b \sim \mathcal{N}(a\mu+b, a^2\sigma^2).$$

In other words, linear functions of normal random variables are normally distributed.

Special case: Suppose $Y \sim \mathcal{N}(\mu,\sigma^2)$ and consider

$$Z = \frac{Y-\mu}{\sigma} = aY + b,$$

where $a = 1/\sigma$ and $b = -\mu/\sigma$. With these values of a and b, note that

$$a\mu + b = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0 \quad \text{and} \quad a^2\sigma^2 = \frac{\sigma^2}{\sigma^2} = 1.$$

This shows $Z \sim \mathcal{N}(0,1)$; i.e., Z has a standard normal distribution. $\square$

Remark: The mgf technique is perhaps most useful when finding the distribution of the sum of independent random variables $Y_1, Y_2, ..., Y_n$. We start with the $n = 2$ case.

Result: Suppose Y_1 and Y_2 are random variables with moment generating functions $m_{Y_1}(t)$ and $m_{Y_2}(t)$, respectively. Define

$$U = Y_1 + Y_2,$$

the sum of Y_1 and Y_2. If Y_1 and Y_2 are independent, then

$$\begin{aligned} m_U(t) = E(e^{tU}) = E[e^{t(Y_1+Y_2)}] = E(e^{tY_1+tY_2}) &= E(e^{tY_1}e^{tY_2}) \\ &\overset{Y_1 \perp\!\!\!\perp Y_2}{=} E(e^{tY_1})E(e^{tY_2}) = m_{Y_1}(t)m_{Y_2}(t). \end{aligned}$$

Recall from STAT 511 that if Y_1 and Y_2 are independent, then functions of Y_1 and Y_2 are too; e.g., e^{tY_1} and e^{tY_2}. Therefore, we have shown *the mgf of the sum of independent random variables is the product of the marginal mgfs.*

Example 6.11. Suppose $Y_1 \sim \text{Poisson}(\lambda_1)$ and $Y_2 \sim \text{Poisson}(\lambda_2)$. If Y_1 and Y_2 are independent, the mgf of $U = Y_1 + Y_2$ is

$$m_U(t) = m_{Y_1}(t)m_{Y_2}(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}.$$

We recognize this as the mgf of a Poisson random variable with mean $\lambda = \lambda_1 + \lambda_2$. Because mgfs are unique, we have shown

$$Y_1 \sim \text{Poisson}(\lambda_1), Y_2 \sim \text{Poisson}(\lambda_2), Y_1 \perp\!\!\!\perp Y_2 \implies U = Y_1 + Y_2 \sim \text{Poisson}(\lambda_1+\lambda_2). \ \square$$

Note: The mgf result on the last page can be generalized. Suppose $Y_1, Y_2, ..., Y_n$ are mutually independent random variables with moment generating functions $m_{Y_1}(t)$, $m_{Y_2}(t)$, ..., $m_{Y_n}(t)$, respectively. Define

$$U = \sum_{i=1}^{n} Y_i = Y_1 + Y_2 + \cdots + Y_n,$$

the sum of $Y_1, Y_2, ..., Y_n$. Then

$$m_U(t) = m_{Y_1}(t)m_{Y_2}(t)\cdots m_{Y_n}(t) = \prod_{i=1}^{n} m_{Y_i}(t).$$

For example, as a generalization of Example 6.11 (last page), if $Y_1, Y_2, ..., Y_n$ were mutually independent Poisson random variables with means $\lambda_1, \lambda_2, ..., \lambda_n$, respectively, then

$$U = \sum_{i=1}^{n} Y_i \sim \text{Poisson}(\lambda_1 + \lambda_2 + \cdots + \lambda_n).$$

Example 6.12. Suppose $Y_1, Y_2, ..., Y_n$ are mutually independent exponential random variables, each with the same mean $\beta > 0$. For $t < 1/\beta$, the mgf of $U = Y_1 + Y_2 + \cdots + Y_n$ is

$$\begin{aligned} m_U(t) &= m_{Y_1}(t)m_{Y_2}(t)\cdots m_{Y_n}(t) \\ &= \left(\frac{1}{1-\beta t}\right) \times \left(\frac{1}{1-\beta t}\right) \times \cdots \times \left(\frac{1}{1-\beta t}\right) = \left(\frac{1}{1-\beta t}\right)^n. \end{aligned}$$

We recognize this as the mgf of a gamma random variable with shape parameter $\alpha = n$ and scale parameter β. Because mgfs are unique, $U = Y_1 + Y_2 + \cdots + Y_n \sim \text{gamma}(n, \beta)$. $\square$

Remark: As wonderful as the mgf technique is (especially when dealing with sums of independent random variables), it is not always helpful. Suppose that in Example 6.12, the random variables $Y_1, Y_2, ..., Y_n$ were mutually independent with exponential distributions, but suppose they had different means, that is, suppose

$$\begin{aligned} Y_1 &\sim \text{exponential}(\beta_1) \\ Y_2 &\sim \text{exponential}(\beta_2) \\ &\vdots \\ Y_n &\sim \text{exponential}(\beta_n). \end{aligned}$$

In this situation, the mgf of $U = Y_1 + Y_2 + \cdots + Y_n$ is

$$\begin{aligned} m_U(t) &= m_{Y_1}(t)m_{Y_2}(t)\cdots m_{Y_n}(t) \\ &= \left(\frac{1}{1-\beta_1 t}\right) \times \left(\frac{1}{1-\beta_2 t}\right) \times \cdots \times \left(\frac{1}{1-\beta_n t}\right) = \prod_{i=1}^{n} \frac{1}{1-\beta_i t}. \end{aligned}$$

This is the mgf of $U = Y_1 + Y_2 + \cdots + Y_n$, but it does not have a form of one we recognize. Therefore, we are unable to conclude what the distribution of U is in this case.

Example 6.13. Suppose $Y_1, Y_2, ..., Y_n$ are mutually independent Bernoulli random variables, each with success probability (mean) p, where $0 < p < 1$. Recall the Bernoulli(p) distribution is another name for the $b(n, p)$ distribution when $n = 1$; i.e., the pmf of $Y \sim$ Bernoulli(p) is

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

The Bernoulli distribution applies when Y has only two outcomes: "success" ($Y = 1$) and "failure" ($Y = 0$). Find the distribution of $U = Y_1 + Y_2 + \cdots + Y_n$.

Solution. The mgf of $Y \sim$ Bernoulli(p) is given by

$$m_Y(t) = E(e^{tY}) = (1-p)e^{t(0)} + pe^{t(1)} = q + pe^t,$$

where $q = 1 - p$. Therefore, the mgf of $U = Y_1 + Y_2 + \cdots + Y_n$ is

$$\begin{aligned} m_U(t) &= m_{Y_1}(t)m_{Y_2}(t)\cdots m_{Y_n}(t) \\ &= (q + pe^t) \times (q + pe^t) \times \cdots \times (q + pe^t) = (q + pe^t)^n. \end{aligned}$$

We recognize this as the mgf of a binomial random variable with number of trials n and success probability p. Because mgfs are unique, $U = Y_1 + Y_2 + \cdots + Y_n \sim b(n, p)$. This example shows a binomial random variable $U \sim b(n, p)$ can always be expressed as the sum of mutually independent Bernoulli random variables, each with the same mean p. $\square$

Example 6.14. Suppose $Y_1, Y_2, ..., Y_n$ are mutually independent normal random variables with means $\mu_1, \mu_2, ..., \mu_n$ and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$, respectively. That is, suppose

$$\begin{aligned} Y_1 &\sim \mathcal{N}(\mu_1, \sigma_1^2) \\ Y_2 &\sim \mathcal{N}(\mu_2, \sigma_2^2) \\ &\vdots \\ Y_n &\sim \mathcal{N}(\mu_n, \sigma_n^2). \end{aligned}$$

Find the distribution of the linear combination

$$U = \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n.$$

Solution. We studied linear combinations in STAT 511. Recall the mean of U is

$$E(U) = E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i) = \sum_{i=1}^{n} a_i \mu_i$$

and the variance of U is

$$V(U) = V\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\underbrace{\sum_{i<j}\sum a_i a_j \text{Cov}(Y_i, Y_j)}_{=\,0} = \sum_{i=1}^{n} a_i^2 V(Y_i) = \sum_{i=1}^{n} a_i^2 \sigma_i^2,$$

because $Y_1, Y_2, ..., Y_n$ are mutually independent (i.e., all the covariances are zero). So, we already know the mean and variance of U. However, what is the distribution of U? The mgf of U is

$$\begin{aligned} m_U(t) = E(e^{tU}) &= E[e^{t(a_1Y_1+a_2Y_2+\cdots+a_nY_n)}] \\ &= E(e^{a_1tY_1}e^{a_2tY_2}\cdots e^{a_ntY_n}) \\ &= E(e^{a_1tY_1})E(e^{a_2tY_2})\cdots E(e^{a_ntY_n}) \\ &= m_{Y_1}(a_1t)m_{Y_2}(a_2t)\cdots m_{Y_n}(a_nt) = \prod_{i=1}^{n} m_{Y_i}(a_it). \end{aligned}$$

Now recall that

$$m_{Y_i}(t) = \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right) \implies m_{Y_i}(a_it) = \exp\left(a_i\mu_i t + \frac{a_i^2\sigma_i^2 t^2}{2}\right).$$

Therefore,

$$m_U(t) = \prod_{i=1}^{n} \exp\left(a_i\mu_i t + \frac{a_i^2\sigma_i^2 t^2}{2}\right) = \exp\left[\left(\sum_{i=1}^{n} a_i\mu_i\right)t + \frac{\left(\sum_{i=1}^{n} a_i^2\sigma_i^2\right)t^2}{2}\right].$$

We recognize this as the mgf of a normal random variable with mean $\sum_{i=1}^{n} a_i\mu_i$ and variance $\sum_{i=1}^{n} a_i^2\sigma_i^2$. Because mgfs are unique, we have shown

$$U = \sum_{i=1}^{n} a_iY_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

In other words, linear combinations of normal random variables are normally distributed.

Remark: Note that if we take

- $a_1 = a_2 = \cdots = a_n = 1$:
$$\sum_{i=1}^{n} Y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$
- $a_1 = a_2 = \cdots = a_n = 1$; $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$ (i.e., common means and variances):
$$\sum_{i=1}^{n} Y_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right).$$
- $a_1 = a_2 = \cdots = a_n = \frac{1}{n}$; $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$ (i.e., common means and variances):
$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Therefore, Example 6.14 has many important special cases (which will be used later). □

Example 6.15. Suppose $Y_1, Y_2, ..., Y_n$ are mutually independent random variables with

$$\begin{aligned} Y_1 &\sim \mathcal{N}(\mu_1, \sigma_1^2) \\ Y_2 &\sim \mathcal{N}(\mu_2, \sigma_2^2) \\ &\vdots \\ Y_n &\sim \mathcal{N}(\mu_n, \sigma_n^2). \end{aligned}$$

Find the distribution of

$$U = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2.$$

Solution. We have already shown (in Example 6.10) that

$$Z_i = \frac{Y_i - \mu_i}{\sigma_i} \sim \mathcal{N}(0,1)$$

and we know (from Example 6.4) that $Z_i \sim \mathcal{N}(0,1) \Longrightarrow Z_i^2 \sim \chi^2(1)$. Therefore, we can write

$$U = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2 = \sum_{i=1}^{n} Z_i^2.$$

Because $Y_1, Y_2, ..., Y_n$ are mutually independent (by assumption), we know that $Z_1^2, Z_2^2, ..., Z_n^2$ are too because functions of mutually independent random variables are also mutually independent. Recall the mgf of each $Z_i^2 \sim \chi^2(1)$ is given by

$$m_{Z_i^2}(t) = \left(\frac{1}{1-2t} \right)^{1/2}, \quad \text{for } t < 1/2.$$

Therefore, the mgf of U is

$$\begin{aligned} m_U(t) &= m_{Z_1^2}(t) m_{Z_2^2}(t) \cdots m_{Z_n^2}(t) \\ &= \left(\frac{1}{1-2t} \right)^{1/2} \times \left(\frac{1}{1-2t} \right)^{1/2} \times \cdots \times \left(\frac{1}{1-2t} \right)^{1/2} = \left(\frac{1}{1-2t} \right)^{n/2}. \end{aligned}$$

We recognize this as the mgf of a χ^2 random variable with n degrees of freedom. Because mgfs are unique, we have shown

$$U = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2(n).$$

Special case: $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$ (i.e., common means and variances):

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 \sim \chi^2(n).$$

In Chapter 7, we will show that if we replace μ above with $\overline{Y}$, we get

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \sim \chi^2(n-1).$$

A degree of freedom is "lost" for "estimating" the common mean μ with $\overline{Y}$. $\square$

6.5 Bivariate transformations

Univariate transformations: Recall if Y is a continuous random variable with pdf $f_Y(y)$, then the pdf of the one-to-one function $U = h(Y)$, where nonzero, is

$$f_U(u) = f_Y(h^{-1}(u)) \left| \frac{d}{du} h^{-1}(u) \right|.$$

We now extend the transformation technique to bivariate distributions.

Setting: Suppose $\mathbf{Y} = (Y_1, Y_2)$ is a continuous random vector with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ and support R_{Y_1,Y_2}; i.e., the region in $\mathbb{R}^2$ where $f_{Y_1,Y_2}(y_1, y_2) > 0$. Define

$$\begin{aligned} U_1 &= h_1(Y_1, Y_2) \\ U_2 &= h_2(Y_1, Y_2) \end{aligned}$$

so that

$$\left(\begin{array}{c} U_1 \\ U_2 \end{array} \right) = h \left(\begin{array}{c} Y_1 \\ Y_2 \end{array} \right) = \left(\begin{array}{c} h_1(Y_1, Y_2) \\ h_2(Y_1, Y_2) \end{array} \right)$$

is a vector-valued mapping from R_{Y_1,Y_2} to

$$R_{U_1,U_2} = \{(u_1, u_2) : u_1 = h_1(y_1, y_2),\ u_2 = h_2(y_1, y_2),\ \text{for } (y_1, y_2) \in R_{Y_1,Y_2}\},$$

the support of $\mathbf{U} = (U_1, U_2)$. In what follows, we require h to be a one-to-one transformation. That is, for each $(u_1, u_2) \in R_{U_1,U_2}$, there is only one $(y_1, y_2) \in R_{Y_1,Y_2}$ satisfying

$$\begin{aligned} u_1 &= h_1(y_1, y_2) \\ u_2 &= h_2(y_1, y_2). \end{aligned}$$

Because h is one-to-one, we can find the **inverse transformation**

$$\begin{aligned} y_1 &= h_1^{-1}(u_1, u_2) \\ y_2 &= h_2^{-1}(u_1, u_2). \end{aligned}$$

The **Jacobian** of the (inverse) transformation is defined as

$$J = \det \left| \begin{array}{cc} \dfrac{\partial h_1^{-1}(u_1, u_2)}{\partial u_1} & \dfrac{\partial h_1^{-1}(u_1, u_2)}{\partial u_2} \\ \dfrac{\partial h_2^{-1}(u_1, u_2)}{\partial u_1} & \dfrac{\partial h_2^{-1}(u_1, u_2)}{\partial u_2} \end{array} \right|,$$

that is, J is the determinant of this 2×2 matrix of partial derivatives. By the Change of Variables Theorem from analysis, we conclude the joint pdf of $\mathbf{U} = (U_1, U_2)$, where nonzero, is given by

$$f_{U_1,U_2}(u_1, u_2) = f_{Y_1,Y_2}(h_1^{-1}(u_1, u_2), h_2^{-1}(u_1, u_2))|J|,$$

where $|J|$ denotes the absolute value of J.

Recall: We calculate the determinant of a 2×2 matrix as follows:

$$\det \left| \begin{array}{cc} a & b \\ c & d \end{array} \right| = ad - bc.$$

Example 6.16. Suppose $Y_1 \sim \text{gamma}(\alpha, 1)$, $Y_2 \sim \text{gamma}(\beta, 1)$, and Y_1 and Y_2 are independent. Use a bivariate transformation to find the joint pdf of $\mathbf{U} = (U_1, U_2)$, where

$$\begin{aligned} U_1 &= h_1(Y_1, Y_2) = Y_1 + Y_2 \\ U_2 &= h_2(Y_1, Y_2) = \frac{Y_1}{Y_1 + Y_2}. \end{aligned}$$

Solution. Because Y_1 and Y_2 are independent, we know the joint pdf of $\mathbf{Y} = (Y_1, Y_2)$ is given by

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &\overset{Y_1 \perp\!\!\!\perp Y_2}{=} f_{Y_1}(y_1) f_{Y_2}(y_2) \\ &= \underbrace{\frac{1}{\Gamma(\alpha)} y_1^{\alpha-1} e^{-y_1}}_{f_{Y_1}(y_1)} \underbrace{\frac{1}{\Gamma(\beta)} y_2^{\beta-1} e^{-y_2}}_{f_{Y_2}(y_2)} = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha-1} y_2^{\beta-1} e^{-(y_1+y_2)}, \end{aligned}$$

for $y_1 > 0$ and $y_2 > 0$. That is, the support of $\mathbf{Y} = (Y_1, Y_2)$ is

$$R_{Y_1,Y_2} = \{(y_1, y_2) : y_1 > 0, y_2 > 0\},$$

the entire first quadrant. What is the support of $\mathbf{U} = (U_1, U_2)$? The transformation

$$\begin{aligned} u_1 &= h_1(y_1, y_2) = y_1 + y_2 \\ u_2 &= h_2(y_1, y_2) = \frac{y_1}{y_1 + y_2} \end{aligned}$$

maps values of $(y_1, y_2) \in R_{Y_1,Y_2}$ to

$$R_{U_1,U_2} = \{(u_1, u_2) : u_1 > 0,\ 0 < u_2 < 1\}.$$

Both support sets are shown in Figure 6.11 (see next page). To verify the transformation above is one-to-one, we show $h(y_1, y_2) = h(y_1^*, y_2^*) \Longrightarrow y_1 = y_1^*$ and $y_2 = y_2^*$, where

$$h \left(\begin{array}{c} y_1 \\ y_2 \end{array} \right) = \left(\begin{array}{c} h_1(y_1, y_2) \\ h_2(y_1, y_2) \end{array} \right) = \left(\begin{array}{c} y_1 + y_2 \\ \dfrac{y_1}{y_1 + y_2} \end{array} \right).$$

Suppose $h(y_1, y_2) = h(y_1^*, y_2^*)$. This means both of these equations hold:

$$y_1 + y_2 = y_1^* + y_2^* \qquad \text{and} \qquad \frac{y_1}{y_1 + y_2} = \frac{y_1^*}{y_1^* + y_2^*}.$$

The two equations together imply that $y_1 = y_1^*$. The first equation then implies $y_2 = y_2^*$. Hence, the transformation $h : R_{Y_1,Y_2} \to R_{U_1,U_2}$ is one-to-one.

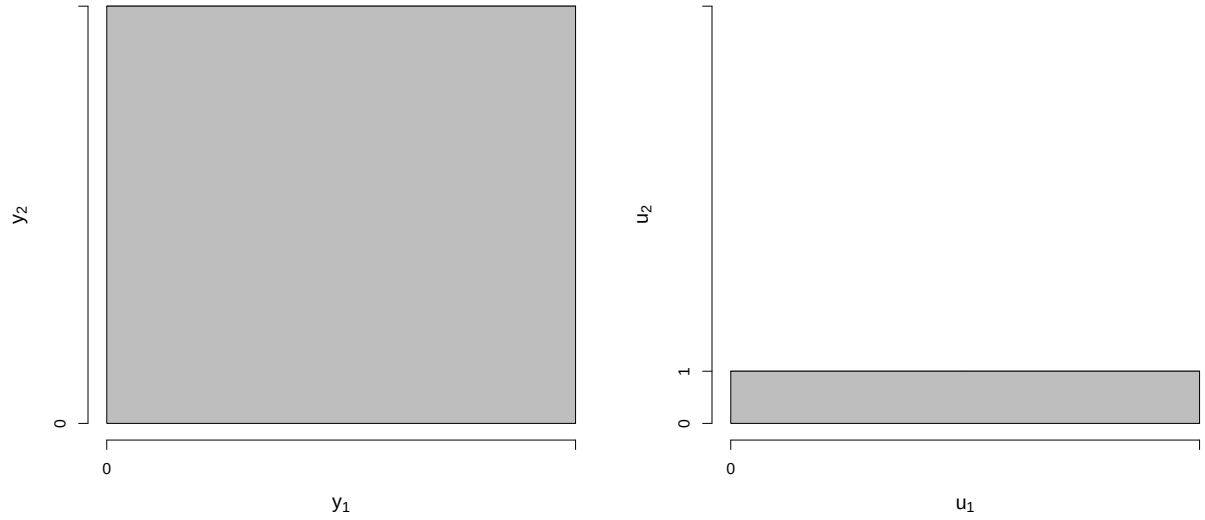

Figure 6.11: Left: The support $R_{Y_1,Y_2} = \{(y_1, y_2) : y_1 > 0, y_2 > 0\}$ in Example 6.16. Right: $R_{U_1,U_2} = \{(u_1, u_2) : u_1 > 0,\ 0 < u_2 < 1\}$. The transformation $h : R_{Y_1,Y_2} \to R_{U_1,U_2}$.

The inverse transformation is found by solving

$$\begin{aligned} u_1 &= y_1 + y_2 \\ u_2 &= \frac{y_1}{y_1 + y_2} \end{aligned}$$

for $y_1 = h_1^{-1}(u_1, u_2)$ and $y_2 = h_2^{-1}(u_1, u_2)$. Straightforward algebra shows

$$\begin{aligned} y_1 &= h_1^{-1}(u_1, u_2) = u_1 u_2 \\ y_2 &= h_2^{-1}(u_1, u_2) = u_1(1 - u_2). \end{aligned}$$

The Jacobian is

$$J = \det \begin{vmatrix} \dfrac{\partial h_1^{-1}(u_1,u_2)}{\partial u_1} & \dfrac{\partial h_1^{-1}(u_1,u_2)}{\partial u_2} \\ \dfrac{\partial h_2^{-1}(u_1,u_2)}{\partial u_1} & \dfrac{\partial h_2^{-1}(u_1,u_2)}{\partial u_2} \end{vmatrix} = \det \begin{vmatrix} u_2 & u_1 \\ 1 - u_2 & -u_1 \end{vmatrix} = -u_1 u_2 - u_1(1 - u_2) = -u_1.$$

Therefore, the joint pdf of $\mathbf{U} = (U_1, U_2)$, where nonzero, is

$$\begin{aligned} f_{U_1,U_2}(u_1, u_2) &= f_{Y_1,Y_2}(h_1^{-1}(u_1, u_2), h_2^{-1}(u_1, u_2))|J| \\ &= f_{Y_1,Y_2}(u_1 u_2, u_1(1 - u_2))| - u_1| \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}(u_1 u_2)^{\alpha-1}[u_1(1 - u_2)]^{\beta-1} e^{-u_1 u_2 - u_1(1-u_2)} \times u_1 \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1} u_2^{\alpha-1}(1 - u_2)^{\beta-1} e^{-u_1}. \end{aligned}$$

Summarizing,

$$f_{U_1,U_2}(u_1,u_2) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1} u_2^{\alpha-1}(1-u_2)^{\beta-1} e^{-u_1}, & u_1 > 0,\ 0 < u_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Note: We are done performing the bivariate transformation. We started with the joint pdf of $\mathbf{Y} = (Y_1, Y_2)$, and now we have the joint pdf of $\mathbf{U} = (U_1, U_2)$. We now make further observations.

Q: What is the marginal pdf of U_1?
A: To find the marginal pdf of U_1, we take the joint pdf $f_{U_1,U_2}(u_1,u_2)$ and integrate over u_2, that is,

$$\begin{aligned} f_{U_1}(u_1) &\overset{u_1>0}{=} \int_{u_2=0}^{1} \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1} u_2^{\alpha-1}(1-u_2)^{\beta-1} e^{-u_1} du_2 \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1} e^{-u_1} \int_{u_2=0}^{1} u_2^{\alpha-1}(1-u_2)^{\beta-1} du_2 \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1} e^{-u_1} \times \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \frac{1}{\Gamma(\alpha+\beta)} u_1^{\alpha+\beta-1} e^{-u_1}. \end{aligned}$$

Summarizing,

$$f_{U_1}(u_1) = \begin{cases} \dfrac{1}{\Gamma(\alpha+\beta)} u_1^{\alpha+\beta-1} e^{-u_1}, & u_1 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this as a gamma pdf with shape parameter $\alpha+\beta$ and scale parameter 1. That is, $U_1 \sim \text{gamma}(\alpha+\beta, 1)$.

Q: What is the marginal pdf of U_2?
A: To find the marginal pdf of U_2, we take the joint pdf $f_{U_1,U_2}(u_1,u_2)$ and integrate over u_1, that is,

$$\begin{aligned} f_{U_2}(u_2) &\overset{0<u_2<1}{=} \int_{u_1=0}^{\infty} \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1} u_2^{\alpha-1}(1-u_2)^{\beta-1} e^{-u_1} du_1 \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u_2^{\alpha-1}(1-u_2)^{\beta-1} \int_{u_1=0}^{\infty} u_1^{\alpha+\beta-1} e^{-u_1} du_1 \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u_2^{\alpha-1}(1-u_2)^{\beta-1} \times \Gamma(\alpha+\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u_2^{\alpha-1}(1-u_2)^{\beta-1}. \end{aligned}$$

Summarizing,

$$f_{U_2}(u_2) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u_2^{\alpha-1}(1-u_2)^{\beta-1}, & 0 < u_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this as a beta pdf with parameters α and β. That is, $U_2 \sim \text{beta}(\alpha, \beta)$.

Note: We make an additional observation in Example 6.16. Note that we can write

$$\begin{aligned} f_{U_1,U_2}(u_1,u_2) &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1} u_2^{\alpha-1}(1-u_2)^{\beta-1} e^{-u_1} \\ &= \underbrace{\frac{1}{\Gamma(\alpha+\beta)} u_1^{\alpha+\beta-1} e^{-u_1}}_{f_{U_1}(u_1)} \times \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u_2^{\alpha-1}(1-u_2)^{\beta-1}}_{f_{U_2}(u_2)} = f_{U_1}(u_1) f_{U_2}(u_2). \end{aligned}$$

Therefore, we have shown that $U_1 \sim \text{gamma}(\alpha+\beta, 1)$, $U_2 \sim \text{beta}(\alpha, \beta)$, and U_1 and U_2 are independent.

Example 6.17. Suppose $Y_1 \sim$ exponential(1), $Y_2 \sim$ exponential(1), and Y_1 and Y_2 are independent. The joint pdf of $\mathbf{Y} = (Y_1, Y_2)$ is therefore

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} e^{-(y_1+y_2)}, & y_1 > 0,\ y_2 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

In Example 6.5, we used the cdf technique to (rather painstakingly) show

$$U_1 = Y_1 + Y_2 \sim \text{gamma}(2, 1).$$

This is much easier to show by using the mgf technique; note that for $t < 1$, we have

$$m_{U_1}(t) = m_{Y_1}(t) m_{Y_2}(t) = \frac{1}{1-t} \times \frac{1}{1-t} = \left(\frac{1}{1-t}\right)^2,$$

which is the gamma$(2, 1)$ mgf.

Q: What is the distribution of $U_2 = Y_1 - Y_2$?
A: The random variable U_2 does have a "named distribution," but it is much less well known. One thing we might try to do first is to derive the moment generating function of U_2. From first principles,

$$\begin{aligned} m_{U_2}(t) = E(e^{tU_2}) = E[e^{t(Y_1-Y_2)}] &= E(e^{tY_1} e^{-tY_2}) \\ &\overset{Y_1 \perp\!\!\!\perp Y_2}{=} E(e^{tY_1}) E(e^{-tY_2}) \\ &= m_{Y_1}(t) m_{Y_2}(-t) = \frac{1}{1-t} \times \frac{1}{1+t} = \frac{1}{1-t^2}. \end{aligned}$$

Note that this mgf is valid for $t < 1$ and $-t < 1 \Longleftrightarrow -1 < t < 1$. This is the mgf of a **double exponential** random variable; the pdf of U_2 is

$$f_{U_2}(u_2) = \begin{cases} \frac{1}{2} e^{-|u_2|}, & -\infty < u_2 < \infty \\ 0, & \text{otherwise.} \end{cases}$$

The double exponential distribution is also known as the **LaPlace distribution**. The pdf of U_2 is shown in Figure 6.12 (see next page).

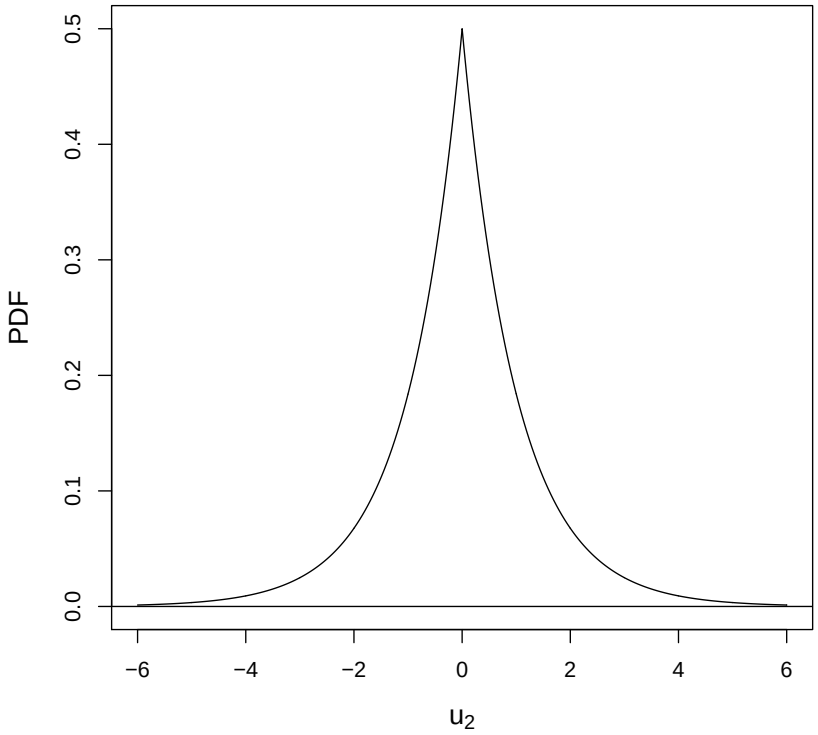

Figure 6.12: The standard double exponential (LaPlace) pdf.

Derivation: Starting with the joint pdf

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} e^{-(y_1+y_2)}, & y_1 > 0,\ y_2 > 0 \\ 0, & \text{otherwise,} \end{cases}$$

let's derive the pdf of $U_2 = Y_1 - Y_2$ by using a bivariate transformation. Of course, we cannot perform a bivariate transformation with only 1 random variable, so let's use

$$\begin{aligned} U_1 &= h_1(Y_1,Y_2) = Y_1 + Y_2 \\ U_2 &= h_2(Y_1,Y_2) = Y_1 - Y_2. \end{aligned}$$

Our strategy will be to use a bivariate transformation to derive the joint pdf $f_{U_1,U_2}(u_1,u_2)$. We will then integrate $f_{U_1,U_2}(u_1,u_2)$ over u_1 to derive the (marginal) pdf of U_2.

The bivariate support of $\mathbf{Y} = (Y_1,Y_2)$ is $R_{Y_1,Y_2} = \{(y_1,y_2) : y_1 > 0,\ y_2 > 0\}$, the entire first quadrant; see Figure 6.13 (left) on the next page. We need to determine the support of $\mathbf{U} = (U_1,U_2)$. Clearly,

$$y_1 > 0,\ y_2 > 0 \implies u_1 = y_1 + y_2 > 0.$$

In addition,

$$u_2 = y_1 - y_2 < y_1 + y_2 = u_1 \implies u_2 < u_1.$$

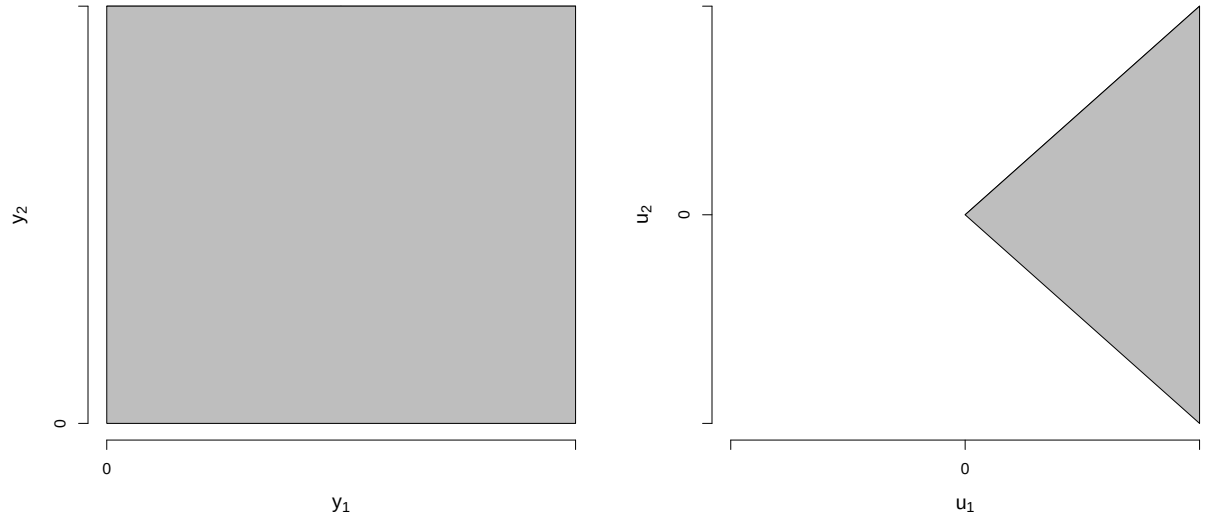

Figure 6.13: Left: The support $R_{Y_1,Y_2} = \{(y_1, y_2) : y_1 > 0,\ y_2 > 0\}$ in Example 6.17. Right: $R_{U_1,U_2} = \{(u_1, u_2) : u_1 > 0,\ -u_1 < u_2 < u_1\}$. The upper boundary line is $u_2 = u_1$; the lower is $u_2 = -u_1$.

Finally,

$$u_2 = y_1 - y_2 > -y_1 - y_2 = -u_1 \implies -u_1 < u_2.$$

Therefore, the support of $\mathbf{U} = (U_1, U_2)$ is

$$R_{U_1,U_2} = \{(u_1, u_2) : u_1 > 0,\ -u_1 < u_2 < u_1\}.$$

This set is shown above in Figure 6.13 (right). The joint pdf $f_{U_1,U_2}(u_1, u_2)$, which we are about to derive, is nonzero over this region.

We next have to verify the transformation defined by

$$\begin{aligned} u_1 &= h_1(y_1, y_2) = y_1 + y_2 \\ u_2 &= h_2(y_1, y_2) = y_1 - y_2 \end{aligned}$$

is one-to-one. Note that this is a **linear transformation**; i.e., we have

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} y_1 + y_2 \\ y_1 - y_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \mathbf{A}\mathbf{y},$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

We know this (linear) transformation is one-to-one because $\mathbf{A}^{-1}$ exists; e.g., $\det(\mathbf{A}) \neq 0$, the columns of $\mathbf{A}$ are linearly independent, $\text{rank}(\mathbf{A}) = 2$; i.e., $\mathbf{A}$ is full rank, etc.

Because the transformation

$$\begin{aligned} u_1 &= h_1(y_1, y_2) = y_1 + y_2 \\ u_2 &= h_2(y_1, y_2) = y_1 - y_2 \end{aligned}$$

is one-to-one, we can find the (unique) inverse transformation. Straightforward algebra shows

$$\begin{aligned} y_1 &= h_1^{-1}(u_1, u_2) = \frac{u_1 + u_2}{2} \\ y_2 &= h_2^{-1}(u_1, u_2) = \frac{u_1 - u_2}{2}. \end{aligned}$$

The Jacobian is

$$J = \det \begin{vmatrix} \underline{\frac{\partial h_1^{-1}(u_1,u_2)}{\partial u_1}} & \underline{\frac{\partial h_1^{-1}(u_1,u_2)}{\partial u_2}} \\ \underline{\frac{\partial h_2^{-1}(u_1,u_2)}{\partial u_1}} & \underline{\frac{\partial h_2^{-1}(u_1,u_2)}{\partial u_2}} \end{vmatrix} = \det \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = \frac{1}{2}\left(-\frac{1}{2}\right) - \frac{1}{2}\left(\frac{1}{2}\right) = -\frac{1}{2}.$$

Therefore, the joint pdf of $\mathbf{U} = (U_1, U_2)$, where nonzero, is

$$\begin{aligned} f_{U_1,U_2}(u_1, u_2) &= f_{Y_1,Y_2}(h_1^{-1}(u_1, u_2), h_2^{-1}(u_1, u_2))|J| \\ &= f_{Y_1,Y_2}\left(\frac{u_1 + u_2}{2}, \frac{u_1 - u_2}{2}\right)\left|-\frac{1}{2}\right| \\ &= \frac{1}{2}\exp\left[-\left(\frac{u_1 + u_2}{2} + \frac{u_1 - u_2}{2}\right)\right] = \frac{1}{2}e^{-u_1}. \end{aligned}$$

Summarizing,

$$f_{U_1,U_2}(u_1, u_2) = \begin{cases} \frac{1}{2}e^{-u_1}, & u_1 > 0, -u_1 < u_2 < u_1 \\ 0, & \text{otherwise.} \end{cases}$$

Finally, to derive the $f_{U_2}(u_2)$, the marginal pdf of $U_2 = Y_2 - Y_1$, we integrate the joint pdf $f_{U_1,U_2}(u_1, u_2)$ over u_1. From Figure 6.13 (right; last page), we should quickly see that how we integrate over u_1 depends on whether $u_2 < 0$ or $u_2 \geq 0$.

Case 1: $u_2 < 0$. The marginal pdf of U_2 is

$$f_{U_2}(u_2) \stackrel{u_2<0}{=} \int_{u_1=-u_2}^{\infty} \frac{1}{2}e^{-u_1}du_1 = \frac{1}{2}\left(-e^{-u_1}\Big|_{u_1=-u_2}^{\infty}\right) = \frac{1}{2}e^{u_2}.$$

Case 2: $u_2 \geq 0$. The marginal pdf of U_2 is

$$f_{U_2}(u_2) \stackrel{u_2\geq 0}{=} \int_{u_1=u_2}^{\infty} \frac{1}{2}e^{-u_1}du_1 = \frac{1}{2}\left(-e^{-u_1}\Big|_{u_1=u_2}^{\infty}\right) = \frac{1}{2}e^{-u_2}.$$

Combining both cases, we have

$$f_{U_2}(u_2) = \begin{cases} \frac{1}{2}e^{-|u_2|}, & -\infty < u_2 < \infty \\ 0, & \text{otherwise.} \end{cases}$$

We have shown the difference of two independent exponential(1) random variables follows a standard double exponential (LaPlace) distribution.

Q: How could we find the mean and variance of U_2?
A: To find $E(U_2)$, we could calculate

$$E(U_2) = \int_{\mathbb{R}} u_2 f_{U_2}(u_2) du_2 = \int_{-\infty}^{\infty} \frac{u_2}{2} e^{-|u_2|} du_2.$$

We could then calculate

$$E(U_2^2) = \int_{\mathbb{R}} u_2^2 f_{U_2}(u_2) du_2 = \int_{-\infty}^{\infty} \frac{u_2^2}{2} e^{-|u_2|} du_2$$

and get $V(U_2)$ using the variance computing formula. Alternatively, we could recall the mgf of U_2 is

$$m_{U_2}(t) = \frac{1}{1-t^2}, \quad \text{for } -1 < t < 1,$$

and get the moments of U_2 from $m_{U_2}(t)$. Ultimately, it is easiest to use the Law of the Unconscious Statistician. Because Y_1 and Y_2 are independent exponential(1) random variables, we have

$$E(U_2) = E(Y_1 - Y_2) = E(Y_1) - E(Y_2) = 1 - 1 = 0$$

and

$$V(U_2) = V(Y_1 - Y_2) \stackrel{Y_1 \perp\!\!\!\perp Y_2}{=} V(Y_1) + V(Y_2) = 1 + 1 = 2. \ \square$$

Remark: We now generalize the exercise of performing bivariate transformations (in two dimensions) to that of performing transformations in higher dimensions.

Setting: Suppose that $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ is a continuous random vector with joint pdf $f_{\mathbf{Y}}(y_1, y_2, ..., y_n)$, which is nonzero over the support $R_{\mathbf{Y}} \subset \mathbb{R}^n$. Define

$$\begin{aligned} U_1 &= h_1(Y_1, Y_2, ..., Y_n) \\ U_2 &= h_2(Y_1, Y_2, ..., Y_n) \\ &\vdots \\ U_n &= h_n(Y_1, Y_2, ..., Y_n). \end{aligned}$$

Assume this is a one-to-one transformation from $R_{\mathbf{Y}}$ to

$$R_{\mathbf{U}} = \{(u_1, u_2, ..., u_n) : u_i = h_i(y_1, y_2, ..., y_n), i = 1, 2, ..., n, \text{ for } (y_1, y_2, ..., y_n) \in R_{\mathbf{Y}}\},$$

the support of $\mathbf{U} = (U_1, U_2, ..., U_n)$. Because the transformation is one-to-one, the inverse transformation exists and is

$$\begin{aligned} y_1 &= h_1^{-1}(u_1, u_2, ..., u_n) \\ y_2 &= h_2^{-1}(u_1, u_2, ..., u_n) \\ &\vdots \\ y_n &= h_n^{-1}(u_1, u_2, ..., u_n). \end{aligned}$$

With $\mathbf{u} = (u_1, u_2, ..., u_n)$, the Jacobian of the inverse transformation is

$$J = \det \begin{pmatrix} \frac{\partial h_1^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial h_1^{-1}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h_1^{-1}(\mathbf{u})}{\partial u_n} \\ \frac{\partial h_2^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial h_2^{-1}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h_2^{-1}(\mathbf{u})}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_n^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial h_n^{-1}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h_n^{-1}(\mathbf{u})}{\partial u_n} \end{pmatrix};$$

i.e., J is the determinant of this $n \times n$ matrix of partial derivatives. The pdf of $\mathbf{U} = (U_1, U_2, ..., U_n)$, where nonzero, is given by

$$f_{\mathbf{U}}(u_1, u_2, ..., u_n) = f_{\mathbf{Y}}(h_1^{-1}(u_1, u_2, ..., u_n), h_2^{-1}(u_1, u_2, ..., u_n), ..., h_n^{-1}(u_1, u_2, ..., u_n))|J|.$$

This result generalizes our discussion on bivariate transformations (in two dimensions).

Example 6.18. Suppose $\mathbf{Y} = (Y_1, Y_2, Y_3)$ is a continuous random vector with joint pdf

$$f_{\mathbf{Y}}(y_1, y_2, y_3) = \begin{cases} 48y_1y_2y_3, & 0 < y_1 < y_2 < y_3 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Define

$$\begin{aligned} U_1 &= h_1(Y_1, Y_2, Y_3) = \frac{Y_1}{Y_2} \\ U_2 &= h_2(Y_1, Y_2, Y_3) = \frac{Y_2}{Y_3} \\ U_3 &= h_3(Y_1, Y_2, Y_3) = Y_3. \end{aligned}$$

Perform a trivariate transformation to derive the joint pdf of $\mathbf{U} = (U_1, U_2, U_3)$.

Solution. Note that the support of $\mathbf{Y} = (Y_1, Y_2, Y_3)$ is

$$R_{\mathbf{Y}} = \{(y_1, y_2, y_3) : 0 < y_1 < y_2 < y_3 < 1\},$$

the upper orthant of the unit cube in $\mathbb{R}^3$. The support of $\mathbf{U}$ is

$$R_{\mathbf{U}} = \{(u_1, u_2, u_3) : 0 < u_1 < 1,\ 0 < u_2 < 1,\ 0 < u_3 < 1\},$$

the entire unit cube. It is easy to show the transformation defined by

$$\begin{aligned} u_1 &= h_1(y_1, y_2, y_3) = \frac{y_1}{y_2} \\ u_2 &= h_2(y_1, y_2, y_3) = \frac{y_2}{y_3} \\ u_3 &= h_3(y_1, y_2, y_3) = y_3 \end{aligned}$$

is one-to-one. Suppose $h_i(y_1, y_2, y_3) = h_i(y_1^*, y_2^*, y_3^*)$, for $i = 1, 2, 3$. The third equation implies $y_3 = y_3^*$. The second equation implies $y_2 = y_2^*$. The first equation implies $y_1 = y_1^*$.

Because the transformation is one-to-one, the inverse transformation exists and is given by

$$\begin{aligned} y_1 &= h_1^{-1}(u_1, u_2, u_3) = u_1u_2u_3 \\ y_2 &= h_2^{-1}(u_1, u_2, u_3) = u_2u_3 \\ y_3 &= h_3^{-1}(u_1, u_2, u_3) = u_3. \end{aligned}$$

With $\mathbf{u} = (u_1, u_2, u_3)$, the Jacobian of the inverse transformation is

$$J = \det \begin{pmatrix} \dfrac{\partial h_1^{-1}(\mathbf{u})}{\partial u_1} & \dfrac{\partial h_1^{-1}(\mathbf{u})}{\partial u_2} & \dfrac{\partial h_1^{-1}(\mathbf{u})}{\partial u_3} \\ \dfrac{\partial h_2^{-1}(\mathbf{u})}{\partial u_1} & \dfrac{\partial h_2^{-1}(\mathbf{u})}{\partial u_2} & \dfrac{\partial h_2^{-1}(\mathbf{u})}{\partial u_3} \\ \dfrac{\partial h_3^{-1}(\mathbf{u})}{\partial u_1} & \dfrac{\partial h_3^{-1}(\mathbf{u})}{\partial u_2} & \dfrac{\partial h_3^{-1}(\mathbf{u})}{\partial u_3} \end{pmatrix} = \det \begin{pmatrix} u_2u_3 & u_1u_3 & u_1u_2 \\ 0 & u_3 & u_2 \\ 0 & 0 & 1 \end{pmatrix} = u_2u_3^2.$$

Therefore, the joint pdf of $\mathbf{U} = (U_1, U_2, U_3)$, where nonzero, is given by

$$\begin{aligned} f_{\mathbf{U}}(u_1, u_2, u_3) &= f_{\mathbf{Y}}(h_1^{-1}(u_1, u_2, u_3), h_2^{-1}(u_1, u_2, u_3), h_3^{-1}(u_1, u_2, u_3))|J| \\ &= 48(u_1u_2u_3)(u_2u_3)(u_3) \times u_2u_3^2 \\ &= 48u_1u_2^3u_3^5. \end{aligned}$$

Summarizing,

$$f_{\mathbf{U}}(u_1, u_2, u_3) = \begin{cases} 48u_1u_2^3u_3^5, & 0 < u_1 < 1,\ 0 < u_2 < 1,\ 0 < u_3 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Note: We are done performing the trivariate transformation. We started with the joint pdf of $\mathbf{Y} = (Y_1, Y_2, Y_3)$, and now we have the joint pdf of $\mathbf{U} = (U_1, U_2, U_3)$.

Note: We make additional observations in Example 6.18. Note that we can write

$$\begin{aligned} f_{U_1,U_2,U_3}(u_1, u_2, u_3) &= 48u_1u_2^3u_3^5 \\ &= 2u_1 \times 4u_2^3 \times 6u_3^5 \ = \ f_{U_1}(u_1)f_{U_2}(u_2)f_{U_3}(u_3), \end{aligned}$$

where the marginal pdfs are

$$f_{U_1}(u_1) = \begin{cases} 2u_1, & 0 < u_1 < 1 \\ 0, & \text{otherwise,} \end{cases} \qquad f_{U_2}(u_2) = \begin{cases} 4u_2^3, & 0 < u_2 < 1 \\ 0, & \text{otherwise,} \end{cases}$$

and

$$f_{U_3}(u_3) = \begin{cases} 6u_3^5, & 0 < u_3 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, $U_1 \sim \text{beta}(2, 1)$, $U_2 \sim \text{beta}(4, 1)$, $U_3 \sim \text{beta}(6, 1)$, and U_1, U_2, and U_3 are mutually independent. $\square$

6.6 Order statistics

Remark: We encounter order statistics on a daily basis. Phrases like "minimum temperature," "maximum tolerable dose," "highest test score," "lowest barometric pressure," and "median salary" all refer to order statistics. Mathematically, order statistics are the ordered values of random variables.

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ are random variables. The **order statistics** are the ordered values of $Y_1, Y_2, ..., Y_n$; i.e.,

$$\begin{aligned} Y_{(1)} &= \min\{Y_1, Y_2, ..., Y_n\} \\ Y_{(2)} &= \text{second smallest of } Y_1, Y_2, ..., Y_n \\ &\vdots \\ Y_{(n)} &= \max\{Y_1, Y_2, ..., Y_n\}. \end{aligned}$$

Therefore, order statistics satisfy $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$.

Example 6.19. Non-small cell lung cancer (NSCLC) is the most common type of lung cancer in humans (roughly 85% of all cases). A recent study in *Japanese Journal of Clinical Oncology* examined a small group of NSCLC patients who had been treated with both gefitinib and erlotinib (two cancer drugs). Here were the times until treatment failure (TTF, in months) for $n = 14$ patients:

0.8 7.5 13.4 1.4 0.5 68.9 16.1 20.4 15.6 4.2 2.4 8.2 5.3 14.0

The authors described how "treatment failure" could mean disease progression to a higher stage, withdrawal from treatment due to adverse reaction, or death. Here are the order statistics:

0.5 0.8 1.4 2.4 4.2 5.3 7.5 8.2 13.4 14.0 15.6 16.1 20.4 68.9

In this example, we see $y_1 = 0.8 = y_{(2)}$, $y_2 = 7.5 = y_{(7)}$, and so on. The minimum and maximum order statistics are

$$y_{(1)} = 0.5 \quad \text{and} \quad y_{(14)} = 68.9,$$

respectively. Many other familiar quantities are either order statistics themselves or functions of order statistics. For example, the **median** of these observations is

$$m = \frac{y_{(7)} + y_{(8)}}{2} = \frac{7.5 + 8.2}{2} = 7.85.$$

The **range** is $r = y_{(14)} - y_{(1)} = 68.9 - 0.5 = 68.4$ and the **interquartile range** is IQR $= y_{(11)} - y_{(4)} = 15.6 - 2.4 = 13.2$. □

Setting: Suppose $Y_1, Y_2, ..., Y_n$ are mutually independent continuous random variables, with common cdf $F_Y(y)$ and pdf $f_Y(y)$. Our goal is to derive the distributions of the order statistics $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$.

Preview: Most of the time, but certainly not always, we will be interested in minimums and maximums. We will show the pdf of the **minimum order statistic** $Y_{(1)}$, where nonzero, is

$$f_{Y_{(1)}}(y) = nf_Y(y)[1 - F_Y(y)]^{n-1}.$$

We will also show the pdf of the **maximum order statistic** $Y_{(n)}$, where nonzero, is

$$f_{Y_{(n)}}(y) = nf_Y(y)[F_Y(y)]^{n-1}.$$

These formulas are important and should be committed to memory. In general, the pdf of the kth order statistic $Y_{(k)}$, where nonzero, is

$$f_{Y_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!}\,[F_Y(y)]^{k-1} f_Y(y)[1 - F_Y(y)]^{n-k}.$$

Note that when $k = 1$ (minimum), this formula reduces to the one for $f_{Y_{(1)}}(y)$. Similarly, when $k = n$ (maximum), this formula reduces to the one for $f_{Y_{(n)}}(y)$.

Remark: It is important to discuss the assumptions stated above. We are assuming the random variables $Y_1, Y_2, ..., Y_n$

- are continuous
- are mutually independent
- all have the same probability distribution described by $F_Y(y)$ and $f_Y(y)$.

It makes perfect sense to think about order statistics when $Y_1, Y_2, ..., Y_n$ are discrete; however, allowing for the possibility of "ties" among the observations makes the mathematics more difficult. Note that when $Y_1, Y_2, ..., Y_n$ are truly continuous, then, theoretically, ties are not possible; i.e.,

$$Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$$

with probability one. The second and third assumptions also simplify the mathematics, enabling us to get the closed-form expressions above.

Important: Going forward, when random variables $Y_1, Y_2, ..., Y_n$ are (a) mutually independent and (b) have the same or "identical" distribution, we will streamline this by saying

"$Y_1, Y_2, ..., Y_n$ are **i**ndependent and **i**dentically **d**istributed."

This can be streamlined further by saying "$Y_1, Y_2, ..., Y_n$ are **iid**." We usually describe the common distribution of $Y_1, Y_2, ..., Y_n$ by referencing the pdf $f_Y(y)$, the cdf $F_Y(y)$, or even the mgf $m_Y(t)$.

Derivations: Suppose $Y_1, Y_2, ..., Y_n$ are iid with cdf $F_Y(y)$ and pdf $f_Y(y)$. The cdf of the minimum order statistic $Y_{(1)}$ is

$$\begin{aligned} F_{Y_{(1)}}(y) = P(Y_{(1)} \leq y) &= 1 - P(Y_{(1)} > y) \\ &= 1 - P(Y_1 > y, Y_2 > y, ..., Y_n > y) \\ &= 1 - P(Y_1 > y)P(Y_2 > y) \cdots P(Y_n > y) \\ &= 1 - [P(Y > y)]^n \\ &= 1 - [1 - P(Y \leq y)]^n \\ &= 1 - [1 - F_Y(y)]^n. \end{aligned}$$

Therefore, the pdf of $Y_{(1)}$, where nonzero, is given by

$$\begin{aligned} f_{Y_{(1)}}(y) &= \frac{d}{dy} F_{Y_{(1)}}(y) \\ &= \frac{d}{dy} \{1 - [1 - F_Y(y)]^n\} \\ &= -n[1 - F_Y(y)]^{n-1}[-f_Y(y)] = nf_Y(y)[1 - F_Y(y)]^{n-1}. \end{aligned}$$

This is our closed-form expression for the pdf of the minimum order statistic. Now, the maximum. The cdf of the maximum order statistic $Y_{(n)}$ is

$$\begin{aligned} F_{Y_{(n)}}(y) = P(Y_{(n)} \leq y) &= P(Y_1 \leq y, Y_2 \leq y, ..., Y_n \leq y) \\ &= P(Y_1 \leq y)P(Y_2 \leq y) \cdots P(Y_n \leq y) \\ &= [P(Y \leq y)]^n \\ &= [F_Y(y)]^n. \end{aligned}$$

Therefore, the pdf of $Y_{(n)}$, where nonzero, is given by

$$\begin{aligned} f_{Y_{(n)}}(y) &= \frac{d}{dy} F_{Y_{(n)}}(y) \\ &= \frac{d}{dy} \{[F_Y(y)]^n\} = nf_Y(y)[F_Y(y)]^{n-1}. \end{aligned}$$

This is our closed-form expression for the pdf of the maximum order statistic.

Example 6.20. Suppose $Y_1, Y_2, ..., Y_n$ are iid exponential with mean $\beta > 0$. Recall the exponential(β) pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\beta} e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise} \end{cases}$$

and the exponential(β) cdf is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y/\beta}, & y > 0. \end{cases}$$

In this example, we will find the pdf of $Y_{(1)}$, the minimum order statistic, and the pdf of $Y_{(n)}$, the maximum order statistic.

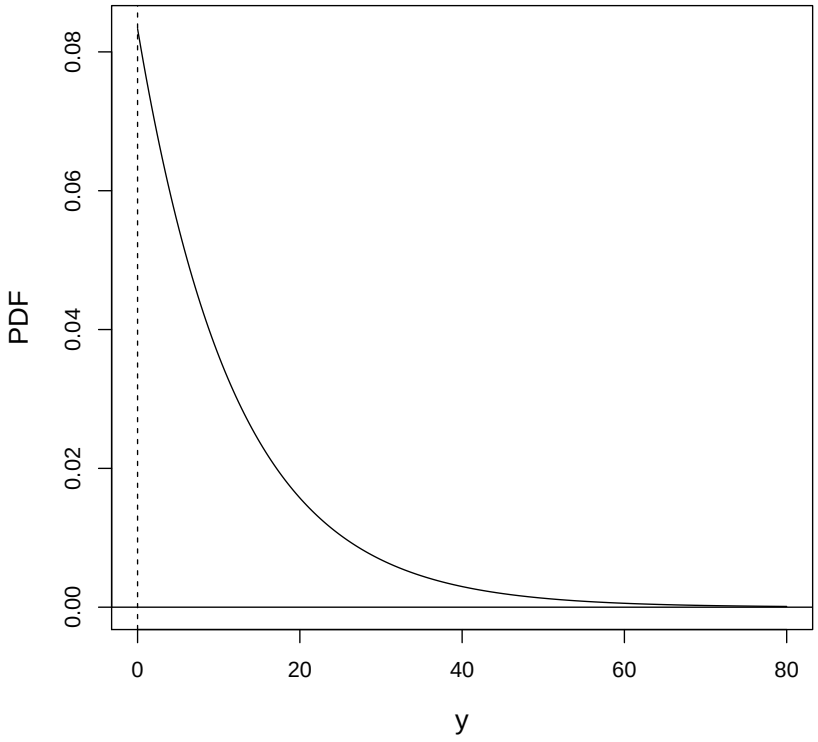

Figure 6.14: Exponential pdf with mean $\beta = 12$.

Solutions. The pdf of $Y_{(1)}$, for $y > 0$, is given by

$$\begin{aligned} f_{Y_{(1)}}(y) = nf_Y(y)[1 - F_Y(y)]^{n-1} &= n\left(\frac{1}{\beta}e^{-y/\beta}\right)[1 - (1 - e^{-y/\beta})]^{n-1} \\ &= \frac{n}{\beta}e^{-y/\beta}(e^{-y/\beta})^{n-1} \\ &= \frac{n}{\beta}(e^{-y/\beta})^n = \frac{n}{\beta}e^{-ny/\beta}. \end{aligned}$$

Summarizing,

$$f_{Y_{(1)}}(y) = \begin{cases} \frac{n}{\beta}e^{-ny/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Note that the nonzero part of this pdf can be written as

$$\frac{n}{\beta}e^{-ny/\beta} = \frac{1}{\left(\frac{\beta}{n}\right)}e^{-y/\left(\frac{\beta}{n}\right)},$$

which we recognize as an exponential pdf with mean β/n. Therefore,

$$Y_1, Y_2, ..., Y_n \sim \text{iid exponential}(\beta) \implies Y_{(1)} \sim \text{exponential}(\beta/n).$$

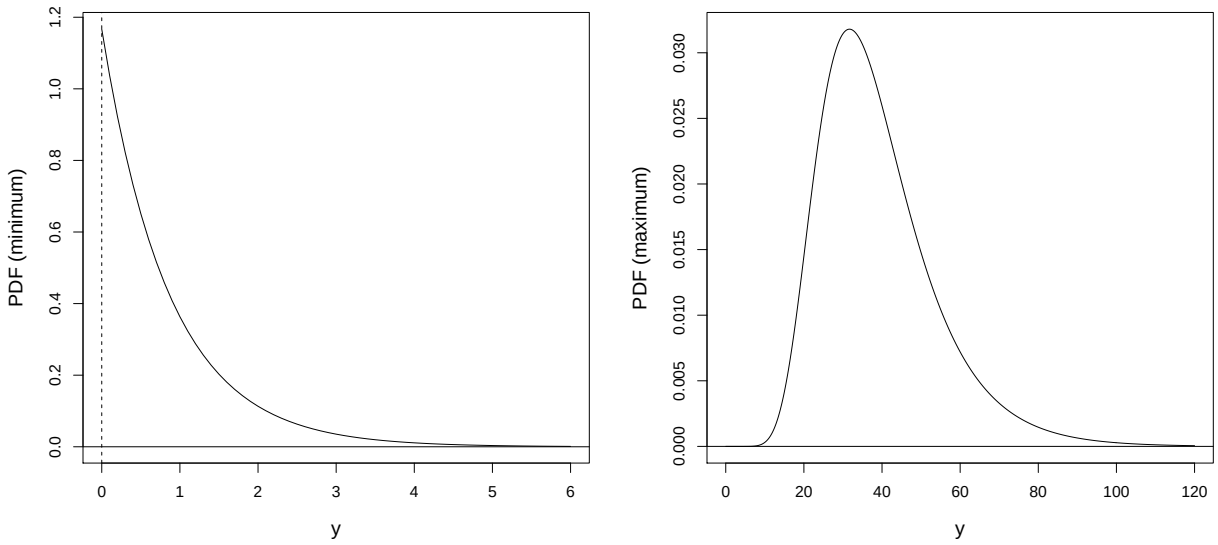

Figure 6.15: Left: Pdf of the minimum order statistic $Y_{(1)}$ in Example 6.19 assuming an exponential distribution for TTF with mean $\beta = 12$ months. Right: Pdf of $Y_{(14)}$, the maximum order statistic. Note that the horizontal axes are different in the two figures.

The pdf of $Y_{(n)}$, for $y > 0$, is given by

$$f_{Y_{(n)}}(y) = nf_Y(y)[F_Y(y)]^{n-1} = n\left(\frac{1}{\beta}e^{-y/\beta}\right)(1-e^{-y/\beta})^{n-1} = \frac{n}{\beta}e^{-y/\beta}(1-e^{-y/\beta})^{n-1}.$$

Summarizing,

$$f_{Y_{(n)}}(y) = \begin{cases} \frac{n}{\beta}e^{-y/\beta}(1-e^{-y/\beta})^{n-1}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is not one of a "named" distribution, but it is a valid pdf nonetheless.

Application: Suppose the time until treatment failure (TTF) for the $n = 14$ cancer patients in Example 6.19 follows an exponential distribution with mean $\beta = 12$ months. This pdf is shown in Figure 6.14 (see previous page). Under this assumption, the pdf of the minimum and maximum order statistics are given by

$$f_{Y_{(1)}}(y) = \begin{cases} \frac{14}{12}e^{-14y/12}, & y > 0 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_{(14)}}(y) = \begin{cases} \frac{14}{12}e^{-y/12}(1-e^{-y/12})^{13}, & y > 0 \\ 0, & \text{otherwise,} \end{cases}$$

respectively. These pdfs are shown in Figure 6.15 above. □

Derivation: Suppose $Y_1, Y_2, ..., Y_n$ are iid with cdf $F_Y(y)$ and pdf $f_Y(y)$. We now argue the pdf of the kth order statistic $Y_{(k)}$, where nonzero, is

$$f_{Y_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!}\,[F_Y(y)]^{k-1} f_Y(y)[1-F_Y(y)]^{n-k}.$$

Our argument is heuristic. Think of each of $Y_1, Y_2, ..., Y_n$ as a "trial," and consider the trinomial distribution with the following categories:

Category	Description	Probability	# Observations
1	Less than y	$p_1 = P(Y < y) = F_Y(y)$	$k-1$
2	Equal to y	$p_2 = P(Y = y) =$ "$f_Y(y)$"	1
3	Greater than y	$p_3 = P(Y > y) = 1 - F_Y(y)$	$n-k$

Now, utilize the trinomial pmf with these category probabilities and counts (# observations) to get

$$\frac{n!}{(k-1)!\;1!\;(n-k)!}\,p_1^{k-1}p_2^1\,p_3^{n-k} \;=\; \frac{n!}{(k-1)!(n-k)!}\,[F_Y(y)]^{k-1} f_Y(y)[1-F_Y(y)]^{n-k}.$$

Example 6.21. Suppose $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{U}(0,1)$; i.e., a uniform distribution from 0 to 1. Recall the $\mathcal{U}(0,1)$ pdf is given by

$$f_Y(y) = \begin{cases} 1, & 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

and the $\mathcal{U}(0,1)$ cdf is

$$F_Y(y) = \begin{cases} 0, & y \le 0 \\ y, & 0 < y < 1 \\ 1, & y \ge 1. \end{cases}$$

Find the pdf of $Y_{(k)}$, the kth order statistic.

Solution. The pdf of $Y_{(k)}$, for $0 < y < 1$, is given by

$$\begin{aligned} f_{Y_{(k)}}(y) &= \frac{n!}{(k-1)!(n-k)!}\,[F_Y(y)]^{k-1} f_Y(y)[1-F_Y(y)]^{n-k} \\ &= \frac{n!}{(k-1)!(n-k)!}\,y^{k-1}\,(1)\,(1-y)^{n-k}. \end{aligned}$$

Writing $n! = \Gamma(n+1)$, $(k-1)! = \Gamma(k)$ and $(n-k)! = \Gamma(n-k+1)$, note that

$$f_{Y_{(k)}}(y) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\,y^{k-1}(1-y)^{(n-k+1)-1},$$

for $0 < y < 1$. We recognize this as a beta pdf with parameters $\alpha = k$ and $\beta = n-k+1$. Therefore, $Y_{(k)} \sim \text{beta}(k, n-k+1)$. See Figure 6.16 (next page). $\square$

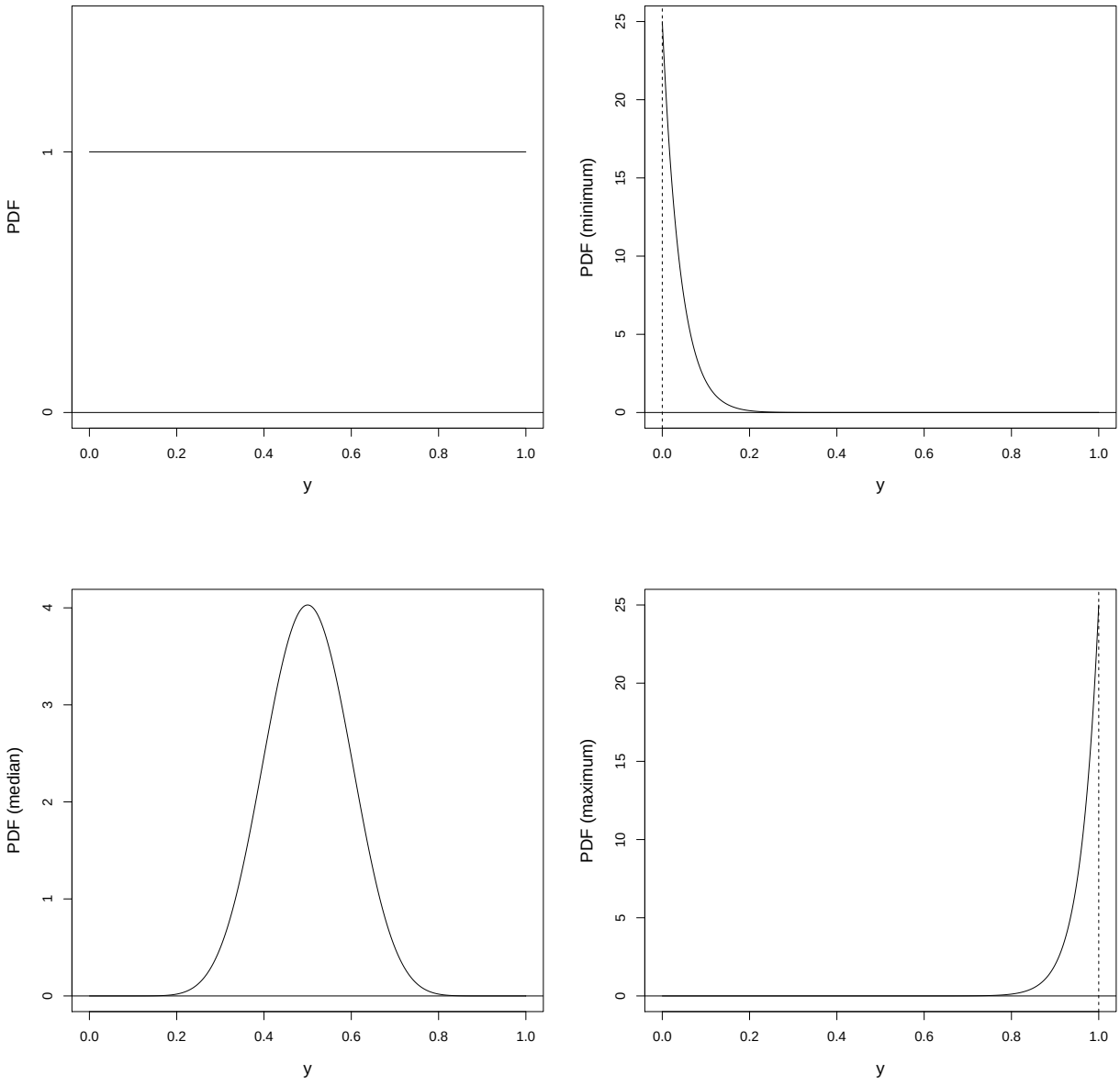

Figure 6.16: Upper left: $\mathcal{U}(0,1)$ pdf in Example 6.21. All order statistic distributions in the remaining figures are based on $n = 25$ observations from the $\mathcal{U}(0,1)$ distribution. Upper right: Pdf of the minimum order statistic $Y_{(1)} \sim \text{beta}(1, 25)$. Lower left: Pdf of the median $Y_{(13)} \sim \text{beta}(13, 13)$. Lower right: Pdf of the maximum order statistic $Y_{(25)} \sim \text{beta}(25, 1)$.

Note: Many common statistics are functions of two order statistics; e.g., the range, the interquartile range, etc. We now describe the **joint pdf** of any two order statistics.

Result: Suppose $Y_1, Y_2, ..., Y_n$ are iid with cdf $F_Y(y)$ and pdf $f_Y(y)$. The joint pdf of $Y_{(j)}$ and $Y_{(k)}$, $1 \leq j < k \leq n$, where nonzero, is given by

$$f_{Y_{(j)},Y_{(k)}}(y_j, y_k) = \frac{n!}{(j-1)!(k-1-j)!(n-k)!} [F_Y(y_j)]^{j-1} f_Y(y_j)[F_Y(y_k) - F_Y(y_j)]^{k-1-j} \times f_Y(y_k)[1 - F_Y(y_k)]^{n-k}.$$

Analogous to before (when finding the marginal pdf of $Y_{(k)}$), we can see where this formula comes from by appealing to the multinomial distribution with the following categories and probabilities:

Category	Description	Probability	# Observations
1	Less than y_j	$p_1 = F_Y(y_j)$	$j-1$
2	Equal to y_j	$p_2 =$ "$f_Y(y_j)$"	1
3	Between y_j and y_k	$p_3 = F_Y(y_k) - F_Y(y_j)$	$k-1-j$
4	Equal to y_k	$p_4 =$ "$f_Y(y_k)$"	1
5	Greater than y_k	$p_5 = 1 - F_Y(y_k)$	$n-k$

Note that the multinomial coefficient

$$\frac{n!}{(j-1)!\ 1!\ (k-1-j)!\ 1!\ (n-k)!} = \frac{n!}{(j-1)!(k-1-j)!(n-k)!}$$

counts the number of ways the n observations $Y_1, Y_2, ..., Y_n$ ("trials") can fall into these 5 categories.

Special case: The formula for $f_{Y_{(j)},Y_{(k)}}(y_j, y_k)$ above simplifies substantially when $j = 1$ (minimum) and $k = n$ (maximum). The joint pdf of the minimum and maximum order statistics, $Y_{(1)}$ and $Y_{(n)}$, where nonzero, is

$$f_{Y_{(1)},Y_{(n)}}(y_1, y_n) = n(n-1)f_Y(y_1)[F_Y(y_n) - F_Y(y_1)]^{n-2} f_Y(y_n).$$

Example 6.22. Suppose $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{U}(0,1)$. Recall the $\mathcal{U}(0,1)$ pdf is given by

$$f_Y(y) = \begin{cases} 1, & 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

and the $\mathcal{U}(0,1)$ cdf is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ y, & 0 < y < 1 \\ 1, & y \geq 1. \end{cases}$$

(a) Find the joint pdf of $Y_{(1)}$ and $Y_{(n)}$.
(b) Find the pdf of the range $Y_{(n)} - Y_{(1)}$.

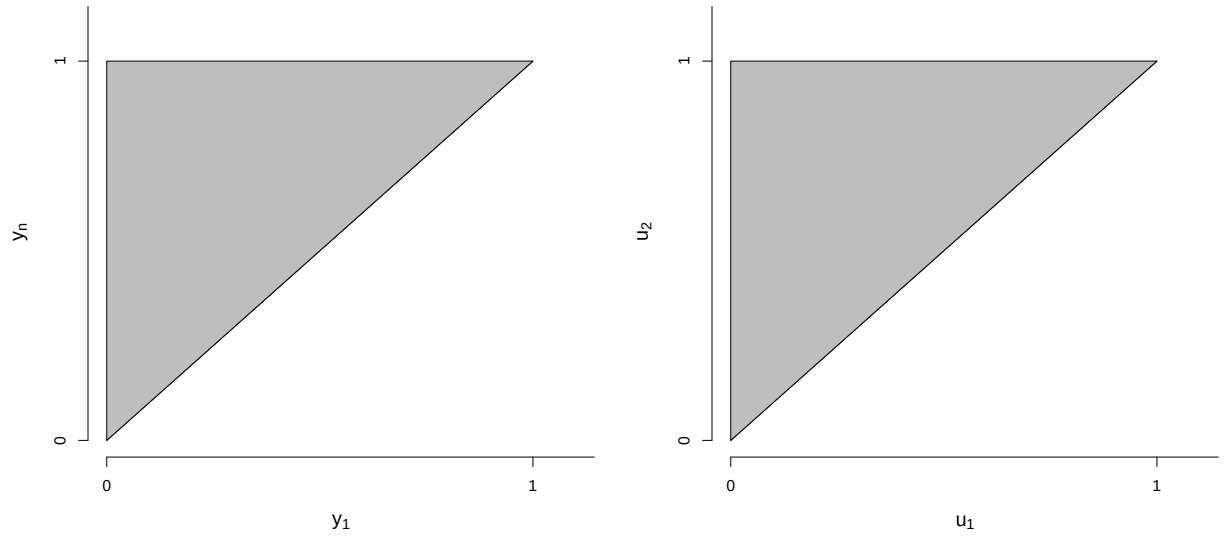

Figure 6.17: Left: The support $R_{Y_{(1)},Y_{(n)}} = \{(y_1, y_n) : 0 < y_1 < y_n < 1\}$ in Example 6.22. Right: $R_{U_1,U_2} = \{(u_1, u_2) : 0 < u_1 < u_2 < 1\}$.

Solutions. (a) Note that the support of $(Y_{(1)}, Y_{(n)})$ is $R_{Y_{(1)},Y_{(n)}} = \{(y_1, y_n) : 0 < y_1 < y_n < 1\}$, which makes sense because the minimum $Y_{(1)}$ cannot be greater than the maximum $Y_{(n)}$. Therefore, for $0 < y_1 < y_n < 1$, the joint pdf of $Y_{(1)}$ and $Y_{(n)}$ is

$$\begin{aligned} f_{Y_{(1)},Y_{(n)}}(y_1, y_n) &= n(n-1)f_Y(y_1)[F_Y(y_n) - F_Y(y_1)]^{n-2} f_Y(y_n) \\ &= n(n-1)(y_n - y_1)^{n-2}. \end{aligned}$$

Summarizing,

$$f_{Y_{(1)},Y_{(n)}}(y_1, y_n) = \begin{cases} n(n-1)(y_n - y_1)^{n-2}, & 0 < y_1 < y_n < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The support of $(Y_{(1)}, Y_{(n)})$ is shown in Figure 6.17 above (left). The joint pdf $f_{Y_{(1)},Y_{(n)}}(y_1, y_n)$ is a three-dimensional function which takes the value $n(n-1)(y_n - y_1)^{n-2}$ over this region (and equals zero, otherwise).

(b) To find the pdf of the range, let's use a bivariate transformation with

$$\begin{aligned} U_1 &= h_1(Y_{(1)}, Y_{(n)}) = Y_{(n)} - Y_{(1)} \\ U_2 &= h_2(Y_{(1)}, Y_{(n)}) = Y_{(n)}. \end{aligned}$$

The formula for the joint pdf $f_{Y_{(1)},Y_{(n)}}(y_1, y_n)$ is above. We will perform the bivariate transformation to derive $f_{U_1,U_2}(u_1, u_2)$, the joint pdf of $\mathbf{U} = (U_1, U_2)$. We will then integrate $f_{U_1,U_2}(u_1, u_2)$ over u_2 to obtain the (marginal) pdf of $U_1 = Y_{(n)} - Y_{(1)}$.

Let's first determine the support of $\mathbf{U} = (U_1, U_2)$. Note that

$$0 < y_{(1)} < y_{(n)} < 1 \implies u_1 = y_{(n)} - y_{(1)} \in (0, 1).$$

Also, $u_2 = y_{(n)} > y_{(n)} - y_{(1)} = u_1$. Therefore, the support of $\mathbf{U} = (U_1, U_2)$ is $R_{U_1,U_2} = \{(u_1, u_2) : 0 < u_1 < u_2 < 1\}$; see Figure 6.17 (previous page; right). We next have to verify the transformation defined by

$$\begin{aligned} u_1 &= h_1(y_{(1)}, y_{(n)}) = y_{(n)} - y_{(1)} \\ u_2 &= h_2(y_{(1)}, y_{(n)}) = y_{(n)}. \end{aligned}$$

is one-to-one. Note that this is a **linear transformation**; i.e., we have

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} y_{(n)} - y_{(1)} \\ y_{(n)} \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_{(1)} \\ y_{(n)} \end{pmatrix} = \mathbf{A} \begin{pmatrix} y_{(1)} \\ y_{(n)} \end{pmatrix},$$

where

$$\mathbf{A} = \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}.$$

We know this (linear) transformation is one-to-one because $\mathbf{A}^{-1}$ exists. Therefore, the inverse transformation exists and is given by

$$\begin{aligned} y_{(1)} &= h_1^{-1}(u_1, u_2) = u_2 - u_1 \\ y_{(n)} &= h_2^{-1}(u_1, u_2) = u_2. \end{aligned}$$

The Jacobian of the (inverse) transformation is

$$J = \det \begin{vmatrix} \dfrac{\partial h_1^{-1}(u_1, u_2)}{\partial u_1} & \dfrac{h_1^{-1}(u_1, u_2)}{\partial u_2} \\ \dfrac{\partial h_2^{-1}(u_1, u_2)}{\partial u_1} & \dfrac{\partial h_2^{-1}(u_1, u_2)}{\partial u_2} \end{vmatrix} = \det \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = -1.$$

Therefore, the joint pdf of $\mathbf{U} = (U_1, U_2)$ is

$$\begin{aligned} f_{U_1,U_2}(u_1, u_2) &= f_{Y_{(1)},Y_{(n)}}(h_1^{-1}(u_1, u_2), h_2^{-1}(u_1, u_2))|J| \\ &= n(n-1)[u_2 - (u_2 - u_1)]^{n-2} = n(n-1)u_1^{n-2}, \end{aligned}$$

for $0 < u_1 < u_2 < 1$; see Figure 6.17 (last page; right). Therefore, the (marginal) pdf of $U_1 = Y_{(n)} - Y_{(1)}$, for $0 < u_1 < 1$, is

$$\begin{aligned} f_{U_1}(u_1) = \int_{\mathbb{R}} f_{U_1,U_2}(u_1, u_2) du_2 &= \int_{u_2=u_1}^{1} n(n-1)u_1^{n-2} du_2 \\ &= n(n-1)u_1^{n-2}(1 - u_1) \\ &= \frac{\Gamma(n+1)}{\Gamma(n-1)\Gamma(2)} u_1^{(n-1)-1}(1 - u_1)^{2-1}, \end{aligned}$$

a beta pdf with parameters $\alpha = n - 1$ and $\beta = 2$; i.e., $U_1 = Y_{(n)} - Y_{(1)} \sim \text{beta}(n - 1, 2)$. $\square$

7 Sampling Distributions and the Central Limit Theorem

7.1 Introduction

Preview: Suppose $Y_1, Y_2, ..., Y_n$ are random variables. In most of the problems we encounter going forward, $Y_1, Y_2, ..., Y_n$ will be regarded as "iid" random variables from some common probability distribution. We will denote this probability distribution by $p_Y(y)$ or $f_Y(y)$, depending on whether the random variables are discrete or continuous, respectively. Recall that the acronym "iid" means "**i**ndependent and **i**dentically **d**istributed." That is, the random variables $Y_1, Y_2, ..., Y_n$

- are mutually independent
- all have the same (or identical) probability distribution described by $p_Y(y)$ or $f_Y(y)$.

Important: In statistical applications, one often envisions $Y_1, Y_2, ..., Y_n$ as being observations on n individuals which have been sampled from a large population of individuals. Under this conceptualization, the common probability distribution $p_Y(y)$ or $f_Y(y)$ is called the **population distribution**. The population distribution describes the distribution of the random variable Y for each individual in the population. In other words, $p_Y(y)$ or $f_Y(y)$ serve as probability models for a population.

Terminology: A **random sample** $Y_1, Y_2, ..., Y_n$ measures the value of Y for a sample of n individuals drawn from the population and is viewed as n iid replicates of the random variable Y. We call n the **sample size**. For our purposes, the phrase "random sample" and "iid sample" will mean the same thing.

Examples:

1. The body mass index Y is measured for $n = 328$ fourth-grade children. Suppose the measurements $Y_1, Y_2, ..., Y_{328}$ are iid from a gamma(α, β) population distribution.
2. The number of days spent in a neonatal intensive care unit Y is observed for $n = 127$ premature infants. Suppose the number of days $Y_1, Y_2, ..., Y_{127}$ are iid from a Poisson(λ) population distribution.
3. The claim amount Y is recorded for a sample of $n = 111$ car accidents. Suppose the amounts $Y_1, Y_2, ..., Y_{111}$ are iid from a lognormal(μ, σ^2) population distribution.
4. The disease status Y (diseased/not) is observed for $n = 42$ USC students who visit the Student Health Center in a given week. Suppose the statuses $Y_1, Y_2, ..., Y_{42}$ are iid from a Bernoulli(p) population distribution.

Remark: A common format for a first sequence in mathematical statistics (like STAT 511-512-513) is to accept a given parametric family of distributions (e.g., normal, Poisson, gamma, etc.) as being appropriate for the population and then proceed to develop

what is exclusively model-dependent, parametric statistical inference. This is in contrast to **nonparametric inference**, which may make few or no assumptions about the underlying population distribution.

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ is a random sample (iid sample) from a population distribution described by $p_Y(y)$ or $f_Y(y)$. In mathematical terms, a **statistic** T is a function of $Y_1, Y_2, ..., Y_n$, that is,

$$T = T(\mathbf{Y}) = T(Y_1, Y_2, ..., Y_n).$$

In other words, a statistic is a function of the sample $Y_1, Y_2, ..., Y_n$.

Examples: Here are some common statistics:

1. Sample sum:
$$T(\mathbf{Y}) = \sum_{i=1}^{n} Y_i$$
2. Sample mean:
$$T(\mathbf{Y}) = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
3. Sample variance:
$$T(\mathbf{Y}) = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$
4. Minimum order statistic: $T(\mathbf{Y}) = Y_{(1)} = \min\{Y_1, Y_2, ..., Y_n\}$
5. Sample range: $T(\mathbf{Y}) = Y_{(n)} - Y_{(1)}$.

Remark: The definition of a statistic is very broad. For example, even something highly nonstandard like

$$T(\mathbf{Y}) = \ln(S^2 + 5) - 12.08e^{-\tan\left(\sum_{i=1}^{n}|Y_i^3|\right)} + 9.44$$

satisfies the definition. It is a function of the sample $Y_1, Y_2, ..., Y_n$.

Restriction: There is one important restriction in the definition. A statistic $T = T(\mathbf{Y})$ cannot depend on any population-level parameters that are unknown. For example, if the population mean $E(Y) = \mu$ and the population variance $V(Y) = \sigma^2$ are unknown, then

- $\overline{Y}$ is a statistic, but $\overline{Y} - \mu$ is not.
- S^2 is a statistic, but S^2/σ^2 is not.

In other words, we have to be able to calculate the value of T once the random variables' values $Y_1 = y_1$, $Y_2 = y_2$, ..., $Y_n = y_n$ have been observed. We cannot calculate T if it depends on (population-level) quantities that are not known.

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ is a random sample (iid sample) from a population distribution described by $p_Y(y)$ or $f_Y(y)$. Suppose $T = T(\mathbf{Y})$ is a statistic. The probability distribution of T is called its **sampling distribution**.

Revelation: Because $T = T(\mathbf{Y})$, a function of $Y_1, Y_2, ..., Y_n$, a statistic T is itself a random variable. Therefore, T has its own probability distribution! This distribution is called the sampling distribution of T. In notation,

$$\begin{aligned} Y_1, Y_2, ..., Y_n &\sim p_Y(y),\ f_Y(y) \longleftarrow \text{population distribution} \\ T = T(\mathbf{Y}) = T(Y_1, Y_2, ..., Y_n) &\sim p_T(t),\ f_T(t) \longleftarrow \text{sampling distribution of } T. \end{aligned}$$

Common goals: For a statistic $T = T(\mathbf{Y})$, we may want to find its pmf $p_T(t)$ or pdf $f_T(t)$, its cdf $F_T(t)$, or perhaps its mgf $m_T(t)$. These functions identify the (sampling) distribution of T. We might also want to calculate $E(T)$ or $V(T)$. These quantities describe characteristics of T's (sampling) distribution.

Importance: In mathematical statistics, being able to derive sampling distributions (or characteristics of them) is critical; this enables us to understand the underlying mathematical characteristics of inference procedures that are common in statistical practice. This includes confidence intervals and hypothesis tests that are used with single populations, multiple populations (e.g., ANOVA, etc.), regression analysis, time-to-event (survival) analysis, and elsewhere.

7.2 Sample sums and averages

Preview: Sums and averages are common statistics. We start by examining characteristics of well known statistics and then present an important result (involving mgfs) that makes obtaining sampling distributions of sums and averages fairly routine.

Result: Suppose $Y_1, Y_2, ..., Y_n$ is a random sample (iid sample) from a population distribution with mean $E(Y) = \mu$ and variance $V(Y) = \sigma^2$. Then

(a) $E(\overline{Y}) = \mu$

(b) $V(\overline{Y}) = \sigma^2/n$

(c) $E(S^2) = \sigma^2$.

Remark: We already proved parts (a) and (b) in STAT 511; see Example 5.22 (notes, pp 157-158). Recall that

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu.$$

In addition,

$$V(\overline{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

To prove part (c), first note that

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2.$$

Therefore,

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] = \frac{1}{n-1}E\left(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2\right) = \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n} Y_i^2\right) - nE(\overline{Y}^2)\right].$$

Now,

$$E\left(\sum_{i=1}^{n} Y_i^2\right) = \sum_{i=1}^{n} E(Y_i^2) = \sum_{i=1}^{n}\{V(Y_i) + [E(Y_i)]^2\} = \sum_{i=1}^{n}(\sigma^2 + \mu^2) = n(\sigma^2 + \mu^2).$$

In addition,

$$E(\overline{Y}^2) = V(\overline{Y}) + [E(\overline{Y})]^2 = \frac{\sigma^2}{n} + \mu^2.$$

Combining the last two calculations, we have

$$E(S^2) = \frac{1}{n-1}\left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] = \sigma^2. \ \square$$

Important: These results hold for any population distribution (e.g., normal, Poisson, Weibull, etc.). The only restriction is that $E(Y) = \mu < \infty$ for part (a) and $V(Y) = \sigma^2 < \infty$ for parts (b) and (c).

Curiosity: How would we find $V(S^2)$? This in general is a much harder calculation.

Result: Suppose $Y_1, Y_2, ..., Y_n$ is a random sample (iid sample) from a population distribution with moment generating function (mgf) $m_Y(t)$. Let

$$T = \sum_{i=1}^{n} Y_i \quad \text{and} \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

denote the sample sum and the sample mean, respectively. The mgf of the sample sum T is

$$m_T(t) = [m_Y(t)]^n.$$

The mgf of the sample mean $\overline{Y}$ is

$$m_{\overline{Y}}(t) = [m_Y(t/n)]^n.$$

Remark: These results can be helpful in determining the sampling distribution of sums of averages. Recall that mgf are unique; i.e., they uniquely identify a probability distribution. Therefore, if we calculate $m_T(t)$ or $m_{\overline{Y}}(t)$ and recognize it as one that we know (e.g., normal, Poisson, gamma, etc.), then we know T or $\overline{Y}$, respectively, must have that sampling

distribution. Obviously, for this result to be useful, we must know the population-level mgf $m_Y(t)$ or be able to derive it.

Proof. The mgf of $T = \sum_{i=1}^n Y_i$ is given by

$$\begin{aligned} m_T(t) = E(e^{tT}) = E[e^{t(Y_1+Y_2+\cdots+Y_n)}] &= E(e^{tY_1+tY_2+\cdots+tY_n}) \\ &= E(e^{tY_1}e^{tY_2}\cdots e^{tY_n}) \\ &\stackrel{(*)}{=} E(e^{tY_1})E(e^{tY_2})\cdots E(e^{tY_n}) \\ &\stackrel{(**)}{=} m_Y(t)m_Y(t)\cdots m_Y(t) = [m_Y(t)]^n. \end{aligned}$$

The equality $(*)$ is true because $Y_1, Y_2, ..., Y_n$ are mutually independent. The equality $(**)$ is true because $Y_1, Y_2, ..., Y_n$ are identically distributed; i.e., each Y_i has the same population-level mgf $m_Y(t)$. The mgf of the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ is given by

$$m_{\overline{Y}}(t) = E(e^{t\overline{Y}}) = E[e^{(t/n)T}] = m_T(t/n) = [m_Y(t/n)]^n. \ \square$$

Example 7.1. Suppose $Y_1, Y_2, ..., Y_n$ is a random sample (iid sample) from a Poisson(λ) population distribution. Find the sampling distribution of $T = \sum_{i=1}^n Y_i$.

Solution. Recall the Poisson(λ) mgf is

$$m_Y(t) = e^{\lambda(e^t-1)}.$$

Therefore, the mgf of $T = \sum_{i=1}^n Y_i$ is

$$m_T(t) = [m_Y(t)]^n = [e^{\lambda(e^t-1)}]^n = e^{n\lambda(e^t-1)}.$$

We recognize this as the mgf of a Poisson random variable with mean $n\lambda$. Because mgfs are unique, it follows that

$$T \sim \text{Poisson}(n\lambda). \ \square$$

Example 7.2. Suppose $Y_1, Y_2, ..., Y_n$ is a random sample (iid sample) from an exponential(β) population distribution. Find the sampling distribution of the sample sum $T = \sum_{i=1}^n Y_i$ and the sample mean $\overline{Y}$.

Solution. Recall the exponential(β) mgf is

$$m_Y(t) = \frac{1}{1-\beta t}, \quad \text{for } t < \frac{1}{\beta}.$$

Therefore, the mgf of $T = \sum_{i=1}^n Y_i$ is

$$m_T(t) = [m_Y(t)]^n = \left(\frac{1}{1-\beta t}\right)^n,$$

for $t < 1/\beta$. We recognize this as the mgf of a gamma random variable with shape parameter $\alpha = n$ and scale parameter β. Because mgfs are unique, it follows that

$$T \sim \text{gamma}(n, \beta).$$

The mgf of $\overline{Y}$ is

$$m_{\overline{Y}}(t) = [m_Y(t/n)]^n = \left[\frac{1}{1-\beta\left(\frac{t}{n}\right)}\right]^n = \left[\frac{1}{1-\left(\frac{\beta}{n}\right)t}\right]^n,$$

for $t < n/\beta$. We recognize this as the mgf of a gamma random variable with shape parameter $\alpha = n$ and scale parameter β/n. Because mgfs are unique, it follows that

$$\overline{Y} \sim \text{gamma}(n, \beta/n). \ \square$$

Remark: As we have just seen, there are problems for which deriving the sampling distribution of $T = \sum_{i=1}^n Y_i$ or $\overline{Y}$ is very easy by using mgfs. However, there are also problems where using mgfs is not helpful. For example, suppose $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{U}(0,1)$ population distribution; recall the $\mathcal{U}(0,1)$ mgf is given by

$$m_Y(t) = \begin{cases} \dfrac{e^t - 1}{t}, & t \neq 0 \\ 1, & t = 0. \end{cases}$$

The mgf of $T = \sum_{i=1}^n Y_i$ is therefore

$$m_T(t) = [m_Y(t)]^n = \begin{cases} \left(\dfrac{e^t - 1}{t}\right)^n, & t \neq 0 \\ 1, & t = 0. \end{cases}$$

This is the mgf of T, but we do not recognize this mgf as one that we know. Therefore, we cannot use it to determine the sampling distribution of T.

7.3 Sampling distributions arising from the normal distribution

Preview: The $\mathcal{N}(\mu, \sigma^2)$ distribution is the most commonly assumed population distribution in statistical applications. This section is dedicated to deriving sampling distribution results that arise when $Y_1, Y_2, ..., Y_n$ are iid from a $\mathcal{N}(\mu, \sigma^2)$ population distribution.

Result 1: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. The sampling distribution of the sample sum

$$T = \sum_{i=1}^n Y_i \sim \mathcal{N}(n\mu, n\sigma^2).$$

The sampling distribution of the sample mean is

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

and therefore

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Proof. Recall the mgf of $Y \sim \mathcal{N}(\mu, \sigma^2)$ is given by

$$m_Y(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Therefore, the mgf of the sample sum $T = \sum_{i=1}^n Y_i$ is

$$m_T(t) = [m_Y(t)]^n = \left[\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)\right]^n = \exp\left(n\mu t + \frac{n\sigma^2 t^2}{2}\right).$$

We recognize this as the mgf of a normal random variable with mean $n\mu$ and variance $n\sigma^2$. Because mgfs are unique, it follows that

$$T = \sum_{i=1}^n Y_i \sim \mathcal{N}(n\mu, n\sigma^2).$$

The mgf of the sample mean $\overline{Y}$ is

$$\begin{aligned} m_{\overline{Y}}(t) = [m_Y(t/n)]^n &= \left\{\exp\left[\mu\left(\frac{t}{n}\right) + \frac{\sigma^2(t/n)^2}{2}\right]\right\}^n \\ &= \exp\left[n\mu\left(\frac{t}{n}\right) + \frac{n\sigma^2(t/n)^2}{2}\right] = \exp\left[\mu t + \frac{(\sigma^2/n)t^2}{2}\right]. \end{aligned}$$

We recognize this as the mgf of a normal random variable with mean μ and variance σ^2/n. Because mgfs are unique, it follows that

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Finally, we can calculate the mgf of

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

directly. Note that

$$\begin{aligned} m_Z(t) = E[\exp(tZ)] &= E\left\{\exp\left[t\left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right)\right]\right\} \\ &= E\left[\exp\left(\frac{-\mu t}{\sigma/\sqrt{n}}\right)\exp\left(\frac{t\overline{Y}}{\sigma/\sqrt{n}}\right)\right] = \exp\left(\frac{-\mu t}{\sigma/\sqrt{n}}\right) m_{\overline{Y}}\left(\frac{t}{\sigma/\sqrt{n}}\right). \end{aligned}$$

Note that

$$m_{\overline{Y}}\left(\frac{t}{\sigma/\sqrt{n}}\right) = \exp\left[\mu\left(\frac{t}{\sigma/\sqrt{n}}\right) + \frac{(\sigma^2/n)\left(\frac{t}{\sigma/\sqrt{n}}\right)^2}{2}\right] = \exp\left(\frac{\mu t}{\sigma/\sqrt{n}}\right) e^{t^2/2}.$$

Therefore,

$$m_Z(t) = \exp\left(\frac{-\mu t}{\sigma/\sqrt{n}}\right)\exp\left(\frac{\mu t}{\sigma/\sqrt{n}}\right)e^{t^2/2} = e^{t^2/2}.$$

We recognize this as the mgf of a standard normal random variable. Because mgfs are unique, it follows that

$$Z = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1). \ \square$$

Result 2: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. Then

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2 = \sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2 \sim \chi^2(n).$$

Proof. This is a special case of Example 6.15 (pp 23, notes). $\square$

Result 3: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. Then the sample mean

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and the sample variance

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

are **independent** statistics; i.e., $\overline{Y} \perp\!\!\!\perp S^2$.

Remark: There are many ways to prove this result, but unfortunately they all require more advanced mathematical statistics. Note that the authors of your textbook prove $\overline{Y} \perp\!\!\!\perp S^2$ in the $n = 2$ case; see WMS (pp 358).

Result 4: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. Then

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Remark: It is interesting to compare this result with Result 2 above; in particular, noting the effect of replacing the population mean μ in Result 2 with the sample mean $\overline{Y}$ in Result 4. A degree of freedom is "lost" when making this replacement; i.e., $\chi^2(n)$ versus $\chi^2(n-1)$.

Proof. Write

$$\begin{aligned} W_1 = \sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2 &= \sum_{i=1}^{n}\left(\frac{Y_i-\overline{Y}+\overline{Y}-\mu}{\sigma}\right)^2 \\ &= \sum_{i=1}^{n}\left(\frac{Y_i-\overline{Y}}{\sigma}\right)^2 + \underbrace{2\sum_{i=1}^{n}\left(\frac{Y_i-\overline{Y}}{\sigma}\right)\left(\frac{\overline{Y}-\mu}{\sigma}\right)}_{=\,0} + \sum_{i=1}^{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right)^2. \end{aligned}$$

It is easy to show the cross product term is zero because

$$\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0.$$

Therefore, we have

$$W_1 = \sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2 = \underbrace{\sum_{i=1}^{n}\left(\frac{Y_i-\overline{Y}}{\sigma}\right)^2}_{=W_2} + \underbrace{n\left(\frac{\overline{Y}-\mu}{\sigma}\right)^2}_{=W_3} = W_2 + W_3.$$

From Result 2, we know $W_1 \sim \chi^2(n)$. From Result 1, we know

$$W_3 = n\left(\frac{\overline{Y}-\mu}{\sigma}\right)^2 = \left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1).$$

Also note that

$$W_2 = \sum_{i=1}^{n}\left(\frac{Y_i-\overline{Y}}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 = \frac{(n-1)S^2}{\sigma^2}.$$

Furthermore, we know that $W_2 \perp\!\!\!\perp W_3$ because $\overline{Y} \perp\!\!\!\perp S^2$ and functions of independent random variables are independent. The mgf of $W_1 \sim \chi^2(n)$ is, for $t < 1/2$,

$$\begin{aligned}
\left(\frac{1}{1-2t}\right)^{n/2} = m_{W_1}(t) = E(e^{tW_1}) &= E[e^{t(W_2+W_3)}] \\
&= E(e^{tW_2}e^{tW_3}) \\
&\overset{W_2 \perp\!\!\!\perp W_3}{=} E(e^{tW_2})E(e^{tW_3}) \\
&= m_{W_2}(t)m_{W_3}(t) = m_{W_2}(t)\left(\frac{1}{1-2t}\right)^{1/2},
\end{aligned}$$

because $W_3 \sim \chi^2(1)$. This shows

$$m_{W_2}(t) = \frac{\left(\frac{1}{1-2t}\right)^{n/2}}{\left(\frac{1}{1-2t}\right)^{1/2}} = \left(\frac{1}{1-2t}\right)^{(n-1)/2},$$

which, for $t < 1/2$, we recognize as the mgf of a $\chi^2(n-1)$ random variable. Because mgfs are unique,

$$W_2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1). \ \square$$

Note: This sampling distribution result is critical in deriving inference procedures (i.e., confidence intervals and hypothesis tests) for a normal mean and variance. It is important to emphasize this result depends on the underlying population distribution being normal.

Result 5: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. From Result 4 (and recalling the mean and variance of a χ^2 random variable), we get the following "for free:"

$$\begin{aligned} E\left[\frac{(n-1)S^2}{\sigma^2}\right] &= n-1 \\ V\left[\frac{(n-1)S^2}{\sigma^2}\right] &= 2(n-1). \end{aligned}$$

From the first equation above, note that

$$\frac{n-1}{\sigma^2}E(S^2) = n-1 \implies E(S^2) = \sigma^2.$$

Of course, this is nothing new. We proved $E(S^2) = \sigma^2$ in general; i.e., for any population distribution with finite variance. The second equation above implies

$$\frac{(n-1)^2}{\sigma^4}V(S^2) = 2(n-1) \implies V(S^2) = \frac{2\sigma^4}{n-1}.$$

This is a new result. However, it only applies when the underlying population distribution is normal.

Q: Is there a general formula for $V(S^2)$ that applies for any population distribution?
A: Yes, when $Y_1, Y_2, ..., Y_n$ are iid with $E(Y^4) < \infty$; i.e., the fourth population moment is finite, then

$$V(S^2) = \frac{1}{n}\left[\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right],$$

where

$$\mu_4 = E[(Y-\mu)^4].$$

This result is hard to derive in general. As an exercise, show this expression for $V(S^2)$ reduces to $2\sigma^4/(n-1)$ in the normal case.

Result 6: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. The sampling distribution of the sample variance S^2 is

$$S^2 \sim \text{gamma}\left(\frac{n-1}{2}, \frac{2\sigma^2}{n-1}\right).$$

Proof. Apply the result from Example 6.9 (pp 18, notes) with $c = \sigma^2/(n-1)$. We know

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

Therefore,

$$S^2 = \frac{\sigma^2}{n-1}\left[\frac{(n-1)S^2}{\sigma^2}\right] \sim \text{gamma}\left(\frac{n-1}{2}, \frac{2\sigma^2}{n-1}\right). \square$$

7.4 t and F distributions

Preview: In this section, we introduce two additional distributions that often arise in statistical inference: the t and F distributions. We will derive the t pdf and examine certain functions of t and F random variables.

Student's t distribution: Suppose $Z \sim \mathcal{N}(0,1)$, $W \sim \chi^2(\nu)$, and $Z \perp\!\!\!\perp W$. The random variable

$$T = \frac{Z}{\sqrt{W/\nu}} \sim t(\nu),$$

a t distribution with ν degrees of freedom. The pdf of T is given by

$$f_T(t) = \begin{cases} \dfrac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \dfrac{1}{(1+\frac{t^2}{\nu})^{(\nu+1)/2}}, & -\infty < t < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Note: Before we derive the pdf of T (given above), we make the following observations:

- The $t(\nu)$ pdf is continuous and symmetric about zero; see Figure 7.1 (next page).
- When compared to the $\mathcal{N}(0,1)$ pdf, the $t(\nu)$ pdf is less peaked in the center and has more probability in the tails (i.e., leptokurtic).
- As ν increases, the $t(\nu)$ pdf looks more and more like the $\mathcal{N}(0,1)$ pdf. In fact, the sequence of $t(\nu)$ pdfs (in ν) converges pointwise to the $\mathcal{N}(0,1)$ pdf as $\nu \to \infty$.
- The $t(\nu)$ cdf does not exist in closed form for general ν; probabilities and quantiles associated with the $t(\nu)$ distribution can be calculated in R using the `pt` and `qt` functions, respectively.

Application: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. We already know

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

If we replace the population standard deviation σ above with the sample standard deviation $S = \sqrt{S^2}$, then the new quantity

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

To see why this is true, note that

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} = \frac{\sigma}{S}\left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right) = \frac{\dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}\Big/(n-1)}} \sim \frac{\text{“}\mathcal{N}(0,1)\text{”}}{\sqrt{\dfrac{\text{“}\chi^2(n-1)\text{”}}{n-1}}}.$$

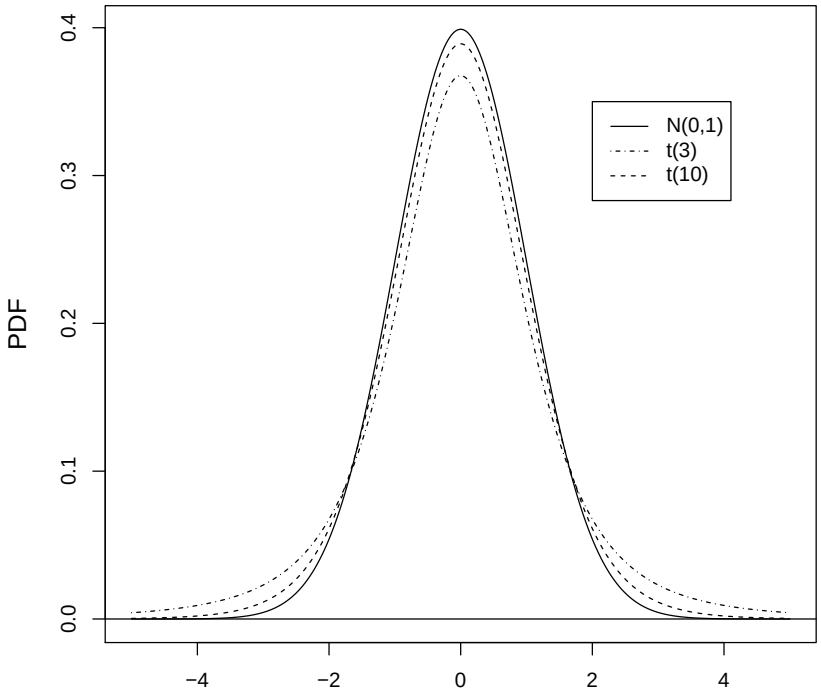

Figure 7.1: t pdfs with $\nu = 3$ and $\nu = 10$ degrees of freedom. The $\mathcal{N}(0,1)$ pdf is also shown.

Because $\overline{Y} \perp\!\!\!\perp S^2$ under a normal population distribution assumption,

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad W = \frac{(n-1)S^2}{\sigma^2}$$

are also independent. Therefore,

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

This sampling distribution result is critical to derive one-sample statistical inference procedures for a normal mean μ (e.g., confidence intervals and hypothesis tests). $\square$

Derivation: Suppose $Z \sim \mathcal{N}(0,1)$, $W \sim \chi^2(\nu)$, and $Z \perp\!\!\!\perp W$. Therefore, the joint pdf of (Z, W) is

$$f_{Z,W}(z,w) = f_Z(z) f_W(w) = \underbrace{\frac{1}{\sqrt{2\pi}} e^{-z^2/2}}_{\mathcal{N}(0,1) \text{ pdf}} \underbrace{\frac{1}{\Gamma(\frac{\nu}{2}) 2^{\nu/2}} w^{\frac{\nu}{2}-1} e^{-w/2}}_{\chi^2(\nu) \text{ pdf}},$$

for $-\infty < z < \infty$ and $w > 0$. That is, the support of (Z, W) is

$$R_{Z,W} = \{(z,w) : -\infty < z < \infty,\ w > 0\}.$$

Consider the bivariate transformation

$$\begin{aligned} T = h_1(Z,W) &= \frac{Z}{\sqrt{W/\nu}} \\ U = h_2(Z,W) &= W. \end{aligned}$$

Our strategy will be to use a bivariate transformation to derive the joint pdf $f_{T,U}(t,u)$. We will then integrate $f_{T,U}(t,u)$ over u to derive the (marginal) pdf of T. The transformation defined above is one-to-one from $R_{Z,W}$ to

$$R_{T,U} = \{(t,u) : -\infty < t < \infty,\ u > 0\}.$$

Therefore, the inverse transformation exists and is given by

$$\begin{aligned} z = h_1^{-1}(t,u) &= t\sqrt{u/\nu} \\ w = h_2^{-1}(t,u) &= u. \end{aligned}$$

The Jacobian of the (inverse) transformation is

$$J = \det \begin{vmatrix} \frac{\partial h_1^{-1}(t,u)}{\partial t} & \frac{\partial h_1^{-1}(t,u)}{\partial u} \\ \frac{\partial h_2^{-1}(t,u)}{\partial t} & \frac{\partial h_2^{-1}(t,u)}{\partial u} \end{vmatrix} = \det \begin{vmatrix} \sqrt{u/\nu} & \frac{t}{\sqrt{\nu}}\frac{1}{2\sqrt{u}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{u}{\nu}}.$$

Therefore, for $-\infty < t < \infty$ and $u > 0$, the joint pdf of (T,U) is

$$\begin{aligned} f_{T,U}(t,u) &= f_{Z,W}(h_1^{-1}(t,u), h_2^{-1}(t,u))|J| \\ &= \frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{u/\nu})^2/2} \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} u^{\frac{\nu}{2}-1} e^{-u/2} \left|\sqrt{\frac{u}{\nu}}\right| \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\nu}} e^{-(t\sqrt{u/\nu})^2/2} \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} u^{\frac{\nu+1}{2}-1} e^{-u/2} \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\nu}} \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} u^{\frac{\nu+1}{2}-1} e^{-u(1+\frac{t^2}{\nu})/2}. \end{aligned}$$

Therefore, the marginal pdf of T, for $-\infty < t < \infty$, is

$$f_T(t) = \int_{u=0}^{\infty} f_{T,U}(t,u)du = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\nu}} \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} \int_{u=0}^{\infty} u^{\frac{\nu+1}{2}-1} e^{-u(1+\frac{t^2}{\nu})/2} du.$$

Note that the integrand in

$$\int_{u=0}^{\infty} u^{\frac{\nu+1}{2}-1} e^{-u(1+\frac{t^2}{\nu})/2} du$$

is a gamma(a,b) kernel with $a = (\nu+1)/2$ and $b = 2\left(1+\frac{t^2}{\nu}\right)^{-1}$. Therefore, the last integral equals

$$\Gamma(a)b^a = \Gamma\left(\frac{\nu+1}{2}\right)\left[2\left(1+\frac{t^2}{\nu}\right)^{-1}\right]^{(\nu+1)/2}.$$

Therefore, for $-\infty < t < \infty$, the pdf of T is

$$f_T(t) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\nu}}\frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}}\,\Gamma\left(\frac{\nu+1}{2}\right)\left[2\left(1+\frac{t^2}{\nu}\right)^{-1}\right]^{(\nu+1)/2} = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\frac{1}{(1+\frac{t^2}{\nu})^{(\nu+1)/2}},$$

as claimed. □

Mean/Variance: If $T \sim t(\nu)$, then $E(T) = 0$, provided that $\nu > 1$. In addition,

$$V(T) = \frac{\nu}{\nu-2},$$

provided that $\nu > 2$. If these conditions on ν are not satisfied, then the corresponding quantities do not exist. For example, if $\nu = 1$, then

$$f_T(t) = \frac{\Gamma(\frac{1+1}{2})}{\sqrt{\pi}\,\Gamma(\frac{1}{2})}\frac{1}{(1+t^2)^{(1+1)/2}} = \frac{1}{\pi(1+t^2)},$$

which we recognize as a standard Cauchy pdf; see Example 6.7 (notes, pp 13-14). In other words, when $\nu = 1$, the t pdf reduces to the standard Cauchy. We know $E(T)$ does not exist when T has this pdf.

Derivation: To show $E(T) = 0$ when $\nu > 1$, we use the definition of a t random variable, namely,

$$T = \frac{Z}{\sqrt{W/\nu}},$$

where $Z \sim \mathcal{N}(0,1)$, $W \sim \chi^2(\nu)$, and $Z \perp\!\!\!\perp W$. Write

$$E(T) = E\left(\frac{Z}{\sqrt{W/\nu}}\right) \stackrel{Z \perp\!\!\!\perp W}{=} E(Z)E\left(\frac{1}{\sqrt{W/\nu}}\right).$$

Because $E(Z) = 0$, this last expression equals 0 provided the second expectation is finite. Therefore, let's investigate the second expectation, and we will see why the $\nu > 1$ condition is needed. Recall that $W \sim \chi^2(\nu) \stackrel{d}{=} \text{gamma}(\frac{\nu}{2}, 2)$. Therefore,

$$\begin{aligned} E\left(\frac{1}{\sqrt{W/\nu}}\right) = \sqrt{\nu}E\left(\frac{1}{\sqrt{W}}\right) &= \sqrt{\nu}\int_0^\infty \frac{1}{\sqrt{w}}\,\frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}}w^{\frac{\nu}{2}-1}e^{-w/2}dw \\ &= \frac{\sqrt{\nu}}{\Gamma(\frac{\nu}{2})2^{\nu/2}}\int_0^\infty w^{\frac{\nu-1}{2}-1}e^{-w/2}dw. \end{aligned}$$

In the last integral, we recognize

$$w^{\frac{\nu-1}{2}-1}e^{-w/2}$$

as a gamma kernel with shape parameter $a = (\nu-1)/2$ and scale parameter $b = 2$. The last integral is finite as long as

$$\frac{\nu-1}{2} > 0 \iff \nu > 1.$$

Showing $V(T) = \nu/(\nu-2)$ when $\nu > 2$ is done similarly. Because $E(T) = 0$, we have

$$V(T) = E(T^2) - [E(T)]^2 = E(T^2).$$

We can calculate the second moment of T as follows:

$$E(T^2) = E\left[\left(\frac{Z}{\sqrt{W/\nu}}\right)^2\right] = E\left(\frac{Z^2}{W/\nu}\right) \stackrel{Z \perp\!\!\!\perp W}{=} E(Z^2)E\left(\frac{1}{W/\nu}\right) = \nu E\left(\frac{1}{W}\right),$$

because $E(Z^2) = 1$. It therefore suffices to show

$$E\left(\frac{1}{W}\right) = \frac{1}{\nu-2},$$

provided that $\nu > 2$, which I will leave as an exercise. $\square$

Snedecor's F distribution: Suppose $W_1 \sim \chi^2(\nu_1)$, $W_2 \sim \chi^2(\nu_2)$, and $W_1 \perp\!\!\!\perp W_2$. The random variable

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F(\nu_1, \nu_2),$$

an F distribution with (numerator) ν_1 and (denominator) ν_2 degrees of freedom. If $U \sim F(\nu_1, \nu_2)$, then the pdf of U is

$$f_U(u) = \begin{cases} \dfrac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\dfrac{\nu_1}{\nu_2}\right)^{\nu_1/2} \dfrac{u^{\frac{\nu_1}{2}-1}}{[1+(\frac{\nu_1}{\nu_2})u]^{(\nu_1+\nu_2)/2}}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Note: We make the following observations:

- The $F(\nu_1, \nu_2)$ pdf is continuous and skewed to the right; see Figure 7.2 (next page).
- The $F(\nu_1, \nu_2)$ pdf can be derived in the same way as the t pdf was derived. Start with $W_1 \sim \chi^2(\nu_1)$, $W_2 \sim \chi^2(\nu_2)$, and $W_1 \perp\!\!\!\perp W_2$. Define

$$\begin{aligned} U_1 = h_1(W_1, W_2) &= \frac{W_1/\nu_1}{W_2/\nu_2} \\ U_2 = h_2(W_1, W_2) &= W_1. \end{aligned}$$

  Perform a bivariate transformation to find $f_{U_1,U_2}(u_1, u_2)$ and then integrate over u_2.
- The $F(\nu_1, \nu_2)$ cdf does not exist in closed form; probabilities and quantiles associated with the $F(\nu_1, \nu_2)$ distribution can be calculated in R using the `pf` and `qf` functions, respectively.

Mean/Variance: If $F \sim F(\nu_1, \nu_2)$, then

$$\begin{aligned} E(F) &= \frac{\nu_2}{\nu_2-2}, \quad \text{if } \nu_2 > 2 \\ V(F) &= 2\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)}, \quad \text{if } \nu_2 > 4. \end{aligned}$$

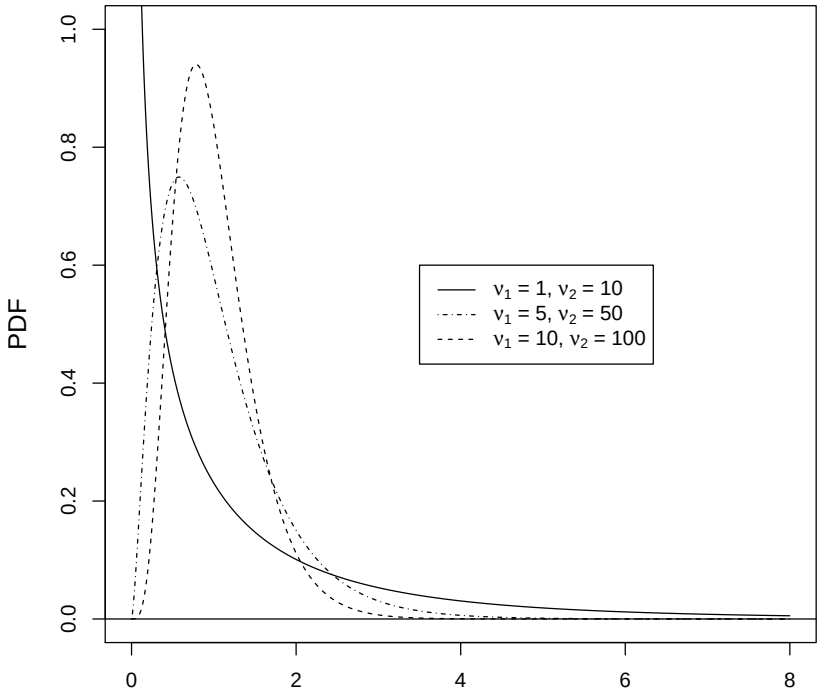

Figure 7.2: $F(\nu_1, \nu_2)$ pdfs for different combinations of ν_1 and ν_2.

Application: Suppose we have two independent random samples:

- $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an iid sample from a $\mathcal{N}(\mu_1, \sigma_1^2)$ population distribution
- $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an iid sample from a $\mathcal{N}(\mu_2, \sigma_2^2)$ population distribution.

Define the sample means

$$\overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad \overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j}$$

and the sample variances

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(Y_{1j} - \overline{Y}_{1+})^2 \quad \text{and} \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(Y_{2j} - \overline{Y}_{2+})^2.$$

We know

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \quad \text{and} \quad \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1).$$

Furthermore,

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \perp\!\!\!\perp \frac{(n_2-1)S_2^2}{\sigma_2^2}$$

because the two samples are independent. Therefore,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}\Big/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}\Big/(n_2-1)} \sim \frac{\text{“}\chi^2(n_1-1)\text{”}/(n_1-1)}{\text{“}\chi^2(n_2-1)\text{”}/(n_2-1)} \sim F(n_1-1, n_2-1).$$

This sampling distribution result is critical to derive statistical inference procedures which compare the variances of two normal populations. Note further that if the two populations have the same variance; i.e., if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

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

Note: We have the following results concerning functions of t and F random variables.

- If $Y \sim F(\nu_1, \nu_2)$, then
$$U = \frac{1}{Y} \sim F(\nu_2, \nu_1).$$
- If $Y \sim t(\nu)$, then
$$U = Y^2 \sim F(1, \nu).$$
- If $Y \sim F(\nu_1, \nu_2)$, then
$$U = \frac{(\nu_1/\nu_2)Y}{1+(\nu_1/\nu_2)Y} \sim \text{beta}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right).$$

Each of these results can be proven by performing a univariate transformation.

7.5 Central Limit Theorem

Preview: The Central Limit Theorem (CLT) is one of the most important results in statistics. It describes the approximate sampling distribution of sample means (or sums). Under very mild conditions, these approximate distributions turn out to be normal.

Recall: To set our ideas, recall Result 1 on pp 50 (notes). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. The sampling distribution of the sum

$$T = \sum_{i=1}^{n} Y_i \sim \mathcal{N}(n\mu, n\sigma^2).$$

The sampling distribution of the sample mean

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

In other words, when the population distribution is normal, the sampling distribution of both T and $\overline{Y}$ is also normal.

Q: What happens when the population distribution is not normal? For example, what if $Y_1, Y_2, ..., Y_n$ are iid Poisson? iid uniform? iid Bernoulli? iid exponential?
A: Regardless of what the population distribution is (for the most part), the sampling distribution of T and $\overline{Y}$ is still **approximately normal**, that is,

$$T = \sum_{i=1}^{n} Y_i \sim \mathcal{AN}(n\mu, n\sigma^2)$$

and

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right),$$

where μ and σ^2 are the population mean and population variance, respectively. The symbol $\mathcal{AN}$ is read "approximately normal."

Remark: The approximate sampling distributions above represent what is conferred by the CLT. The quality of the approximation depends primarily on these two factors:

- sample size−the larger the sample size n, the better the approximation
- symmetry/skewness of the population distribution−the more symmetric the population distribution, the better the approximation.

Terminology: The **skewness** of a random variable Y is given by

$$\xi = \frac{E[(Y-\mu)^3]}{\sigma^3},$$

where $\mu = E(Y)$ and $\sigma^2 = V(Y)$. Note that

- $\xi = 0 \Longrightarrow$ the population distribution $p_Y(y)$ or $f_Y(y)$ is symmetric about μ
- $\xi > 0 \Longrightarrow$ the population distribution $p_Y(y)$ or $f_Y(y)$ is skewed right
- $\xi < 0 \Longrightarrow$ the population distribution $p_Y(y)$ or $f_Y(y)$ is skewed left.

Example 7.3. An insurance company issues 250 vision care insurance policies. The number of claims filed by a policyholder under a vision care insurance policy during one year is a Poisson random variable Y with mean 2. Assume the numbers of claims filed by different policyholders are mutually independent.

Q: What is the probability there is a total of between 475 and 550 claims during a one-year period?

Solution. Denote the $n = 250$ claim counts by $Y_1, Y_2, ..., Y_{250}$ and assume $Y_1, Y_2, ..., Y_{250}$ are iid from a Poisson($\lambda = 2$) population distribution; see Figure 7.3 (next page, left). Let $T = \sum_{i=1}^{250} Y_i$ denote the total number of claims filed. We want to calculate $P(475 \leq T \leq 550)$.

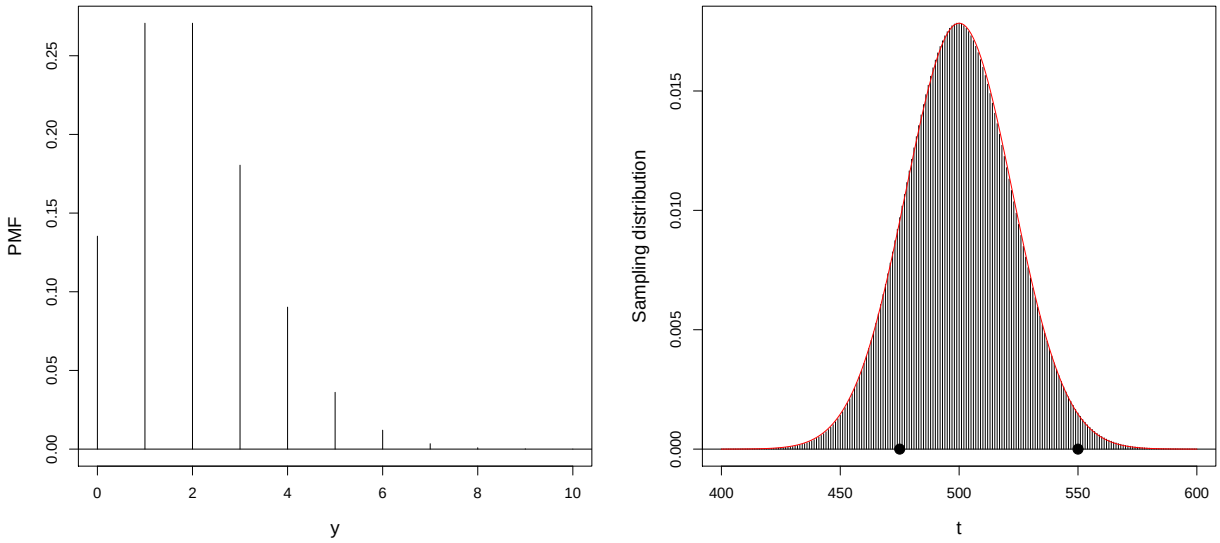

Figure 7.3: Left: Population distribution of $Y \sim \text{Poisson}(\lambda = 2)$ in Example 7.3. Right: Exact sampling distribution of $T = \sum_{i=1}^{250} Y_i \sim \text{Poisson}(500)$. The smooth red curve is the $\mathcal{N}(500, 500)$ pdf, which represents the approximate sampling distribution of T conferred by the CLT. Solid dark circles are shown at $t = 475$ and $t = 550$.

Exact calculation: From Example 7.1, we know the (exact) sampling distribution of T is $T \sim \text{Poisson}(500)$. This sampling distribution is shown in Figure 7.3 (above, right). Therefore,

$$P(475 \leq T \leq 550) = \sum_{t=475}^{550} \frac{500^t e^{-500}}{t!} \approx 0.8605.$$

```
> sum(dpois(475:550,500))
[1] 0.8605368
```

CLT approximation: The population distribution is $\text{Poisson}(\lambda = 2)$, so $\mu = 2$ and $\sigma^2 = 2$. From the CLT, the approximate sampling distribution of T is $T \sim \mathcal{AN}(500, 500)$. This approximate sampling distribution is shown in Figure 7.3 (above, right). Therefore,

$$P(475 \leq T \leq 550) \approx \int_{475}^{550} \frac{1}{\sqrt{2\pi}\sqrt{500}} e^{-(t-500)^2/2(500)} dt \approx 0.8556.$$

```
> pnorm(550,500,sqrt(500))-pnorm(475,500,sqrt(500))
[1] 0.8555501
```

As we can see, approximating the sampling distribution of $T = \sum_{i=1}^{250} Y_i$ by using the CLT is very accurate, despite the population distribution of Y being discrete and also skewed to the right. The reason for the high accuracy is the large sample size ($n = 250$). $\square$

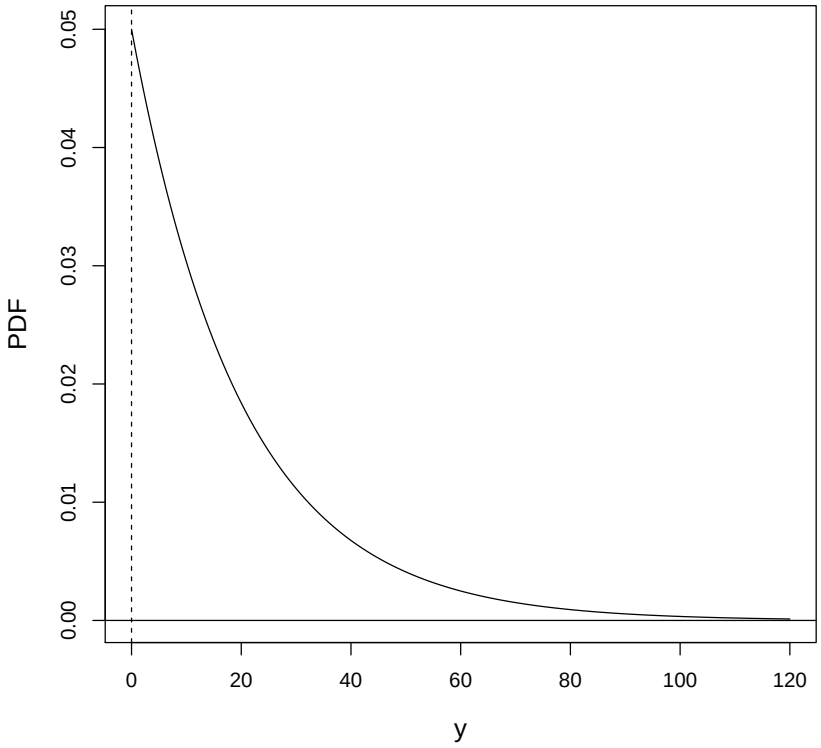

Figure 7.4: Population distribution of $Y \sim$ exponential($\beta = 20$) in Example 7.4.

Example 7.4. The time Y (in days) to recruit patients for a clinical trial follows an exponential distribution with mean $\beta = 20$. What is the probability the average of $n = 10$ patients' recruiting times will exceed 30 days? Assume the times of different patients are mutually independent.

Solution. Denote the $n = 10$ times by $Y_1, Y_2, ..., Y_{10}$ and assume $Y_1, Y_2, ..., Y_{10}$ are iid from an exponential($\beta = 20$) population distribution. This population distribution is shown in Figure 7.4 above. Let $\overline{Y}$ denote the sample mean time among the 10 patients. We want to calculate $P(\overline{Y} > 30)$.

Exact calculation: From Example 7.2, we know the (exact) sampling distribution of $\overline{Y}$ is $\overline{Y} \sim$ gamma$(10, 2)$. This sampling distribution is shown in Figure 7.5 (next page, left). Therefore,

$$P(\overline{Y} > 30) = \int_{30}^{\infty} \frac{1}{\Gamma(10)2^{10}} y^{10-1} e^{-y/2} dy \approx 0.0699.$$

```
> 1-pgamma(30,10,1/2)
[1] 0.06985366
```

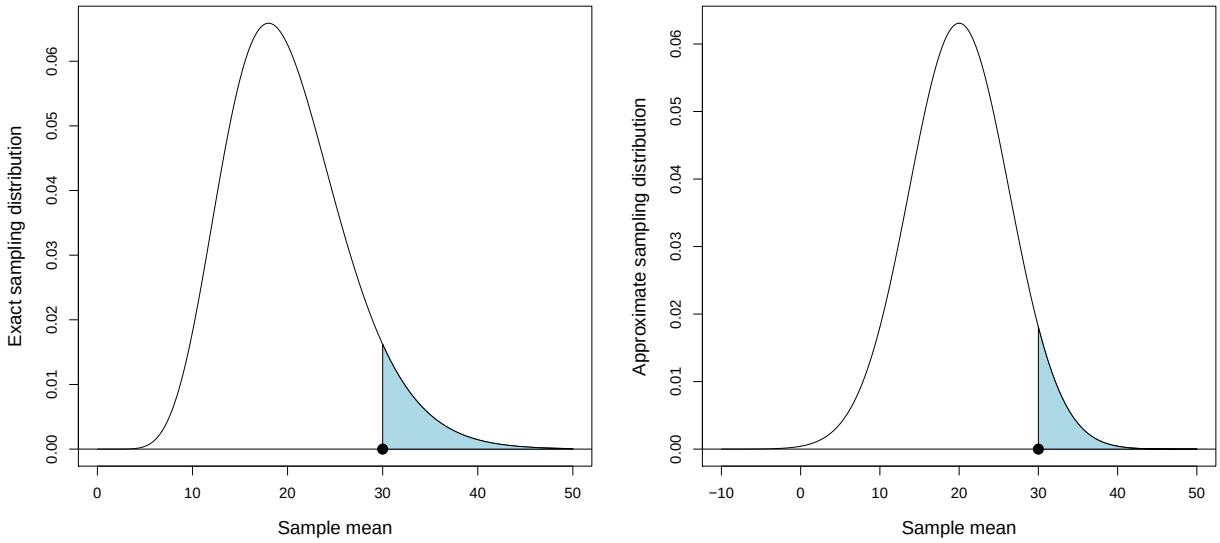

Figure 7.5: Left: Exact sampling distribution of $\overline{Y} \sim$ gamma$(10, 2)$ in Example 7.4. Right: Approximate sampling distribution of $\overline{Y} \sim \mathcal{AN}(20, 40)$ conferred by the CLT. In both figures, the probability $P(\overline{Y} > 30)$ is shown shaded.

CLT approximation: The population distribution is exponential$(\beta = 20)$, so $\mu = 20$ and $\sigma^2 = 400$. From the CLT, the approximate sampling distribution of $\overline{Y}$ is $\overline{Y} \sim \mathcal{AN}(20, 40)$. This approximate sampling distribution is shown in Figure 7.5 (above, right). Therefore,

$$P(\overline{Y} > 30) \approx \int_{30}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{40}} e^{-(y-20)^2/2(40)} dy \approx 0.0569.$$

```
> 1-pnorm(30,20,sqrt(40))
[1] 0.05692315
```

The approximation is not terrible, but there are notable discrepancies between the exact sampling distribution of $\overline{Y}$ and the approximation conferred by the CLT. This is because the exponential$(\beta = 20)$ population distribution is highly skewed and the sample size $n = 10$ is small. □

Q: Why rely on the CLT to approximate the sampling distribution of the sum $T = \sum_{i=1}^{n} Y_i$ or the sample mean $\overline{Y}$ if we can just work with the exact sampling distributions (like in Example 7.3 and Example 7.4)?
A: Because in many problems, the exact sampling distribution of T or $\overline{Y}$ may not be known or it may be impossible to derive in closed form. In this situation, using a CLT approximation may be our only option. This is illustrated in the next example.

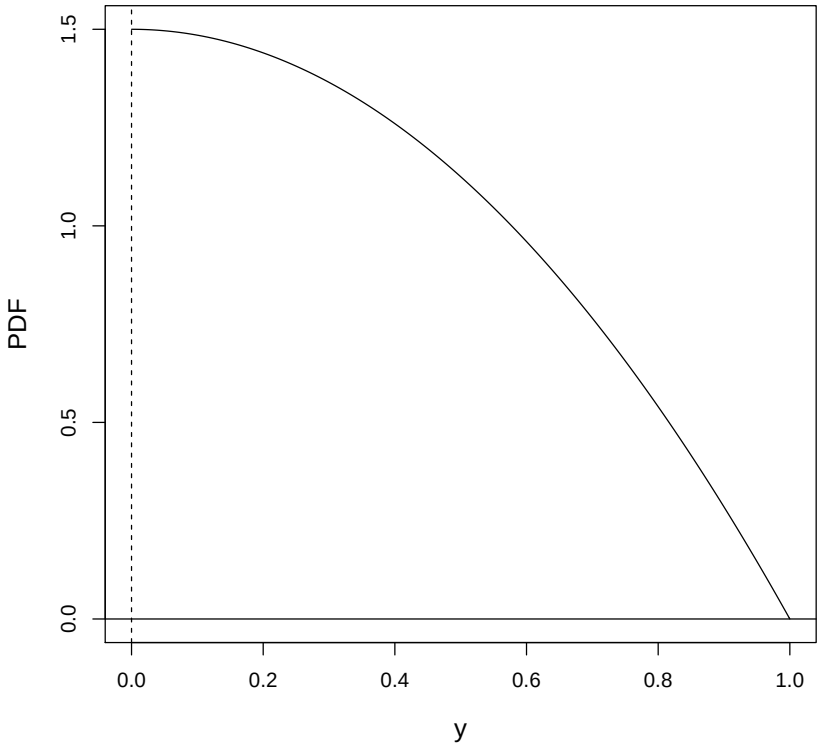

Figure 7.6: Population pdf of Y in Example 7.5.

Example 7.5. The amount of gravel (in 1000s of tons) sold by a construction company in a given week is a continuous random variable Y with the population pdf

$$f_Y(y) = \begin{cases} \frac{3}{2}(1-y^2), & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Assuming weekly sales are mutually independent, find the probability the total sales for a given year exceeds 22,500 tons.

Solution. Denote the $n = 52$ weekly sales amounts by $Y_1, Y_2, ..., Y_{52}$ and assume $Y_1, Y_2, ..., Y_{52}$ are iid from a population described by $f_Y(y)$. This population distribution is shown in Figure 7.6 above. Let $T = \sum_{i=1}^{52} Y_i$ denote the total sales for the year. We want to find $P(T > 22.5)$.

Remark: It is not clear how one would go about deriving the exact sampling distribution of $T = \sum_{i=1}^{52} Y_i$ in this example. Without this distribution, we cannot calculate $P(T > 22.5)$ exactly. However, we can approximate $P(T > 22.5)$ by using the CLT. To do this, we need to find the population mean $E(Y) = \mu$ and the population variance $V(Y) = \sigma^2$. The CLT approximation to the sampling distribution of T requires only these values.

The population mean is

$$\mu = E(Y) = \int_{\mathbb{R}} y f_Y(y) dy = \frac{3}{2}\int_0^1 y(1-y^2) = \frac{3}{2}\left(\frac{y^2}{2} - \frac{y^4}{4}\right)\Big|_0^1 = \frac{3}{8}.$$

The population second moment is

$$E(Y^2) = \int_{\mathbb{R}} y^2 f_Y(y) dy = \frac{3}{2}\int_0^1 y^2(1-y^2) = \frac{3}{2}\left(\frac{y^3}{3} - \frac{y^5}{5}\right)\Big|_0^1 = \frac{1}{5}.$$

Therefore, the population variance is

$$\sigma^2 = V(Y) = E(Y^2) - [E(Y)]^2 = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = \frac{19}{320}.$$

Applying the CLT, the sampling distribution of $T = \sum_{i=1}^{52} Y_i$ is approximately normal with mean

$$n\mu = 52\left(\frac{3}{8}\right) = 19.5$$

and variance

$$n\sigma^2 = 52\left(\frac{19}{320}\right) = 3.0875;$$

i.e., $T \sim \mathcal{AN}(19.5, 3.0875)$. This approximate sampling distribution is shown in Figure 7.7 (next page). Therefore,

$$P(T > 22.5) \approx \int_{22.5}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{3.0875}} e^{-(t-19.5)^2/2(3.0875)} dt \approx 0.0439. \ \square$$

```
> 1-pnorm(22.5,19.5,sqrt(3.0875))
[1] 0.04388026
```

Theory: We would like to officially state the CLT and prove it rigorously. To do so, we need this theoretical result. Suppose $Z_1, Z_2, Z_3, ...$ is a sequence of random variables, where Z_n has mgf $m_{Z_n}(t)$. Suppose that $m_{Z_n}(t) \to m_Z(t)$, as $n \to \infty$ for all $t \in (-h, h)$ $\exists h > 0$; i.e., the sequence of functions $m_{Z_n}(t)$ converges pointwise for all t in an open neighborhood about $t = 0$. Then

1. There exists a unique cdf $F_Z(z)$ whose moments are determined by $m_Z(t)$.
2. The sequence of cdfs
$$F_{Z_n}(z) \to F_Z(z),$$
as $n \to \infty$, for all $z \in C_{F_Z}$, the set of points $z \in \mathbb{R}$ where $F_Z(\cdot)$ is continuous.

In other words, *convergence of mgfs implies convergence of cdfs.* We write $Z_n \xrightarrow{d} Z$, as $n \to \infty$, and say that "Z_n converges in distribution to Z."

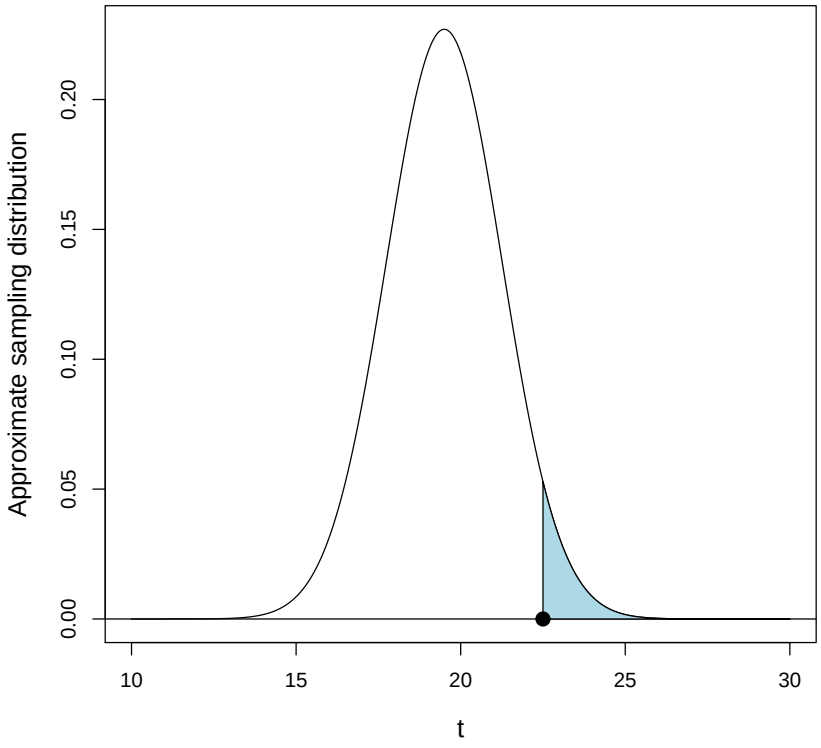

Figure 7.7: Approximate sampling distribution of $T = \sum_{i=1}^{52} Y_i \sim \mathcal{AN}(19.5, 3.0875)$ in Example 7.5. The probability $P(T > 22.5) \approx 0.0439$ is shown shaded.

Central Limit Theorem: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2 < \infty$. Define

$$Z_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sqrt{n}\sigma}.$$

Then

$$P(Z_n \leq z) = F_{Z_n}(z) \to \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = F_Z(z) = P(Z \leq z),$$

as $n \to \infty$. That is, the sequence of cdfs $F_{Z_n}(z)$ converges pointwise to the cdf of a standard normal random variable $F_Z(z)$. We write $Z_n \xrightarrow{d} \mathcal{N}(0, 1)$.

Discussion: Before we prove the CLT, we make the following remarks:

- It is important to appreciate how modest the assumptions are for the CLT to "work." All we need is (a) $Y_1, Y_2, ..., Y_n$ is an iid sample and (b) the population variance $\sigma^2 < \infty$. Of course, although the finite variance assumption covers most population distributions, it does not cover all of them; e.g., Cauchy, etc.

- Convergence in distribution is a mathematical concept that investigates the stochastic behavior of a sequence of random variables, here,

$$Z_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n Y_i - n\mu}{\sqrt{n}\sigma}$$

when $n \to \infty$. In practical terms, this notion is somewhat fanciful because the sample size n is always finite and may also be small (e.g., $n = 10$). This is why we use the phrase "approximately normal" in applications. We will interpret the statement

$$Z_n = \frac{\sum_{i=1}^n Y_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{to mean} \quad \sum_{i=1}^n Y_i \sim \mathcal{AN}(n\mu, n\sigma^2) \text{ for } n \text{ large.}$$

Similarly, we will interpret the statement

$$Z_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{to mean} \quad \overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right) \text{ for } n \text{ large.}$$

The "n large" approximate sampling distributions above are what we have been using all along; see Examples 7.3-7.5.

Proof. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2 < \infty$. Define

$$Z_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}.$$

It suffices to show $m_{Z_n}(t)$, the mgf of Z_n, converges pointwise to $m_Z(t) = e^{t^2/2}$, the mgf of $Z \sim \mathcal{N}(0,1)$. Define

$$U_i = \frac{Y_i - \mu}{\sigma},$$

for $i = 1, 2, ..., n$. Note that

$$E(U_i) = E\left(\frac{Y_i - \mu}{\sigma}\right) = \frac{1}{\sigma}E(Y_i - \mu) = \frac{1}{\sigma}(\mu - \mu) = 0$$

and

$$V(U_i) = V\left(\frac{Y_i - \mu}{\sigma}\right) = \frac{1}{\sigma^2}V(Y_i) = \frac{\sigma^2}{\sigma^2} = 1.$$

That is, by construction, $U_1, U_2, ..., U_n$ are iid random variables with mean 0 and variance 1. Let $m_U(t)$ denote the common mgf of $U_1, U_2, ..., U_n$. Simple algebra yields

$$Z_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\frac{1}{n}\sum_{i=1}^n Y_i - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n Y_i - n\mu}{\sqrt{n}\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i.$$

Therefore, the mgf of Z_n is given by

$$\begin{aligned} m_{Z_n}(t) = E(e^{tZ_n}) &= E\left[e^{\frac{t}{\sqrt{n}}(U_1+U_2+\cdots+U_n)}\right] \\ &= E\left(e^{\frac{t}{\sqrt{n}}U_1}e^{\frac{t}{\sqrt{n}}U_2}\cdots e^{\frac{t}{\sqrt{n}}U_n}\right) \\ &= E\left(e^{\frac{t}{\sqrt{n}}U_1}\right)E\left(e^{\frac{t}{\sqrt{n}}U_2}\right)\cdots E\left(e^{\frac{t}{\sqrt{n}}U_n}\right) = [E\left(e^{\frac{t}{\sqrt{n}}U}\right)]^n = \left[m_U(t/\sqrt{n})\right]^n. \end{aligned}$$

Now write $m_U(t/\sqrt{n})$ in its McLaurin series expansion:

$$m_U(t/\sqrt{n}) = \sum_{k=0}^{\infty} m_U^{(k)}(0)\frac{\left(\frac{t}{\sqrt{n}}-0\right)^k}{k!},$$

where

$$m_U^{(k)}(0) = \frac{d^k}{dt^k}m_U(t)\bigg|_{t=0}.$$

Note that

$$\begin{aligned} m_U^{(0)}(0) &= m_U(0) = 1 \\ m_U^{(1)}(0) &= E(U) = 0 \\ m_U^{(2)}(0) &= E(U^2) = 1. \end{aligned}$$

Therefore, the expansion above becomes

$$m_U(t/\sqrt{n}) = 1 + \frac{(t/\sqrt{n})^2}{2!} + r_U(t/\sqrt{n}),$$

where the remainder term

$$r_U(t/\sqrt{n}) = \sum_{k=3}^{\infty} m_U^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!}.$$

To summarize, we have written

$$m_{Z_n}(t) = \left[m_U(t/\sqrt{n})\right]^n = \left[1 + \frac{t^2/2}{n} + r_U(t/\sqrt{n})\right]^n.$$

It therefore suffices to argue

$$\lim_{n\to\infty} nr_U(t/\sqrt{n}) = 0$$

for all $t \in \mathbb{R}$. If we can do this, then

$$\lim_{n\to\infty} m_{Z_n}(t) = \lim_{n\to\infty}\left[1 + \frac{t^2/2}{n} + r_U(t/\sqrt{n})\right]^n = \lim_{n\to\infty}\left(1 + \frac{t^2/2}{n}\right)^n = e^{t^2/2},$$

and we will be done. However, note that

$$\begin{aligned} nr_U(t/\sqrt{n}) &= n\sum_{k=3}^{\infty} m_U^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!} \\ &= m_U^{(3)}(0)\frac{t^3}{3!}\frac{1}{\sqrt{n}} + m_U^{(4)}(0)\frac{t^4}{4!}\frac{1}{n} + m_U^{(5)}(0)\frac{t^5}{5!}\frac{1}{n\sqrt{n}} + m_U^{(6)}(0)\frac{t^6}{6!}\frac{1}{n^2} + \cdots. \end{aligned}$$

For any $t \in \mathbb{R}$, each term on the RHS above converges to 0 as $n \to \infty$. Therefore, $\lim_{n\to\infty} nr_U(t/\sqrt{n}) = 0$. $\square$

Remark: One of the most important applications of the CLT arises when the population distribution of Y is Bernoulli(p). Recall the Bernoulli(p) pmf is given by

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

The Bernoulli population distribution is applicable when we measure "success/failure" (1/0) outcomes on each individual in the population; e.g., diseased/not, respond to treatment/not, defective/not, etc. The parameter p satisfies $0 < p < 1$ and is called "the probability of success" for any individual in the population. Recall the mean and variance of $Y \sim$ Bernoulli(p) are given by

$$\begin{aligned} E(Y) &= p \\ V(Y) &= p(1-p). \end{aligned}$$

Application: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(p) population distribution. Because the population variance $V(Y) = p(1-p) < \infty$, the CLT applies and therefore

$$T = \sum_{i=1}^{n} Y_i \sim \mathcal{AN}(np, np(1-p))$$

and

$$\widehat{p} = \overline{Y} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right)$$

for n large. The symbol $\widehat{p}$ is used to denote the proportion of "1's" in the sample; i.e., the **sample proportion**.

Note: From Example 6.13 (notes, pp 21), we know the exact sampling distribution of $T = \sum_{i=1}^{n} Y_i \sim b(n, p)$. Recall that

$$m_T(t) = [m_Y(t)]^n = (q + pe^t)^n,$$

where $q = 1 - p$. We recognize the mgf of T as the mgf of a $b(n, p)$ random variable. Therefore, from the CLT, it must be true that the $b(n, p)$ pmf can be approximated by a normal pdf with mean np and variance $np(1-p)$. Recall the approximation is best when the sample size n is large and the skewness in the population distribution is close to zero. The skewness of $Y \sim$ Bernoulli(p) is given by

$$\xi = \frac{E[(Y-\mu)^3]}{\sigma^3} = \frac{E[(Y-p)^3]}{[p(1-p)]^{3/2}} = \frac{1-2p}{\sqrt{p(1-p)}}.$$

This means the CLT approximation will be best when n is large and p is close to 0.5. See Figure 7.8 (next page).

Example 7.6. PRAMS, the Pregnancy Risk Assessment Monitoring System, is a surveillance project of the Centers for Disease Control and Prevention and state health departments.

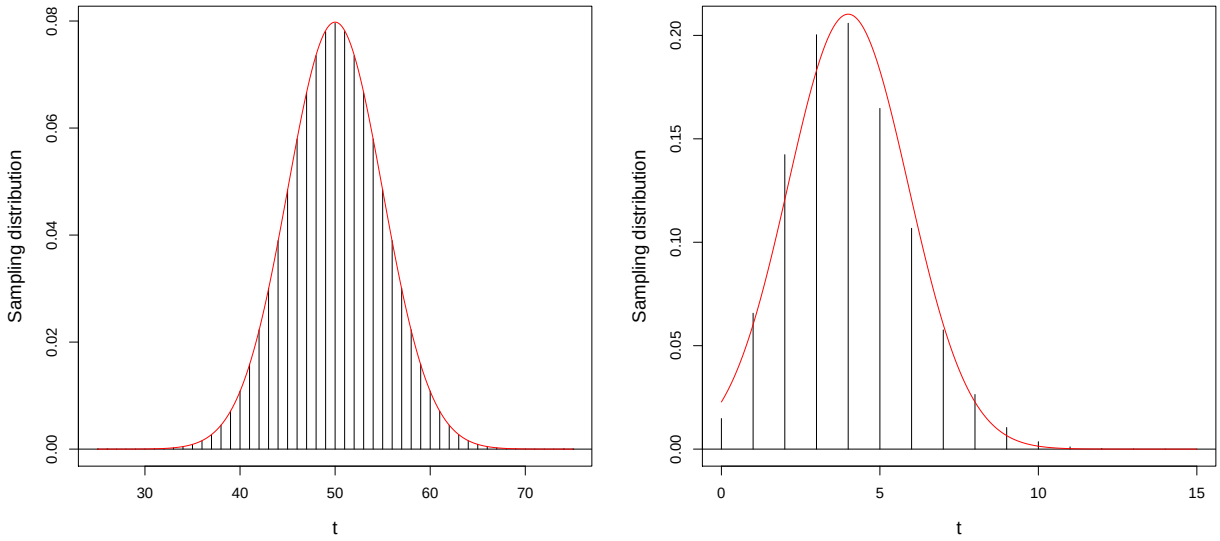

Figure 7.8: $b(n,p)$ pmfs with $\mathcal{N}(np, np(1-p))$ pdfs overlaid. Left: $n = 100$ and $p = 0.5$. Right: $n = 40$ and $p = 0.1$.

In a recent PRAMS survey, $n = 999$ women who had recently given birth were asked about their smoking habits. Investigators observed

$$Y_i = \begin{cases} 1, & i\text{th woman smoked during last 3 months of pregnancy} \\ 0, & \text{otherwise.} \end{cases}$$

Assuming $Y_1, Y_2, ..., Y_{999}$ are iid from a Bernoulli$(p = 0.1)$ population distribution, calculate the probability at least 125 women smoked during the last 3 months of pregnancy.

Exact calculation: We know $T = \sum_{i=1}^{999} Y_i \sim b(n = 999, p = 0.1)$, so

$$P(T \geq 125) = \sum_{t=125}^{999} \binom{999}{t} (0.1)^t (0.9)^{999-t} \approx 0.0058.$$

```
> 1-pbinom(124,999,0.10)
[1] 0.005833989
```

CLT approximation: From the CLT, the approximate sampling distribution of T is $T \sim \mathcal{AN}(99.9, 89.91)$. Therefore,

$$P(T \geq 125) \approx \int_{125}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{89.91}} e^{-(t-99.9)^2/2(89.91)} dt \approx 0.0041.$$

```
> 1-pnorm(125,99.9,sqrt(89.91))
[1] 0.004059313
```

8 Estimation

8.1 Introduction

Preview: In this chapter, we transition to one of the most important concepts in statistics, namely, how to **estimate** population-level parameters by using a sample of observations drawn from the population. Intuitively, if $Y_1, Y_2, ..., Y_n$ is a sample from a population distribution described by $p_Y(y)$ or $f_Y(y)$, then the observations $Y_1, Y_2, ..., Y_n$ contain valuable information about characteristics of the population distribution; e.g., the population mean $\mu = E(Y)$, the population variance $\sigma^2 = V(Y)$, and so on.

Importance: The reason the estimation question emerges as relevant is that parameters associated with a population distribution (or distributions) are usually **unknown**. For example, suppose an epidemiologist observes a random sample of $n = 10$ USC undergraduate students and records

$$Y = \text{ the number of sexual partners within the last six months}$$

on each student. As a population-level model, he decides to use $Y \sim \text{Poisson}(\lambda)$, where $\lambda = E(Y)$, the mean of the population. Now, there are over 26,000 undergraduate students at USC. Therefore, the only way the epidemiologist can determine the value of λ is to observe all 26,000+ students. Because it is generally not possible to "sample the entire population" in real life evaluations (especially in larger populations which may number in the millions or billions), we turn to the problem of parameter **estimation**.

Problem: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution described by $p_Y(y)$ or $f_Y(y)$, and let $\theta \in \mathbb{R}$ denote a population-level parameter that is unknown.

- The use of a generic symbol like θ for a population-level parameter allows us to generalize our discussion.
- In the last example, we could write $Y \sim \text{Poisson}(\theta)$, where $\theta = E(Y)$.
- The population-level parameter θ is unknown but also **fixed**; i.e., it is not random. **Note:** When we are introduced to the Bayesian paradigm later, this assumption will no longer be true. Bayesians treat population-level parameters as random and assign them their own probability distributions.

Q: How should we use the sample $Y_1, Y_2, ..., Y_n$ to estimate θ?

Terminology: A **point estimator** of θ is any statistic; i.e.,

$$\widehat{\theta} = T(Y_1, Y_2, ..., Y_n),$$

that estimates θ. Because a point estimator $\widehat{\theta}$ is a statistic, it is **random** and has its own (sampling) distribution.

Illustration: What point estimator should we use in the USC student example? Suppose $Y_1, Y_2, ..., Y_{10}$ is regarded as an iid sample from a Poisson distribution with mean $\theta > 0$. One obvious point estimator of θ is the **sample mean**

$$\widehat{\theta} = \overline{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i \quad \longleftarrow \text{ function of } Y_1, Y_2, ..., Y_{10}.$$

We proved in the last chapter that $E(\overline{Y}) = \mu = \theta$, so at least on average (i.e., across many samples) the value of $\overline{Y}$ will correctly pin down the true value of the population mean. Another candidate point estimator, interestingly, is the **sample variance**; i.e.,

$$\widehat{\theta} = S^2 = \frac{1}{10-1}\sum_{i=1}^{10}(Y_i - \overline{Y})^2 \quad \longleftarrow \text{ function of } Y_1, Y_2, ..., Y_{10}.$$

Recall that in the Poisson distribution, the population mean and population variance are the same; i.e., $\mu = \sigma^2 = \theta$. In the last chapter, we proved the sample variance S^2 satisfies $E(S^2) = \sigma^2$, which is θ under the Poisson assumption. Therefore, on average (i.e., across many samples), the value of S^2 will also correctly pin down the true value of θ.

Questions: Which point estimator should we use: $\overline{Y}$ or S^2? Maybe another point estimator is "better," say, the **sample median**

$$\widehat{\theta} = \frac{Y_{(5)} + Y_{(6)}}{2}.$$

The point is that in any estimation problem, there may be many point estimators to consider. Therefore, we should have a way to evaluate the quality of a point estimator so that we can judge whether or not it does a good job at estimation. This leads us to the next section, where we quantify how *accurate* and how *precise* a point estimator $\widehat{\theta}$ is.

Aside: A **point estimator** $\widehat{\theta} = T(Y_1, Y_2, ..., Y_n)$ is random because its value depends on $Y_1, Y_2, ..., Y_n$ which are random themselves. However, after the values $Y_1 = y_1$, $Y_2 = y_2$, ..., $Y_n = y_n$ have been observed, we can calculate the value of the **point estimate**

$$\widehat{\theta} = T(y_1, y_2, ..., y_n).$$

This is a numerical value because it is based on the observed values $y_1, y_2, ..., y_n$; i.e., "the observed data." To illustrate, in the USC example, a point estimator of θ based on the $n = 10$ students is

$$\overline{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i,$$

the sample mean. If the observed data are

$$y_1 = 4,\ y_2 = 2,\ y_3 = 1,\ y_4 = 3,\ y_5 = 2,\ y_6 = 5,\ y_7 = 0,\ y_8 = 1,\ y_9 = 0,\ y_{10} = 0,$$

then the point estimate is the realized value of the point estimator; i.e.,

$$\overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 1.8.$$

Note that a point estimate is no longer random (i.e., it's a fixed number).

8.2 Bias and mean-squared error

Terminology: Suppose $\widehat{\theta} = T(Y_1, Y_2, ..., Y_n)$ is a point estimator for the population-level parameter θ. We call $\widehat{\theta}$ an **unbiased estimator** of θ if

$$E(\widehat{\theta}) = \theta.$$

In other words, the mean of the sampling distribution of $\widehat{\theta}$ is equal to θ; see Figure 8.1 below (left). If

$$E(\widehat{\theta}) \neq \theta,$$

then we say that $\widehat{\theta}$ is biased. The sampling distribution in Figure 8.1 (right) describes a point estimator $\widehat{\theta}$ that is biased. The **bias** of a point estimator $\widehat{\theta}$ is

$$B(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

Note that if $\widehat{\theta}$ is an unbiased estimator, then $B(\widehat{\theta}) = 0$.

Note: Bias deals with *accuracy*; i.e., how accurate a point estimator is at estimating the population-level parameter θ. Unbiased estimators are perfectly accurate.

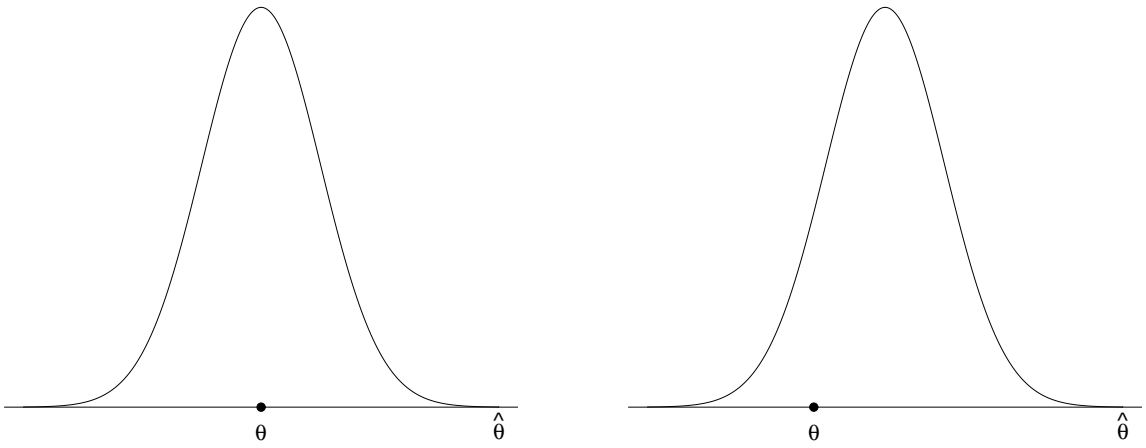

Figure 8.1: Sampling distribution of the point estimator $\widehat{\theta}$. Left: $\widehat{\theta}$ is an unbiased estimator because $E(\widehat{\theta}) = \theta$. Right: $\widehat{\theta}$ is biased because $E(\widehat{\theta}) > \theta$.

Example 8.1. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ population distribution, where $\theta > 0$ is unknown. Consider the two point estimators

$$\begin{aligned} \widehat{\theta}_1 &= 2\overline{Y} \\ \widehat{\theta}_2 &= Y_{(n)}. \end{aligned}$$

Determine whether these point estimators are unbiased.

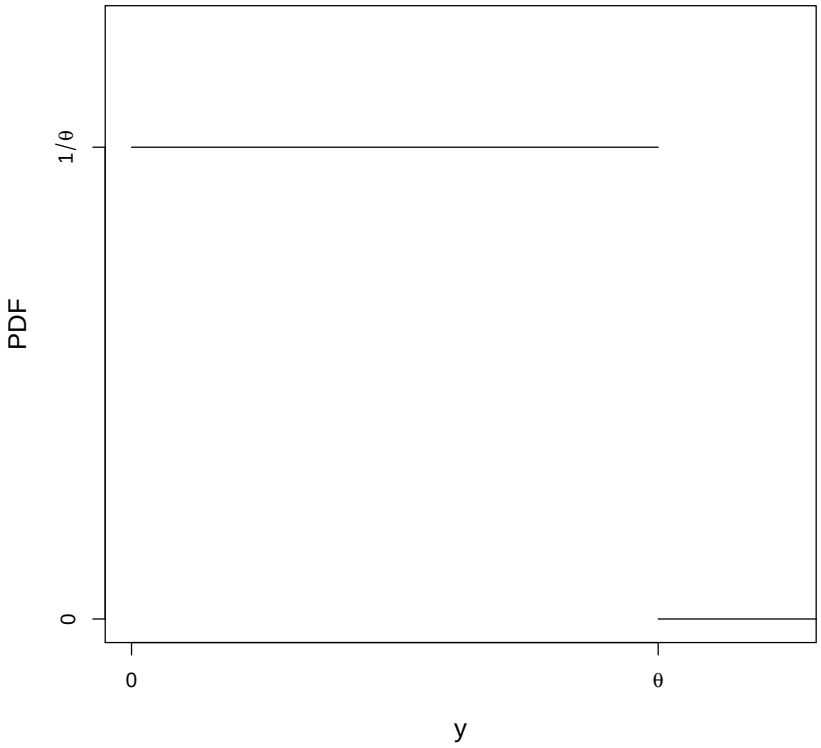

Figure 8.2: Population pdf of $Y \sim \mathcal{U}(0, \theta)$ in Example 8.1.

Solution. The population pdf of $Y \sim \mathcal{U}(0, \theta)$ is shown in Figure 8.2 above. Consider the first point estimator $\widehat{\theta}_1 = 2\overline{Y}$. Note that

$$E(\widehat{\theta}_1) = E(2\overline{Y}) = 2E(\overline{Y}) = 2\left(\frac{\theta}{2}\right) = \theta.$$

Therefore, $\widehat{\theta}_1 = 2\overline{Y}$ is an unbiased estimator of θ.

Note: In the last calculation, we used the fact that $E(\overline{Y}) = \mu$, the population mean. The population mean of $Y \sim \mathcal{U}(0, \theta)$ is $\mu = E(Y) = \theta/2$.

To find $E(\widehat{\theta}_2) = E(Y_{(n)})$, we need to first find the pdf of $Y_{(n)}$, the maximum order statistic. Recall that in general,

$$f_{Y_{(n)}}(y) = n f_Y(y)[F_Y(y)]^{n-1}.$$

The population pdf of $Y \sim \mathcal{U}(0, \theta)$ is

$$f_Y(y) = \begin{cases} \dfrac{1}{\theta}, & 0 < y < \theta \\ 0, & \text{otherwise} \end{cases}$$

and the population cdf is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ \dfrac{y}{\theta}, & 0 < y < \theta \\ 1, & y \geq \theta. \end{cases}$$

Therefore, for $0 < y < \theta$, the pdf of $Y_{(n)}$ is

$$f_{Y_{(n)}}(y) = nf_Y(y)[F_Y(y)]^{n-1} = n\left(\frac{1}{\theta}\right)\left(\frac{y}{\theta}\right)^{n-1} = \frac{ny^{n-1}}{\theta^n}.$$

Summarizing,

$$f_{Y_{(n)}}(y) = \begin{cases} \dfrac{ny^{n-1}}{\theta^n}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

This is called the **power family** pdf; see Exercise 6.17 (pp 309-310, WMS). We can now calculate $E(\widehat{\theta}_2) = E(Y_{(n)})$; note that

$$\begin{aligned} E(Y_{(n)}) = \int_{\mathbb{R}} y f_{Y_{(n)}}(y)dy &= \int_0^\theta y\,\frac{ny^{n-1}}{\theta^n}dy \\ &= \frac{n}{\theta^n}\int_0^\theta y^n dy = \frac{n}{\theta^n}\left(\frac{y^{n+1}}{n+1}\right)\Big|_0^\theta = \frac{n}{\theta^n}\left(\frac{\theta^{n+1}}{n+1}\right) = \left(\frac{n}{n+1}\right)\theta. \end{aligned}$$

Therefore, $\widehat{\theta}_2 = Y_{(n)}$ is a biased estimator of θ. It underestimates θ on average because $n/(n+1) < 1$. $\square$

Exercise: Use the $n = 10$ observed values below to calculate the value of both point estimates:

```
> round(runif(10,0,10),2)
 [1] 0.48 0.07 5.32 5.77 4.52 7.45 7.23 0.31 3.42 4.85
```

The point estimates are

$$\widehat{\theta}_1 = 2\overline{y} = 2(3.94) = 7.88 \quad \text{and} \quad \widehat{\theta}_2 = y_{(10)} = 7.45.$$

Q: Suppose we have two point estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$? Which one should we use? How should we compare them?
A: If both point estimators are unbiased; i.e., if $E(\widehat{\theta}_1) = E(\widehat{\theta}_2) = \theta$, then we would prefer the estimator with the smaller variance.

Note: Whereas bias deals with *accuracy*, the variance of a point estimator describes its *precision*. Small variance means high precision; see Figure 8.3 (next page).

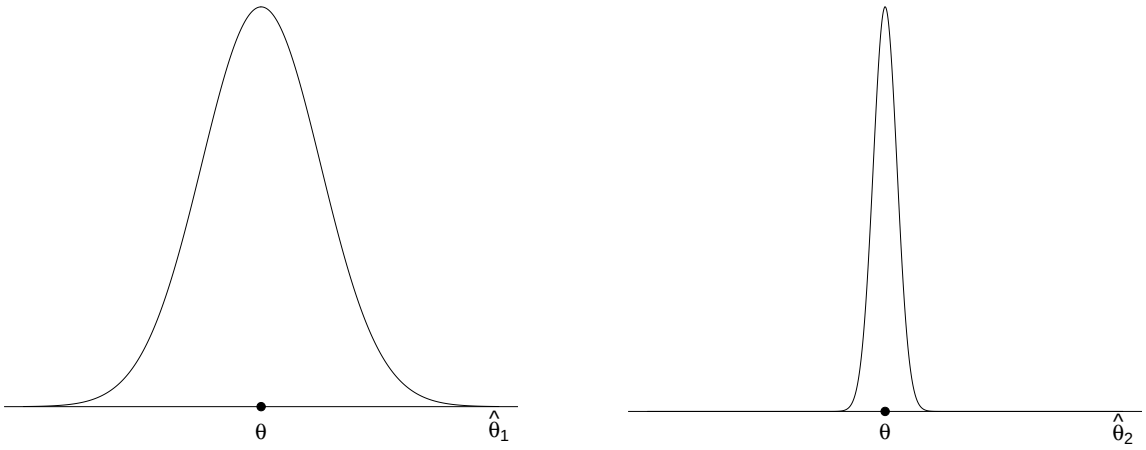

Figure 8.3: Sampling distributions of two unbiased point estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$. The second point estimator (right) has smaller variance (i.e., is more precise).

Example 8.2. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(θ) population distribution, where $\theta > 0$ is unknown. Consider the two point estimators

$$\begin{aligned} \widehat{\theta}_1 &= \overline{Y} \\ \widehat{\theta}_2 &= nY_{(1)}. \end{aligned}$$

Show that both point estimators are unbiased and determine which one has the smaller variance.

Solution. We know that $\widehat{\theta}_1 = \overline{Y}$ is an unbiased estimator of the population mean, which is $\mu = E(Y) = \theta$ under the exponential(θ) model. Therefore, $\widehat{\theta}_1$ is an unbiased estimator of θ; i.e.,

$$E(\widehat{\theta}_1) = E(\overline{Y}) = \theta.$$

To show that $\widehat{\theta}_2$ is unbiased, recall that in Example 6.20 (notes, pp 37-39), we showed

$$Y_1, Y_2, ..., Y_n \sim \text{iid exponential}(\theta) \implies Y_{(1)} \sim \text{exponential}(\theta/n).$$

Therefore,

$$E(\widehat{\theta}_2) = E(nY_{(1)}) = nE(Y_{(1)}) = n\left(\frac{\theta}{n}\right) = \theta.$$

This shows $\widehat{\theta}_2$ is also an unbiased estimator of θ.

Note: Because both point estimators are unbiased, we prefer to use the one with the smaller variance. The variance of $\widehat{\theta}_1 = \overline{Y}$ is given by

$$V(\widehat{\theta}_1) = V(\overline{Y}) = \frac{\theta^2}{n}.$$

Recall: In general, if $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution with variance σ^2, then

$$V(\overline{Y}) = \frac{\sigma^2}{n};$$

we proved this in the last chapter. The population variance of $Y \sim$ exponential(θ) is $\sigma^2 = \theta^2$. Finally, because $Y_{(1)} \sim$ exponential(θ/n), the variance of $\widehat{\theta}_2 = nY_{(1)}$ is

$$V(\widehat{\theta}_2) = V(nY_{(1)}) = n^2V(Y_{(1)}) = n^2\left(\frac{\theta}{n}\right)^2 = \theta^2.$$

Therefore, for all $n > 1$,

$$\frac{\theta^2}{n} = V(\widehat{\theta}_1) < V(\widehat{\theta}_2) = \theta^2.$$

This shows $\widehat{\theta}_1 = \overline{Y}$ is a more precise point estimator than $\widehat{\theta}_2 = nY_{(1)}$. $\square$

Q: How should we compare point estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ if one of them is biased (or perhaps both are biased)?
A: We would prefer the estimator with the smaller mean-squared error.

Terminology: Suppose $\widehat{\theta} = T(Y_1, Y_2, ..., Y_n)$ is a point estimator for the population-level parameter θ. The **mean-squared error** (**MSE**) of $\widehat{\theta}$ is

$$\text{MSE}(\widehat{\theta}) \;=\; E[(\widehat{\theta} - \theta)^2] \;=\; V(\widehat{\theta}) + [B(\widehat{\theta})]^2,$$

where $B(\widehat{\theta}) = E(\widehat{\theta}) - \theta$ is the **bias** of $\widehat{\theta}$ as an estimator of θ. Note that if $\widehat{\theta}$ is an unbiased estimator of θ, then $B(\widehat{\theta}) = 0$ and

$$\text{MSE}(\widehat{\theta}) = V(\widehat{\theta}).$$

Note: In general, the MSE incorporates two components:

- $V(\widehat{\theta})$; this measures **precision**
- $B(\widehat{\theta})$; this measures **accuracy**.

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

Example 8.3. Suppose $Y_1, Y_2, ..., Y_n$ are iid Bernoulli observations with mean p, and let $X = \sum_{i=1}^{n} Y_i$, the sum of the observations; i.e., "the number of successes." Consider the two point estimators of p:

$$\widehat{p}_1 = \frac{X}{n} \quad \text{and} \quad \widehat{p}_2 = \frac{X+2}{n+4}.$$

The point estimator $\widehat{p}_1$ is the usual "sample proportion." The second point estimator arises from "adding two successes and two failures." Compare the estimators on the basis of MSE.

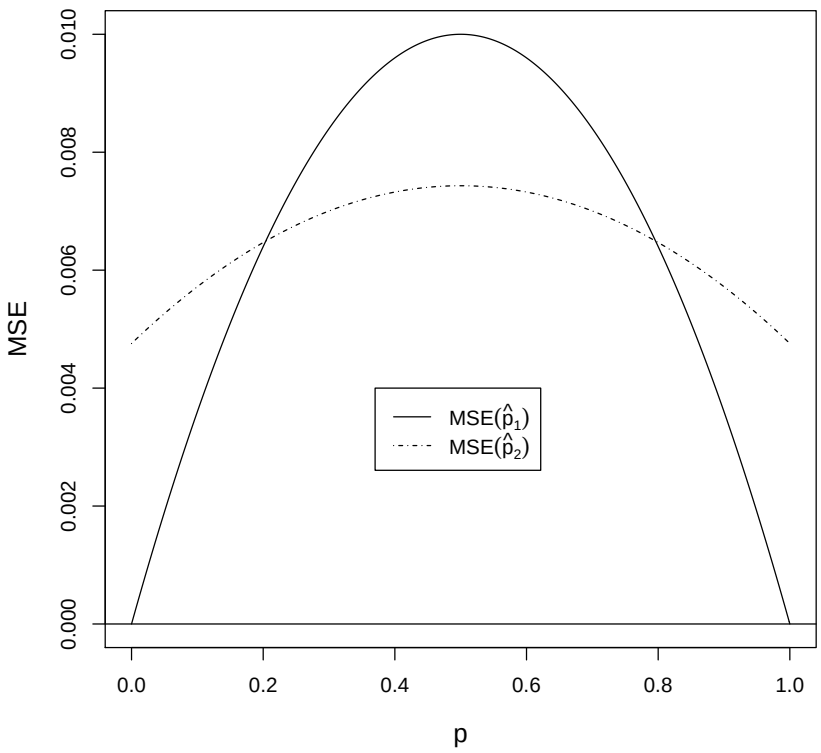

Figure 8.4: Graph of MSE($\widehat{p}_1$) and MSE($\widehat{p}_2$) in Example 8.3 when $n = 25$.

Solution. We know $X \sim b(n, p)$ so $E(X) = np$ and $V(X) = np(1-p)$. The first point estimator $\widehat{p}_1$ is unbiased; note that

$$E(\widehat{p}_1) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{np}{n} = p.$$

Therefore, $B(\widehat{p}_1) = 0$ and the MSE of $\widehat{p}_1$ is equal to the variance of $\widehat{p}_1$; i.e.,

$$\text{MSE}(\widehat{p}_1) = V(\widehat{p}_1) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

The second point estimator $\widehat{p}_2$ is biased; note that

$$E(\widehat{p}_2) = E\left(\frac{X+2}{n+4}\right) = \frac{1}{n+4}[E(X)+2] = \frac{np+2}{n+4}$$

and therefore the bias of $\widehat{p}_2$ is

$$B(\widehat{p}_2) = E(\widehat{p}_2) - p = \frac{np+2}{n+4} - p.$$

The variance of $\widehat{p}_2$ is

$$V(\widehat{p}_2) = V\left(\frac{X+2}{n+4}\right) = \frac{1}{(n+4)^2}V(X+2) = \frac{1}{(n+4)^2}V(X) = \frac{np(1-p)}{(n+4)^2}.$$

Therefore,

$$\text{MSE}(\widehat{p}_2) = V(\widehat{p}_2) + [B(\widehat{p}_2)]^2 = \frac{np(1-p)}{(n+4)^2} + \left(\frac{np+2}{n+4} - p\right)^2.$$

Plots of MSE versus p when $n = 25$ are given in Figure 8.4 (last page); note that $\widehat{p}_2$ is a "better" point estimator when p closer to 0.5. If p is near the extremes (i.e., closer to 0 or 1), then the usual sample proportion $\widehat{p}_1$ is a better point estimator on the basis of MSE. □

Example 8.4. Suppose $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, where both μ and σ^2 are unknown. Consider the two point estimators of σ^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad \text{and} \quad S_b^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Compare the estimators on the basis of MSE.

Solution. From Chapter 7, we know that S^2 is an unbiased estimator of the population variance σ^2 in any population distribution (provided that $\sigma^2 < \infty$). From Result 5 in Chapter 7 (pp 54), we know

$$V(S^2) = \frac{2\sigma^4}{n-1}.$$

Therefore,

$$\text{MSE}(S^2) = V(S^2) = \frac{2\sigma^4}{n-1}.$$

The estimator S_b^2 is biased; note that

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

The bias of S_b^2 is

$$B(S_b^2) = E(S_b^2) - \sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

The variance of S_b^2 is

$$V(S_b^2) = V\left[\left(\frac{n-1}{n}\right)S^2\right] = \left(\frac{n-1}{n}\right)^2 V(S^2) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}.$$

Therefore, the MSE of S_b^2 is

$$\text{MSE}(S_b^2) = V(S_b^2) + [B(S_b^2)]^2 = \frac{2(n-1)\sigma^4}{n^2} + \left(-\frac{\sigma^2}{n}\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4.$$

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

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

Note that the ratio

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

for all $n \geq 2$. Therefore,

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

showing that S_b^2 is a "better" estimator of σ^2 on the basis of MSE. $\square$

Exercise: In Example 8.4, consider point estimators of σ^2 of the form

$$\widehat{\sigma}^2 = c\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

where $c > 0$. Find the value of c that minimizes $\text{MSE}(\widehat{\sigma}^2)$. *Ans:* $c = 1/(n+1)$.

Remark: In some problems, we want to estimate a **function** of a population-level parameter θ, say $\tau(\theta)$. The next two examples illustrate this.

Example 8.5. The number of weekly breakdowns Y at a manufacturing plant follows a Poisson distribution with mean $\theta > 0$. Suppose a random sample of $Y_1, Y_2, ..., Y_n$ of weekly data are available; i.e., $Y_1, Y_2, ..., Y_n$ are iid Poisson(θ). The weekly cost associated with repairing these breakdowns is $C = 3Y + Y^2$ which has expected value

$$\begin{aligned} E(C) = E(3Y + Y^2) &= 3E(Y) + E(Y^2) \\ &= 3E(Y) + \{V(Y) + [E(Y)]^2\} = 3\theta + \theta + \theta^2 = 4\theta + \theta^2. \end{aligned}$$

Find an unbiased estimator of $E(C)$ using the observations $Y_1, Y_2, ..., Y_n$.

Solution. We know $E(\overline{Y}) = \theta$, so it is natural to first try using $4\overline{Y} + \overline{Y}^2$. However, although

$$E(4\overline{Y}) = 4E(\overline{Y}) = 4\theta,$$

unfortunately $\overline{Y}^2$ is not unbiased for θ^2. Note that

$$E(\overline{Y}^2) = V(\overline{Y}) + [E(\overline{Y})]^2 = \frac{\theta}{n} + \theta^2.$$

Therefore, use $\overline{Y}$ as an unbiased estimator of θ and write

$$E\left(\overline{Y}^2 - \frac{\overline{Y}}{n}\right) = E(\overline{Y}^2) - E\left(\frac{\overline{Y}}{n}\right) = \frac{\theta}{n} + \theta^2 - \frac{\theta}{n} = \theta^2.$$

Finally,

$$E\left(4\overline{Y} + \overline{Y}^2 - \frac{\overline{Y}}{n}\right) = E(4\overline{Y}) + E\left(\overline{Y}^2 - \frac{\overline{Y}}{n}\right) = 4\theta + \theta^2.$$

This shows $4\overline{Y} + \overline{Y}^2 - \overline{Y}/n$ is an unbiased estimator of $E(C)$. $\square$

Example 8.6. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from

$$f_Y(y) = \begin{cases} \frac{1}{\theta^2} y e^{-y/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find an unbiased estimator of θ.
(b) Find an unbiased estimator of $\tau(\theta) = 1/\theta$.

Solutions. (a) We recognize $f_Y(y)$ as a gamma pdf with shape parameter $\alpha = 2$ and scale parameter θ; i.e., the population distribution is $Y \sim \text{gamma}(2, \theta)$ so that $\mu = E(Y) = 2\theta$. Therefore,

$$E(\overline{Y}) = 2\theta \implies E\left(\frac{\overline{Y}}{2}\right) = \frac{2\theta}{2} = \theta.$$

This shows $\overline{Y}/2$ is an unbiased estimator of θ.
(b) Because $\overline{Y}/2$ is an unbiased estimator of θ, it is natural to try using $2/\overline{Y}$ as an estimator for $1/\theta$. Note that

$$E\left(\frac{2}{\overline{Y}}\right) = E\left(\frac{2}{\sum_{i=1}^n Y_i/n}\right) = 2nE\left(\frac{1}{T}\right),$$

where $T = \sum_{i=1}^n Y_i$. Now recall $T \sim \text{gamma}(2n, \theta)$; to see why, note that the mgf of T is

$$m_T(t) = [m_Y(t)]^n = \left[\left(\frac{1}{1-\theta t}\right)^2\right]^n = \left(\frac{1}{1-\theta t}\right)^{2n},$$

for $t < 1/\theta$. We recognize $m_T(t)$ as the mgf of $T \sim \text{gamma}(2n, \theta)$. Therefore, the first inverse moment of T is

$$\begin{aligned} E\left(\frac{1}{T}\right) = \int_{\mathbb{R}} \frac{1}{t} f_T(t) dt &= \int_0^\infty \frac{1}{t} \underbrace{\frac{1}{\Gamma(2n)\theta^{2n}} t^{2n-1} e^{-t/\theta}}_{\text{gamma}(2n,\theta)\text{ pdf}} dt \\ &= \frac{1}{\Gamma(2n)\theta^{2n}} \int_0^\infty t^{(2n-1)-1} e^{-t/\theta} dt \\ &= \frac{1}{\Gamma(2n)\theta^{2n}} \Gamma(2n-1)\theta^{2n-1} = \frac{\Gamma(2n-1)\theta^{2n-1}}{(2n-1)\Gamma(2n-1)\theta^{2n}} = \frac{1}{(2n-1)\theta}. \end{aligned}$$

Therefore,

$$E\left(\frac{2}{\overline{Y}}\right) = 2nE\left(\frac{1}{T}\right) = \left(\frac{2n}{2n-1}\right)\frac{1}{\theta},$$

showing that $2/\overline{Y}$ is biased. However, note that

$$E\left(\frac{2}{\overline{Y}}\right) = \left(\frac{2n}{2n-1}\right)\frac{1}{\theta} \implies E\left[\left(\frac{2n-1}{2n}\right)\frac{2}{\overline{Y}}\right] = \left(\frac{2n-1}{2n}\right)\left(\frac{2n}{2n-1}\right)\frac{1}{\theta} = \frac{1}{\theta}.$$

This shows

$$\left(\frac{2n-1}{2n}\right)\frac{2}{\overline{Y}} = \frac{2n-1}{n\overline{Y}} = \frac{2n-1}{T}$$

is an unbiased estimator of $\tau(\theta) = 1/\theta$. $\square$

8.3 Common point estimators and their standard errors

Preview: We examine four settings where the goal to estimate population means and population proportions. We consider one and two populations. In each setting, the resulting point estimators are averages, so we will also be able to describe the approximate (large-sample) sampling distributions of each by using the Central Limit Theorem.

Terminology: Suppose $\widehat{\theta}$ is a point estimator for the population-level parameter θ. The **standard error** of $\widehat{\theta}$ is the standard deviation of $\widehat{\theta}$. We denote the standard error of $\widehat{\theta}$ by

$$\sigma_{\widehat{\theta}} = \sqrt{V(\widehat{\theta})}.$$

The standard error is a measure of variability. It describes numerically how variable the point estimator $\widehat{\theta}$ is in its attempt to estimate θ.

Note: Every point estimator has a standard error. Although we consider simple point estimators in this section, the notion of standard error is important in all statistical analyses, including regression, ANOVA, survival analysis, and others (i.e., basically, anytime we estimate a population-level model).

8.3.1 One population mean

Setting: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution with mean μ and variance $\sigma^2 < \infty$. An unbiased point estimator of μ is

$$\widehat{\mu} = \overline{Y}.$$

The variance of $\overline{Y}$ is

$$V(\overline{Y}) = \frac{\sigma^2}{n} \implies \underbrace{\sigma_{\overline{Y}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}}_{\text{standard error of } \overline{Y}}.$$

Recall from the Central Limit Theorem (CLT) that when the sample size n is large,

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right) \implies Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{AN}(0, 1).$$

From the Empirical Rule (68-95-99.7% Rule), we know

$$P(-2 < Z < 2) = P\left(-2 < \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} < 2\right) = P\left(-2\frac{\sigma}{\sqrt{n}} < \overline{Y} - \mu < 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95.$$

Implication: Suppose the sample size n is large enough for the CLT to apply. When we use $\overline{Y}$ as a point estimator for the population mean μ, it is very likely (i.e., probability approximately 0.95) that the **error in estimation**

$$\epsilon = \overline{Y} - \mu$$

will be within 2 standard errors; see Figure 8.5 (next page).

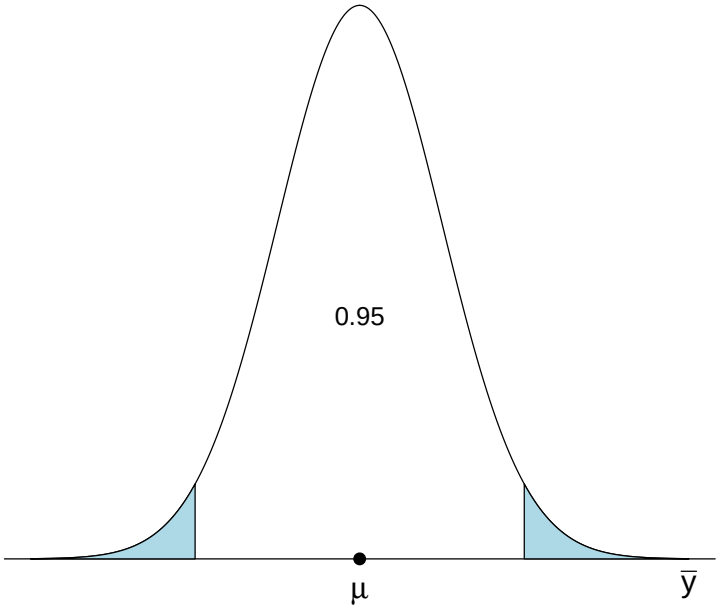

Figure 8.5: (Approximate) sampling distribution of $\overline{Y}$ conferred by the CLT. The middle 95% of the sampling distribution (i.e., $\pm 2\sigma/\sqrt{n}$) is unshaded.

Limitation: The previous result

$$P\left(-2\frac{\sigma}{\sqrt{n}} < \overline{Y} - \mu < 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

is interesting. Unfortunately, if the population standard deviation σ is unknown, then this result is not helpful because we can't calculate the upper and lower bounds

$$\pm 2\sigma_{\overline{Y}} = \pm 2\frac{\sigma}{\sqrt{n}}$$

on the error in estimation. This illustrates a common problem in point estimation, namely, that the standard error of a point estimator (like $\overline{Y}$) depends on population-level parameters that are unknown.

Work-around: If the population standard deviation σ is unknown, then we can not calculate the standard error of the sample mean $\overline{Y}$; i.e.,

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}.$$

However, we can *estimate* it by using

$$\widehat{\sigma}_{\overline{Y}} = \frac{S}{\sqrt{n}},$$

where

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

is the sample standard deviation. We call $\widehat{\sigma}_{\overline{Y}} = S/\sqrt{n}$ the **estimated standard error** of the sample mean $\overline{Y}$. *In general, the estimated standard error is a point estimator of the standard error of a point estimator.*

8.3.2 One population proportion

Setting: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(p) population; i.e., the population pmf of Y is

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

From Example 8.3, an unbiased point estimator of p is the sample proportion

$$\widehat{\theta} = \widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{X}{n},$$

where $X = \sum_{i=1}^{n} Y_i \sim b(n, p)$. The variance of $\widehat{p}$ is

$$V(\widehat{p}) = \frac{p(1-p)}{n} \quad \Longrightarrow \quad \underbrace{\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}}_{\text{standard error of } \widehat{p}}.$$

Furthermore, from the CLT, we know

$$\widehat{p} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right) \quad \Longrightarrow \quad Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{AN}(0, 1).$$

for large n. From the Empirical Rule, we have

$$\begin{aligned} P(-2 < Z < 2) &= P\left(-2 < \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < 2\right) \\ &= P\left(-2\sqrt{\frac{p(1-p)}{n}} < \widehat{p} - p < 2\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95. \end{aligned}$$

Implication: Suppose the sample size n is large enough for the CLT to apply. When we use $\widehat{p}$ as a point estimator for the population proportion p, it is very likely (i.e., probability approximately 0.95) that the **error in estimation**

$$\epsilon = \widehat{p} - p$$

will be within 2 standard errors; see Figure 8.6 (next page).

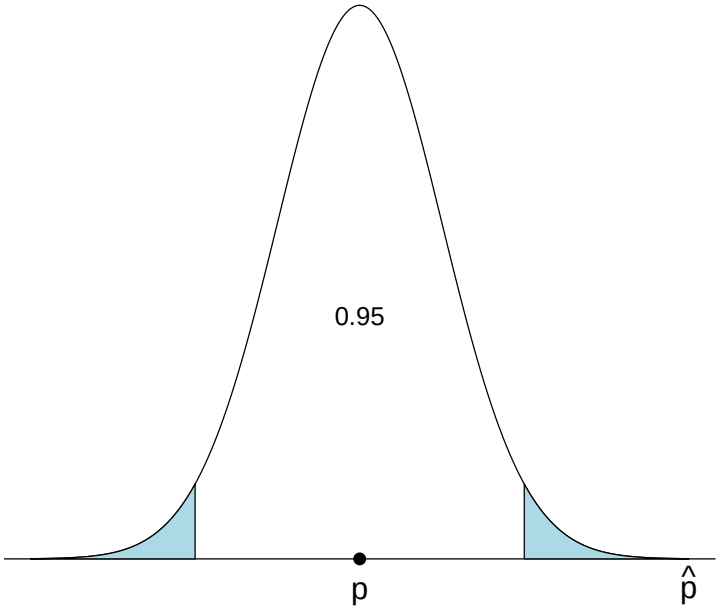

Figure 8.6: (Approximate) sampling distribution of $\widehat{p}$ conferred by the CLT. The middle 95% of the sampling distribution (i.e., $\pm 2\sqrt{p(1-p)/n}$) is unshaded.

Note: We encounter the same problem with the standard error as we did in the one population mean problem. The standard error of $\widehat{p}$ is

$$\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}},$$

which is unknown because it depends on the population-level parameter p. The **estimated standard error** of $\widehat{p}$ as an estimator of p is

$$\widehat{\sigma}_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

8.3.3 Difference of two population means (independent samples)

Setting: Suppose we have two independent random samples:

- $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an iid sample from a population with mean μ_1 and variance σ_1^2
- $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an iid sample from a population with mean μ_2 and variance σ_2^2.

The goal is to estimate the parameter $\theta = \mu_1 - \mu_2$, the difference of the population means.

Define the sample means

$$\overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad \overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j}$$

and the sample variances

$$S_1^2 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{1j} - \overline{Y}_{1+})^2 \quad \text{and} \quad S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_{2j} - \overline{Y}_{2+})^2.$$

An unbiased point estimator of $\theta = \mu_1 - \mu_2$ is

$$\widehat{\theta} = \overline{Y}_{1+} - \overline{Y}_{2+},$$

the difference of the sample means. The variance of $\widehat{\theta}$ is

$$V(\widehat{\theta}) = V(\overline{Y}_{1+} - \overline{Y}_{2+}) = V(\overline{Y}_{1+}) + V(\overline{Y}_{2+}) - 2\underbrace{\text{Cov}(\overline{Y}_{1+}, \overline{Y}_{2+})}_{=0} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\Longrightarrow \quad \underbrace{\sigma_{\overline{Y}_{1+}-\overline{Y}_{2+}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}_{\text{standard error of } \overline{Y}_{1+} - \overline{Y}_{2+}}.$$

To estimate the standard error, we use

$$\widehat{\sigma}_{\overline{Y}_{1+}-\overline{Y}_{2+}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

This estimated standard error uses the sample variances S_1^2 and S_2^2 as (unbiased) point estimators of the population variances σ_1^2 and σ_2^2, respectively.

8.3.4 Difference of two population proportions (independent samples)

Setting: Suppose we have two independent random samples:

- $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an iid sample from a Bernoulli(p_1) population
- $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an iid sample from a Bernoulli(p_2) population.

The goal is to estimate the parameter $\theta = p_1 - p_2$, the difference of the population proportions. Define

$$X_1 = \sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad X_2 = \sum_{j=1}^{n_2} Y_{2j}$$

so that $X_1 \sim b(n_1, p_1)$, $X_2 \sim b(n_2, p_2)$, and $X_1 \perp\!\!\!\perp X_2$ (because the samples are independent).

An unbiased point estimator of $\theta = p_1 - p_2$ is

$$\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2},$$

the difference of the sample proportions. The variance of $\widehat{\theta}$ is

$$V(\widehat{\theta}) = V(\widehat{p}_1 - \widehat{p}_2) = V(\widehat{p}_1) + V(\widehat{p}_2) - 2\underbrace{\text{Cov}(\widehat{p}_1, \widehat{p}_2)}_{=0} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

$$\Longrightarrow \quad \underbrace{\sigma_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}_{\text{standard error of } \widehat{p}_1 - \widehat{p}_2}.$$

To estimate the standard error, we use

$$\widehat{\sigma}_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

This estimated standard error uses the sample proportions $\widehat{p}_1$ and $\widehat{p}_2$ as (unbiased) point estimators of the population proportions p_1 and p_2, respectively.

8.3.5 Summary

Note: Below is a summary of the point estimators $\widehat{\theta}$ in this section, their standard errors $\sigma_{\widehat{\theta}}$, and their estimated standard errors $\widehat{\sigma}_{\widehat{\theta}}$.

Parameter θ	Estimator $\widehat{\theta}$	Standard error $\sigma_{\widehat{\theta}}$	Estimated standard error $\widehat{\sigma}_{\widehat{\theta}}$
μ	$\overline{Y}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{S}{\sqrt{n}}$
p	$\widehat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
$\mu_1 - \mu_2$	$\overline{Y}_{1+} - \overline{Y}_{2+}$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
$p_1 - p_2$	$\widehat{p}_1 - \widehat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$

Point estimators in this section are "averages" or functions of averages so the CLT applies when the sample size(s) is (are) large; i.e.,

$$\widehat{\theta} \sim \mathcal{AN}(\theta, \sigma^2_{\widehat{\theta}}).$$

This means the point estimator $\widehat{\theta}$ should be within two standard errors (i.e., $\pm 2\sigma_{\widehat{\theta}}$) of the population parameter θ with probability approximately equal to 0.95. This is conferred by the Empirical (68-95-99.7%) Rule.

8.4 Confidence intervals

Recall: A point estimator of the population-level parameter θ is any statistic

$$\widehat{\theta} = T(Y_1, Y_2, ..., Y_n)$$

that estimates θ. After the values of $Y_1, Y_2, ..., Y_n$ have been observed, the point estimate $\widehat{\theta} = T(y_1, y_2, ..., y_n)$ is a single number; i.e., it is a "one-shot guess" at the true value of θ.

Example 8.7. Cox *and others* (2015) describe a retrospective study that observed $n = 615$ newborns who were admitted to the neonatal intensive care unit at Richland County Hospital in Columbia, SC. The random variable

$$Y = \text{ birth weight (measured in grams)}$$

was observed on each newborn. Suppose $Y_1, Y_2, ..., Y_{615}$ are regarded as iid from a population distribution with mean μ and variance σ^2. The observed data $y_1, y_2, ..., y_{615}$ are shown in Figure 8.7 (next page). To estimate the population mean μ with the observed data, we can use the **sample mean**

$$\overline{y} = \frac{1}{615}\sum_{i=1}^{615} y_i \approx 2137 \text{ grams}.$$

To estimate the population variance σ^2, we can use the **sample variance**

$$s^2 = \frac{1}{615-1}\sum_{i=1}^{615}(y_i - \overline{y})^2 \approx 981173 \text{ (grams)}^2.$$

Both of these point estimates can be calculated in R as follows:

```
> mean(birth.weights)
[1] 2137.237
> var(birth.weights)
[1] 981173
```

Discussion: Point estimates like $\overline{y}$ and s^2 do not account for variability; their values are single numbers. Therefore, although these estimates are both unbiased, we do not know how variable these estimates are. What we would like to know is how "close" these estimates are to the population-level values μ and σ^2, respectively. Unfortunately, without a measure of variability attached to either estimate, it is impossible to tell.

Interesting: If Y is a continuous random variable, then $\overline{Y}$ is too and therefore

$$P(\overline{Y} = \mu) = 0.$$

This is true because continuous random variables assign zero probability to specific values. Similarly, $P(S^2 = \sigma^2) = 0$. Mathematically, this illustrates the futility of using only point estimators in our quest to learn about population-level parameters.

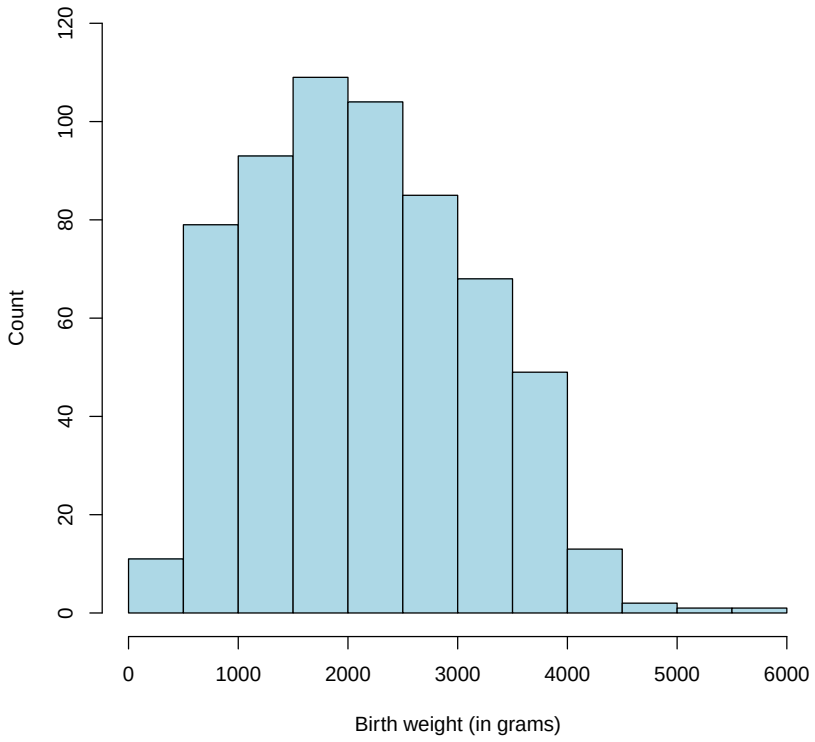

Figure 8.7: Newborn data. Birth weight (in grams) measured for $n = 615$ newborns in Richland County Hospital.

Terminology: A $1-\alpha$ **interval estimator** is an interval (θ_L, θ_U) that contains a population-level parameter θ with probability $1-\alpha$; i.e.,

$$P(\theta_L \leq \theta \leq \theta_U) = 1 - \alpha.$$

A $1-\alpha$ interval estimator is also called a $100(1-\alpha)\%$ **confidence interval**. The probability $1-\alpha$ is called the **confidence coefficient** associated with the interval.

Remark: It is important to understand that when we write $P(\theta_L \leq \theta \leq \theta_U) = 1 - \alpha$ in the definition above, it is the endpoints of θ_L and θ_U that are random quantities; not θ. As usual, we assume the population-level parameter θ is an unknown quantity; however, it is fixed (not random). We will discuss this point more in the examples.

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution described by $p_Y(y)$ or $f_Y(y)$, and let θ denote a population-level parameter. We call

$$Q = Q(Y_1, Y_2, ..., Y_n; \theta)$$

a **pivotal quantity** (or **pivot**) if the distribution of Q does not depend on θ (or any other population-level parameters that are unknown).

Remark: Pivotal quantities are useful when deriving confidence intervals. In the continuous case, suppose $Q = Q(Y_1, Y_2, ..., Y_n; \theta)$ is a pivot whose distribution is described by the pdf $f_Q(q)$, which is free of θ. Define

$$\begin{aligned} q_{1-\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } f_Q(q) \\ q_{\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } f_Q(q). \end{aligned}$$

Because $Q \sim f_Q(q)$, we can first write

$$P(q_{1-\alpha/2} \leq Q \leq q_{\alpha/2}) = 1 - \alpha.$$

We can then rewrite the event $\{q_{1-\alpha/2} \leq Q \leq q_{\alpha/2}\}$, often using straightforward algebra, to derive a $100(1-\alpha)\%$ confidence interval (θ_L, θ_U); i.e., we can write

$$1 - \alpha = P(q_{1-\alpha/2} \leq Q \leq q_{\alpha/2}) = P(\theta_L \leq \theta \leq \theta_U),$$

where the endpoints θ_L and θ_U will depend on the quantiles $q_{1-\alpha/2}$ and $q_{\alpha/2}$ and the sample $Y_1, Y_2, ..., Y_n$ through well-known statistics. We illustrate this technique using examples.

Example 8.8. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma_0^2)$ population distribution, where the population variance σ_0^2 is known. Our goal is to derive a $100(1-\alpha)\%$ confidence interval for the population mean μ. We start with

$$Q = \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Note that Q is a pivot because its distribution is free of μ. Define

$$\begin{aligned} -z_{\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } \mathcal{N}(0,1) \\ z_{\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } \mathcal{N}(0,1) \end{aligned}$$

and refer to Figure 8.8 (next page). Because $Q \sim \mathcal{N}(0,1)$, we can write

$$\begin{aligned} 1 - \alpha = P\left(-z_{\alpha/2} < \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} < z_{\alpha/2}\right) &= P\left(-z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} < \overline{Y} - \mu < z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\ &= P\left(z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} > \mu - \overline{Y} > -z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\ &= P\left(\overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} > \mu > \overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\ &= P\Bigg(\underbrace{\overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}}_{\mu_L} < \mu < \underbrace{\overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}}_{\mu_U}\Bigg). \end{aligned}$$

Therefore,

$$\left(\overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}},\ \overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for μ. $\square$

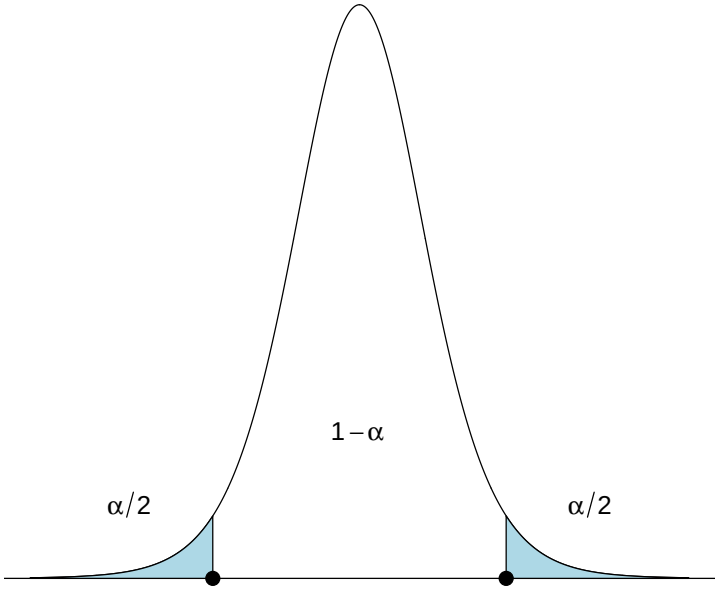

Figure 8.8: $\mathcal{N}(0,1)$ pdf. The lower $\alpha/2$ quantile $-z_{\alpha/2}$ and the upper $\alpha/2$ quantile $z_{\alpha/2}$ are shown by using dark circles.

Example 8.9. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(β) population distribution. Our goal is to derive a $100(1-\alpha)\%$ confidence interval for β. We start with

$$Q = \frac{2T}{\beta} \sim \chi^2(2n)$$

where $T = \sum_{i=1}^{n} Y_i$ is the sample sum. Note that Q is a pivot because its distribution is free of β. Define

$$\begin{aligned} \chi^2_{2n,1-\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } \chi^2(2n) \\ \chi^2_{2n,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } \chi^2(2n) \end{aligned}$$

and refer to Figure 8.9 (next page). Because $Q \sim \chi^2(2n)$, we can write

$$\begin{aligned} 1-\alpha = P\left(\chi^2_{2n,1-\alpha/2} < \frac{2T}{\beta} < \chi^2_{2n,\alpha/2}\right) &= P\left(\frac{1}{\chi^2_{2n,1-\alpha/2}} > \frac{\beta}{2T} > \frac{1}{\chi^2_{2n,\alpha/2}}\right) \\ &= P\left(\frac{2T}{\chi^2_{2n,1-\alpha/2}} > \beta > \frac{2T}{\chi^2_{2n,\alpha/2}}\right) \\ &= P\Bigg(\underbrace{\frac{2T}{\chi^2_{2n,\alpha/2}}}_{\beta_L} < \beta < \underbrace{\frac{2T}{\chi^2_{2n,1-\alpha/2}}}_{\beta_U}\Bigg). \end{aligned}$$

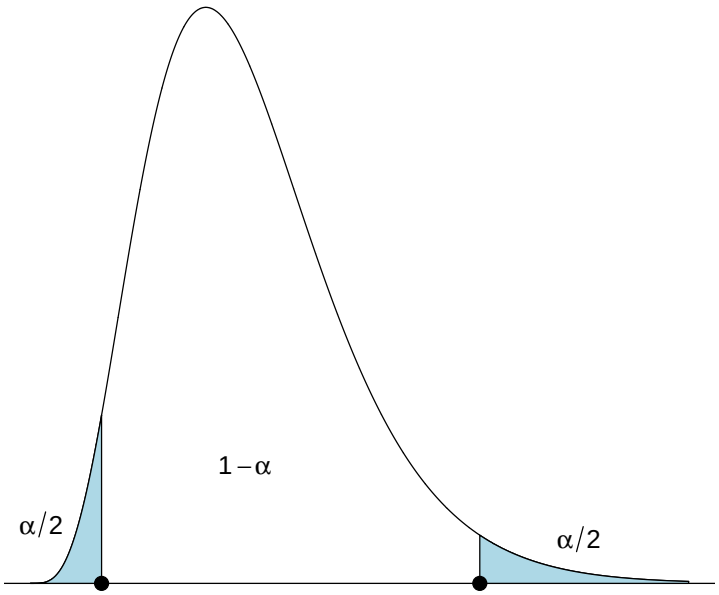

Figure 8.9: $\chi^2(2n)$ pdf. The lower $\alpha/2$ quantile $\chi^2_{2n,1-\alpha/2}$ and the upper $\alpha/2$ quantile $\chi^2_{2n,\alpha/2}$ are shown by using dark circles.

Therefore,

$$\left(\frac{2T}{\chi^2_{2n,\alpha/2}}, \frac{2T}{\chi^2_{2n,1-\alpha/2}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for β.

Illustration: Recall the TTF data in Example 6.19 (notes, pp 35) for $n = 14$ patients:

0.8 7.5 13.4 1.4 0.5 68.9 16.1 20.4 15.6 4.2 2.4 8.2 5.3 14.0

Suppose these data are modeled as iid observations from an exponential(β) population distribution. I calculated a 95% confidence interval for β in R:

```
> ttf = c(0.8,7.5,13.4,1.4,0.5,68.9,16.1,20.4,15.6,4.2,2.4,8.2,5.3,14.0)
> ci.lower = 2*sum(ttf)/qchisq(0.975,28)
> ci.upper = 2*sum(ttf)/qchisq(0.025,28)
> round(c(ci.lower,ci.upper),1)
[1]  8.0 23.3
```

Interpretation: We are 95% confident that the population mean TTF β is between 8.0 and 23.3 months. □

Discussion: In Example 8.9, we derived a $100(1-\alpha)\%$ confidence interval for an exponential (population) mean β to be

$$\left(\frac{2T}{\chi^2_{2n,\alpha/2}}, \frac{2T}{\chi^2_{2n,1-\alpha/2}} \right)$$

from the probability equation

$$1-\alpha = P\left(\frac{2T}{\chi^2_{2n,\alpha/2}} < \beta < \frac{2T}{\chi^2_{2n,1-\alpha/2}} \right),$$

where $T = \sum_{i=1}^{n} Y_i$. The probability above is a bona fide probability because $Y_1, Y_2, ..., Y_n$ are random variables and hence T is also random. Therefore,

$$\left\{ \frac{2T}{\chi^2_{2n,\alpha/2}} < \beta < \frac{2T}{\chi^2_{2n,1-\alpha/2}} \right\}$$

is a random event, one to which we can assign probability. However, once we used the realizations $y_1, y_2, ..., y_{14}$ in the TTF example to calculate $(8.0, 23.3)$ as a 95% confidence interval, it is no longer mathematically appropriate to write

$$0.95 = P(8.0 < \beta < 23.3).$$

The population parameter β is regarded as **fixed**; therefore, the event $\{8.0 < \beta < 23.3\}$ is not random and we do not assign probability to events that are not random.

Q: How do we interpret confidence intervals that are calculated from observed data?
A: We are left with the following relative frequency interpretation:

> "Over the long run, that is, provided that we could sample from the population distribution over and over again, each time by using the same sample size, we would expect $100(1-\alpha)\%$ of the calculated intervals to include the population parameter. The observed interval is just one of these many hypothetical intervals."

Example 8.10. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ population distribution, where $\theta > 0$ is unknown. Our goal is to derive a $100(1-\alpha)\%$ confidence interval for θ. In Example 8.1 (notes), we derived the pdf of the maximum order statistic $Y_{(n)}$ to be

$$f_{Y_{(n)}}(y) = \begin{cases} \dfrac{ny^{n-1}}{\theta^n}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, $Y_{(n)}$ is not a pivotal quantity because its distribution depends on θ. However,

$$Q = \frac{Y_{(n)}}{\theta}$$

is a pivotal quantity. To see why, let's first find the cdf of $Y_{(n)}$. For $0 < y < \theta$, we have

$$F_{Y_{(n)}}(y) = P(Y_{(n)} \leq y) = \int_0^y f_{Y_{(n)}}(t)dt = \int_0^y \frac{nt^{n-1}}{\theta^n}dt = \frac{1}{\theta^n}\left(t^n\big|_0^y\right) = \left(\frac{y}{\theta}\right)^n.$$

Summarizing,

$$F_{Y_{(n)}}(y) = \begin{cases} 0 & y \leq 0 \\ \left(\frac{y}{\theta}\right)^n, & 0 < y < \theta \\ 1, & y \geq \theta. \end{cases}$$

We now use the cdf technique to find the distribution of Q. Note that

$$0 < y_{(n)} < \theta \iff q = h(y_{(n)}) = \frac{y_{(n)}}{\theta} \in (0, 1).$$

Therefore, the support of $Q = h(Y_{(n)}) = Y_{(n)}/\theta$ is

$$R_Q = \{q : 0 < q < 1\}.$$

For $0 < q < 1$, the cdf of Q is

$$F_Q(q) = P(Q \leq q) = P\left(\frac{Y_{(n)}}{\theta} \leq q\right) = P(Y_{(n)} \leq q\theta) = F_{Y_{(n)}}(q\theta) = \left(\frac{q\theta}{\theta}\right)^n = q^n.$$

Summarizing,

$$F_Q(q) = \begin{cases} 0 & q \leq 0 \\ q^n, & 0 < q < 1 \\ 1, & q \geq 1. \end{cases}$$

Therefore, Q is a pivotal quantity because its distribution does not depend on θ. Taking derivatives, the pdf of Q is

$$f_Q(q) = \begin{cases} nq^{n-1}, & 0 < q < 1 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize $f_Q(q)$ as a beta pdf with $\alpha = n$ (the sample size) and $\beta = 1$. Define

$$\begin{aligned} b_{n,1,1-\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of beta}(n, 1) \\ b_{n,1,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of beta}(n, 1) \end{aligned}$$

and refer to Figure 8.10 (next page). Because $Q \sim \text{beta}(n, 1)$, we can write

$$\begin{aligned} 1 - \alpha = P\left(b_{n,1,1-\alpha/2} < \frac{Y_{(n)}}{\theta} < b_{n,1,\alpha/2}\right) &= P\left(\frac{1}{b_{n,1,1-\alpha/2}} > \frac{\theta}{Y_{(n)}} > \frac{1}{b_{n,1,\alpha/2}}\right) \\ &= P\left(\frac{Y_{(n)}}{b_{n,1,1-\alpha/2}} > \theta > \frac{Y_{(n)}}{b_{n,1,\alpha/2}}\right) \\ &= P\Bigg(\underbrace{\frac{Y_{(n)}}{b_{n,1,\alpha/2}}}_{\theta_L} < \theta < \underbrace{\frac{Y_{(n)}}{b_{n,1,1-\alpha/2}}}_{\theta_U}\Bigg). \end{aligned}$$

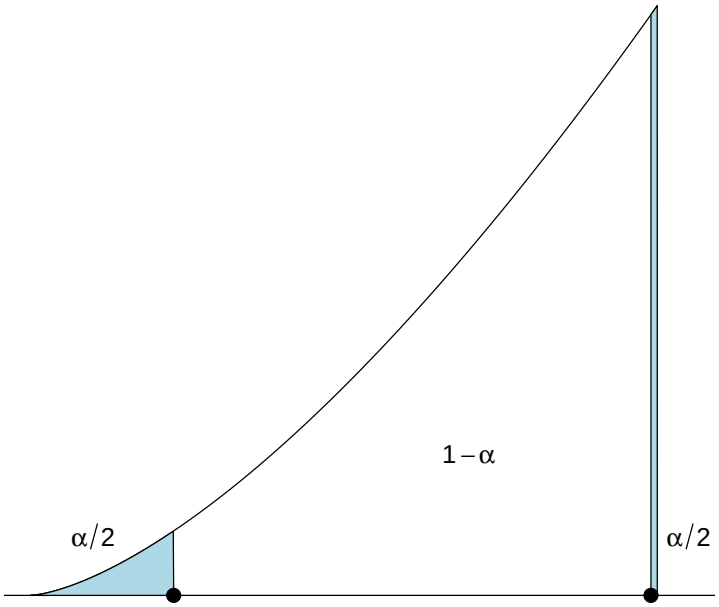

Figure 8.10: Beta$(n, 1)$ pdf. The lower $\alpha/2$ quantile $b_{n,1,1-\alpha/2}$ and the upper $\alpha/2$ quantile $b_{n,1,\alpha/2}$ are shown by using dark circles.

Therefore,

$$\left(\frac{Y_{(n)}}{b_{n,1,\alpha/2}}, \frac{Y_{(n)}}{b_{n,1,1-\alpha/2}} \right)$$

is a $100(1-\alpha)\%$ confidence interval for θ.

8.5 Large-sample confidence intervals

Note: We now revisit the four settings described in Section 8.3 (notes); i.e., estimating population means and population proportions for one and two populations. Our goal now is to write $100(1-\alpha)\%$ confidence intervals for each population parameter:

$$\begin{aligned} \mu &\longleftarrow \text{population mean} \\ p &\longleftarrow \text{population proportion} \\ \mu_1 - \mu_2 &\longleftarrow \text{difference of two population means (independent samples)} \\ p_1 - p_2 &\longleftarrow \text{difference of two population proportions (independent samples).} \end{aligned}$$

In each setting, we presented an unbiased estimator $\widehat{\theta}$ that satisfied

$$\widehat{\theta} \sim \mathcal{AN}(\theta, \sigma^2_{\widehat{\theta}})$$

for large sample sizes; this was conferred by the CLT. We can use this result to construct **large-sample** $100(1-\alpha)\%$ **confidence intervals**. By "large-sample," we mean intervals whose confidence coefficient is approximately $1-\alpha$ when n (or n_1 and n_2) is (are) large.

Derivation: Suppose θ is a population-level parameter. If the point estimator $\widehat{\theta}$ satisfies $\widehat{\theta} \sim \mathcal{AN}(\theta, \sigma^2_{\widehat{\theta}})$, then

$$Q = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{AN}(0,1).$$

The large-sample distribution of Q is free of θ; therefore, Q is a **large-sample pivot**. Define

$$\begin{aligned} -z_{\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } \mathcal{N}(0,1) \\ z_{\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } \mathcal{N}(0,1) \end{aligned}$$

and refer to Figure 8.8 (notes, pp 93). Because $Q \sim \mathcal{AN}(0,1)$, we can write

$$\begin{aligned} 1-\alpha \approx P\left(-z_{\alpha/2} < \frac{\widehat{\theta}-\theta}{\sigma_{\widehat{\theta}}} < z_{\alpha/2}\right) &= P\left(-z_{\alpha/2}\sigma_{\widehat{\theta}} < \widehat{\theta} - \theta < z_{\alpha/2}\sigma_{\widehat{\theta}}\right) \\ &= P\left(z_{\alpha/2}\sigma_{\widehat{\theta}} > \theta - \widehat{\theta} > -z_{\alpha/2}\sigma_{\widehat{\theta}}\right) \\ &= P\left(\widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}} > \theta > \widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}}\right) \\ &= P\left(\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}} < \theta < \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}\right). \end{aligned}$$

This argument shows

$$\left(\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}},\ \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}\right)$$

is a large-sample $100(1-\alpha)\%$ confidence interval for θ.

Problem: Although the preceding interval is a bona fide large-sample interval, the problem is the endpoints depend on the standard error $\sigma_{\widehat{\theta}}$, which depends on unknown population parameters. Therefore, this interval cannot be calculated. Recall the four settings discussed in Section 8.3 (notes) and the associated standard errors:

Parameter θ	Estimator $\widehat{\theta}$	Standard error $\sigma_{\widehat{\theta}}$	Estimated standard error $\widehat{\sigma}_{\widehat{\theta}}$
μ	$\overline{Y}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{S}{\sqrt{n}}$
p	$\widehat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
$\mu_1 - \mu_2$	$\overline{Y}_{1+} - \overline{Y}_{2+}$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
$p_1 - p_2$	$\widehat{p}_1 - \widehat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$

Work-around: In the confidence interval

$$\left(\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}},\ \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}\right),$$

replace the standard error $\sigma_{\widehat{\theta}}$ with the **estimated standard error** $\widehat{\sigma}_{\widehat{\theta}}$ and use

$$\left(\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}},\ \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}\right)$$

instead. This "work-around" has theoretical justification, but only when the sample size(s) is (are) large. We will understand why when we discuss the theoretical notion of **consistency** in Chapter 9. Informally, consistency guarantees

$$\sigma_{\widehat{\theta}} \approx \widehat{\sigma}_{\widehat{\theta}},$$

that is, the standard error and estimated standard error are approximately equal when the sample size(s) is (are) large. Therefore, the confidence coefficient of $(\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}},\ \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}})$ and the one of $(\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}},\ \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}})$ are approximately equal.

Example 8.11. Consider the newborn data in Example 8.7 (notes, pp 90-91). A large-sample 95% confidence interval for the population mean birth weight μ is

$$\overline{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \longrightarrow 2137 \pm 1.96\left(\frac{991}{\sqrt{615}}\right) \longrightarrow (2059, 2215).$$

```
> mean(birth.weights) # sample mean
[1] 2137.237
> sd(birth.weights) # sample standard deviation
[1] 990.5418
> qnorm(0.975,0,1) # upper 0.025 quantile from N(0,1)
[1] 1.959964
```

Interpretation: We are (approximately) 95% confident the population mean birth weight μ is between 2059 and 2215 grams. □

Example 8.12. Consider the PRAMS surveillance project data in Example 7.6 (notes, pp 71-72) where 125 women (out of $n = 999$) smoked during the last 3 months of their pregnancy. A large-sample 95% confidence interval for p, the population proportion of women who smoked during the last 3 months of their pregnancy, is given by

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \longrightarrow \frac{125}{999} \pm 1.96\sqrt{\frac{\frac{125}{999}(1-\frac{125}{999})}{999}} \longrightarrow (0.105, 0.146).$$

Interpretation: We are (approximately) 95% confident the population proportion of women who smoked during the last 3 months of their pregnancy is between 0.105 and 0.146. □

8.6 Sample size determination

Remark: We now discuss one of the most basic fundamental questions when designing a study, namely, *how many individuals need to be sampled*? We will answer this question in the context of writing large-sample confidence intervals for population means and proportions.

Setting: Suppose θ is a population-level parameter. If the point estimator $\widehat{\theta}$ satisfies $\widehat{\theta} \sim \mathcal{AN}(\theta, \sigma^2_{\widehat{\theta}})$, then we showed in the last section

$$1 - \alpha \approx P\left(\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}} < \theta < \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}\right).$$

Write the endpoints as

$$\widehat{\theta} \pm z_{\alpha/2}\sigma_{\widehat{\theta}}$$

and set

$$B = z_{\alpha/2}\sigma_{\widehat{\theta}},$$

the **margin of error** of the interval $(\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}},\ \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}})$. The last equation suggests that if we

- specify the confidence coefficient $1 - \alpha$
- specify the margin of error B,

we can determine the sample size n that satisfies $B = z_{\alpha/2}\sigma_{\widehat{\theta}}$. Note that for this approach to work, we will need to elicit "guesses" for any population-level parameters in the standard error $\sigma_{\widehat{\theta}}$. We now illustrate this approach.

Population mean: We set

$$B = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Solving this equation for n, we get

$$n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2.$$

Note that for this formula to be useful, we must elicit a "guess" of the population-level standard deviation σ (which, in general, will be unknown).

Example 8.13. In a biomedical experiment, we would like to estimate the population mean remaining lifetime μ of healthy rats given a high dose of a toxic substance. We would like to write a 95% confidence interval for μ with a margin of error of $B = 2$ days. Suppose the population standard deviation of the remaining lifetime distribution is $\sigma = 8$ days. How many rats should we use for the experiment?

Solution. From the specifications elicited, we have $z_{0.05/2} = z_{0.025} \approx 1.96$, $B = 2$, and $\sigma = 8$. These specifications lead to the sample size

$$n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2 = \left(\frac{1.96 \times 8}{2}\right)^2 \approx 61.5.$$

Therefore, we would need $n = 62$ rats to achieve these specifications.

Tighter specifications: 99% confidence $\rightarrow z_{0.01/2} = z_{0.005} \approx 2.58$, margin of error $B = 1$, and $\sigma = 8$. These specifications lead to the sample size

$$n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2 = \left(\frac{2.58 \times 8}{1}\right)^2 \approx 426.0.$$

We would need $n = 426$ rats to achieve these specifications.

Weaker specifications: 90% confidence $\rightarrow z_{0.10/2} = z_{0.05} \approx 1.65$, margin of error $B = 3$, and $\sigma = 8$. These specifications lead to the sample size

$$n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2 = \left(\frac{1.65 \times 8}{3}\right)^2 \approx 19.4.$$

We would need $n = 20$ rats to achieve these specifications. □

Population proportion: We set

$$B = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Here, we see the margin of error depends on the population proportion p (which is the parameter we want to write the confidence interval for). Therefore, to determine the necessary sample size, we must first elicit a "guess" for p, say p_0. We then set

$$B = z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} \quad \Longrightarrow \quad n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p_0(1-p_0).$$

Note: If there is no sensible guess for p available, it is best to use $p_0 = 0.5$. The resulting value for n will be as large as possible because

$$n = n(p_0) = \left(\frac{z_{\alpha/2}}{B}\right)^2 p_0(1-p_0),$$

when viewed as a function of p_0, is maximized when $p_0 = 0.5$.

Example 8.14. In a Phase II clinical trial, it is estimated the proportion of patients responding to a drug is $p_0 = 0.35$. Physicians would like to know how many patients they should recruit for a larger Phase III trial. Their resulting 95% confidence interval for p, the population proportion of patients responding to the drug, should have a margin of error no greater than $B = 0.03$. What sample size do they need for the Phase III trial?

Solution. From the specifications elicited, we have $z_{0.05/2} = z_{0.025} \approx 1.96$, $B = 0.03$, and $p_0 = 0.35$. These specifications lead to the sample size

$$n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p_0(1-p_0) = \left(\frac{1.96}{0.03}\right)^2 (0.35)(1-0.35) \approx 971.1.$$

Therefore, their Phase III trial should recruit $n = 972$ patients. □

8.7 Confidence intervals arising from normal populations

Preview: We derive confidence intervals for means and variances when the population distribution is $\mathcal{N}(\mu, \sigma^2)$. We consider one and two populations. To derive the intervals, we use sampling distribution results in Section 7.3 and the t and F distributions in Section 7.4.

8.7.1 Population mean μ

Recall: In Example 8.8 (notes, pp 92), we showed that if $Y_1, Y_2, ..., Y_n$ was an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution (σ^2 **known**), then

$$\left(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for the population mean μ. We derived this interval from the pivotal quantity

$$Q = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

When the population variance σ^2 is unknown, we use

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

instead and derive the confidence interval for μ in the analogous way. Define

$$\begin{aligned} -t_{n-1,\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } t(n-1) \\ t_{n-1,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } t(n-1) \end{aligned}$$

and refer to Figure 8.11 (next page). Because $T \sim t(n-1)$, we can write

$$\begin{aligned} 1-\alpha = P\left(-t_{n-1,\alpha/2} < \frac{\overline{Y}-\mu}{S/\sqrt{n}} < t_{n-1,\alpha/2}\right) &= P\left(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \overline{Y} - \mu < t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\ &= P\left(t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} > \mu - \overline{Y} > -t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\ &= P\left(\overline{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} > \mu > \overline{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\ &= P\Bigg(\underbrace{\overline{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}}_{\mu_L} < \mu < \underbrace{\overline{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}}_{\mu_U}\Bigg). \end{aligned}$$

Therefore,

$$\left(\overline{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \overline{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for μ. $\square$

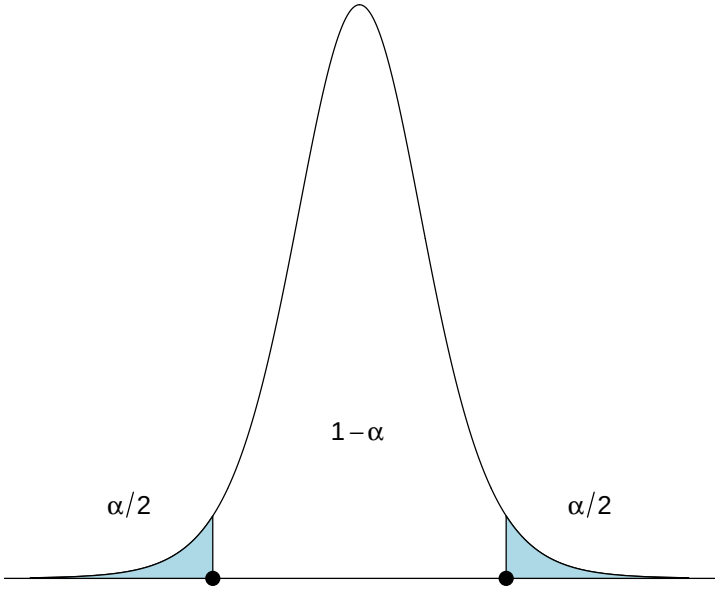

Figure 8.11: $t(n-1)$ pdf. The lower $\alpha/2$ quantile $-t_{n-1,\alpha/2}$ and the upper $\alpha/2$ quantile $t_{n-1,\alpha/2}$ are shown by using dark circles.

Remark: When $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, the confidence intervals

$$\left(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \qquad (\sigma^2 \text{ known})$$

and

$$\left(\overline{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \overline{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \qquad (\sigma^2 \text{ unknown})$$

are not large-sample intervals. They are **exact**. This means the confidence coefficient is exactly equal to $1-\alpha$ regardless of the sample size. Of course, this assumes $Y_1, Y_2, ..., Y_n$ are truly iid and the population distribution $\mathcal{N}(\mu, \sigma^2)$ is correctly specified.

Remark: It is well-known that the performance of the t confidence interval above is fairly **robust** to the underlying $\mathcal{N}(\mu, \sigma^2)$ population distribution assumption.

- In practice, this means that if the population distribution is mildly non-normal, the t confidence interval above can generally still be used to estimate μ.
- However, if there is strong evidence that the population distribution is grossly non-normal, then one should be cautious about using the interval, especially if the sample size n is small.

8.7.2 Population variance σ^2

Setting: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, where both μ and σ^2 are unknown. Our goal is to derive a $100(1-\alpha)\%$ confidence interval for the population variance σ^2. We start with

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

where S^2 is the sample variance. Note that Q is a pivot because its distribution is free of μ and σ^2. Define

$$\begin{aligned} \chi^2_{n-1,1-\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } \chi^2(n-1) \\ \chi^2_{n-1,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } \chi^2(n-1) \end{aligned}$$

and refer to Figure 8.12 (next page). Because $Q \sim \chi^2(n-1)$, we can write

$$\begin{aligned} 1-\alpha = P\left(\chi^2_{n-1,1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,\alpha/2}\right) &= P\left(\frac{1}{\chi^2_{n-1,1-\alpha/2}} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi^2_{n-1,\alpha/2}}\right) \\ &= P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} > \sigma^2 > \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right) \\ &= P\Bigg(\underbrace{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}}_{\sigma^2_L} < \sigma^2 < \underbrace{\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}}_{\sigma^2_U}\Bigg). \end{aligned}$$

Therefore,

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for σ^2. $\square$

Remark: Note that the following two events are equal:

$$\left\{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right\} = \left\{\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}}\right\}.$$

This is true because the square-root function is increasing and both endpoints are positive. Therefore, the latter event also has probability $1-\alpha$; i.e.,

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for σ.

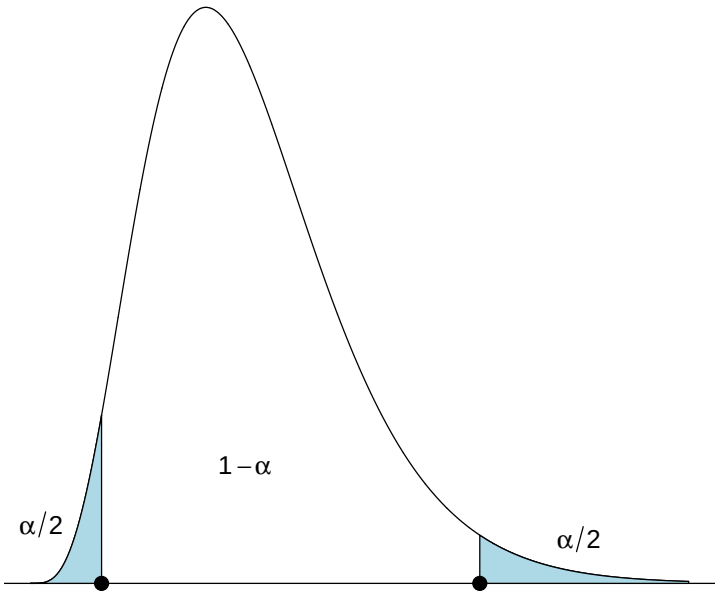

Figure 8.12: $\chi^2(n-1)$ pdf. The lower $\alpha/2$ quantile $\chi^2_{n-1,1-\alpha/2}$ and the upper $\alpha/2$ quantile $\chi^2_{n-1,\alpha/2}$ are shown by using dark circles.

8.7.3 Difference of two population means $\mu_1 - \mu_2$ (independent samples)

Setting: Suppose we have two independent random samples:

- $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an iid sample from a $\mathcal{N}(\mu_1, \sigma_1^2)$ population distribution
- $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an iid sample from a $\mathcal{N}(\mu_2, \sigma_2^2)$ population distribution,

where all population parameters are unknown. Our goal is to derive a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$, the difference of the two population means.

Note: Writing a confidence interval for $\mu_1 - \mu_2$ will allow us to *compare* the population means by noting where the interval resides; i.e., does the interval contain values entirely greater than 0? less than 0? does the interval contain 0?

Derivation: As before, define the sample means

$$\overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad \overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j}$$

and the sample variances

$$S_1^2 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{1j} - \overline{Y}_{1+})^2 \quad \text{and} \quad S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_{2j} - \overline{Y}_{2+})^2.$$

We know

$$\overline{Y}_{1+} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \overline{Y}_{2+} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Because $\overline{Y}_{1+}$ and $\overline{Y}_{2+}$ are both normally distributed, the difference $\overline{Y}_{1+} - \overline{Y}_{2+}$ is too (i.e., the difference is a simple linear combination). Therefore, because the two samples are independent,

$$\overline{Y}_{1+} - \overline{Y}_{2+} \sim \mathcal{N}\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Standardizing, we get

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

We also know

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1) \quad \text{and} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1).$$

Therefore, because the two samples are independent,

$$W = \frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_1+n_2-2).$$

Because $Z \perp\!\!\!\perp W$ (why?), we have

$$T = \frac{\dfrac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}}{\sqrt{\dfrac{\dfrac{(n_1-1)S_1^2}{\sigma_1^2} + \dfrac{(n_2-1)S_2^2}{\sigma_2^2}}{n_1+n_2-2}}} \sim t(n_1+n_2-2).$$

Remark: Although T is a pivotal quantity, it is unhelpful because T depends on the population variances σ_1^2 and σ_2^2, which are unknown. In this context, the population variances are **nuisance parameters** in the sense they are not the population parameters of interest. We want to write a confidence interval for $\mu_1 - \mu_2$, but how we do so depends on the assumption we make regarding σ_1^2 and σ_2^2. We consider two cases:

Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$; i.e., the population variances are **equal**. Under this assumption,

$$T = \frac{\dfrac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}}}}{\sqrt{\dfrac{\dfrac{(n_1-1)S_1^2}{\sigma^2} + \dfrac{(n_2-1)S_2^2}{\sigma^2}}{n_1+n_2-2}}} = \frac{\dfrac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}}{\dfrac{1}{\sigma}\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}} = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}},$$

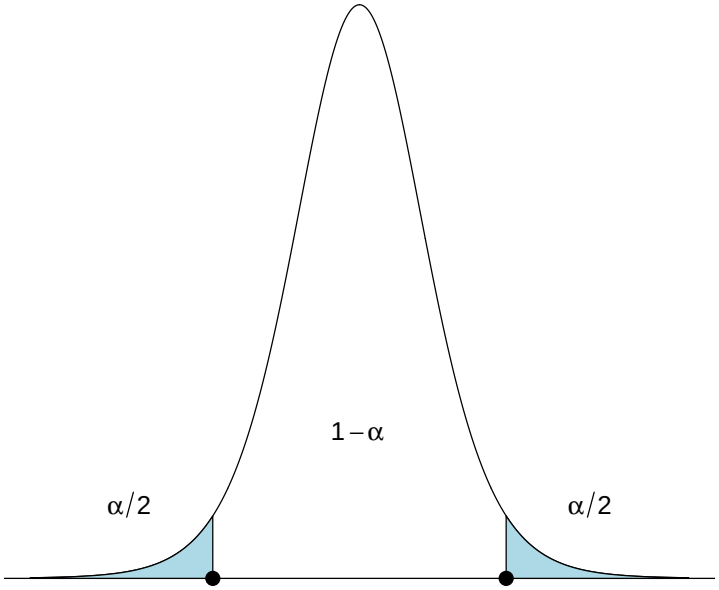

Figure 8.13: $t(n_1 + n_2 - 2)$ pdf. The lower $\alpha/2$ quantile $-t_{n_1+n_2-2,\alpha/2}$ and the upper $\alpha/2$ quantile $t_{n_1+n_2-2,\alpha/2}$ are shown by using dark circles.

where

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

is the **pooled sample variance estimator** of the common population variance σ^2.

Note: Under our modeling assumptions, the pooled sample variance estimator S_p^2 is an **unbiased estimator** of the common population variance σ^2; i.e.,

$$E(S_p^2) = \sigma^2.$$

Proof. Exercise.

Define

$$\begin{aligned} -t_{n_1+n_2-2,\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } t(n_1 + n_2 - 2) \\ t_{n_1+n_2-2,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } t(n_1 + n_2 - 2) \end{aligned}$$

and refer to Figure 8.13 (above). Because

$$T = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

we can write

$$1-\alpha = P\left(-t_{n_1+n_2-2,\alpha/2} < \frac{(\overline{Y}_{1+}-\overline{Y}_{2+})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < t_{n_1+n_2-2,\alpha/2}\right).$$

After performing the usual algebra; i.e., to isolate $\mu_1-\mu_2$ in the center of the inequality, we conclude

$$(\overline{Y}_{1+}-\overline{Y}_{2+}) \pm t_{n_1+n_2-2,\alpha/2} S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$. $\square$

Case 2: $\sigma_1^2 \neq \sigma_2^2$; i.e., the population variances are **unequal**. Under this assumption, we have few options. The reason is that there is no easy-to-use pivotal quantity that arises under this case. An approximate $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$ is given by

$$(\overline{Y}_{1+}-\overline{Y}_{2+}) \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},$$

where the degrees of freedom ν is calculated as

$$\nu = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)}+\frac{S_2^4}{n_2^2(n_2-1)}}.$$

This is called **Satterthwaite's formula**.

Remark: We have formulated two confidence intervals for $\mu_1-\mu_2$ under the normal independent sample assumption: one that assumes equal population variances $\sigma_1^2=\sigma_2^2$ and one that does not.

- If there is doubt on which assumption is more reasonable, my advice is to use the unequal variance interval. The penalty for using it when $\sigma_1^2=\sigma_2^2$ is much smaller than the penalty for using the equal variance interval when $\sigma_1^2 \neq \sigma_2^2$.
- I usually calculate both intervals (easy to do quickly with R) and then determine whether the intervals lead to drastically different conclusions.

Example 8.15. In a study conducted in the Department of Zoology at Virginia Tech University, data were collected on density measurements (i.e., the number of organisms per m^2) at two different locations; see Table 8.1 (next page). Let μ_1 and μ_2 denote the population mean density measurements at location 1 and location 2, respectively. Estimate $\mu_1-\mu_2$ with a 95% confidence interval.

Location 1		Location 2	
5030	4980	2800	2810
13700	11910	4670	1330
10730	8130	6890	3320
11400	26850	7720	1230
860	17660	7030	2130
2200	22800	7330	2190
4250	1130		
15040	1690		

Table 8.1: Example 8.15. Density measurements at two locations.

Solution. We can calculate both the equal and unequal variance confidence intervals for $\mu_1 - \mu_2$ (on the preceding page) by using R:

```
> t.test(loc.1,loc.2,conf.level=0.95,var.equal=TRUE)$conf.int
[1]   914.0939 10639.2394

> t.test(loc.1,loc.2,conf.level=0.95,var.equal=FALSE)$conf.int
[1]  1389.003 10164.331
```

Both intervals support the conjecture (hypothesis) that $\mu_1 - \mu_2 > 0$; i.e., the population mean density measurement at location 1 is larger than the population mean density measurement at location 2. □

Remark: Like the one-sample t confidence interval for one population mean μ, the two-sample t confidence intervals are **robust** to mild departures from normality. This means we can feel reasonably comfortable using the intervals even if the underlying population distributions are not perfectly normal.

8.7.4 Ratio of two population variances σ_2^2/σ_1^2 (independent samples)

Setting: Suppose we have two independent random samples:

- $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an iid sample from a $\mathcal{N}(\mu_1, \sigma_1^2)$ population distribution
- $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an iid sample from a $\mathcal{N}(\mu_2, \sigma_2^2)$ population distribution,

where all population parameters are unknown. Our goal is to derive a $100(1-\alpha)\%$ confidence interval for σ_2^2/σ_1^2, the ratio of the two population variances.

Derivation: We know

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1) \quad \text{and} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1).$$

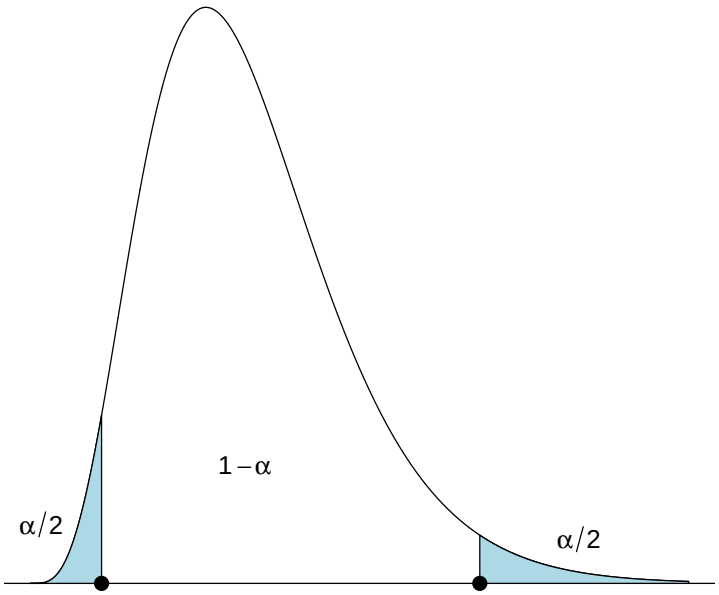

Figure 8.14: $F(n_1-1, n_2-1)$ pdf. The lower $\alpha/2$ quantile $F_{n_1-1,n_2-1,1-\alpha/2}$ and the upper $\alpha/2$ quantile $F_{n_1-1,n_2-1,\alpha/2}$ are shown by using dark circles.

Therefore, because the two samples are independent,

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}\Big/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}\Big/(n_2-1)} = \left(\frac{S_1^2}{S_2^2}\right)\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1-1, n_2-1).$$

Note that F is a pivot because its distribution is free of all unknown population parameters. Define

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } F(n_1-1, n_2-1) \\ F_{n_1-1,n_2-1,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } F(n_1-1, n_2-1) \end{aligned}$$

and refer to Figure 8.14 (above). Because $F \sim F(n_1-1, n_2-1)$, we can write

$$\begin{aligned} 1-\alpha &= P\left(F_{n_1-1,n_2-1,1-\alpha/2} < \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} < F_{n_1-1,n_2-1,\alpha/2}\right) \\ &= P\left(\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,1-\alpha/2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,\alpha/2}\right). \end{aligned}$$

Therefore,

$$\left(\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,1-\alpha/2},\ \frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,\alpha/2}\right)$$

is a $100(1-\alpha)\%$ confidence interval for σ_2^2/σ_1^2. $\square$

9 Properties of Point Estimators and Methods of Estimation

9.1 Introduction

Preview: This chapter presents additional concepts related to point estimation. In particular, we discuss

- sufficiency and its role in determining "best" point estimators
- mathematical methods of point estimation (e.g., method of moments, maximum likelihood, etc.)
- asymptotic concepts which enable us to describe large-sample distributions of point estimators (critical for statistical inference).

We begin by revisiting our discussion from the last chapter on comparing two unbiased estimators. We then generalize this discussion to find the "best" unbiased estimator.

9.2 Relative efficiency

Recall: In the last chapter, we posed this question:

Q: Suppose we have two point estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for the population-level parameter θ. Which one should we use? How should we compare them?
A: If both point estimators are unbiased; i.e., if $E(\widehat{\theta}_1) = E(\widehat{\theta}_2) = \theta$, then we would prefer the estimator with the smaller variance.

Terminology: Suppose $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are two unbiased point estimators of the population-level parameter θ. The **relative efficiency** of $\widehat{\theta}_1$ to $\widehat{\theta}_2$ is given by

$$\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) = \frac{V(\widehat{\theta}_2)}{V(\widehat{\theta}_1)}.$$

We interpret $\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2)$ as follows:

- if $\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) < 1$, then $\widehat{\theta}_2$ is more efficient (i.e., smaller variance) than $\widehat{\theta}_1$
- if $\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) > 1$, then $\widehat{\theta}_2$ is less efficient (i.e., larger variance) than $\widehat{\theta}_1$
- if $\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) = 1$, then $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are equally efficient.

This measure should only used if both point estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are **unbiased** (or at least asymptotically unbiased; see Example 9.3).

Remark: The definition of relative efficiency is somewhat unexciting; after all, we have already compared the variances of unbiased estimators in the last chapter. However, by taking the ratio of the variances, we can statements about the **efficiency** of one unbiased estimator to another. For example, if

$$\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) = \frac{V(\widehat{\theta}_2)}{V(\widehat{\theta}_1)} = 1.25,$$

then we would say "$\widehat{\theta}_1$ is 25% more efficient than $\widehat{\theta}_2$." In other words, using $\widehat{\theta}_2$, one would have to collect 25% more observations to get the same efficiency as one would obtain by using $\widehat{\theta}_1$.

Example 9.1. Suppose $Y_1, Y_2, .., Y_n$ is an iid sample from a Poisson(θ) population distribution, where $\theta > 0$ is unknown. Consider the two point estimators of θ:

$$\begin{aligned} \widehat{\theta}_1 &= \overline{Y} \\ \widehat{\theta}_2 &= S^2. \end{aligned}$$

We know that both estimators are unbiased because $E(\overline{Y}) = \mu = \theta$ and $E(S^2) = \sigma^2 = \theta$; i.e., the population mean and population variance are both equal to θ. Let's calculate

$$\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) = \frac{V(\widehat{\theta}_2)}{V(\widehat{\theta}_1)} = \frac{V(S^2)}{V(\overline{Y})}.$$

Recall that

$$V(\overline{Y}) = \frac{\sigma^2}{n} = \frac{\theta}{n}.$$

Calculating $V(S^2)$ is harder. Recall that, in general, if $Y_1, Y_2, ..., Y_n$ are iid with $E(Y^4) < \infty$; i.e., the fourth population moment is finite, then

$$V(S^2) = \frac{1}{n}\left[\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right],$$

where

$$\mu_4 = E[(Y-\mu)^4]$$

is the fourth central moment. For $Y \sim$ Poisson(θ), we can show

$$\mu_4 = \theta(1+3\theta) \implies V(S^2) = \frac{1}{n}\left[\theta(1+3\theta) - \left(\frac{n-3}{n-1}\right)\theta^2\right] = \frac{\theta}{n} + \frac{2\theta^2}{n-1}.$$

Therefore, the relative efficiency of $\widehat{\theta}_1$ to $\widehat{\theta}_2$ is

$$\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) = \frac{V(\widehat{\theta}_2)}{V(\widehat{\theta}_1)} = \frac{V(S^2)}{V(\overline{Y})} = \frac{\frac{\theta}{n} + \frac{2\theta^2}{n-1}}{\frac{\theta}{n}} = 1 + \left(\frac{2n}{n-1}\right)\theta.$$

Note that eff($\widehat{\theta}_1$ to $\widehat{\theta}_2$) > 1 whenever $n \geq 2$, establishing that $\overline{Y}$ is more efficient than S^2 as a point estimator of θ. $\square$

Example 9.2. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ population distribution, where $\theta > 0$ is unknown. In Example 8.1 (notes, pp 75-77), we showed

$$E(\widehat{\theta}_1) = E(2\overline{Y}) = \theta,$$

that is, $\widehat{\theta}_1 = 2\overline{Y}$ is an unbiased estimator of θ. We also showed

$$E(Y_{(n)}) = \left(\frac{n}{n+1}\right)\theta \implies E\left[\left(\frac{n+1}{n}\right)Y_{(n)}\right] = \left(\frac{n+1}{n}\right)\left(\frac{n}{n+1}\right)\theta = \theta,$$

that is,

$$\widehat{\theta}_2 = \left(\frac{n+1}{n}\right)Y_{(n)}$$

is also an unbiased estimator of θ. Let's calculate

$$\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) = \frac{V(\widehat{\theta}_2)}{V(\widehat{\theta}_1)}.$$

First, recall that $V(Y) = \sigma^2 = \theta^2/12$; i.e., this is the population variance of $Y \sim \mathcal{U}(0, \theta)$. Therefore,

$$V(\widehat{\theta}_1) = V(2\overline{Y}) = 4V(\overline{Y}) = 4\left(\frac{\sigma^2}{n}\right) = 4\left(\frac{\theta^2/12}{n}\right) = \frac{\theta^2}{3n}.$$

Second, note that

$$V\left(\left(\frac{n+1}{n}\right)Y_{(n)}\right) = \left(\frac{n+1}{n}\right)^2 V(Y_{(n)}),$$

where, by the variance computing formula,

$$V(Y_{(n)}) = E(Y_{(n)}^2) - [E(Y_{(n)})]^2.$$

Therefore, we need to calculate the second moment $E(Y_{(n)}^2)$. Recall the pdf of $Y_{(n)}$ is

$$f_{Y_{(n)}}(y) = \begin{cases} \dfrac{ny^{n-1}}{\theta^n}, & 0 < y < \theta \\ 0, & \text{otherwise,} \end{cases}$$

which we derived in Example 8.1. Therefore,

$$\begin{aligned} E(Y_{(n)}^2) = \int_{\mathbb{R}} y^2 f_{Y_{(n)}}(y)dy &= \int_0^\theta y^2 \, \frac{ny^{n-1}}{\theta^n} dy \\ &= \frac{n}{\theta^n}\int_0^\theta y^{n+1} dy = \frac{n}{\theta^n}\left(\frac{y^{n+2}}{n+2}\right)\Big|_0^\theta = \frac{n}{\theta^n}\left(\frac{\theta^{n+2}}{n+2}\right) = \left(\frac{n}{n+2}\right)\theta^2 \end{aligned}$$

and

$$\begin{aligned} V(Y_{(n)}) &= E(Y_{(n)}^2) - [E(Y_{(n)})]^2 \\ &= \left(\frac{n}{n+2}\right)\theta^2 - \left[\left(\frac{n}{n+1}\right)\theta\right]^2 = \left[\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right]\theta^2. \end{aligned}$$

Therefore,

$$\begin{aligned} V(\widehat{\theta}_2) = V\left(\left(\frac{n+1}{n}\right)Y_{(n)}\right) &= \left(\frac{n+1}{n}\right)^2 V(Y_{(n)}) \\ &= \left(\frac{n+1}{n}\right)^2 \left[\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right]\theta^2 = \frac{\theta^2}{n(n+2)}. \end{aligned}$$

Finally, the relative efficiency of $\widehat{\theta}_1$ to $\widehat{\theta}_2$ is

$$\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) = \frac{V(\widehat{\theta}_2)}{V(\widehat{\theta}_1)} = \frac{V\left(\left(\frac{n+1}{n}\right)Y_{(n)}\right)}{V(2\overline{Y})} = \frac{\frac{\theta^2}{n(n+2)}}{\frac{\theta^2}{3n}} = \frac{3}{n+2} < 1,$$

for $n \geq 2$. For example, if $n = 10$, then

$$\text{eff}(\widehat{\theta}_1 \text{ to } \widehat{\theta}_2) = 0.25;$$

i.e., $\widehat{\theta}_1$ is only 25 percent as efficient as $\widehat{\theta}_2$. $\square$

Example 9.3. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, where both parameters μ and σ^2 are unknown. Consider the two estimators of the population mean μ:

$$\begin{aligned} \widehat{\mu}_1 &= \overline{Y} \\ \widehat{\mu}_2 &= \widehat{\phi}_{0.5}, \end{aligned}$$

where

$$\widehat{\phi}_{0.5} = \begin{cases} Y_{((n+1)/2)}, & \text{if } n \text{ is odd} \\ (Y_{(n/2)} + Y_{(n/2+1)})/2, & \text{if } n \text{ is even;} \end{cases}$$

i.e., $\widehat{\phi}_{0.5}$ is the **sample median**. Let's calculate

$$\text{eff}(\widehat{\mu}_1 \text{ to } \widehat{\mu}_2) = \frac{V(\widehat{\mu}_2)}{V(\widehat{\mu}_1)} = \frac{V(\widehat{\phi}_{0.5})}{V(\overline{Y})}.$$

We know

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies V(\overline{Y}) = \frac{\sigma^2}{n}.$$

Whenever I teach STAT 823, I use asymptotic results for sample quantiles to prove

$$\widehat{\phi}_{0.5} \sim \mathcal{AN}\left(\mu, \frac{\pi}{2}\frac{\sigma^2}{n}\right) \implies V(\widehat{\phi}_{0.5}) \approx \frac{\pi}{2}\frac{\sigma^2}{n},$$

for large n. Therefore, the relative efficiency of $\widehat{\mu}_1$ to $\widehat{\mu}_2$ is

$$\text{eff}(\widehat{\mu}_1 \text{ to } \widehat{\mu}_2) = \frac{V(\widehat{\mu}_2)}{V(\widehat{\mu}_1)} = \frac{V(\widehat{\phi}_{0.5})}{V(\overline{Y})} \approx \frac{\frac{\pi}{2}\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}} = \frac{\pi}{2} \approx 1.57.$$

Therefore, when estimating the population mean μ, the sample mean $\widehat{\mu}_1 = \overline{Y}$ is about 57% more efficient than the sample median $\widehat{\mu}_2 = \widehat{\phi}_{0.5}$. $\square$

9.3 Sufficient statistics

Remark: Suppose $Y_1, Y_2, ..., Y_n$ is a random sample from a population distribution, denoted by $p_Y(y)$ or $f_Y(y)$. Intuitively, the sample $Y_1, Y_2, ..., Y_n$ contains valuable information about a population-level parameter θ, and we have already discussed different criteria on how to evaluate the statistic $T = T(Y_1, Y_2, ..., Y_n)$ as a point estimator of θ (e.g., bias, variance, MSE, etc.). More generally, **statistical inference** deals with using the information in the sample to make a statement about population-level parameters. In practice, this is done by using confidence intervals to estimate parameters or by performing hypothesis tests about these parameters.

Remark: Sufficiency plays an important role in statistical inference. Informally, sufficiency is a mathematical concept dealing with **data reduction** and addresses this question when attempting to estimate a population-level parameter θ:

> *"Instead of keeping track of the entire sample, can we reduce the sample to a small number of statistics that contain the same information as the entire sample?"*

If we can find statistics that accomplish this (i.e., retain all the information about θ), then there is no harm in restricting our attention to these statistics when performing statistical inference.

Definition: Suppose $Y_1, Y_2, ..., Y_n$ is a sample from a population distribution with unknown parameter θ. The statistic $T = T(Y_1, Y_2, ..., Y_n)$ is **sufficient** for θ if the conditional distribution of the sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, given T, does not depend on θ.

Informally: If the process of conditioning the sample on T removes all information about θ, then T must contain the same information about θ that the entire sample has.

Implementation: Showing $T = T(Y_1, Y_2, ..., Y_n)$ is sufficient by using the definition above involves showing the following. In the discrete case, we show

$$p_{\mathbf{Y}|T}(\mathbf{y}|t) = \frac{p_{\mathbf{Y}}(\mathbf{y})}{p_T(t)} \quad \text{is free of } \theta.$$

In the continuous case, we show

$$f_{\mathbf{Y}|T}(\mathbf{y}|t) = \frac{f_{\mathbf{Y}}(\mathbf{y})}{f_T(t)} \quad \text{is free of } \theta.$$

Example 9.4. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population distribution, where $\theta > 0$ is unknown. Recall the Poisson(θ) pmf is given by

$$p_Y(y) = \begin{cases} \dfrac{\theta^y e^{-\theta}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Show $T = \sum_{i=1}^n Y_i$ is a sufficient statistic for θ.

Solution. The joint pmf of the sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ is

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= p_Y(y_1) \times p_Y(y_2) \times \cdots \times p_Y(y_n) \\ &= \frac{\theta^{y_1}e^{-\theta}}{y_1!} \times \frac{\theta^{y_2}e^{-\theta}}{y_2!} \times \cdots \times \frac{\theta^{y_n}e^{-\theta}}{y_n!} = \frac{\theta^{\sum_{i=1}^n y_i}e^{-n\theta}}{y_1!y_2!\cdots y_n!} = \frac{\theta^{\sum_{i=1}^n y_i}e^{-n\theta}}{\prod_{i=1}^n y_i!}. \end{aligned}$$

What is the (sampling) distribution of $T = \sum_{i=1}^n Y_i$? In case you have forgotten, we can derive it quickly. The mgf of T is

$$m_T(t) = [m_Y(t)]^n = [e^{\theta(e^t-1)}]^n = e^{n\theta(e^t-1)}.$$

This is the mgf of a Poisson random variable with mean $n\theta$. Because mgfs are unique, it must be true that $T \sim \text{Poisson}(n\theta)$. Therefore, for $t = 0, 1, 2, ...$, the pmf of T is

$$p_T(t) = \frac{(n\theta)^t e^{-n\theta}}{t!},$$

where $t = \sum_{i=1}^n y_i$. Therefore, the conditional pmf of the sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, given $T = t$, is

$$p_{\mathbf{Y}|T}(\mathbf{y}|t) = \frac{p_{\mathbf{Y}}(\mathbf{y})}{p_T(t)} = \frac{\frac{\theta^{\sum_{i=1}^n y_i}e^{-n\theta}}{\prod_{i=1}^n y_i!}}{\frac{(n\theta)^t e^{-n\theta}}{t!}} = \frac{t!}{n^t \prod_{i=1}^n y_i!},$$

which does not depend on θ. Therefore, $T = \sum_{i=1}^n Y_i$ is a sufficient statistic for θ. □

Analogy: Suppose two researchers have collected observations on n individuals which are modeled as iid Poisson(θ) counts; e.g., number of sexual partners, number of claims made, number of missed classes, etc.

- Researcher 1 has the entire sample of observations $Y_1, Y_2, ..., Y_n$
- Researcher 2 has lost the sample but s/he has the value of $T = \sum_{i=1}^n Y_i$.

If the goal is to perform statistical inference for the population-level parameter θ (i.e., the mean of the population), then Researcher 1 and Researcher 2 have the same information! In other words, Researcher 2 has lost nothing by reducing the entire sample to the sample sum $T = \sum_{i=1}^n Y_i$.

Example 9.5. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Rayleigh(θ) population distribution, where $\theta > 0$ is unknown. Recall the Rayleigh pdf is given by

$$f_Y(y) = \begin{cases} \frac{2y}{\theta}e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Show that $T = \sum_{i=1}^n Y_i^2$ is a sufficient statistic for θ.

Solution. The joint pdf of the sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ is

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= f_Y(y_1) \times f_Y(y_2) \times \cdots \times f_Y(y_n) \\ &= \frac{2y_1}{\theta}e^{-y_1^2/\theta} \times \frac{2y_2}{\theta}e^{-y_2^2/\theta} \times \cdots \times \frac{2y_n}{\theta}e^{-y_n^2/\theta} = \left(\frac{2}{\theta}\right)^n \left(\prod_{i=1}^n y_i\right) e^{-\sum_{i=1}^n y_i^2/\theta}. \end{aligned}$$

What is the (sampling) distribution of $T = \sum_{i=1}^n Y_i^2$? In Exercise 6.34 (WMS, pp 318), we showed

$$Y \sim \text{Rayleigh}(\theta) \implies U = Y^2 \sim \text{exponential}(\theta).$$

Therefore,

$$T = \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n U_i \sim \text{gamma}(n, \theta).$$

This is true because $U_1, U_2, ..., U_n$ are iid exponential(θ) and

$$m_T(t) = [m_U(t)]^n = \left(\frac{1}{1-\theta t}\right)^n,$$

which we recognize as the gamma(n, θ) mgf. Therefore, the pdf of T, for $t > 0$, is

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

Therefore, the conditional pdf of the sample $Y_1, Y_2, ..., Y_n$, given $T = t$, is given by

$$f_{\mathbf{Y}|T}(\mathbf{y}|t) = \frac{f_{\mathbf{Y}}(\mathbf{y})}{f_T(t)} = \frac{\left(\frac{2}{\theta}\right)^n \prod_{i=1}^n y_i \; e^{-\sum_{i=1}^n y_i^2/\theta}}{\frac{1}{\Gamma(n)\theta^n}t^{n-1}e^{-t/\theta}} = \frac{2^n\Gamma(n)\prod_{i=1}^n y_i}{t^{n-1}},$$

which does not depend on θ. Therefore, $T = \sum_{i=1}^n Y_i^2$ is a sufficient statistic for θ. □

Remark: The approach we have outlined to show a statistic T is sufficient appeals to the definition of sufficiency. That is, we show directly the conditional distribution of the sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, given T, does not depend on θ.

- If I ask you to show that T is sufficient by "using the definition," then this is the approach I want you to take.
- What if we don't know which statistic is sufficient? Then the approach we have just outlined is not practical to implement. For example, imagine trying different statistics T and for each one attempting to show that $p_{\mathbf{Y}|T}(\mathbf{y}|t)$ or $f_{\mathbf{Y}|T}(\mathbf{y}|t)$ is free of θ. This might involve a large amount of trial and error and you would have to derive the sampling distribution of T each time.
- The Factorization Theorem (to be discussed shortly) makes getting sufficient statistics easy. To prepare for this theorem, we first introduce the likelihood function.

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution denoted by $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where θ is an unknown population parameter.

- **Note:** Going forward, we now write $p_Y(y|\theta)$ in place of $p_Y(y)$ to emphasize the population pmf depends on θ. Similarly, we write $f_Y(y|\theta)$ in place of $f_Y(y)$.

The **likelihood function**, which is denoted by $L(\theta|\mathbf{y}) = L(\theta|y_1, y_2, ..., y_n)$, is determined as follows:

- In the discrete case,

$$L(\theta|\mathbf{y}) = p_{\mathbf{Y}}(\mathbf{y}|\theta) = p_Y(y_1|\theta) \times p_Y(y_2|\theta) \times \cdots \times p_Y(y_n|\theta) = \prod_{i=1}^{n} p_Y(y_i|\theta).$$

- In the continuous case,

$$L(\theta|\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}|\theta) = f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) = \prod_{i=1}^{n} f_Y(y_i|\theta).$$

Note: Mathematically, the likelihood function $L(\theta|\mathbf{y})$ is the same function as the joint pmf $p_{\mathbf{Y}}(\mathbf{y}|\theta)$ in the discrete case and the joint pdf $f_{\mathbf{Y}}(\mathbf{y}|\theta)$ in the continuous case. The only difference is in how we interpret each function.

- The joint distributions $p_{\mathbf{Y}}(\mathbf{y}|\theta)$ or $f_{\mathbf{Y}}(\mathbf{y}|\theta)$ describe the random behavior of the sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ when θ is fixed.
- The likelihood function $L(\theta|\mathbf{y})$ is viewed as a function of θ with the sample data $\mathbf{y} = (y_1, y_2, ..., y_n)$ held fixed.

Example 9.4 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population distribution, where $\theta > 0$ is unknown. Recall the Poisson(θ) pmf is given by

$$p_Y(y|\theta) = \begin{cases} \dfrac{\theta^y e^{-\theta}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

$$\begin{aligned} L(\theta|\mathbf{y}) = \prod_{i=1}^{n} p_Y(y_i|\theta) &= p_Y(y_1|\theta) \times p_Y(y_2|\theta) \times \cdots \times p_Y(y_n|\theta) \\ &= \frac{\theta^{y_1} e^{-\theta}}{y_1!} \times \frac{\theta^{y_2} e^{-\theta}}{y_2!} \times \cdots \times \frac{\theta^{y_n} e^{-\theta}}{y_n!} = \frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{y_1! y_2! \cdots y_n!} = \frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{\prod_{i=1}^n y_i!}. \end{aligned}$$

Application: Suppose actuaries have an iid sample of $n = 84$ policies and on each one they observe

$$Y = \text{ the number of accidents in a given year.}$$

The observed data (in tabular form) are given on the next page:

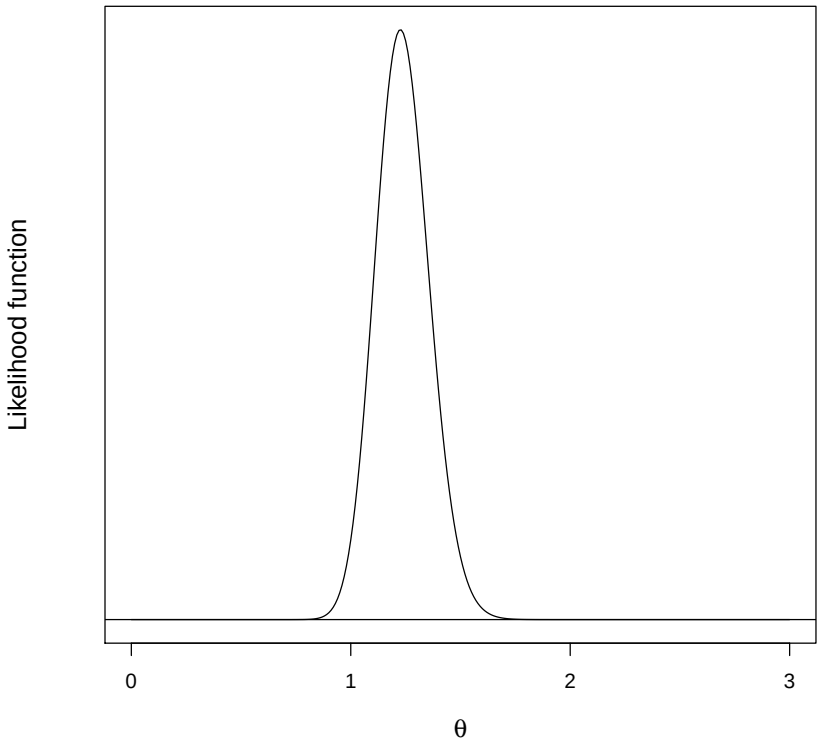

Figure 9.1: Accident data. Poisson likelihood function $L(\theta|\mathbf{y})$ in Example 9.4.

Number of accidents	Number of policies
0	32
1	26
2	12
3	7
4	4
5	2
6	1

Suppose the observations $Y_1, Y_2, ..., Y_{84}$ are modeled as iid Poisson counts with mean $\theta > 0$. We calculate

$$\sum_{i=1}^{84} y_i = 103 \quad \text{and} \quad \prod_{i=1}^{84} y_i! = 3944197523094110208000.$$

Therefore, the likelihood function based on these observations is given by

$$L(\theta|\mathbf{y}) = \frac{\theta^{\sum_{i=1}^{84} y_i} e^{-84\theta}}{\prod_{i=1}^{84} y_i!} = \left(\frac{1}{3944197523094110208000}\right) \theta^{103} e^{-84\theta} \propto \theta^{103} e^{-84\theta}.$$

This function is shown in Figure 9.1 above. □

Example 9.5 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Rayleigh(θ) population distribution, where $\theta > 0$ is unknown. Recall the Rayleigh pdf is given by

$$f_Y(y|\theta) = \begin{cases} \frac{2y}{\theta}e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

$$\begin{aligned} L(\theta|\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i|\theta) &= f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) \\ &= \frac{2y_1}{\theta}e^{-y_1^2/\theta} \times \frac{2y_2}{\theta}e^{-y_2^2/\theta} \times \cdots \times \frac{2y_n}{\theta}e^{-y_n^2/\theta} \\ &= \left(\frac{2}{\theta}\right)^n \left(\prod_{i=1}^{n} y_i\right) e^{-\sum_{i=1}^{n} y_i^2/\theta}. \end{aligned}$$

Application: A light bulb company manufactures filaments that are not expected to wear out during an extended period of "intense use." With the goal of guaranteeing bulb reliability in these conditions, engineers sample $n = 30$ bulbs, simulate their long term use, and record

$$Y = \text{ time until failure (in 100s hours)}$$

for each bulb. Here are the lifetimes:

4.43	5.93	3.74	5.82	5.90	2.90	2.64	6.49	5.31	8.49
1.01	1.07	1.41	3.42	1.22	4.01	0.57	1.47	2.81	8.52
0.52	4.77	0.85	2.21	6.85	3.43	1.87	5.15	2.02	10.58

Suppose the observations $Y_1, Y_2, ..., Y_{30}$ are modeled as iid Rayleigh(θ), where $\theta > 0$. We calculate

$$\sum_{i=1}^{30} y_i^2 = 645.0 \quad \text{and} \quad \prod_{i=1}^{30} y_i = 87086335417057.6.$$

Therefore, the likelihood function based on these observations is given by

$$L(\theta|\mathbf{y}) = \left(\frac{2}{\theta}\right)^{30} \left(\prod_{i=1}^{30} y_i\right) e^{-\sum_{i=1}^{30} y_i^2/\theta} = \left(2^{30} \times 87086335417057.6\right) \frac{e^{-645.0/\theta}}{\theta^{30}} \propto \frac{e^{-645.0/\theta}}{\theta^{30}}.$$

This function is shown in Figure 9.2 (next page). □

Remark: In a sense, constructing the likelihood function $L(\theta|\mathbf{y})$ is a form of **data reduction**, that is, we are reducing the sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ to a function of the population-level parameter θ. With the notion of a likelihood function in hand, we now return to sufficiency and the Factorization Theorem.

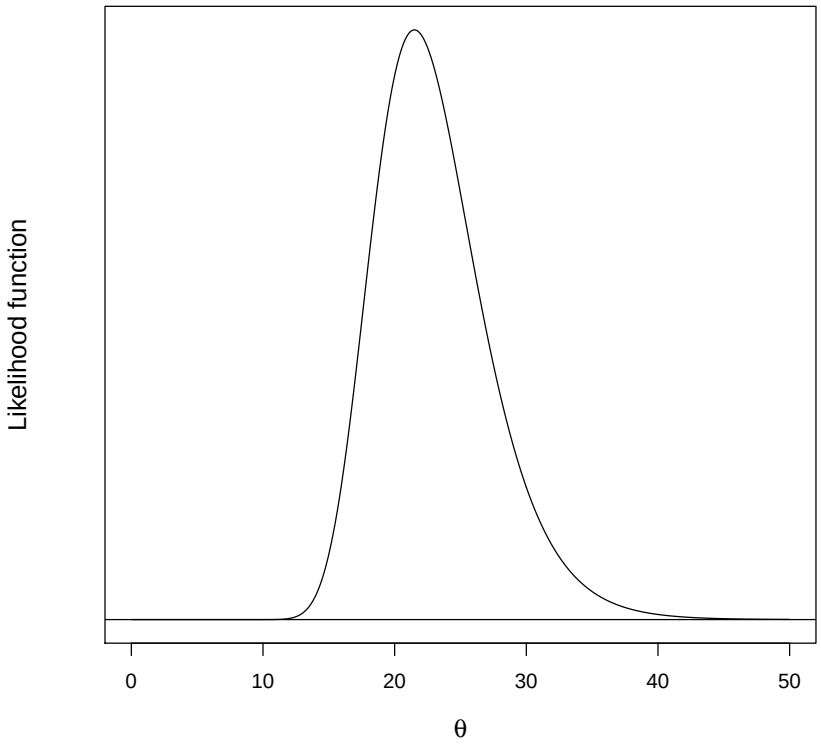

Figure 9.2: Light bulb data. Rayleigh likelihood function $L(\theta|\mathbf{y})$ in Example 9.5.

Factorization Theorem: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution denoted by $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where θ is an unknown population parameter. The statistic $T = T(Y_1, Y_2, ..., Y_n)$ is a sufficient statistic if and only if we can write the likelihood function as follows:

$$L(\theta|\mathbf{y}) = L(\theta|y_1, y_2, ..., y_n) = g(t, \theta)h(y_1, y_2, ..., y_n),$$

where g and h are nonnegative functions and

- $g(t, \theta)$ is a function of $t = t(y_1, y_2, ..., y_n)$ and θ only
- $h(y_1, y_2, ..., y_n)$ is a function of $y_1, y_2, ..., y_n$ only; i.e., the function $h(y_1, y_2, ..., y_n)$ cannot depend on θ.

Remark: The Factorization Theorem makes getting sufficient statistics easy!

- There is no need to appeal to the definition of sufficiency to demonstrate a particular statistic is sufficient; all we have to do is work with the likelihood function directly.
- In most cases, a sufficient statistic presents itself immediately.

Example 9.4 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population distribution, where $\theta > 0$ is unknown. Recall the Poisson(θ) pmf is given by

$$p_Y(y|\theta) = \begin{cases} \dfrac{\theta^y e^{-\theta}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Note that we can write the likelihood function as

$$L(\theta|\mathbf{y}) = \frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{\prod_{i=1}^n y_i!} = \underbrace{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}_{g(t,\theta)} \times \underbrace{\frac{1}{\prod_{i=1}^n y_i!}}_{h(y_1,y_2,...,y_n)},$$

where $t = \sum_{i=1}^n y_i$. By the Factorization Theorem, it follows that $T = \sum_{i=1}^n Y_i$ is a sufficient statistic for θ. $\square$

Example 9.5 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Rayleigh(θ) population distribution, where $\theta > 0$ is unknown. Recall the Rayleigh pdf is given by

$$f_Y(y|\theta) = \begin{cases} \dfrac{2y}{\theta} e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Note that we can write the likelihood function as

$$L(\theta|\mathbf{y}) = \left(\frac{2}{\theta}\right)^n \left(\prod_{i=1}^n y_i\right) e^{-\sum_{i=1}^n y_i^2/\theta} = \underbrace{\frac{e^{-\sum_{i=1}^n y_i^2/\theta}}{\theta^n}}_{g(t,\theta)} \times \underbrace{2^n \prod_{i=1}^n y_i}_{h(y_1,y_2,...,y_n)},$$

where $t = \sum_{i=1}^n y_i^2$. By the Factorization Theorem, it follows that $T = \sum_{i=1}^n Y_i^2$ is a sufficient statistic for θ. $\square$

Example 9.6. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a beta($\theta, 1$) population distribution, where $\theta > 0$ is unknown. Recall the beta($\theta, 1$) pdf is given by

$$f_Y(y|\theta) = \begin{cases} \theta y^{\theta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

$$\begin{aligned} L(\theta|\mathbf{y}) = \prod_{i=1}^n f_Y(y_i|\theta) &= f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) \\ &= \theta y_1^{\theta-1} \times \theta y_2^{\theta-1} \times \cdots \times \theta y_n^{\theta-1} = \theta^n \left(\prod_{i=1}^n y_i\right)^{\theta-1}. \end{aligned}$$

Note that we can write the likelihood function as

$$L(\theta|\mathbf{y}) = \theta^n \left(\prod_{i=1}^n y_i\right)^{\theta-1} = \underbrace{\theta^n \left(\prod_{i=1}^n y_i\right)^{\theta}}_{g(t,\theta)} \times \underbrace{\frac{1}{\prod_{i=1}^n y_i}}_{h(y_1,y_2,\ldots,y_n)},$$

where $t = \prod_{i=1}^n y_i$. By the Factorization Theorem, it follows that $T = \prod_{i=1}^n Y_i$ is a sufficient statistic for θ. $\square$

Example 9.7. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ population distribution, where $\theta > 0$ is unknown. Recall the $\mathcal{U}(0, \theta)$ pdf is given by

$$f_Y(y|\theta) = \begin{cases} \frac{1}{\theta}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

Important: In this example, unlike Examples 9.4-9.6, it is important to note that the support of the random variable Y depends on the unknown parameter θ. When this is the case, we need to be careful in how we apply the Factorization Theorem. Because the $\mathcal{U}(0, \theta)$ pdf is nonzero only when $0 < y < \theta$, let's write

$$f_Y(y|\theta) = \frac{1}{\theta}\, I(0 < y < \theta),$$

where $I(\cdot)$ is the **indicator function**; i.e.,

$$I(0 < y < \theta) = \begin{cases} 1, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

$$\begin{aligned} L(\theta|\mathbf{y}) = \prod_{i=1}^n f_Y(y_i|\theta) &= f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) \\ &= \frac{1}{\theta}\, I(0 < y_1 < \theta) \times \frac{1}{\theta}\, I(0 < y_2 < \theta) \times \cdots \times \frac{1}{\theta}\, I(0 < y_n < \theta) \\ &= \frac{1}{\theta^n} \prod_{i=1}^n I(0 < y_i < \theta). \end{aligned}$$

A sufficient statistic is "hiding" in the

$$\prod_{i=1}^n I(0 < y_i < \theta)$$

term. To see why, note that

$$\prod_{i=1}^n I(0 < y_i < \theta) = 1 \iff I(0 < y_{(n)} < \theta) = 1.$$

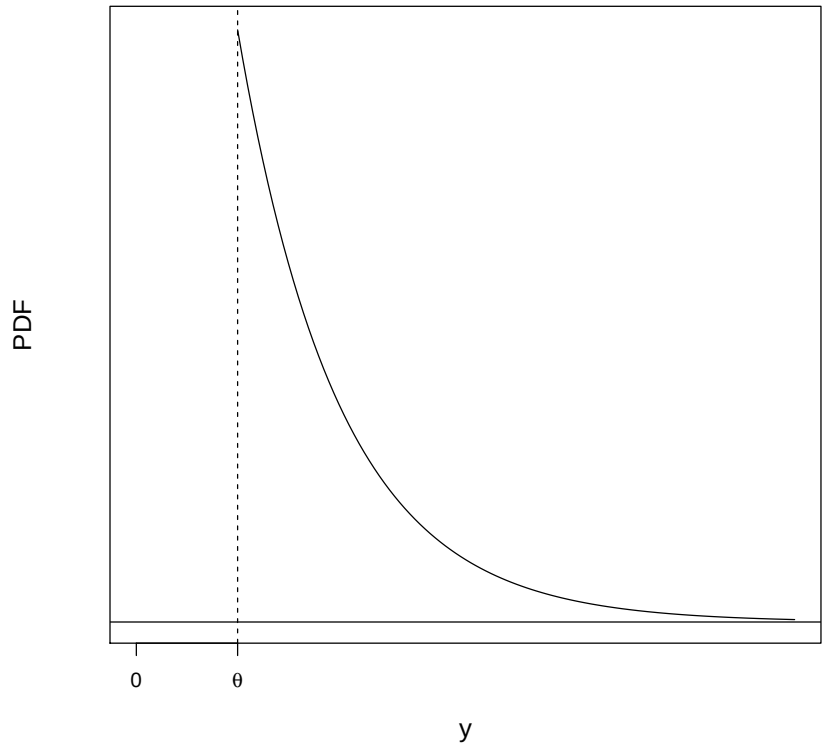

Figure 9.3: Shifted exponential pdf in Example 9.8.

Therefore, we can write the likelihood function as

$$L(\theta|\mathbf{y}) = \frac{1}{\theta^n} I(0 < y_{(n)} < \theta) = \underbrace{\frac{1}{\theta^n} I(0 < y_{(n)} < \theta)}_{g(t,\theta)} \times \underbrace{1}_{h(y_1,y_2,\ldots,y_n)},$$

where $t = y_{(n)}$. By the Factorization Theorem, it follows that $T = Y_{(n)}$ is a sufficient statistic for θ. $\square$

Remark: Whenever the population pmf/pdf has support that depends on the parameter θ, a sufficient statistic will usually be an order statistic (or a collection of order statistics).

Example 9.8. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from

$$f_Y(y|\theta) = \begin{cases} e^{-(y-\theta)}, & y > \theta \\ 0, & \text{otherwise.} \end{cases}$$

This is the (population) pdf of a **shifted exponential** random variable; i.e., an exponential(1) pdf shifted to the right by θ units; see Figure 9.3 above. In this model, θ represents the smallest value Y can be.

The likelihood function is given by

$$\begin{aligned} L(\theta|\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i|\theta) &= f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) \\ &= e^{-(y_1-\theta)} I(y_1 > \theta) \times e^{-(y_2-\theta)} I(y_2 > \theta) \times \cdots \times e^{-(y_n-\theta)} I(y_n > \theta) \\ &= e^{-\sum_{i=1}^{n}(y_i-\theta)} \prod_{i=1}^{n} I(y_i > \theta). \end{aligned}$$

Note that

$$\prod_{i=1}^{n} I(y_i > \theta) = 1 \iff I(y_{(1)} > \theta) = 1.$$

Therefore, we can write the likelihood function as

$$L(\theta|\mathbf{y}) = e^{-\sum_{i=1}^{n}(y_i-\theta)} I(y_{(1)} > \theta) = \underbrace{e^{n\theta}\ I(y_{(1)} > \theta)}_{g(t,\theta)} \times \underbrace{e^{-\sum_{i=1}^{n} y_i}}_{h(y_1,y_2,...,y_n)},$$

where $t = y_{(1)}$. By the Factorization Theorem, it follows that $T = Y_{(1)}$ is a sufficient statistic for θ. $\square$

Important: If $T = T(Y_1, Y_2, ..., Y_n)$ is a sufficient statistic for θ, then any 1:1 function of T is also a sufficient statistic. The function need only be 1:1 over the **parameter space**; i.e., over the possible values of θ.

Example 9.9. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(θ) population distribution, where $\theta > 0$ is unknown. Recall the exponential(θ) pdf is given by

$$f_Y(y|\theta) = \begin{cases} \frac{1}{\theta} e^{-y/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

$$\begin{aligned} L(\theta|\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i|\theta) &= f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) \\ &= \frac{1}{\theta} e^{-y_1/\theta} \times \frac{1}{\theta} e^{-y_2/\theta} \times \cdots \times \frac{1}{\theta} e^{-y_n/\theta} = \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} y_i/\theta}. \end{aligned}$$

Note that we can write the likelihood function as

$$L(\theta|\mathbf{y}) = \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} y_i/\theta} = \underbrace{\frac{1}{\theta^n} e^{-\sum_{i=1}^{n} y_i/\theta}}_{g(t,\theta)} \times \underbrace{1}_{h(y_1,y_2,...,y_n)},$$

where $t = \sum_{i=1}^{n} y_i$. By the Factorization Theorem, it follows that $T = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for θ. $\square$

Note: In Example 9.9, the parameter space is $\{\theta : \theta > 0\}$. Therefore, any 1:1 function of $T = \sum_{i=1}^{n} Y_i$ over $(0, \infty)$ is also a sufficient statistic; for example,

$$\begin{aligned}
\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i &\longleftarrow \text{ 1:1 function of } T \text{ over } (0,\infty) \\
\exp\left(\sum_{i=1}^{n} Y_i\right) &\longleftarrow \text{ 1:1 function of } T \text{ over } (0,\infty) \\
\left(\sum_{i=1}^{n} Y_i\right)^2 &\longleftarrow \text{ 1:1 function of } T \text{ over } (0,\infty).
\end{aligned}$$

Remark: The Factorization Theorem can also be applied in population models with more than one parameter.

Factorization Theorem (Extension): Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution denoted by $p_Y(y|\boldsymbol{\theta})$ or $f_Y(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$ is an unknown population parameter. The statistic

$$\mathbf{T} = \mathbf{T}(Y_1, Y_2, ..., Y_n) = \begin{pmatrix} T_1(Y_1, Y_2, ..., Y_n) \\ T_2(Y_1, Y_2, ..., Y_n) \\ \vdots \\ T_k(Y_1, Y_2, ..., Y_n) \end{pmatrix}$$

is a sufficient statistic if and only if we can write the likelihood function as follows:

$$L(\boldsymbol{\theta}|\mathbf{y}) = L(\boldsymbol{\theta}|y_1, y_2, ..., y_n) = g(t_1, t_2, ..., t_k, \boldsymbol{\theta})h(y_1, y_2, ..., y_n),$$

where g and h are nonnegative functions and

- $g(t_1, t_2, ..., t_k, \boldsymbol{\theta})$ is a function of $t_1 = t_1(y_1, y_2, ..., y_n), t_2 = t_2(y_1, y_2, ..., y_n), ..., t_k = t_k(y_1, y_2, ..., y_n)$ and $\boldsymbol{\theta}$ only
- $h(y_1, y_2, ..., y_n)$ is a function of $y_1, y_2, ..., y_n$ only.

Example 9.10. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a gamma(α, β) population distribution, where $\alpha > 0$ and $\beta > 0$ are unknown. Recall the gamma(α, β) pdf is given by

$$f_Y(y|\boldsymbol{\theta}) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Note that the population-level parameter $\boldsymbol{\theta} = (\alpha, \beta)$ is two-dimensional; i.e., $d = 2$. Find a sufficient statistic for $\boldsymbol{\theta}$.

Solution. The likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i|\boldsymbol{\theta}) &= f_Y(y_1|\boldsymbol{\theta}) \times f_Y(y_2|\boldsymbol{\theta}) \times \cdots \times f_Y(y_n|\boldsymbol{\theta}) \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} y_1^{\alpha-1} e^{-y_1/\beta} \times \frac{1}{\Gamma(\alpha)\beta^\alpha} y_2^{\alpha-1} e^{-y_2/\beta} \times \cdots \times \frac{1}{\Gamma(\alpha)\beta^\alpha} y_n^{\alpha-1} e^{-y_n/\beta} \\ &= \left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\right]^n \left(\prod_{i=1}^{n} y_i\right)^{\alpha-1} e^{-\sum_{i=1}^{n} y_i/\beta}. \end{aligned}$$

Note that we can write the likelihood function as

$$L(\boldsymbol{\theta}|\mathbf{y}) = \underbrace{\left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\right]^n \left(\prod_{i=1}^{n} y_i\right)^{\alpha} e^{-\sum_{i=1}^{n} y_i/\beta}}_{g(t_1,t_2,\boldsymbol{\theta})} \times \underbrace{\frac{1}{\prod_{i=1}^{n} y_i}}_{h(y_1,y_2,\ldots,y_n)},$$

where $\mathbf{t} = (t_1, t_2) = (\prod_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i)$. By the Factorization Theorem, it follows that

$$\mathbf{T} = \begin{pmatrix} \prod_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i \end{pmatrix}$$

is a sufficient statistic for $\boldsymbol{\theta} = (\alpha, \beta)$. $\square$

Remark: The Factorization Theorem can also be applied in population models where the observations $Y_1, Y_2, ..., Y_n$ are not iid.

Example 9.11. Consider the simple linear regression model

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

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

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

Therefore, the pdf of Y_i is

$$f_{Y_i}(y_i|\boldsymbol{\theta}) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2}, & -\infty < y_i < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Note: The random variables $Y_1, Y_2, ..., Y_n$ are mutually independent (functions of the independent ϵ_i's are independent). However, $Y_1, Y_2, ..., Y_n$ are *not* identically distributed because

$E(Y_i) = \beta_0 + \beta_1 x_i$ changes as i does. The parameter $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$ is three-dimensional; i.e., $d = 3$. The likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i|\boldsymbol{\theta}) &= f_{Y_1}(y_1|\boldsymbol{\theta}) \times f_{Y_2}(y_2|\boldsymbol{\theta}) \times \cdots \times f_{Y_n}(y_n|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}. \end{aligned}$$

It is easy to show that

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

where $\mathbf{t} = (t_1, t_2, t_3) = (\sum_{i=1}^n y_i, \sum_{i=1}^n y_i^2, \sum_{i=1}^n x_i y_i)$. Because we can write

$$L(\boldsymbol{\theta}|\mathbf{y}) = \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} g^*(t_1, t_2, t_3, \beta_0, \beta_1)\right\}}_{g(t_1, t_2, t_3, \boldsymbol{\theta})} \times \underbrace{1}_{h(y_1, y_2, \ldots, y_n)},$$

it follows from the Factorization Theorem that

$$\mathbf{T} = \left(\sum_{i=1}^{n} Y_i, \ \sum_{i=1}^{n} Y_i^2, \ \sum_{i=1}^{n} x_i Y_i \right)$$

is a sufficient statistic for $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$. $\square$

9.4 Minimum variance unbiased estimators (MVUEs)

Problem: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution with unknown parameter $\theta \in \mathbb{R}$. In this subsection, we describe how to find the best point estimator $\widehat{\theta}$ for θ. Of course, it is important to define what "best" means. Consider the class of point estimators

$$\mathcal{C} = \{\widehat{\theta} : E(\widehat{\theta}) = \theta\}.$$

That is, $\mathcal{C}$ is the collection of *all* unbiased point estimators of θ. Our goal is to find the (unbiased) point estimator $\widehat{\theta} \in \mathcal{C}$ that has the smallest variance. We call this point estimator the **minimum variance unbiased estimator (MVUE)** of θ.

Remark: On the surface, finding the best estimator seems insurmountable because $\mathcal{C}$ is a large collection. For example, suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ)

distribution, where $\theta > 0$ is unknown. From Example 9.1 (notes, pp 112), we know

$$\begin{aligned} E(\overline{Y}) &= \theta \\ E(S^2) &= \theta, \end{aligned}$$

that is, both $\widehat{\theta}_1 = \overline{Y}$ and $\widehat{\theta}_2 = S^2$ are unbiased estimators of θ. Therefore, both $\widehat{\theta}_1 \in \mathcal{C}$ and $\widehat{\theta}_2 \in \mathcal{C}$. How many other (unbiased) point estimators are in $\mathcal{C}$? Note that

$$\widehat{\theta}_a = a\overline{Y} + (1-a)S^2 \in \mathcal{C}$$

for any $a \in (0, 1)$ because

$$E(\widehat{\theta}_a) = E[a\overline{Y} + (1-a)S^2] = aE(\overline{Y}) + (1-a)E(S^2) = a\theta + (1-a)\theta = \theta.$$

Therefore, we are faced with the task of finding the unbiased point estimator with the smallest variance from an (uncountably) infinite collection of point estimators!

Remark: To make this problem tractable (i.e., finding the unbiased point estimator with the smallest variance), we introduce helpful theory that will allow us to solve it. We will soon learn the critical role sufficiency plays in solving the problem.

Rao-Blackwell Theorem: Suppose $\widehat{\theta}$ is an unbiased estimator of θ; i.e.,

$$E(\widehat{\theta}) = \theta,$$

and suppose $T = T(Y_1, Y_2, ..., Y_n)$ is a sufficient statistic for θ. Define

$$\widehat{\theta}^* = E(\widehat{\theta}|T),$$

the conditional expectation of $\widehat{\theta}$ given T. Then,

$$\begin{aligned} E(\widehat{\theta}^*) &= \theta \\ V(\widehat{\theta}^*) &\leq V(\widehat{\theta}). \end{aligned}$$

Discussion: The Rao-Blackwell Theorem says one thing: we can always improve an unbiased point estimator $\widehat{\theta}$ by conditioning it on a sufficient statistic T. By "improve," we mean that we can reduce the variance of $\widehat{\theta}$ (or, at least, not increase it). Now, what is $\widehat{\theta}^* = E(\widehat{\theta}|T)$ exactly? Recall that conditional expectations (STAT 511, CH 5) are always functions of the random variable on which you are conditioning, here, T, a sufficient statistic. Therefore, whatever $\widehat{\theta}^* = E(\widehat{\theta}|T)$ is, we know *it is a function of a sufficient statistic.*

Remark: To use the Rao-Blackwell Theorem (towards finding the MVUE), some students think they have to

1. Find an unbiased estimator $\widehat{\theta}$.
2. Find a sufficient statistic T.

3. Derive the conditional distribution $\widehat{\theta}$ given T.

4. Find the mean $E(\widehat{\theta}|T)$ of this conditional distribution.

This is not the case at all! Because $\widehat{\theta}^* = E(\widehat{\theta}|T)$ is a function of a sufficient statistic T, the Rao-Blackwell Theorem simply convinces us that in our search for the MVUE, we can restrict attention to those point estimators that are functions of a sufficient statistic.

Proof. Suppose $\widehat{\theta}$ is an unbiased estimator of θ, and suppose $T = T(Y_1, Y_2, ..., Y_n)$ is a sufficient statistic. We first point out that $\widehat{\theta}^* = E(\widehat{\theta}|T)$ is a point estimator and it does not depend on θ. Because T is sufficient, we know the conditional distribution of $\widehat{\theta}$, given T, does not depend on θ. Therefore, the conditional mean $E(\widehat{\theta}|T)$ does not depend on θ either. Using our iterated rule for expectations (STAT 511, CH5), we have

$$E(\widehat{\theta}^*) = E[E(\widehat{\theta}|T)] = E(\widehat{\theta}) = \theta,$$

because $\widehat{\theta}$ is an unbiased estimator by assumption. Using our iterated rule for variances ("Adam's Rule," STAT 511, CH5), we have

$$V(\widehat{\theta}) = E[V(\widehat{\theta}|T)] + V[E(\widehat{\theta}|T)] = \underbrace{E[V(\widehat{\theta}|T)]}_{\geq\, 0} + V(\widehat{\theta}^*).$$

Because variances are nonnegative, $V(\widehat{\theta}|T)$ is a nonnegative random variable. The mean of a nonnegative random variable is nonnegative so $E[V(\widehat{\theta}|T)] \geq 0$. This shows $V(\widehat{\theta}) \geq V(\widehat{\theta}^*)$ so we are done. □

Recipe for finding MVUEs: Suppose we want to find the MVUE for θ.

1. Start by finding a sufficient statistic $T = T(Y_1, Y_2, ..., Y_n)$. *The Rao-Blackwell Theorem guarantees us that the MVUE must depend on T.*

2. Find a function of T that is an unbiased estimator of θ. This is usually accomplished by calculating $E(T)$ directly and then adjusting T to "make it unbiased."

We now illustrate this recipe by using numerous examples. Before we do, we have to make two important remarks.

Remark: For this recipe to work, we need a sufficient statistic T to possess one additional theoretical characteristic, namely, we need T to satisfy the following condition:

> **Completeness:** $E[\phi(T)] = 0$ for all $\theta \Longrightarrow \phi(T) = 0$ for all θ, with probability 1; i.e., the only function $\phi(T)$ that is an unbiased estimator of 0 is $\phi(T) = 0$.

The completeness requirement is technical, but it is crucial in ensuring our recipe for finding MVUEs is valid. The authors of your textbook make virtually no mention of this requirement because it is viewed as too advanced for an undergraduate sequence in mathematical statistics. For this reason, numerous reviewers of WMS (and those of us that teach from

it) have criticized the authors for leaving out this crucial requirement. Personally, I tend to side with the authors' approach because all of the distributions and sufficient statistics presented in the text enjoy the completeness property. However, you should know there are distributions which give rise to sufficient statistics that are *not* complete, in which case the recipe we have just presented does not work.

Remark: We also need to know the following result:

> **Uniqueness:** If a MVUE exists, then it is unique.

The uniqueness property of MVUEs guarantees that once we have found an unbiased estimator of θ that is a function of a sufficient statistic T, then this unbiased estimator is *the* MVUE. There cannot be more than one function of T which estimates θ unbiasedly, and, of course, Rao-Blackwell guarantees that no estimator that does not depend on a sufficient statistic can be MVUE.

Example 9.4 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population distribution, where $\theta > 0$ is unknown. We have already shown $T = \sum_{i=1}^n Y_i$ is a sufficient statistic for θ. Therefore, the MVUE must be a function of T. Note that

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

Therefore,

$$E(\overline{Y}) = E\left(\frac{T}{n}\right) = \frac{n\theta}{n} = \theta.$$

This shows $\widehat{\theta} = \overline{Y}$ is the MVUE of θ. It is a function of a sufficient statistic $T = \sum_{i=1}^n Y_i$ and is unbiased. $\square$

Example 9.5 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Rayleigh(θ) population distribution, where $\theta > 0$ is unknown. Recall the Rayleigh pdf is given by

$$f_Y(y|\theta) = \begin{cases} \frac{2y}{\theta} e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We have already shown $T = \sum_{i=1}^n Y_i^2$ is a sufficient statistic for θ. Therefore, the MVUE must be a function of T. Note that

$$E(T) = E\left(\sum_{i=1}^n Y_i^2\right) = \sum_{i=1}^n E(Y_i^2) = \sum_{i=1}^n \theta = n\theta,$$

because $U_i = Y_i^2 \sim$ exponential(θ). Therefore,

$$E\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right) = E\left(\frac{T}{n}\right) = \frac{n\theta}{n} = \theta.$$

This shows

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

is the MVUE of θ. It is a function of a sufficient statistic $T = \sum_{i=1}^{n} Y_i^2$ and is unbiased. $\square$

Example 9.6 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a beta$(\theta, 1)$ population distribution, where $\theta > 0$ is unknown. Recall the beta$(\theta, 1)$ pdf is given by

$$f_Y(y) = \begin{cases} \theta y^{\theta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

We have already shown $T = \prod_{i=1}^{n} Y_i$ is a sufficient statistic for θ. Therefore, the MVUE must be a function of T. Consider the function

$$-\ln T = -\ln \prod_{i=1}^{n} Y_i = \sum_{i=1}^{n} -\ln Y_i = \sum_{i=1}^{n} U_i,$$

where $U_i = -\ln Y_i$, for $i = 1, 2, ..., n$.

Result: $Y \sim \text{beta}(\theta, 1) \implies U = -\ln Y \sim \text{exponential}(1/\theta)$.

Proof. We use the transformation method. Note that $h(y) = -\ln y$ is strictly decreasing and hence one-to-one over $R_Y = \{y : 0 < y < 1\}$. To find the support of U, note that

$$0 < y < 1 \iff u = -\ln y > 0.$$

Therefore, $R_U = \{u : u > 0\}$. We now find the inverse transformation:

$$u = h(y) = -\ln y \implies y = h^{-1}(u) = e^{-u}.$$

The derivative of the inverse transformation is

$$\frac{d}{du} h^{-1}(u) = \frac{d}{du} e^{-u} = -e^{-u}.$$

Therefore, for $u > 0$, the pdf of U is

$$\begin{aligned} f_U(u) &= f_Y(h^{-1}(u)) \left| \frac{d}{du} h^{-1}(u) \right| \\ &= \theta(e^{-u})^{\theta-1} \times |-e^{-u}| = \theta(e^{-u})^{\theta} = \theta e^{-\theta u}. \end{aligned}$$

Summarizing, the pdf of $U = h(Y) = -\ln Y$ is

$$f_U(u) = \begin{cases} \theta e^{-\theta u}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this as an exponential pdf with mean $1/\theta$. This establishes the result. $\square$

Returning to the problem at hand, $U_1, U_2, ..., U_n$ are iid exponential with mean $1/\theta$. Therefore,

$$-\ln T = \sum_{i=1}^{n} U_i \sim \text{gamma}\left(n, \frac{1}{\theta}\right) \implies E(-\ln T) = \frac{n}{\theta} \implies E\left(-\frac{\ln T}{n}\right) = \frac{1}{\theta}.$$

We have found a function of a sufficient statistic $T = \prod_{i=1}^{n} Y_i$ whose expectation is $1/\theta$ (not θ). Therefore, we are still not done. Let's set

$$V = -\ln T \sim \text{gamma}\left(n, \frac{1}{\theta}\right)$$

and calculate the first inverse moment of V. Note that

$$\begin{aligned} E\left(\frac{1}{V}\right) = \int_{\mathbb{R}} \frac{1}{v} f_V(v) dv &= \int_0^{\infty} \frac{1}{v} \underbrace{\frac{\theta^n}{\Gamma(n)} v^{n-1} e^{-\theta v}}_{\text{gamma}(n, 1/\theta) \text{ pdf}} dv \\ &= \frac{\theta^n}{\Gamma(n)} \int_0^{\infty} v^{(n-1)-1} e^{-\theta v} dv \\ &= \frac{\theta^n}{\Gamma(n)} \Gamma(n-1) \left(\frac{1}{\theta}\right)^{n-1} = \frac{\Gamma(n-1)\theta^n}{(n-1)\Gamma(n-1)\theta^{n-1}} = \frac{\theta}{n-1}. \end{aligned}$$

Therefore,

$$\frac{\theta}{n-1} = E\left(\frac{1}{V}\right) = E\left(-\frac{1}{\ln T}\right) \implies E\left(-\frac{n-1}{\ln T}\right) = \theta.$$

This shows

$$\widehat{\theta} = -\frac{n-1}{\ln T} = -\frac{n-1}{\ln \prod_{i=1}^{n} Y_i} = -\frac{n-1}{\sum_{i=1}^{n} \ln Y_i}$$

is the MVUE of θ. It is a function of a sufficient statistic $T = \prod_{i=1}^{n} Y_i$ and is unbiased. □

Example 9.7 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ population distribution, where $\theta > 0$ is unknown. We have already shown $T = Y_{(n)}$ is a sufficient statistic for θ. Therefore, the MVUE must be a function of T. In Example 8.1 (notes, pp 75-77), we showed

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

This shows

$$\widehat{\theta} = \left(\frac{n+1}{n}\right) Y_{(n)}$$

is the MVUE of θ. It is a function of a sufficient statistic $T = Y_{(n)}$ and is unbiased. □

Example 9.8 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from

$$f_Y(y|\theta) = \begin{cases} e^{-(y-\theta)}, & y > \theta \\ 0, & \text{otherwise.} \end{cases}$$

We have already shown $T = Y_{(1)}$ is a sufficient statistic for θ. Therefore, the MVUE must be a function of T. Let's find the pdf of $Y_{(1)}$ so we can calculate its expectation. Recall that in general,

$$f_{Y_{(1)}}(y) = nf_Y(y)[1 - F_Y(y)]^{n-1}.$$

The population cdf is

$$F_Y(y) = \begin{cases} 0, & y \leq \theta \\ 1 - e^{-(y-\theta)}, & y > \theta. \end{cases}$$

Therefore, for $y > \theta$, the pdf of $Y_{(1)}$ is

$$f_{Y_{(1)}}(y) = ne^{-(y-\theta)}\left\{1 - [1 - e^{-(y-\theta)}]\right\}^{n-1} = ne^{-(y-\theta)}\left[e^{-(y-\theta)}\right]^{n-1} = n\left[e^{-(y-\theta)}\right]^{n} = ne^{-n(y-\theta)}.$$

Summarizing,

$$f_{Y_{(1)}}(y) = \begin{cases} ne^{-n(y-\theta)}, & y > \theta \\ 0, & \text{otherwise.} \end{cases}$$

The mean of $Y_{(1)}$ is

$$E(Y_{(1)}) = \int_{\mathbb{R}} yf_{Y_{(1)}}(y)dy = \int_{\theta}^{\infty} y \times ne^{-n(y-\theta)}dy.$$

In the last integral, let

$$u = y - \theta \implies du = dy$$

so that

$$E(Y_{(1)}) = \int_0^{\infty} (u + \theta)\, ne^{-nu}du = E(U + \theta),$$

where $U \sim$ exponential$(1/n)$; note that ne^{-nu} is the exponential$(1/n)$ pdf and the last integral is over $(0, \infty)$. Therefore,

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

This shows

$$\widehat{\theta} = Y_{(1)} - \frac{1}{n}$$

is the MVUE of θ. It is a function of a sufficient statistic $T = Y_{(1)}$ and is unbiased. □

Example 9.9 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(θ) population distribution, where $\theta > 0$ is unknown. We have already shown $T = \sum_{i=1}^n Y_i$ is a sufficient statistic for θ. Therefore, the MVUE must be a function of T. Because

$$E(\overline{Y}) = \theta,$$

it follows that $\widehat{\theta} = \overline{Y}$ is the MVUE of θ. It is a function of a sufficient statistic $T = \sum_{i=1}^n Y_i$ and is unbiased. □

Remark: In some problems, we are more interested in estimating a **function** of θ, say $\tau(\theta)$, where $\tau : \mathbb{R} \to \mathbb{R}$. To find the MVUE of $\tau(\theta)$, we adjust our MVUE recipe slightly.

1. Start by finding a sufficient statistic $T = T(Y_1, Y_2, ..., Y_n)$.
2. Find a function of T that is an unbiased estimator of $\tau(\theta)$.

Example 9.9 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(θ) distribution, where $\theta > 0$ is unknown. Find the MVUE of $\tau(\theta) = \theta^2$, the population variance.

Solution. We have already shown $T = \sum_{i=1}^n Y_i$ is a sufficient statistic and that $\overline{Y}$ is the MVUE of θ. Therefore, consider $\overline{Y}^2$. Using the variance computing formula, we have

$$E(\overline{Y}^2) = V(\overline{Y}) + [E(\overline{Y})]^2 = \frac{\theta^2}{n} + \theta^2 = \theta^2\left(\frac{1}{n} + 1\right) = \left(\frac{n+1}{n}\right)\theta^2,$$

showing that $\overline{Y}^2$ is biased. However,

$$E(\overline{Y}^2) = \left(\frac{n+1}{n}\right)\theta^2 \implies E\left[\left(\frac{n}{n+1}\right)\overline{Y}^2\right] = \theta^2.$$

This shows

$$\left(\frac{n}{n+1}\right)\overline{Y}^2$$

is the MVUE of θ^2. It is a function of a sufficient statistic $T = \sum_{i=1}^n Y_i$ and is unbiased. $\square$

Example 9.12. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a normal distribution with mean μ and variance 1. The $\mathcal{N}(\mu, 1)$ pdf is given by

$$f_Y(y|\mu) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the MVUE of μ.
(b) Find the MVUE of $\tau(\mu) = \exp(\mu)$.

Solutions. We start by finding a sufficient statistic T. The likelihood function is given by

$$\begin{aligned} L(\mu|\mathbf{y}) = \prod_{i=1}^n f_Y(y_i|\mu) &= f_Y(y_1|\mu) \times f_Y(y_2|\mu) \times \cdots \times f_Y(y_n|\mu) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_1-\mu)^2} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_2-\mu)^2} \times \cdots \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_n-\mu)^2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2}. \end{aligned}$$

We now write

$$\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i^2 - 2\mu y_i + \mu^2) = \sum_{i=1}^n y_i^2 - 2\mu \sum_{i=1}^n y_i + n\mu^2.$$

Therefore, we can write the likelihood function as

$$\begin{aligned} L(\mu|\mathbf{y}) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\left(\sum_{i=1}^n y_i^2 - 2\mu\sum_{i=1}^n y_i + n\mu^2\right)} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n y_i^2} \times e^{\mu\sum_{i=1}^n y_i} \times e^{-n\mu^2/2} \\ &= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{\mu\sum_{i=1}^n y_i} \times e^{-n\mu^2/2}}_{g(t,\mu)} \times \underbrace{e^{-\frac{1}{2}\sum_{i=1}^n y_i^2}}_{h(y_1,y_2,\ldots,y_n)}, \end{aligned}$$

where $t = \sum_{i=1}^n y_i$. By the Factorization Theorem, it follows that $T = \sum_{i=1}^n Y_i$ is a sufficient statistic for μ.

(a) The MVUE of μ must be a function of T. Because

$$E(\overline{Y}) = \mu,$$

it follows that $\overline{Y}$ is the MVUE of μ. It is a function of a sufficient statistic $T = \sum_{i=1}^n Y_i$ and is unbiased.

(b) We now have to find a function of $T = \sum_{i=1}^n Y_i$ that is an unbiased estimator of $\tau(\mu) = \exp(\mu)$. Because $\overline{Y}$ is MVUE for μ, let's start by working with $\exp(\overline{Y})$. Note that

$$E[\exp(\overline{Y})] = E(e^{\overline{Y}}) = E(e^{t\overline{Y}})\Big|_{t=1} = m_{\overline{Y}}(1),$$

where $m_{\overline{Y}}(t)$ is the mgf of $\overline{Y}$. We know

$$\begin{aligned} \overline{Y} \sim \mathcal{N}\left(\mu, \frac{1}{n}\right) &\implies m_{\overline{Y}}(t) = \exp\left[\mu t + \frac{(\frac{1}{n})t^2}{2}\right] \\ &\implies m_{\overline{Y}}(1) = \exp\left(\mu + \frac{1}{2n}\right) = \exp(\mu)\exp\left(\frac{1}{2n}\right). \end{aligned}$$

Therefore, we have shown

$$E[\exp(\overline{Y})] = \exp(\mu)\exp\left(\frac{1}{2n}\right) \implies E\left[\frac{\exp(\overline{Y})}{\exp\left(\frac{1}{2n}\right)}\right] = \exp(\mu).$$

This shows

$$\frac{\exp(\overline{Y})}{\exp\left(\frac{1}{2n}\right)} = \exp\left(\overline{Y} - \frac{1}{2n}\right) = e^{\overline{Y}-\frac{1}{2n}}$$

is the MVUE of $\tau(\mu) = \exp(\mu)$. It is a function of a sufficient statistic $T = \sum_{i=1}^n Y_i$ and is unbiased. $\square$

9.5 Method of moments

Preview: Having just discussed the notion of a "best" point estimator (i.e., the MVUE), it is important to remember that our treatment of this problem was limited to iid sampling and relatively simple population-level models (i.e., those models involving a single parameter θ). This represents only a small fraction of the possible scenarios we may encounter where estimation is needed. Therefore, we pursue two additional methods which will produce point estimators:

- method of moments
- method of maximum likelihood.

These methods can be applied in a variety of situations and their utility is not limited to iid sampling and/or single parameter population-level models.

Method of moments: Suppose $Y_1, Y_2, ..., Y_n$ is a sample from a population distribution, denoted by $p_Y(y|\boldsymbol{\theta})$ or $f_Y(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$ is an unknown population parameter. The **method of moments (MOM)** approach says to equate population moments to sample moments and solve the resulting system of equations for all unknown parameters. Recall the kth **population moment** is

$$\mu_k' = E(Y^k),$$

and define the kth **sample moment** to be

$$m_k' = \frac{1}{n}\sum_{i=1}^{n} Y_i^k.$$

Let d denote the number of parameters to be estimated; i.e., d is the dimension of $\boldsymbol{\theta}$. The method of moments (MOM) procedure uses the following system of d equations and d unknowns:

$$\begin{aligned} \mu_1' &= m_1' \\ \mu_2' &= m_2' \\ &\vdots \\ \mu_d' &= m_d'. \end{aligned}$$

Estimators are obtained by solving this system algebraically for $\theta_1, \theta_2, ..., \theta_d$. Population moments $\mu_1', \mu_2', ..., \mu_d'$ will usually be functions of $\theta_1, \theta_2, ..., \theta_d$. The resulting estimators are called **method of moments estimators**. If $\boldsymbol{\theta}$ is a scalar (i.e., if $d = 1$), then we only need one equation. If $d = 2$, we need 2 equations, and so on.

Example 9.13. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ population distribution, where $\theta > 0$ is unknown. Recall the $\mathcal{U}(0, \theta)$ pdf is given by

$$f_Y(y|\theta) = \begin{cases} \dfrac{1}{\theta}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

In the $\mathcal{U}(0, \theta)$ population-level model, there is $d = 1$ parameter. The first population moment is

$$E(Y) = \frac{\theta}{2}.$$

The first sample moment is

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}.$$

Therefore, the MOM estimator of θ is found by solving

$$\frac{\theta}{2} \stackrel{\text{set}}{=} \overline{Y} \implies \widehat{\theta} = 2\overline{Y}.$$

The MOM estimator of θ is $\widehat{\theta} = 2\overline{Y}$. $\square$

Example 9.14. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a beta$(\theta, 1)$ population distribution, where $\theta > 0$ is unknown. Recall the beta$(\theta, 1)$ pdf is given by

$$f_Y(y|\theta) = \begin{cases} \theta y^{\theta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

In the beta$(\theta, 1)$ population-level model, there is $d = 1$ parameter. The first population moment is

$$E(Y) = \frac{\theta}{\theta + 1}.$$

The first sample moment is

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}.$$

Therefore, the MOM estimator of θ is found by solving

$$\begin{aligned} \frac{\theta}{\theta+1} \stackrel{\text{set}}{=} \overline{Y} &\implies \theta = (\theta+1)\overline{Y} \\ &\implies \theta = \theta\overline{Y} + \overline{Y} \\ &\implies \theta - \theta\overline{Y} = \overline{Y} \implies \theta(1-\overline{Y}) = \overline{Y} \implies \widehat{\theta} = \frac{\overline{Y}}{1-\overline{Y}}. \end{aligned}$$

The MOM estimator of θ is $\widehat{\theta} = \overline{Y}/(1-\overline{Y})$. $\square$

Example 9.15. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Rayleigh(θ) population distribution, where $\theta > 0$ is unknown. Recall the Rayleigh pdf is given by

$$f_Y(y|\theta) = \begin{cases} \frac{2y}{\theta}e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The first population moment is

$$E(Y) = \int_{\mathbb{R}} y f_Y(y) dy = \int_0^\infty \frac{2y^2}{\theta} e^{-y^2/\theta} dy.$$

In the last integral, let

$$u = y^2 \implies du = 2y\ dy.$$

Therefore, we have

$$E(Y) = \int_0^\infty \frac{2y^2}{\theta} e^{-u/\theta}\ \frac{du}{2y} = \frac{1}{\theta}\int_0^\infty u^{\frac{1}{2}} e^{-u/\theta} du = \frac{1}{\theta}\Gamma\left(\frac{3}{2}\right)\theta^{\frac{3}{2}} = \Gamma\left(\frac{3}{2}\right)\sqrt{\theta}.$$

The first sample moment is

$$\frac{1}{n}\sum_{i=1}^n Y_i = \overline{Y}.$$

Therefore, the MOM estimator of θ is found by solving

$$\Gamma\left(\frac{3}{2}\right)\sqrt{\theta} \stackrel{\text{set}}{=} \overline{Y} \implies \sqrt{\theta} = \frac{\overline{Y}}{\Gamma\left(\frac{3}{2}\right)} \implies \widehat{\theta} = \left[\frac{\overline{Y}}{\Gamma\left(\frac{3}{2}\right)}\right]^2.$$

The MOM estimator of θ is $\widehat{\theta} = \overline{Y}^2/\Gamma^2(3/2)$. $\square$

Observation: In the last three examples, we see that the MOM estimator in each case is *not* a function of a sufficient statistic.

- Example 9.13: $\mathcal{U}(0, \theta)$. The MOM estimator $\widehat{\theta} = 2\overline{Y}$ is not a function of the sufficient statistic $T(Y_1, Y_2, ..., Y_n) = Y_{(n)}$.
- Example 9.14: beta$(\theta, 1)$. The MOM estimator $\widehat{\theta} = \overline{Y}/(1-\overline{Y})$ is not a function of the sufficient statistic $T(Y_1, Y_2, ..., Y_n) = \prod_{i=1}^n Y_i$.
- Example 9.15: Rayleigh(θ). The MOM estimator $\widehat{\theta} = \overline{Y}^2/\Gamma^2(3/2)$ is not a function of the sufficient statistic $T(Y_1, Y_2, ..., Y_n) = \sum_{i=1}^n Y_i^2$.

Discussion: Moments describe only limited aspects of a distribution (population or sample), so it is not surprising that MOM estimators are not necessarily the best estimators. Remember that if a point estimator does not depend on a sufficient statistic, then it cannot be "best" as we have defined it (this is what the Rao-Blackwell Theorem guarantees).

Remark: In some instances, MOM estimators are best. For example, suppose

- $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population distribution, or
- $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(θ) population distribution.

In both instances, the MOM estimator is $\widehat{\theta} = \overline{Y}$, and, in both instances, a sufficient statistic is $T(Y_1, Y_2, ..., Y_n) = \sum_{i=1}^n Y_i$. Therefore, the MOM estimator in each case is (a) unbiased and (b) a function of a sufficient statistic, and hence is the MVUE. However, there is no theory which says MOM estimators are necessarily optimal, as the preceding examples illustrate.

Example 9.16. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a gamma(α, β) population distribution, where both $\alpha > 0$ and $\beta > 0$ are unknown. In this population-level model, there are $d = 2$ parameters, so we will need 2 equations to find the MOM estimators. The first two population moments are

$$\begin{aligned} E(Y) &= \alpha\beta \\ E(Y^2) = V(Y) + [E(Y)]^2 &= \alpha\beta^2 + (\alpha\beta)^2. \end{aligned}$$

The first two sample moments are

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} Y_i &= \overline{Y} \\ \frac{1}{n}\sum_{i=1}^{n} Y_i^2 &= m_2'. \end{aligned}$$

Therefore, the MOM estimators of α and β are found by solving

$$\begin{aligned} \alpha\beta &\stackrel{\text{set}}{=} \overline{Y} \\ \alpha\beta^2 + (\alpha\beta)^2 &\stackrel{\text{set}}{=} m_2'. \end{aligned}$$

Substituting the first equation into the second, we get

$$\alpha\beta^2 + \overline{Y}^2 = m_2' \implies \alpha\beta^2 = m_2' - \overline{Y}^2.$$

Solving for β in the first equation, we get

$$\beta = \frac{\overline{Y}}{\alpha} \implies \alpha\left(\frac{\overline{Y}}{\alpha}\right)^2 = m_2' - \overline{Y}^2 \implies \frac{1}{\alpha} = \frac{m_2' - \overline{Y}^2}{\overline{Y}^2} \implies \widehat{\alpha} = \frac{\overline{Y}^2}{m_2' - \overline{Y}^2}.$$

Substituting $\widehat{\alpha}$ into the original system (the first equation), we get

$$\widehat{\alpha}\beta = \overline{Y} \implies \widehat{\beta} = \frac{\overline{Y}}{\widehat{\alpha}} \implies \widehat{\beta} = \frac{\overline{Y}}{\frac{\overline{Y}^2}{m_2' - \overline{Y}^2}} \implies \widehat{\beta} = \frac{m_2' - \overline{Y}^2}{\overline{Y}}.$$

These are the MOM estimators of α and β, respectively. □

Observation: In Example 9.10 (notes, pp 126-127), we showed a sufficient statistic for $\boldsymbol{\theta} = (\alpha, \beta)$ under the gamma population model was

$$\mathbf{T} = \begin{pmatrix} \prod_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i \end{pmatrix}.$$

Again, we see that the MOM estimators of α and β (in Example 9.16) are not functions of the sufficient statistic.

9.6 Maximum likelihood estimation

Remark: The method of maximum likelihood is, by far, the most common method to use when finding point estimators; i.e., when estimating a population-level model. Maximum likelihood estimators are found by maximizing the likelihood function.

Recall: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution denoted by $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where θ (a scalar) is an unknown population parameter. Recall the **likelihood function**, denoted by $L(\theta|\mathbf{y}) = L(\theta|y_1, y_2, ..., y_n)$, is found as follows:

- In the discrete case,

$$L(\theta|\mathbf{y}) = p_Y(y_1|\theta) \times p_Y(y_2|\theta) \times \cdots \times p_Y(y_n|\theta) = \prod_{i=1}^{n} p_Y(y_i|\theta).$$

- In the continuous case,

$$L(\theta|\mathbf{y}) = f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) = \prod_{i=1}^{n} f_Y(y_i|\theta).$$

Note: In the discrete case, the likelihood function $L(\theta|\mathbf{y})$ actually provides a (joint) probability. Suppressing $p_Y(y|\theta)$'s dependence on θ, note that

$$\begin{aligned} L(\theta|\mathbf{y}) &= p_Y(y_1) \times p_Y(y_2) \times \cdots \times p_Y(y_n) \\ &= P(Y_1 = y_1)P(Y_2 = y_2)\cdots P(Y_n = y_n) = \underbrace{P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n)}_{\text{joint probability of the sample}}. \end{aligned}$$

In this light, a casual (but useful) interpretation of the likelihood function is that it is "the probability of the data." Therefore, one can think of maximum likelihood estimates as those estimates which "maximize the probability of the data."

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution denoted by $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where $\theta \in \mathbb{R}$ is an unknown population parameter (for now, we consider the scalar case). The **maximum likelihood estimator** (**MLE**) of θ is the value of θ that maximizes the likelihood function $L(\theta|\mathbf{y})$; i.e.,

$$\widehat{\theta} = \arg\max_{\theta} L(\theta|\mathbf{y}).$$

Therefore, we can find the MLE of θ by writing out the likelihood function, viewing it as a function of θ (as always), and then maximizing it as a function of θ.

Q: How do we maximize $L(\theta|\mathbf{y})$?
A: For many scenarios, this reduces to a calculus problem. If $L(\theta|\mathbf{y})$ is a differentiable function of θ, then we can take the derivative of $L(\theta|\mathbf{y})$ and set it equal to zero; i.e.,

$$\frac{\partial}{\partial\theta} L(\theta|\mathbf{y}) \stackrel{\text{set}}{=} 0.$$

Calculus trick: Because the natural logarithm function is increasing, the value of θ that maximizes $L(\theta|\mathbf{y})$ is the same as the value of θ that maximizes $\ln L(\theta|\mathbf{y})$; i.e.,

$$\widehat{\theta} = \arg\max_{\theta} L(\theta|\mathbf{y}) = \arg\max_{\theta} \ln L(\theta|\mathbf{y}).$$

Therefore, we can also take the derivative of the **log-likelihood function** $\ln L(\theta|\mathbf{y})$ and set it equal to zero; i.e.,

$$\frac{\partial}{\partial\theta} \ln L(\theta|\mathbf{y}) \stackrel{\text{set}}{=} 0.$$

We then solve this equation (which is called the **score equation**) for θ to get a first-order critical point $\widehat{\theta}$. We then show

$$\frac{\partial^2}{\partial\theta^2} \ln L(\theta|\mathbf{y})\Big|_{\theta=\widehat{\theta}} < 0,$$

which verifies $\widehat{\theta}$ maximizes $\ln L(\theta|\mathbf{y})$ by the Second Derivative Test. If $\widehat{\theta}$ maximizes $\ln L(\theta|\mathbf{y})$, then $\widehat{\theta}$ maximizes $L(\theta|\mathbf{y})$ too.

Example 9.17. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population distribution, where $\theta > 0$ is unknown. Recall the Poisson(θ) pmf is given by

$$p_Y(y|\theta) = \begin{cases} \dfrac{\theta^y e^{-\theta}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

$$L(\theta|\mathbf{y}) = \frac{\theta^{y_1}e^{-\theta}}{y_1!} \times \frac{\theta^{y_2}e^{-\theta}}{y_2!} \times \cdots \times \frac{\theta^{y_n}e^{-\theta}}{y_n!} = \frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{y_1!y_2!\cdots y_n!} = \frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{\prod_{i=1}^n y_i!}.$$

The log-likelihood function is given by

$$\begin{aligned} \ln L(\theta|\mathbf{y}) &= \ln\left(\frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{\prod_{i=1}^n y_i!}\right) \\ &= \ln\left(\theta^{\sum_{i=1}^n y_i}\right) + \ln\left(e^{-n\theta}\right) - \ln\left(\prod_{i=1}^n y_i!\right) = \sum_{i=1}^n y_i \ln\theta - n\theta - \ln\left(\prod_{i=1}^n y_i!\right). \end{aligned}$$

The derivative of the log-likelihood function is given by

$$\begin{aligned} \frac{\partial}{\partial\theta} \ln L(\theta|\mathbf{y}) &= \frac{\sum_{i=1}^n y_i}{\theta} - n \stackrel{\text{set}}{=} 0 \\ &\implies \sum_{i=1}^n y_i - n\theta = 0 \implies \sum_{i=1}^n y_i = n\theta \implies \widehat{\theta} = \frac{1}{n}\sum_{i=1}^n y_i = \overline{y}. \end{aligned}$$

We now show this first-order critical point $\widehat{\theta} = \overline{y}$ maximizes $\ln L(\theta|\mathbf{y})$. The second derivative of the log-likelihood function is given by

$$\frac{\partial^2}{\partial\theta^2} \ln L(\theta|\mathbf{y}) = -\frac{\sum_{i=1}^n y_i}{\theta^2}.$$

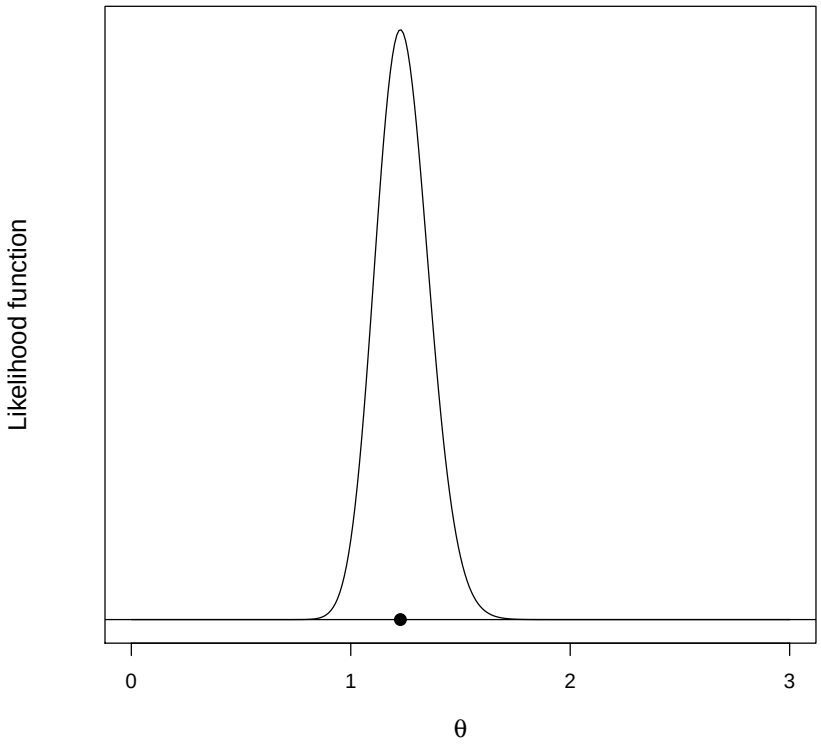

Figure 9.4: Accident data. Poisson likelihood function $L(\theta|\mathbf{y})$ in Example 9.17. The maximum likelihood estimate $\widehat{\theta} = \overline{y} \approx 1.23$ is shown by using a dark circle.

Note that

$$\frac{\partial^2}{\partial\theta^2} \ln L(\theta|\mathbf{y})\Big|_{\theta=\overline{y}} = -\frac{\sum_{i=1}^{n} y_i}{\overline{y}^2} = -\frac{n\overline{y}}{\overline{y}^2} = -\frac{n}{\overline{y}} < 0.$$

Therefore, $\widehat{\theta} = \overline{y}$ maximizes $\ln L(\theta|\mathbf{y})$. The MLE of θ is

$$\widehat{\theta} = \overline{Y}.$$

Application: In Example 9.4 (notes, pp 118-119), we examined an iid sample of $n = 84$ policies and the observed data on

$$Y = \text{ the number of accidents in a given year.}$$

Suppose the observations $Y_1, Y_2, ..., Y_{84}$ are modeled as iid Poisson counts with mean $\theta > 0$. Under this assumption, the maximum likelihood estimate of θ based on these data is

$$\widehat{\theta} = \overline{y} = \frac{1}{84}\sum_{i=1}^{84} y_i = \frac{1}{84}(103) \approx 1.23.$$

The likelihood function $L(\theta|\mathbf{y})$ based on these data is shown in Figure 9.4 (above). □

Example 9.18. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Rayleigh(θ) population distribution, where $\theta > 0$ is unknown. Recall the Rayleigh pdf is given by

$$f_Y(y|\theta) = \begin{cases} \frac{2y}{\theta}e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

$$L(\theta|\mathbf{y}) = \frac{2y_1}{\theta}e^{-y_1^2/\theta} \times \frac{2y_2}{\theta}e^{-y_2^2/\theta} \times \cdots \times \frac{2y_n}{\theta}e^{-y_n^2/\theta} = \left(\frac{2}{\theta}\right)^n \left(\prod_{i=1}^n y_i\right) e^{-\sum_{i=1}^n y_i^2/\theta}.$$

The log-likelihood function is given by

$$\begin{aligned} \ln L(\theta|\mathbf{y}) &= \ln\left[\left(\frac{2}{\theta}\right)^n \left(\prod_{i=1}^n y_i\right) e^{-\sum_{i=1}^n y_i^2/\theta}\right] \\ &= \ln\left[\left(\frac{2}{\theta}\right)^n\right] + \ln\left(\prod_{i=1}^n y_i\right) + \ln\left(e^{-\sum_{i=1}^n y_i^2/\theta}\right) \\ &= n(\ln 2 - \ln\theta) + \ln\left(\prod_{i=1}^n y_i\right) - \frac{\sum_{i=1}^n y_i^2}{\theta}. \end{aligned}$$

The derivative of the log-likelihood function is given by

$$\begin{aligned} \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{y}) &= -\frac{n}{\theta} + \frac{\sum_{i=1}^n y_i^2}{\theta^2} \stackrel{\text{set}}{=} 0 \\ &\implies -n\theta + \sum_{i=1}^n y_i^2 = 0 \implies \sum_{i=1}^n y_i^2 = n\theta \implies \widehat{\theta} = \frac{1}{n}\sum_{i=1}^n y_i^2. \end{aligned}$$

We now show this first-order critical point $\widehat{\theta}$ maximizes $\ln L(\theta|\mathbf{y})$. The second derivative of the log-likelihood function is given by

$$\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\mathbf{y}) = \frac{n}{\theta^2} - \frac{2\sum_{i=1}^n y_i^2}{\theta^3}.$$

Note that

$$\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\mathbf{y})\Big|_{\theta=\widehat{\theta}} = \frac{n}{\widehat{\theta}^2} - \frac{2\sum_{i=1}^n y_i^2}{\widehat{\theta}^3} = \frac{n\widehat{\theta}}{\widehat{\theta}^3} - \frac{2n\widehat{\theta}}{\widehat{\theta}^3} = -\frac{n\widehat{\theta}}{\widehat{\theta}^3} = -\frac{n}{\widehat{\theta}^2} < 0.$$

Therefore, $\widehat{\theta}$ maximizes $\ln L(\theta|\mathbf{y})$. The MLE of θ is

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n Y_i^2.$$

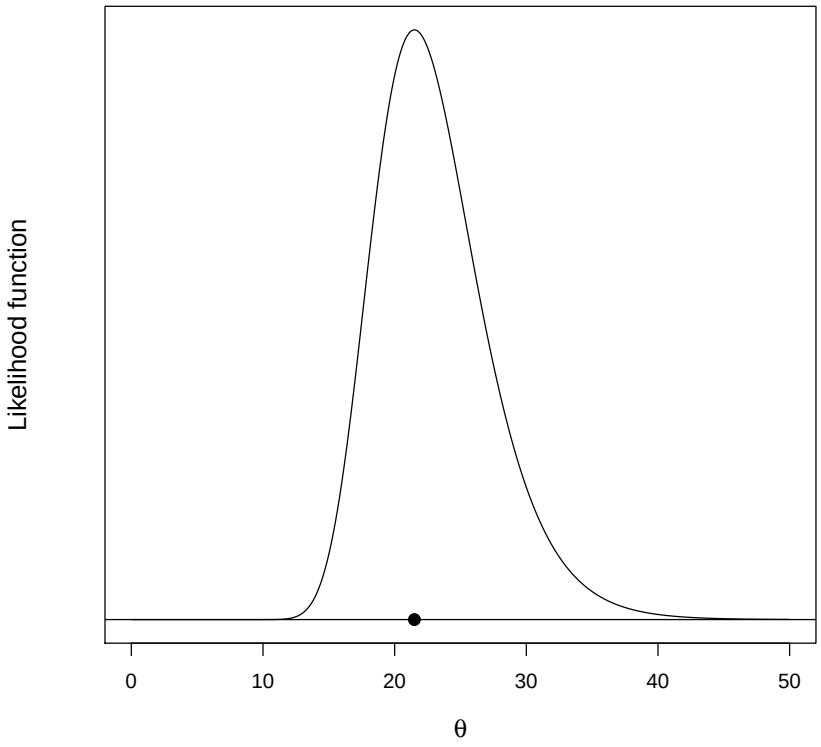

Figure 9.5: Light bulb data. Rayleigh likelihood function $L(\theta|\mathbf{y})$ in Example 9.18. The maximum likelihood estimate $\widehat{\theta} = m_2' = 21.5$ is shown by using a dark circle.

Application: In Example 9.5 (notes, pp 120-121), we examined an iid sample of $n = 30$ bulbs and the observed data on

$$Y = \text{ time until failure (in 100s hours).}$$

Suppose the observations $Y_1, Y_2, ..., Y_{30}$ are modeled as iid Rayleigh(θ), where $\theta > 0$. Under this assumption, the maximum likelihood estimate of θ based on these data is

$$\widehat{\theta} = \frac{1}{30}\sum_{i=1}^{30} y_i^2 = \frac{1}{30}(645.0) = 21.5.$$

The likelihood function $L(\theta|\mathbf{y})$ based on these data is shown in Figure 9.5 (above). □

Remark: In population-level models where the support depends on an unknown parameter, we have to be careful in how we find the MLE.

Example 9.19. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ population distribution, where $\theta > 0$ is unknown. Recall the $\mathcal{U}(0, \theta)$ pdf is given by

$$f_Y(y|\theta) = \begin{cases} \dfrac{1}{\theta}, & 0 \leq y \leq \theta \\ 0, & \text{otherwise.} \end{cases}$$

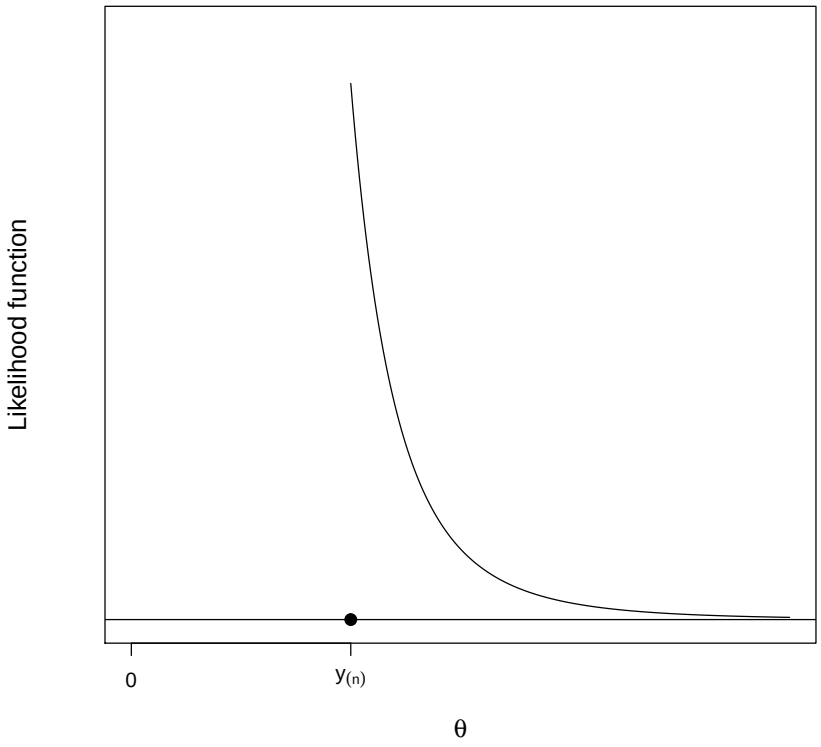

Figure 9.6: Uniform likelihood function $L(\theta|\mathbf{y})$ in Example 9.19. The maximum likelihood estimate $\widehat{\theta} = y_{(n)}$ is shown by using a dark circle.

The likelihood function is given by

$$\begin{aligned} L(\theta|\mathbf{y}) &= f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) \\ &= \frac{1}{\theta} I(0 \leq y_1 \leq \theta) \times \frac{1}{\theta} I(0 \leq y_2 \leq \theta) \times \cdots \times \frac{1}{\theta} I(0 \leq y_n \leq \theta) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 \leq y_i \leq \theta) \\ &= \frac{1}{\theta^n} I(0 \leq y_{(n)} \leq \theta). \end{aligned}$$

The likelihood function $L(\theta|\mathbf{y})$ is shown in Figure 9.6 (above). Note that $L(\theta|\mathbf{y})$ is not differentiable for all θ; therefore, we cannot use a calculus argument. However, note that

- For $\theta \geq y_{(n)}$, $L(\theta|\mathbf{y}) = 1/\theta^n$, which is a decreasing function of θ (see above).
- For $\theta < y_{(n)}$, $L(\theta|\mathbf{y}) = 0$.

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

Observation: In the last three examples, we see that the MLE in each case *is* a function of a sufficient statistic.

- Example 9.17: Poisson(θ). The MLE $\widehat{\theta} = \overline{Y}$ is a function of the sufficient statistic $T(Y_1, Y_2, ..., Y_n) = \sum_{i=1}^n Y_i$.
- Example 9.18: Rayleigh(θ). The MLE $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n Y_i^2$ is a function of the sufficient statistic $T(Y_1, Y_2, ..., Y_n) = \sum_{i=1}^n Y_i^2$.
- Example 9.19: $\mathcal{U}(0, \theta)$. The MLE $\widehat{\theta} = Y_{(n)}$ is a function of the sufficient statistic $T(Y_1, Y_2, ..., Y_n) = Y_{(n)}$.

Discussion: These examples illustrate the link that exists between maximum likelihood estimation and sufficiency. That is, if a sufficient statistic $T = T(Y_1, Y_2, ..., Y_n)$ exists, then the MLE $\widehat{\theta}$ will depend on T. This is easy to show. If T is sufficient, then by the Factorization Theorem we can write

$$L(\theta|\mathbf{y}) = L(\theta|y_1, y_2, ..., y_n) = g(t, \theta)h(y_1, y_2, ..., y_n) \propto g(t, \theta).$$

Therefore, when we maximize $L(\theta|\mathbf{y})$ or $\ln L(\theta|\mathbf{y})$ with respect to θ to find the MLE, this will only depend on $t = T(y_1, y_2, ..., y_n)$ through $g(t, \theta)$; i.e., the term $h(y_1, y_2, ..., y_n)$ is simply a proportionality constant, so it will not affect the maximization. Compare this to MOM estimators which are not guaranteed to be functions of sufficient statistics.

Remark: We now discuss situations where a population-level model includes more than one parameter; i.e., the population-level parameter to be estimated is $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$.

Multiparameter setting: Suppose $Y_1, Y_2, ..., Y_n$ is a sample from a population distribution, denoted by $p_Y(y|\boldsymbol{\theta})$ or $f_Y(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$ is an unknown population parameter. Let $L(\boldsymbol{\theta}|\mathbf{y})$ denote the likelihood function. The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}|\mathbf{y}).$$

That is, the MLE (as before in the scalar case) is the value $\widehat{\boldsymbol{\theta}}$ that maximizes the likelihood function. Maximizing $L(\boldsymbol{\theta}|\mathbf{y})$ or $\ln L(\boldsymbol{\theta}|\mathbf{y})$ is basically a multivariable calculus problem. If $\ln L(\boldsymbol{\theta}|\mathbf{y})$ is a differentiable function, then the MLE can be found by solving the system of equations

$$\begin{aligned}
\frac{\partial}{\partial \theta_1} \ln L(\boldsymbol{\theta}|\mathbf{y}) &\stackrel{\text{set}}{=} 0 \\
\frac{\partial}{\partial \theta_2} \ln L(\boldsymbol{\theta}|\mathbf{y}) &\stackrel{\text{set}}{=} 0 \\
&\vdots \\
\frac{\partial}{\partial \theta_d} \ln L(\boldsymbol{\theta}|\mathbf{y}) &\stackrel{\text{set}}{=} 0
\end{aligned}$$

for $\theta_1, \theta_2, ..., \theta_d$. The equations above are called the **score equations**.

Q: How can we verify a solution to the score equations is a maximizer?
A: Mathematically, for a d-dimensional maximization problem, we can calculate the Hessian matrix

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

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

Example 9.20. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, where both parameters $-\infty < \mu < \infty$ and $\sigma^2 > 0$ are unknown. Recall the $\mathcal{N}(\mu, \sigma^2)$ pdf is given by

$$f_Y(y|\mu, \sigma^2) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Set $\boldsymbol{\theta} = (\mu, \sigma^2)$. The likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) = L(\mu, \sigma^2|\mathbf{y}) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_1-\mu)^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_2-\mu)^2} \times \cdots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_n-\mu)^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2}. \end{aligned}$$

The log-likelihood function is given by

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

The score equations are

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

We now want to solve the score equations. Note that in the first equation, we have

$$\frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mu) = 0 \implies \sum_{i=1}^n (y_i - \mu) = 0 \implies \widehat{\mu} = \overline{y}.$$

Plugging this solution into the second equation, we get

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i - \overline{y})^2 = 0 \implies n\sigma^2 = \sum_{i=1}^n (y_i - \overline{y})^2 \implies \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2.$$

Now, let's show the first order critical point

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \widehat{\mu} \\ \widehat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \end{pmatrix}$$

is a maximizer. The Hessian matrix is

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2}{\partial \mu^2}\ln L(\boldsymbol{\theta}|\mathbf{y}) & \frac{\partial^2}{\partial \mu \partial \sigma^2}\ln L(\boldsymbol{\theta}|\mathbf{y}) \\ \frac{\partial^2}{\partial \sigma^2 \partial \mu}\ln L(\boldsymbol{\theta}|\mathbf{y}) & \frac{\partial^2}{\partial (\sigma^2)^2}\ln L(\boldsymbol{\theta}|\mathbf{y}) \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2 \end{pmatrix}.$$

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

$$\begin{aligned} \mathbf{a}'\mathbf{H}\Big|_{\mu=\widehat{\mu},\sigma^2=\widehat{\sigma}^2} &= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} -\frac{n}{\widehat{\sigma}^2} & -\frac{1}{\widehat{\sigma}^4}\sum_{i=1}^{n}(y_i - \widehat{\mu}) \\ -\frac{1}{\widehat{\sigma}^4}\sum_{i=1}^{n}(y_i - \widehat{\mu}) & \frac{n}{2\widehat{\sigma}^4} - \frac{1}{\widehat{\sigma}^6}\sum_{i=1}^{n}(y_i - \widehat{\mu})^2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\ &= -\frac{na_1^2}{\widehat{\sigma}^2} - \frac{na_2^2}{2\widehat{\sigma}^4} < 0. \end{aligned}$$

Therefore, the first-order critical point we found above *is* a maximizer and hence the MLE of $\boldsymbol{\theta} = (\mu, \sigma^2)$ is

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \overline{Y} \\ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \end{pmatrix}.$$

Observation: Note that the MLE of σ^2 in this example is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = S_b^2 \quad \text{and not} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

This demonstrates that MLEs may be biased estimators; i.e., there is no guarantee that MLEs will be unbiased. □

Exercise: Revisit Example 9.20 and determine the MLE of σ^2 when $\mu = \mu_0$ is known. In this case, there is only one unknown population parameter (σ^2) and the likelihood function is

$$L(\sigma^2|\mathbf{y}) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_0)^2}.$$

The MLE of σ^2 when $\mu = \mu_0$ is known turns out to be

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_0)^2.$$

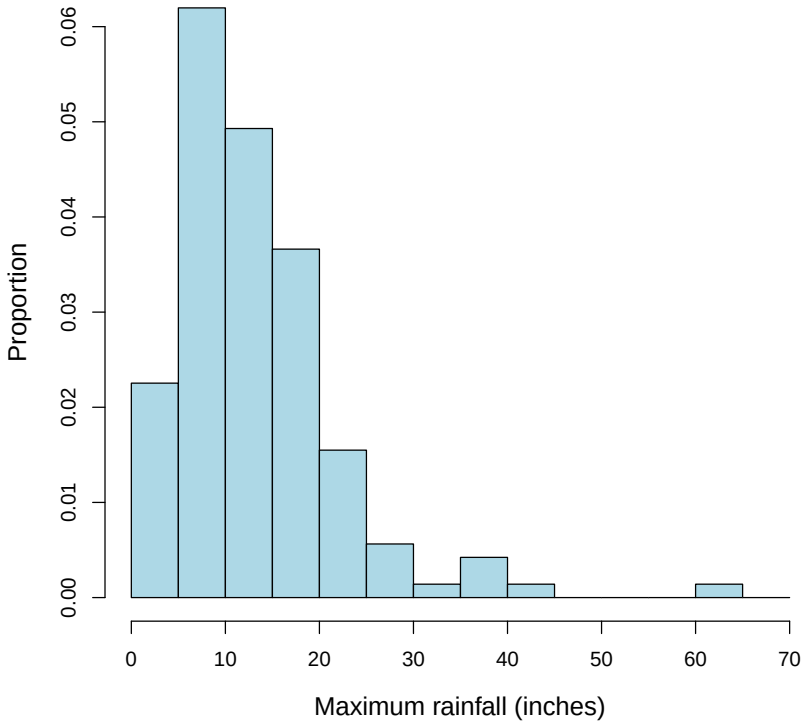

Figure 9.7: Rainfall data. Histogram of maximum rainfall amounts corresponding to $n = 142$ recent major weather events in the United States. Time period: 2000-2018.

Discussion: When $\mu = \mu_0$, the solutions

$$\begin{pmatrix} \overline{Y} \\ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \mu_0 \\ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_0)^2 \end{pmatrix}$$

should be "close" to each other in distance. What if they are not "close?" What might be true if these two solutions are far away from each other?

Remark: In most "real" problems involving data analysis, the likelihood is a complicated function which must be maximized by using numerical optimization methods. This is usually carried out by using statistical software (e.g., R, etc.).

Example 9.21. Using resources from the National Oceanic and Atmospheric Administration, I recorded the maximum rainfall amount (in inches) for the most recent $n = 142$ major weather events (e.g., hurricanes, cyclones, etc.) in the United States. Note that I only included the event when the maximum rainfall amount occurred in the United States or in one of its territories. I restricted attention to those events that occurred since 2000. A histogram of the data is shown in Figure 9.7 (above).

Analysis: Let's assume $Y_1, Y_2, ..., Y_{142}$ are iid observations from a gamma(α, β) population distribution, where $\alpha > 0$ and $\beta > 0$ are unknown, and estimate α and β by using maximum likelihood. Recall the gamma(α, β) pdf is

$$f_Y(y|\alpha,\beta) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1}e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

$$\begin{aligned} L(\alpha,\beta|\mathbf{y}) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} y_1^{\alpha-1}e^{-y_1/\beta} \times \frac{1}{\Gamma(\alpha)\beta^\alpha} y_2^{\alpha-1}e^{-y_2/\beta} \times \cdots \times \frac{1}{\Gamma(\alpha)\beta^\alpha} y_{142}^{\alpha-1}e^{-y_{142}/\beta} \\ &= \left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\right]^{142} \left(\prod_{i=1}^{142} y_i\right)^{\alpha-1} e^{-\sum_{i=1}^{142} y_i/\beta}. \end{aligned}$$

The log-likelihood function is given by

$$\ln L(\alpha,\beta|\mathbf{y}) = -142\ln\Gamma(\alpha) - 142\alpha\ln\beta + (\alpha-1)\sum_{i=1}^{142}\ln y_i - \frac{\sum_{i=1}^{142} y_i}{\beta}.$$

This function cannot be maximized analytically. If you take the derivative of $\ln L(\alpha, \beta|\mathbf{y})$ with respect to α, then you will be forced to deal with the derivative of the gamma function $\Gamma(\alpha)$, which is not easy to work with. Therefore, let's maximize $\ln L(\alpha, \beta|\mathbf{y})$ numerically (see R code online). I've done this using the Nelder-Mead numerical optimization routine using R's `optim` function. I used the MOM estimates as starting values; see Example 9.16 (pp 140, notes).

```
> mle = optim(par=c(alpha.mom,beta.mom),fn=loglike,method="Nelder-Mead")
> c(alpha.mom,beta.mom) # MOM estimates
[1] 2.316696 5.625545

> mle$par # MLEs
[1] 2.516750 5.178724
```

Therefore, the maximum likelihood estimates of α and β for the rainfall data are

$$\begin{aligned} \widehat{\alpha} &\approx 2.52 \\ \widehat{\beta} &\approx 5.18. \end{aligned}$$

I superimposed the gamma$(\widehat{\alpha} = 2.52, \widehat{\beta} = 5.18)$ density onto the histogram of the rainfall data; see Figure 9.8 (next page, left). A quantile-quantile (qq) plot of the data is shown in Figure 9.8 (right). The linear pattern in the qq plot suggests the gamma distribution is a reasonably good fit. Note that the outlier on the high side (60.58 inches) is Hurricane Harvey which hit Texas in 2017. $\square$

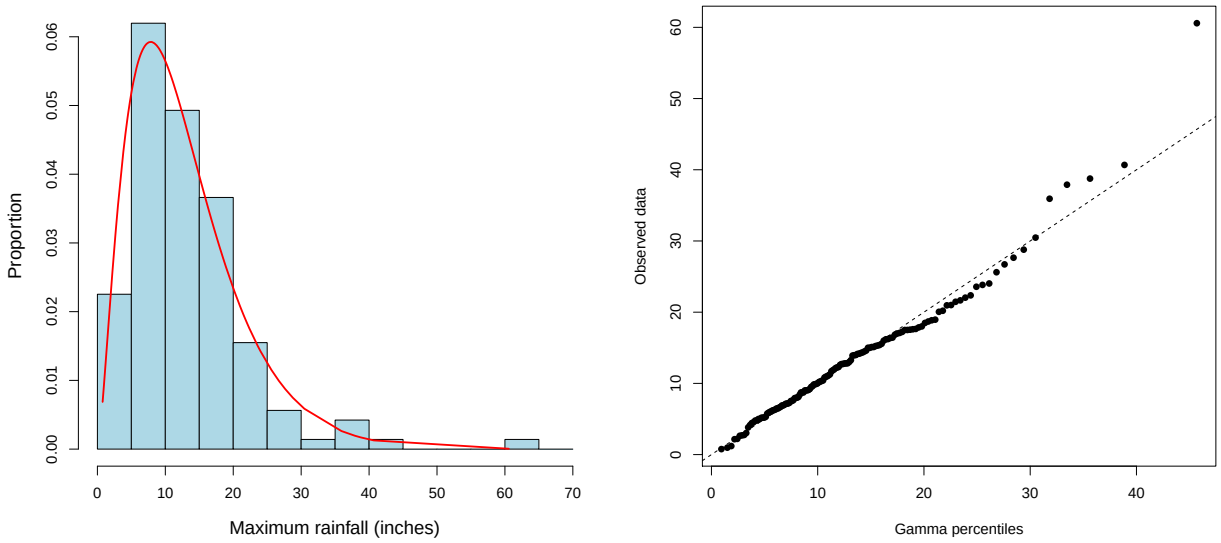

Figure 9.8: Left: Histogram of the rainfall data in Example 9.21 with a gamma$(2.52, 5.18)$ density superimposed. Right: Quantile-quantile (qq) plot for the data under a gamma population distribution assumption.

Example 9.22. *Logistic regression.* Suppose $Y_1, Y_2, ..., Y_n$ are independent Bernoulli random variables; specifically, $Y_i \sim$ Bernoulli(p_i), where

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

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

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

Taking logarithms and simplifying gives

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

Closed-form expressions for the maximizers $\widehat{\beta}_0$ and $\widehat{\beta}_1$ do not exist except in very simple situations. Again, numerical optimization methods are needed to maximize $\ln L(\beta_0, \beta_1|\mathbf{y})$. For example, R's `glm` (generalized linear model) function does this using a technique known as "iteratively re-weighted least squares." $\square$

Invariance: One of the nicest results in mathematical statistics is the invariance property of maximum likelihood estimators. Succinctly put, the invariance property says that if $\widehat{\theta}$ is the MLE of θ, then $\tau(\widehat{\theta})$ is the MLE of $\tau(\theta)$, where τ is *any* function. For example,

- $\widehat{\theta}^2$ is the MLE of θ^2
- $\sin\widehat{\theta}$ is the MLE of $\sin\theta$
- $e^{-\widehat{\theta}}$ is the MLE of $e^{-\theta}$.

The invariance property also holds in the multiparameter setting when τ is a vector-valued function. For example, in Example 9.21, the maximum likelihood estimate of $E(Y) = \alpha\beta$, the population mean maximum rainfall amount (under a gamma assumption), is

$$\widehat{\alpha}\widehat{\beta} \approx 2.52(5.18) \approx 13.05 \text{ inches.}$$

The maximum likelihood estimate of the population variance $V(Y) = \alpha\beta^2$ is

$$\widehat{\alpha}\widehat{\beta}^2 \approx 2.52(5.18)^2 \approx 67.62 \text{ (inches)}^2.$$

Example 9.23. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(β) population distribution, where $\beta > 0$ is unknown. Find the MLE of ϕ_p, the pth quantile of the distribution of Y.

Solution. Recall the pth quantile of a (continuous) probability distribution is the value ϕ_p which solves

$$p = P(Y \leq \phi_p) = F_Y(\phi_p),$$

where $F_Y(y)$ is the cdf of Y. Recall for $Y \sim$ exponential(β), the cdf is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y/\beta}, & y > 0. \end{cases}$$

Therefore, the pth quantile of $Y \sim$ exponential(β) is found by solving

$$p = 1 - e^{-\phi_p/\beta} \implies \phi_p = -\beta\ln(1-p).$$

By the invariance property of MLEs, the MLE of $\tau(\beta) = -\beta\ln(1-p)$ is $-\widehat{\beta}\ln(1-p)$, where $\widehat{\beta}$ is the MLE of β. The likelihood function is given by

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

It is easy to show (verify) the MLE of β is $\widehat{\beta} = \overline{Y}$. Therefore, the MLE of the pth quantile is

$$\widehat{\phi}_p = -\overline{Y}\ln(1-p). \ \square$$

9.7 Large-sample (asymptotic) considerations

Recall: In Chapter 8, we described important characteristics we would like a point estimator $\widehat{\theta} = T(Y_1, Y_2, ..., Y_n)$ to possess. In particular, we discussed bias, variance, and MSE of $\widehat{\theta}$, and we quantified the merit of a point estimator by using these characteristics. In this chapter, we even addressed the question of finding the "best" point estimator; i.e., finding the unbiased estimator $\widehat{\theta}$ that had the smallest possible variance (the MVUE). Throughout all of these discussions, whether or not you realized it, we were always making assessments based on a fixed sample size. Therefore, these assessments utilized finite-sample distributional results.

Remark: Many statistical inference procedures we use in practice are not based on finite-sample results. Recall from Chapter 8 when we wrote a $100(1-\alpha)\%$ confidence interval for a population mean μ to be

$$\overline{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

This is an exact (finite-sample) confidence interval, but only when $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. The reason it is exact is that the interval is derived from the pivotal quantity

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

which follows a $t(n-1)$ distribution, regardless of what the sample size n is. Therefore, the confidence coefficient of the interval is exactly $1-\alpha$. On the other hand, when the population distribution is not normal (or maybe not even known), we wrote

$$\overline{Y} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

as an *approximate* $100(1-\alpha)\%$ confidence interval for μ. This interval is not exact. Its confidence coefficient is approximately equal to $1-\alpha$ when the sample size n is large.

Discussion: Because the last interval is only approximate, it is important to see where the approximations arise. First, we are using the CLT to argue that

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{so that} \quad Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{AN}(0,1)$$

when n is large. Then, we are additionally approximating the standard error

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}} \quad \text{by using} \quad \widehat{\sigma}_{\overline{Y}} = \frac{S}{\sqrt{n}}$$

with the hope that the estimated standard error $\widehat{\sigma}_{\overline{Y}}$ will be "close" to the true standard error $\sigma_{\overline{Y}}$ when the sample size n is large. Note that the other large-sample confidence intervals we examined in Section 8.5 (notes); i.e.,

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

and

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

all rely heavily on these same types of large-sample approximations.

Remark: Because large-sample inference procedures are so common in statistics, it is important to have a mathematical understanding of large-sample (or asymptotic) results. Of course, we have already presented one very important large-sample result in Chapter 7, namely, the Central Limit Theorem. This result allowed us to approximate probabilities of events involving sample means (and sample sums).

Central Limit Theorem: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution with mean $E(Y) = \mu$ and variance $V(Y) = \sigma^2 < \infty$. The sequence of random variables

$$Z_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. We say "Z_n converges in distribution to $\mathcal{N}(0,1)$."

Preview: We now discuss additional large-sample (asymptotic) results which will enhance our understanding of why commonly used statistical methods are valid in large samples.

Q: What's the point? Large-sample results are technically valid only under the assumption that $n \to \infty$. This is not realistic.
A: Because finite-sample results are often not available (or they are intractable), and large-sample results can offer a good approximation to them when n is "large." We already saw this in Chapter 7 when we used the CLT to approximate probabilities—often these approximations were very good in finite samples.

9.7.1 Consistency

Terminology: We say an estimator $\widehat{\theta}_n$ is a **consistent** estimator of θ if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \epsilon) = 0.$$

That is, the sequence of probabilities $P(|\widehat{\theta}_n - \theta| > \epsilon) \to 0$, as $n \to \infty$.

Remark: Consistency is a desirable large-sample property for an estimator to possess. If $\widehat{\theta}_n$ is consistent, then the probability the estimator $\widehat{\theta}_n$ differs from the true θ becomes small as the sample size n becomes large. On the other hand, if an estimator is *not* consistent, then the estimator $\widehat{\theta}_n$ may never get close to θ, regardless of how large the sample size n is.

Remark: The statement "$\widehat{\theta}_n$ is a consistent estimator of θ" is often replaced by the statement "$\widehat{\theta}_n$ converges in probability to θ." If the latter terminology is used, then we write

$$\widehat{\theta}_n \xrightarrow{p} \theta, \quad \text{as } n \to \infty.$$

Example 9.24. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ population distribution, where $\theta > 0$ is unknown. Recall the $\mathcal{U}(0, \theta)$ pdf is given by

$$f_Y(y) = \begin{cases} \dfrac{1}{\theta}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

Show that $\widehat{\theta}_n = Y_{(n)}$ is a consistent estimator of θ.

Solution. Suppose $\epsilon > 0$. It suffices to show

$$\lim_{n\to\infty} P(|Y_{(n)} - \theta| > \epsilon) = 0$$

or, equivalently,

$$\lim_{n\to\infty} P(|Y_{(n)} - \theta| \leq \epsilon) = 1.$$

We will first derive an expression for $P(|Y_{(n)} - \theta| \leq \epsilon)$ and then show this sequence of probabilities converges to 1. Because $P(|Y_{(n)} - \theta| \leq \epsilon)$ is a probability involving the maximum order statistic $Y_{(n)}$, we have to use the (sampling) distribution of $Y_{(n)}$ to find the probability. In Chapter 6, we showed the cdf of the maximum order statistic $Y_{(n)}$ is

$$\begin{aligned} F_{Y_{(n)}}(y) = P(Y_{(n)} \leq y) &= P(Y_1 \leq y, Y_2 \leq y, ..., Y_n \leq y) \\ &= P(Y_1 \leq y)P(Y_2 \leq y) \cdots P(Y_n \leq y) \\ &= [P(Y \leq y)]^n \\ &= [F_Y(y)]^n. \end{aligned}$$

The population cdf of $Y \sim \mathcal{U}(0, \theta)$ is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ \dfrac{y}{\theta}, & 0 < y < \theta \\ 1, & y \geq \theta. \end{cases}$$

Therefore, the cdf of $Y_{(n)}$ is

$$F_{Y_{(n)}}(y) = \begin{cases} 0, & y \leq 0 \\ \left(\dfrac{y}{\theta}\right)^n, & 0 < y < \theta \\ 1, & y \geq \theta. \end{cases}$$

Now, we observe that

$$\begin{aligned} P(|Y_{(n)} - \theta| \leq \epsilon) = P(-\epsilon \leq Y_{(n)} - \theta \leq \epsilon) &= P(\theta - \epsilon \leq Y_{(n)} \leq \theta + \epsilon) \\ &= F_{Y_{(n)}}(\theta + \epsilon) - F_{Y_{(n)}}(\theta - \epsilon). \end{aligned}$$

Because $\epsilon > 0$, it follows that $\theta + \epsilon > \theta$ and therefore

$$F_{Y_{(n)}}(\theta + \epsilon) = 1.$$

Similarly, $\theta - \epsilon < \theta$ and therefore

$$F_{Y_{(n)}}(\theta - \epsilon) = \left(\frac{\theta - \epsilon}{\theta}\right)^n.$$

Therefore,

$$P(|Y_{(n)} - \theta| \leq \epsilon) = F_{Y_{(n)}}(\theta + \epsilon) - F_{Y_{(n)}}(\theta - \epsilon) = 1 - \underbrace{\left(\frac{\theta - \epsilon}{\theta}\right)^n}_{\to\, 0} \to 1,$$

as $n \to \infty$. We have shown

$$\lim_{n\to\infty} P(|Y_{(n)} - \theta| \leq \epsilon) = 1,$$

so we are done. $\square$

Remark: The preceding example illustrates how one shows an estimator $\widehat{\theta}_n$ is consistent by appealing to the definition of consistency; i.e., by showing

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \epsilon) = 0 \quad \text{or} \quad \lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| \leq \epsilon) = 1$$

directly. For most problems, this approach is not necessary. The following result presents sufficient conditions for an estimator $\widehat{\theta}_n$ to be consistent; showing these conditions hold is often easier.

Result: Suppose $\widehat{\theta}_n$ is a point estimator of θ. If both the bias

$$B(\widehat{\theta}_n) = E(\widehat{\theta}_n - \theta) = E(\widehat{\theta}_n) - \theta \to 0$$

and the variance

$$V(\widehat{\theta}_n) \to 0$$

as $n \to \infty$, then $\widehat{\theta}_n$ is a consistent estimator of θ.

Remark: It is easy to see why these conditions are sufficient. From Markov's Inequality (STAT 511, CH4), we have

$$P(|\widehat{\theta}_n - \theta| > \epsilon) = P((\widehat{\theta}_n - \theta)^2 > \epsilon^2) \leq \frac{E[(\widehat{\theta}_n - \theta)^2]}{\epsilon^2}.$$

We have already shown

$$E[(\widehat{\theta}_n - \theta)^2] = V(\widehat{\theta}_n) + [B(\widehat{\theta}_n)]^2.$$

Therefore, if both $V(\widehat{\theta}_n) \to 0$ and $B(\widehat{\theta}_n) \to 0$, then $P(|\widehat{\theta}_n - \theta| > \epsilon)$ is less than or equal to something that is converging to zero. Because probabilities are nonnegative (Kolmogorov Axiom 1), then $P(|\widehat{\theta}_n - \theta| > \epsilon) \to 0$ as well.

Weak Law of Large Numbers (WLLN): Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population with mean $E(Y) = \mu$ and variance $V(Y) = \sigma^2 < \infty$. Then

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} E(Y) = \mu,$$

as $n \to \infty$. That is, the sample mean $\overline{Y}$ is a consistent estimator of the population mean.

Proof. The sample mean $\overline{Y}$ is an unbiased estimator of μ so $B(\overline{Y}) = 0$ for all n. The variance

$$V(\overline{Y}) = \frac{\sigma^2}{n} \to 0,$$

as $n \to \infty$, provided that $\sigma^2 < \infty$. Therefore, the sufficient conditions in the previous result are satisfied. □

Generalization: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population and let $h : \mathbb{R} \to \mathbb{R}$ be any real function. Then

$$\frac{1}{n}\sum_{i=1}^{n} h(Y_i) \xrightarrow{p} E[h(Y)],$$

as $n \to \infty$, provided that $V[h(Y)] < \infty$. Note that the WLLN stated above is a special case of this result when $h(Y) = Y$: i.e., h is the identity function.

Example 9.25. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(θ) population distribution, where the population mean $E(Y) = \theta$ is unknown. From the WLLN, we know

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} E(Y) = \theta,$$

as $n \to \infty$. That is, the sample mean is a consistent estimator of θ. Applying the generalized version of the WLLN, the following probability limits also hold when $n \to \infty$:

$$\begin{aligned} m_2' = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 &\xrightarrow{p} E(Y^2) = 2\theta^2 \\ m_3' = \frac{1}{n}\sum_{i=1}^{n} Y_i^3 &\xrightarrow{p} E(Y^3) = 6\theta^3 \\ \frac{1}{n}\sum_{i=1}^{n} \sin Y_i &\xrightarrow{p} E(\sin Y). \end{aligned}$$

Note that

$$E(Y^2) = V(Y) + [E(Y)]^2 = \theta^2 + \theta^2 = 2\theta^2$$

and

$$E(Y^3) = \int_0^\infty \frac{y^3}{\theta}\, e^{-y/\theta} dy = \frac{\Gamma(4)\theta^4}{\theta} = 6\theta^3.$$

I don't know what

$$E(\sin Y) = \int_0^\infty \frac{\sin y}{\theta}\ e^{-y/\theta} dy$$

is, but I know $V(\sin Y) < \infty$ because $\sin Y$ and $(\sin Y)^2$ are bounded random variables. The expectation of any bounded random variable is finite. $\square$

Application: Use the WLLN to approximate the integral

$$\int_0^\infty \sin y\ e^{-y} dy.$$

Note that e^{-y} is the exponential pdf with mean $\theta = 1$. Therefore, if I generate an iid sample $Y_1, Y_2, ..., Y_n$ from an exponential(1) population distribution, the WLLN says

$$\frac{1}{n}\sum_{i=1}^n \sin Y_i \approx E(\sin Y) = \int_0^\infty \sin y\ e^{-y} dy$$

when n is large, where above $Y \sim$ exponential(1). In R,

```
> exp.data = rexp(1000000,1)
> mean(sin(exp.data))
[1] 0.4998616
```

This illustrates a commonly used statistical technique known as **Monte Carlo integration**, where (complicated) integrals are approximated by using simulation. How would you approximate

$$\int_0^\infty 17y^{1.3}\cos(y^2)\ e^{-y} dy?$$

Suppose $Y \sim$ exponential(1) again. If $Y_1, Y_2, ..., Y_n$ are iid exponential(1), then the WLLN says

$$\frac{1}{n}\sum_{i=1}^n 17Y_i^{1.3}\cos(Y_i^2) \approx E[17Y^{1.3}\cos(Y^2)] = \int_0^\infty 17y^{1.3}\cos(y^2)\ e^{-y} dy$$

when n is large. Therefore, simply repeat the simulation:

```
> exp.data = rexp(1000000,1)
> mean(17*exp.data^(1.3)*cos(exp.data^2))
[1] 1.328258
```

Functions of consistent estimators: Suppose we have two consistent point estimators $\widehat{\theta}_n$ and $\widehat{\theta}'_n$ that satisfy

$$\begin{aligned} \widehat{\theta}_n &\xrightarrow{p} \theta \\ \widehat{\theta}'_n &\xrightarrow{p} \theta', \end{aligned}$$

as $n \to \infty$. We get the following results:

1. $c\widehat{\theta}_n \xrightarrow{p} c\theta$, for any constant $c \in \mathbb{R}$
2. $\widehat{\theta}_n \pm \widehat{\theta}'_n \xrightarrow{p} \theta \pm \theta'$
3. $\widehat{\theta}_n\widehat{\theta}'_n \xrightarrow{p} \theta\theta'$
4. $$\frac{\widehat{\theta}_n}{\widehat{\theta}'_n} \xrightarrow{p} \frac{\theta}{\theta'}, \quad \text{provided that } \theta' \neq 0.$$

More generally, if $g : \mathbb{R} \to \mathbb{R}$ is a **continuous** function, then

$$\widehat{\theta}_n \xrightarrow{p} \theta \implies g(\widehat{\theta}_n) \xrightarrow{p} g(\theta),$$

as $n \to \infty$. In other words, convergence in probability (consistency) is preserved under continuous mappings. Proving the results above is analogous to how one proves the corresponding results involving limits of sequences of real numbers.

Example 9.26. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a beta$(\theta, 1)$ population distribution, where $\theta > 0$ is unknown. Recall the beta$(\theta, 1)$ pdf is given by

$$f_Y(y) = \begin{cases} \theta y^{\theta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find a consistent estimator of θ.

Solution. From the WLLN, we know

$$\overline{Y} \xrightarrow{p} E(Y) = \frac{\theta}{\theta+1},$$

as $n \to \infty$. We know

$$1 - \overline{Y} \xrightarrow{p} 1 - \frac{\theta}{\theta+1} = \frac{1}{\theta+1}$$

by continuity because $g(x) = 1 - x$ is a continuous function. Therefore, from (4) above, we have

$$\frac{\overline{Y}}{1-\overline{Y}} \xrightarrow{p} \frac{\frac{\theta}{\theta+1}}{\frac{1}{\theta+1}} = \theta,$$

as $n \to \infty$. Therefore, $\overline{Y}/(1-\overline{Y})$ is a consistent estimator of θ. $\square$

Example 9.27. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population with mean $E(Y) = \mu$, variance $V(Y) = \sigma^2$, and finite fourth moment; i.e., $E(Y^4) < \infty$. Prove

$$S_b^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \xrightarrow{p} \sigma^2.$$

Proof. First note that

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 &= \sum_{i=1}^{n}(Y_i^2-2Y_i\overline{Y}+\overline{Y}^2)\\ &= \sum_{i=1}^{n}Y_i^2-2\overline{Y}\sum_{i=1}^{n}Y_i+n\overline{Y}^2 = \sum_{i=1}^{n}Y_i^2-2n\overline{Y}^2+n\overline{Y}^2 = \sum_{i=1}^{n}Y_i^2-n\overline{Y}^2.\end{aligned}$$

Therefore, write

$$S_b^2=\frac{1}{n}\left(\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2\right)=\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\overline{Y}^2.$$

From the WLLN, we know

$$\frac{1}{n}\sum_{i=1}^{n}Y_i^2 \xrightarrow{p} E(Y^2)=V(Y)+[E(Y)]^2=\sigma^2+\mu^2.$$

Also, from the WLLN, we know

$$\overline{Y}\xrightarrow{p}\mu \implies \overline{Y}^2\xrightarrow{p}\mu^2,$$

by continuity because $g(x)=x^2$ is a continuous function. Therefore,

$$S_b^2=\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\overline{Y}^2\xrightarrow{p}(\sigma^2+\mu^2)-\mu^2=\sigma^2. \ \square$$

Remark: We have just shown that

$$S_b^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$$

is a consistent estimator of the population variance σ^2, which is interesting because S_b^2 is a biased estimator of σ^2 in finite samples; i.e., $E(S_b^2)\neq\sigma^2$. What about our usual (unbiased version of the) sample variance? Note that

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\left(\frac{n}{n-1}\right)\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\left(\frac{n}{n-1}\right)S_b^2.$$

We have already shown $S_b^2\xrightarrow{p}\sigma^2$. Therefore,

$$S^2=\left(\frac{n}{n-1}\right)S_b^2\xrightarrow{p}(1)\sigma^2=\sigma^2,$$

because $n/(n-1)\to 1$, as $n\to\infty$.

Summary: Suppose $Y_1,Y_2,...,Y_n$ is an iid sample from a population with mean $E(Y)=\mu$, variance $V(Y)=\sigma^2$, and finite fourth moment; i.e., $E(Y^4)<\infty$. We have shown

- $\overline{Y}$ is a consistent estimator of μ
- S^2 is a consistent estimator of σ^2.

9.7.2 Slutsky's Theorem

Remark: We now discuss an important theoretical result that helps to explain why many large-sample statistical inference procedures (e.g., confidence intervals, hypothesis tests, etc.) are valid.

Slutsky's Theorem: Suppose the sequence of random variables

$$U_n \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, and suppose $W_n \xrightarrow{p} 1$. Then

$$\frac{U_n}{W_n} \xrightarrow{d} \mathcal{N}(0,1)$$

as well.

Remark: Recall that $U_n \xrightarrow{d} \mathcal{N}(0,1)$ means

$$F_{U_n}(u) \to \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv = \text{cdf of } \mathcal{N}(0,1)$$

for all $u \in \mathbb{R}$, as $n \to \infty$; i.e., the sequence of cdfs $F_{U_1}(u), F_{U_2}(u), F_{U_3}(u), ...$ converges pointwise to the cdf of a $\mathcal{N}(0,1)$ random variable. Slutsky's Theorem says the sequence of cdfs

$$F_{\frac{U_n}{W_n}}(u) \to \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv$$

just like the $F_{U_n}(u)$ sequence does. That is, U_n and U_n/W_n have the same (standard normal) distribution in the limit.

Example 9.28. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population with mean μ and variance $\sigma^2 < \infty$. Prove that

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$.

Proof. We will use Slutsky's Theorem. From the CLT, we know

$$U_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. In Example 9.27 (notes), we just showed

$$S^2 \xrightarrow{p} \sigma^2 \implies \frac{S^2}{\sigma^2} \xrightarrow{p} 1 \implies W_n = \frac{S}{\sigma} \xrightarrow{p} 1;$$

the last two implications follow from continuity. Now, simply note that

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} = \frac{\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} = \frac{U_n}{W_n} \xrightarrow{d} \mathcal{N}(0,1)$$

by Slutsky's Theorem. □

Implication: Because

$$\frac{\overline{Y}-\mu}{S/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$, we can write

$$\begin{aligned} 1-\alpha \approx P\left(-z_{\alpha/2} < \frac{\overline{Y}-\mu}{S/\sqrt{n}} < z_{\alpha/2}\right) &= P\left(-z_{\alpha/2}\frac{S}{\sqrt{n}} < \overline{Y}-\mu < z_{\alpha/2}\frac{S}{\sqrt{n}}\right) \\ &= P\left(z_{\alpha/2}\frac{S}{\sqrt{n}} > \mu-\overline{Y} > -z_{\alpha/2}\frac{S}{\sqrt{n}}\right) \\ &= P\left(\overline{Y}+z_{\alpha/2}\frac{S}{\sqrt{n}} > \mu > \overline{Y}-z_{\alpha/2}\frac{S}{\sqrt{n}}\right) \\ &= P\left(\overline{Y}-z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{Y}+z_{\alpha/2}\frac{S}{\sqrt{n}}\right). \end{aligned}$$

This argument proves

$$\left(\overline{Y}-z_{\alpha/2}\frac{S}{\sqrt{n}},\ \overline{Y}+z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

is a large-sample $100(1-\alpha)\%$ confidence interval for the population mean μ.

Remark: If $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(p) population with $0 < p < 1$, then the same type of Slutsky's Theorem argument can be used to show

$$\frac{\widehat{p}-p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. Therefore,

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

is a large-sample $100(1-\alpha)\%$ confidence interval for the population proportion p.

9.7.3 Large-sample properties of MLEs

Remark: Not only do maximum likelihood estimators (MLEs) have useful interpretations (i.e., they "maximize the probability of the data"), we know they are also functions of sufficient statistics. Therefore, MLEs, or adjusted versions of them, often turn out to be MVUEs in finite samples.

Preview: Most MLEs also enjoy desirable large-sample properties, specifically, they are consistent (C) and they follow asymptotically normal (AN) sampling distributions. Coupling these two large-sample characteristics together, we often say that "MLEs are CAN." Therefore, we can use MLEs as a basis for large-sample inference; in fact, many large-sample inference techniques used in practice are derived from this fact.

Result: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution described by $p_Y(y|\theta)$ or $f_Y(y|\theta)$, and suppose $\widehat{\theta}$ is the MLE for θ. Under certain regularity conditions,

$$\widehat{\theta} \xrightarrow{p} \theta,$$

as $n \to \infty$; i.e., $\widehat{\theta}$ is a **consistent** estimator of θ. In addition,

$$\frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\theta)}{n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, where

$$v(\theta) = \frac{1}{E\left[-\frac{\partial^2}{\partial \theta^2} \ln p_Y(Y|\theta)\right]} \quad \text{(discrete case)}$$

$$v(\theta) = \frac{1}{E\left[-\frac{\partial^2}{\partial \theta^2} \ln f_Y(Y|\theta)\right]} \quad \text{(continuous case)}.$$

Note that the convergence in distribution result above, restated, says

$$\widehat{\theta} \sim \mathcal{AN}\left(\theta, \frac{v(\theta)}{n}\right), \quad \text{for large } n.$$

Remark: The quantity

$$\frac{v(\theta)}{n}$$

is the large-sample (approximate) variance of $\widehat{\theta}$. In a more advanced course, one calls

$$E\left[-\frac{\partial^2}{\partial \theta^2} \ln p_Y(Y|\theta)\right] \quad \text{or} \quad E\left[-\frac{\partial^2}{\partial \theta^2} \ln f_Y(Y|\theta)\right]$$

the **Fisher information** and $v(\theta)/n$ the **Cramér-Rao Lower Bound** (CRLB). Interestingly, for those distributions where suitable regularity conditions hold, the CRLB turns out to be the lower bound on the variance of any unbiased estimator of θ in finite samples; hence, using this lower bound can be helpful when attempting to find MVUEs.

Remark: The "regularity conditions" needed for us to claim that MLEs are consistent and asymptotically normal (CAN) are technical but end up holding for most of the population distributions we examine in this course. However, one condition needed is that the support for Y cannot depend on θ. Therefore, this discussion (that MLEs are CAN) excludes MLEs derived from uniform distributions, shifted exponential distributions, power family distributions; i.e., any population distribution where $y \leq \theta$ or $y \geq \theta$. For example, when $Y_1, Y_2, ..., Y_n$ are iid from a $\mathcal{U}(0, \theta)$ population distribution, the MLE is

$$\widehat{\theta} = Y_{(n)}.$$

However, because the support $R_Y = \{y : 0 \leq y \leq \theta\}$ depends on θ, we cannot say that $\widehat{\theta}$ is approximately normal when n is large (and, in fact, it is not).

Example 9.29. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population distribution, where $\theta > 0$ is unknown. Recall the Poisson(θ) pmf is given by

$$p_Y(y|\theta) = \begin{cases} \dfrac{\theta^y e^{-\theta}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

In Example 9.17 (notes, pp 142-143), we showed that

$$\widehat{\theta} = \overline{Y}$$

is the MLE of θ. From the WLLN, we already know that

$$\overline{Y} \xrightarrow{p} \theta,$$

as $n \to \infty$; i.e., $\widehat{\theta}$ is a **consistent** estimator of θ. In addition, from the CLT (Chapter 7), we already know that

$$\overline{Y} \sim \mathcal{AN}\left(\theta, \frac{\theta}{n}\right), \quad \text{for large } n.$$

We now demonstrate we get the same large-sample distribution by appealing to the large-sample theory for MLEs. The pmf of Y, where nonzero, is

$$p_Y(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}.$$

The natural logarithm of the pmf is

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

The first derivative of the log pmf is

$$\frac{\partial}{\partial \theta} \ln p_Y(y|\theta) = \frac{\partial}{\partial \theta}(y \ln \theta - \theta - \ln y!) = \frac{y}{\theta} - 1.$$

The second derivative of the log pmf is

$$\frac{\partial^2}{\partial \theta^2} \ln p_Y(y|\theta) = -\frac{y}{\theta^2}.$$

Therefore,

$$E\left[-\frac{\partial^2}{\partial \theta^2} \ln p_Y(Y|\theta)\right] = E\left(\frac{Y}{\theta^2}\right) = \frac{E(Y)}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta} \implies v(\theta) = \frac{1}{E\left[-\frac{\partial^2}{\partial \theta^2} \ln p_Y(Y|\theta)\right]} = \theta.$$

Appealing to the large-sample properties of MLEs, we have

$$\overline{Y} \sim \mathcal{AN}\left(\theta, \frac{\theta}{n}\right), \quad \text{for large } n.$$

Example 9.30. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Rayleigh(θ) population distribution, where $\theta > 0$ is unknown. Recall the Rayleigh pdf is given by

$$f_Y(y|\theta) = \begin{cases} \frac{2y}{\theta}e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

In Example 9.18 (notes, pp 144), we showed that

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

is the MLE of θ. From the WLLN, we already know that

$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 \xrightarrow{p} E(Y^2) = \theta,$$

as $n \to \infty$; i.e., $\widehat{\theta}$ is a **consistent** estimator of θ. Note that you can show $E(Y^2) = \theta$ directly or by recalling $Y^2 \sim$ exponential(θ). We now derive the large-sample distribution of $\widehat{\theta}$ by appealing to the large-sample theory for MLEs. The pdf of Y, where nonzero, is

$$f_Y(y|\theta) = \frac{2y}{\theta}e^{-y^2/\theta}.$$

The natural logarithm of the pdf is

$$\ln f_Y(y|\theta) = \ln\left(\frac{2y}{\theta}e^{-y^2/\theta}\right) = \ln(2y) - \ln\theta - \frac{y^2}{\theta}.$$

The first derivative of the log pdf is

$$\frac{\partial}{\partial\theta}\ln f_Y(y|\theta) = \frac{\partial}{\partial\theta}\left[\ln(2y) - \ln\theta - \frac{y^2}{\theta}\right] = -\frac{1}{\theta} + \frac{y^2}{\theta^2}.$$

The second derivative of the log pdf is

$$\frac{\partial^2}{\partial\theta^2}\ln f_Y(y|\theta) = \frac{1}{\theta^2} - \frac{2y^2}{\theta^3}.$$

Therefore,

$$\begin{aligned} E\left[-\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y|\theta)\right] &= E\left(-\frac{1}{\theta^2} + \frac{2Y^2}{\theta^3}\right) \\ &= -\frac{1}{\theta^2} + \frac{2E(Y^2)}{\theta^3} \\ &= -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = -\frac{1}{\theta^2} + \frac{2}{\theta^2} = \frac{1}{\theta^2} \end{aligned}$$

and thus,

$$v(\theta) = \frac{1}{E\left[-\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y|\theta)\right]} = \theta^2.$$

Appealing to the large-sample properties of MLEs, we have

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \sim \mathcal{AN}\left(\theta, \frac{\theta^2}{n}\right), \quad \text{for large } n.$$

Large-sample confidence intervals: Because MLEs are consistent and asymptotically normal (CAN), we can write large-sample confidence intervals for the population-level parameters they estimate. The "AN" part of CAN means

$$U_n = \frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\theta)}{n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. The "C" part of CAN means

$$\widehat{\theta} \xrightarrow{p} \theta, \quad \text{as } n \to \infty.$$

Provided that $v(\theta)$ is a continuous function (which it usually is), we have

$$v(\widehat{\theta}) \xrightarrow{p} v(\theta) \implies \frac{v(\widehat{\theta})}{v(\theta)} \xrightarrow{p} 1 \implies W_n = \sqrt{\frac{v(\widehat{\theta})}{v(\theta)}} \xrightarrow{p} 1.$$

Therefore,

$$\frac{U_n}{W_n} = \frac{\frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\theta)}{n}}}}{\sqrt{\frac{v(\widehat{\theta})}{v(\theta)}}} = \frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\widehat{\theta})}{n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, by Slutsky's Theorem. Therefore, for large n, we can write

$$\begin{aligned}
1 - \alpha \approx P\left(-z_{\alpha/2} < \frac{\widehat{\theta} - \theta}{\sqrt{v(\widehat{\theta})/n}} < z_{\alpha/2}\right) &= P\left(-z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}} < \widehat{\theta} - \theta < z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}}\right) \\
&= P\left(z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}} > \theta - \widehat{\theta} > -z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}}\right) \\
&= P\left(\widehat{\theta} + z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}} > \theta > \widehat{\theta} - z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}}\right) \\
&= P\left(\widehat{\theta} - z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}} < \theta < \widehat{\theta} + z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}}\right).
\end{aligned}$$

This argument proves

$$\left(\widehat{\theta} - z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}},\ \widehat{\theta} + z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}}\right)$$

is a large-sample $100(1-\alpha)\%$ confidence interval for θ.

Remark: The novelty of the general formula above is that it "works" for any MLE (provided the population distribution's regularity conditions are met). This greatly expands our ability to write large-sample confidence intervals for parameters in a variety of population-level models (e.g., Poisson, Rayleigh, geometric, exponential/gamma, beta, etc.). All we have to do is determine the MLE and then appeal to this large-sample result.

Illustration:

- Poisson(θ). MLE: $\widehat{\theta} = \overline{Y}$, $v(\theta) = \theta \Longrightarrow v(\widehat{\theta}) = \overline{Y}$. Therefore,

$$\overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\overline{Y}}{n}}$$

  is a large-sample $100(1-\alpha)\%$ confidence interval for θ.

- Rayleigh(θ). MLE: $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 = m_2'$ (the second sample moment), $v(\theta) = \theta^2 \Longrightarrow v(\widehat{\theta}) = (m_2')^2$. Therefore,

$$m_2' \pm z_{\alpha/2}\sqrt{\frac{(m_2')^2}{n}} \iff m_2' \pm z_{\alpha/2}\left(\frac{m_2'}{\sqrt{n}}\right)$$

  is a large-sample $100(1-\alpha)\%$ confidence interval for θ.

Exercise: Calculate large-sample 95% confidence intervals for θ using the accident data (pp 118-119, notes) under a Poisson(θ) assumption and using the light bulb data (pp 120, notes) under a Rayleigh(θ) assumption. Note $z_{0.025} \approx 1.96$.

- **Accident data** (Poisson): MLe: $\overline{y} = 1.23$, sample size $n = 84$

$$\overline{y} \pm z_{\alpha/2}\sqrt{\frac{\overline{y}}{n}} \implies 1.23 \pm 1.96\sqrt{\frac{1.23}{84}} \implies (0.99, 1.47).$$

- **Light bulb data** (Rayleigh): MLe: $m_2' = 21.5$, sample size $n = 30$

$$m_2' \pm z_{\alpha/2}\left(\frac{m_2'}{\sqrt{n}}\right) \implies 21.5 \pm 1.96\left(\frac{21.5}{\sqrt{30}}\right) \implies (13.8, 29.2).$$

These confidence intervals are interpreted in the usual way; i.e., we are (approximately) 95% confident the corresponding interval contains its population-level parameter θ.