STAT 513
THEORY OF STATISTICAL INFERENCE

Fall 2023

Lecture Notes

Joshua M. Tebbs
Department of Statistics
University of South Carolina

Contents

10 Hypothesis Testing

10.1 Introduction

Preview: 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 parameter. Recall that we use different notation for discrete and continuous population-level models:

- discrete: $p_Y(y|\theta)$ is a probability mass function; e.g., Bernoulli, Poisson, etc.
- continuous: $f_Y(y|\theta)$ is a probability density function; e.g., uniform, normal, exponential, etc.

Our goal in this chapter is to learn about the theory of hypothesis tests for θ (and functions of θ; e.g., a population mean, a population variance, etc.).

Remark: Hypothesis testing is a form of **statistical inference**, which is the process by which we make a decision (or "infer") about the value of an unknown population parameter. Recall in STAT 512 we studied other types statistical inference procedures:

- In Chapter 9, we studied methods of point estimation (MOM and MLE) and we discussed how one might find the "best" point estimator for θ.
- In Chapter 8, we learned how one could develop interval estimators (i.e., confidence intervals) for θ.

In both settings, our goal was to use the observations $Y_1, Y_2, ..., Y_n$ to estimate the value of θ. Point estimators are "one-shot guesses," whereas interval estimators incorporate uncertainty (hence, producing an interval of "plausible values" of θ). In due course, we will learn there is an elegant duality between interval estimation and hypothesis testing.

Importance: Hypothesis tests are widely used in translational areas, including biomedicine, epidemiology, engineering, and the social sciences. They are commonly taught in introductory courses, so most students have at least heard of them. Hypothesis tests are useful because they can shed insight on answers to interesting questions. For example,

- Have starting salaries of USC graduates increased over the last 5 years?
- Is a new drug or intervention superior when compared to the standard method of treatment?
- How do vaccination rates compare among different races/ethnicities?

Not surprisingly, the theoretical foundation of hypothesis testing is often hidden from students in their first exposure. In this course, we will learn the underlying mathematics of how hypothesis tests work.

10.2 Definitions and examples

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 parameter (a scalar, unless otherwise noted). A **hypothesis test** is a statistical inference procedure which pits two competing hypotheses regarding θ against each other. The goal is to determine which hypothesis is more supported by the available information in the sample.

Remark: One can think of any hypothesis test as consisting of four parts:

1. the null hypothesis H_0
2. the alternative hypothesis H_a
3. a test statistic $T = T(Y_1, Y_2, ..., Y_n)$
4. a rejection region (which we will denote by RR).

The null and alternative hypotheses are written in terms of θ. A test statistic T is used to decide between H_0 and H_a. If T falls in the rejection region (RR), then we reject H_0 in favor of H_a. If it does not, then we do not reject H_0.

Example 10.1. Suppose $Y_1, Y_2, ..., Y_{10}$ is an iid sample of size $n = 10$ from a $\mathcal{N}(\theta, \sigma^2 = 100)$ population distribution, where θ is unknown. If θ denotes the population mean starting salary (in \$1000s) among all USC STEM majors (which we model using a normal distribution), then we might be interested in testing

$$\begin{gathered} H_0 : \theta = 50 \\ \text{versus} \\ H_a : \theta > 50. \end{gathered}$$

The null hypothesis H_0 states the population mean salary is \$50,000, whereas the alternative hypothesis H_a states the population mean salary is more than \$50,000.

Q: Which statistic T should we use as a test statistic to decide between H_0 and H_a?
A: If possible, we want to pick a statistic whose (sampling) distribution we know or at least can derive. We also learned in STAT 512 how **sufficient statistics** contained all of the available information about θ. Therefore, we might start by considering statistics (a) which are sufficient and (b) which we can derive the sampling distribution of (or we already know).

Suppose we decide to use the sample mean

$$T = T(Y_1, Y_2, ..., Y_{10}) = \overline{Y} = \frac{1}{10}\sum_{i=1}^{n} Y_i$$

(which is sufficient when σ^2 is known) as a test statistic and a rejection region of the form

$$\text{RR} = \{\overline{y} > 55\}.$$

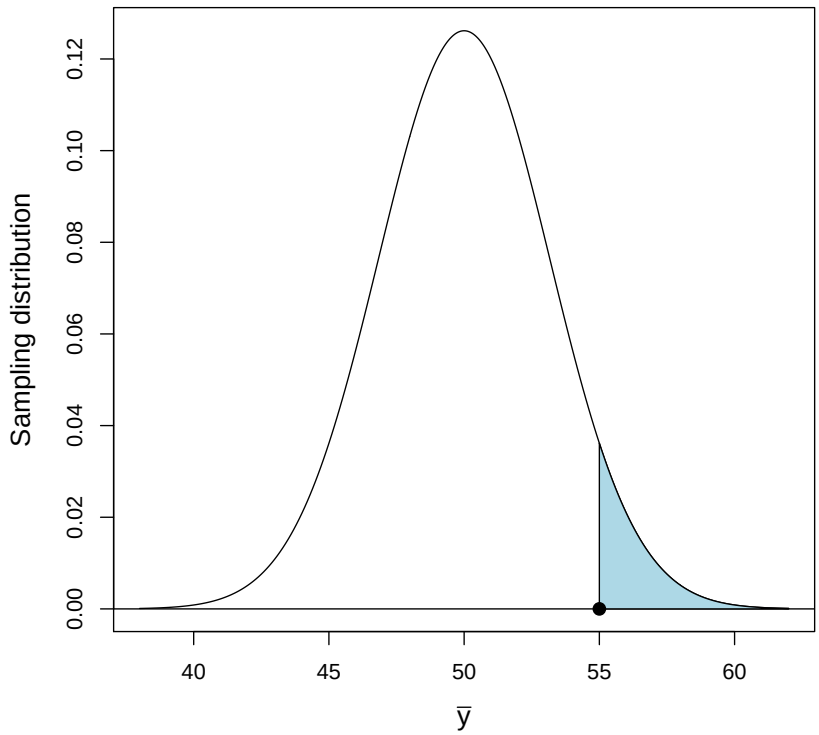

Figure 10.1: Example 10.1. Sampling distribution of $\overline{Y}$ when H_0 is true. The rejection region $\text{RR} = \{\overline{y} \geq 55\}$ is shown shaded.

That is, we will reject $H_0 : \theta = 50$ in favor of $H_a : \theta > 50$ when the sample mean $\overline{Y}$ is larger than 55. What is the sampling distribution of $T = T(Y_1, Y_2, ..., Y_{10}) = \overline{Y}$? From STAT 512, we know

$$\overline{Y} \sim \mathcal{N}\left(\theta,\ \frac{\sigma^2}{10}\right) \Longrightarrow\ \overline{Y} \sim \mathcal{N}\left(\theta, 10\right).$$

Figure 10.1 shows the sampling distribution of $\overline{Y}$ when H_0 is true; i.e., when the population mean is $\theta = 50$. Note that even when H_0 is true, the probability of the rejection region is

$$P_{H_0}(\text{RR}) = P_{H_0}(\overline{Y} > 55) \approx 0.057.$$

Therefore, even when $H_0 : \theta = 50$ is true, there is a chance we will reject it when using this rejection region.

```
> 1-pnorm(55,50,sqrt(10))
[1] 0.05692315
```

Note: For any event A, the notation $P_{H_0}(A)$ means that we are calculating $P(A)$ under the assumption that H_0 is true.

Q: Could we use another test statistic in Example 10.1?
A: Of course, we can. However, calculations are likely to be much harder.

Instead of using the sample mean $\overline{Y}$, suppose we decide to use the maximum order statistic

$$T = T(Y_1, Y_2, ..., Y_{10}) = Y_{(10)}$$

as a test statistic and a rejection region of the form

$$\text{RR} = \{y_{(10)} > 80\}.$$

What is the sampling distribution of $T = Y_{(10)}$? From STAT 512, we recall

$$f_{Y_{(10)}}(y) = 10 f_Y(y)[F_Y(y)]^9,$$

where $f_Y(y)$ is the $\mathcal{N}(\theta, 100)$ pdf and $F_Y(y)$ is the $\mathcal{N}(\theta, 100)$ cdf; i.e.,

$$F_Y(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi(100)}} e^{-(t-\theta)^2/2(100)} dt.$$

Therefore, the sampling distribution of $Y_{(10)}$ cannot be written out in closed form (it depends on an improper integral which has no closed-form answer), but we can graph it in R. Figure 10.2 (next page) shows the sampling distribution of $Y_{(10)}$ when H_0 is true; i.e., when the population mean is 50.

Q: What is the probability of this rejection region when H_0 is true?
A: This is difficult to calculate as we would have to integrate $f_{Y_{(10)}}(y)$. In theory, it is

$$\begin{aligned} P_{H_0}(\text{RR}) &= P_{H_0}(Y_{(10)} > 80) \\ &= \int_{80}^{\infty} f_{Y_{(10)}}(y) dy \\ &= \int_{80}^{\infty} \frac{10}{\sqrt{2\pi(100)}} e^{-(y-50)^2/2(100)} \left[\int_{-\infty}^{y} \frac{1}{\sqrt{2\pi(100)}} e^{-(t-50)^2/2(100)} dt \right]^9 dy. \end{aligned}$$

Instead of attempting this calculation (it is futile), we can use **Monte Carlo simulation** to approximate $P_{H_0}(Y_{(10)} > 80)$. Briefly, this consists of

1. Generating a large number of $\mathcal{N}(50, 100)$ iid samples, each of size 10.
2. Recording the value of the maximum order statistic $y_{(10)}$ in each sample.
3. Calculating the proportion of samples where the maximum $y_{(10)}$ exceeds 80.

By the WLLN (Chapter 9), this proportion will approximate the true probability when the number of samples (say, B) is large. I coded this using $B = 100,000$ samples (see R code online) and obtained

$$P_{H_0}(\text{RR}) = P_{H_0}(Y_{(10)} > 80) \approx 0.014.$$

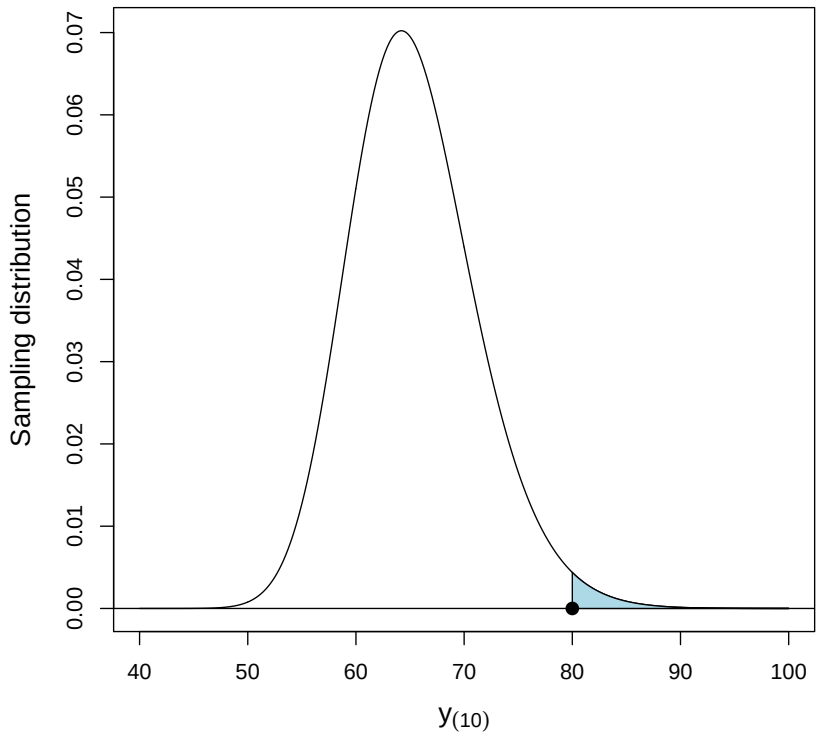

Figure 10.2: Example 10.1. Sampling distribution of $Y_{(10)}$ when H_0 is true. The rejection region $\text{RR} = \{y_{(10)} \geq 80\}$ is shown shaded.

Comparison: It is interesting to compare the two rejection regions:

$$\text{RR}_1 = \{\overline{y} > 55\} \quad \text{and} \quad \text{RR}_2 = \{y_{(10)} > 80\}.$$

The probability of rejecting H_0 when H_0 is true is smaller for the second test (0.057 versus 0.014), so one may think using RR_2 is "better" in some sense. However, what if H_0 is *not* true? Which test has a better chance of rejecting H_0? I constructed the following table:

	Probability of rejecting $H_0 : \theta = 50$				
	$\theta = 50$	$\theta = 55$	$\theta = 60$	$\theta = 65$	$\theta = 70$
$\text{RR}_1 = \{\overline{y} > 55\}$	0.057	0.500	0.943	0.999	0.999+
$\text{RR}_2 = \{y_{(10)} > 80\}$	0.014	0.061	0.207	0.501	0.822

Therefore, although the second rejection region/test does a better job of guarding against one type of error (i.e., rejecting a true H_0), the first one has a much larger probability (power) of rejecting H_0 when H_a is true. $\square$

States of Nature: For any hypothesis test, we can make one of two mistakes:

- Type I Error: Rejecting H_0 when H_0 is true
- Type II Error: Not rejecting H_0 when H_a is true.

Therefore, for any test we perform, there are four possible scenarios. These are summarized in the following table:

		Decision	
		Reject H_0	Do not reject H_0
Truth	H_0	Type I Error	OK
	H_a	OK	Type II Error

Remark: Because H_0 and H_a are written in terms of θ (an unknown population-level parameter), we never get to know for sure which hypothesis is correct. Therefore, the best we can do is to select rejection regions (and sample sizes) that guard against making these errors. In addition, this is why we use language like "Reject H_0" and "Do not reject H_0," because we are simply gauging how much **evidence** we have against H_0. Using the language "Accept H_0" is typically frowned against because this suggests we have accepted H_0 as being true. In reality, it may or may not be.

Terminology: The probability of Type I Error is denoted by α. It is the probability we reject H_0 when H_0 is true; i.e.,

$$\alpha = P(\text{Type I Error}) = P_{H_0}(\text{RR}) = P(\text{Reject } H_0 | H_0 \text{ true}).$$

This is also called the **significance level** (or **level**) of the test. The probability of Type II Error is denoted by β. It is the probability of not rejecting H_0 when H_a is true; i.e.,

$$\beta = P(\text{Type II Error}) = P_{H_a}(\overline{\text{RR}}) = P(\text{Do not reject } H_0 | H_a \text{ true}).$$

Remarks: The definition for α above is straightforward when H_0 specifies a single value for θ; e.g., $H_0 : \theta = 50$. This is an example of a **simple** (or **sharp**) null hypothesis. However, what if in Example 10.1 we chose to write

$$\begin{gathered} H_0 : \theta \leq 50 \\ \text{versus} \\ H_a : \theta > 50 \end{gathered}$$

instead? In this test, H_0 is an example of a **composite** hypothesis because there is more than one value of θ that makes H_0 true. We will learn later how to redefine $\alpha = P(\text{Type I Error})$ in this situation (i.e., for a composite H_0). It should be clear that $\beta = P(\text{Type II Error})$ will be different for different values of θ which satisfy H_a. We have already seen this in Example 10.1 (see the table on the last page).

Example 10.2. Suppose $Y_1, Y_2, ..., Y_{15}$ is an iid sample from an exponential(θ) population, where $\theta > 0$ is unknown. We are interested in testing

$$H_0 : \theta = 10$$
$$\text{versus}$$
$$H_a : \theta < 10.$$

We will use the test statistic $T = T(Y_1, Y_2, ..., Y_{15}) = \sum_{i=1}^{15} Y_i$ and a rejection region of the form

$$\text{RR} = \left\{ t = \sum_{i=1}^{15} y_i < k \right\}.$$

(a) Find the constant k which ensures $\alpha = 0.05$.
(b) For the value of k in part (a), calculate β when $\theta = 5$.

Solution. In part (a), we want to determine k so that

$$0.05 = P_{H_0}(\text{RR}) = P_{H_0}(T < k).$$

What is the sampling distribution of $T = \sum_{i=1}^{15} Y_i$? Recall

$$Y_1, Y_2, ..., Y_n \sim \text{iid exponential}(\theta) \implies T = \sum_{i=1}^{n} Y_i \sim \text{gamma}(n, \theta).$$

Therefore, $T = \sum_{i=1}^{15} Y_i \overset{H_0}{\sim} \text{gamma}(15, 10)$, and thus

$$0.05 = P_{H_0}(T < k) \implies k \approx 92.5;$$

i.e., k is the 0.05 quantile (5th percentile) of a gamma distribution with shape 15 and scale 10; see Figure 10.3 (left, next page).

```
> qgamma(0.05,15,1/10)
[1] 92.4633
```

In part (b), we now calculate

$$\beta = P(\text{Type II Error}) = P(\text{Do not reject } H_0 | \theta = 5) = P(T > k | \theta = 5),$$

where $k \approx 92.5$. When $\theta = 5$, the test statistic $T = \sum_{i=1}^{15} Y_i \sim \text{gamma}(15, 5)$. Therefore,

$$\beta = P(T > k | \theta = 5) \approx 0.178.$$

```
> k = qgamma(0.05,15,1/10) # critical value
> 1-pgamma(k,15,1/5)
[1] 0.1775725
```

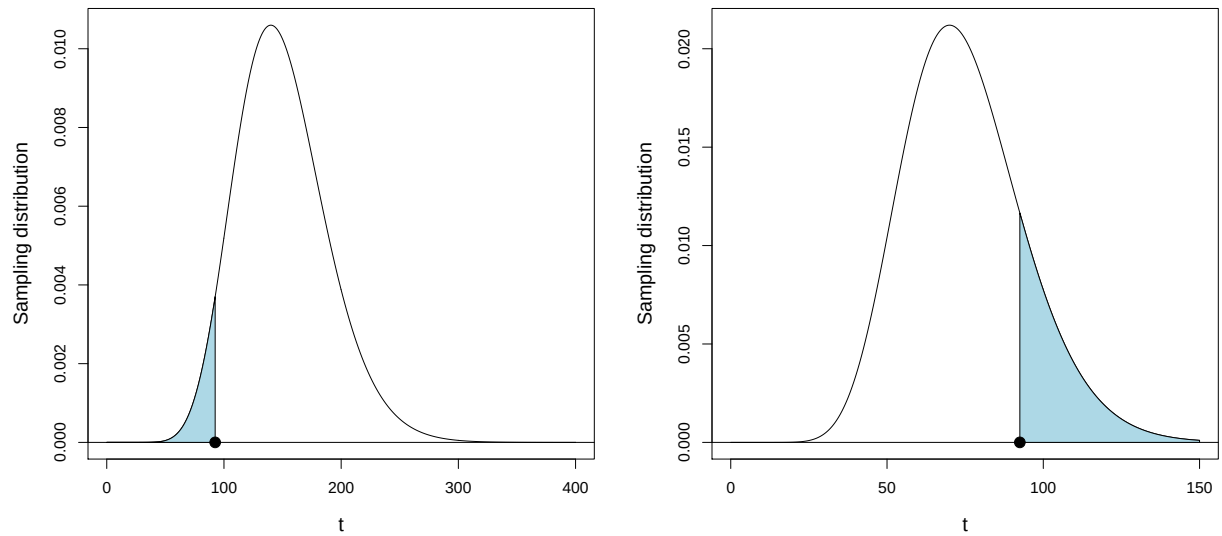

Figure 10.3: Example 10.2. Left: Sampling distribution of $T = \sum_{i=1}^{15} Y_i$ when H_0 is true. The rejection region RR $= \{t = \sum_{i=1}^{15} y_i < k\}$ is shown shaded. Right: Sampling distribution of $T = \sum_{i=1}^{15} Y_i$ when $\theta = 5$; i.e., when H_a is true. The probability of Type II Error (when $\theta = 5$) is shown shaded.

Discussion: This example shows while the probability of Type I Error α may be "acceptable," the probability of Type II Error β may be "unacceptable." How can we remedy this? In practice, attaining acceptable values for α and β simultaneously is a balancing act.

- The reason for the "balancing act" is that α and β are **inversely related**; i.e., as one increases, the other decreases (other things being equal).
- A common approach in hypothesis testing is to first select a value of α which is "acceptable" (e.g., $\alpha = 0.05$, etc.). Then, use a sufficiently large sample size n to attain a desired target probability for β. We will see examples of this later. $\square$

Exercise: Redo Example 10.2 using (a) $\alpha = 0.01$ and $n = 15$ (β will increase) and (b) $\alpha = 0.05$ and $n = 30$ (β will decrease).

Example 10.3. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population, where $\theta > 0$ is unknown. In STAT 512 (Example 9.4, pp 118), we used a Poisson distribution to model the number of accidents per year for a sample of $n = 84$ policies. Suppose we want to test

$$H_0 : \theta = 1$$
$$\text{versus}$$
$$H_a : \theta > 1.$$

We will use the test statistic $T = T(Y_1, Y_2, ..., Y_{84}) = \sum_{i=1}^{84} Y_i$ and a rejection region of the form

$$\text{RR} = \left\{ t = \sum_{i=1}^{84} y_i \geq k \right\}.$$

Find the constant k which ensures $\alpha = 0.01$.

Solution. We want to determine k so that

$$0.01 = P_{H_0}(\text{RR}) = P_{H_0}(T \geq k).$$

What is the (sampling) distribution of $T = \sum_{i=1}^{84} Y_i$? Recall

$$Y_1, Y_2, ..., Y_n \sim \text{iid Poisson}(\theta) \implies T = \sum_{i=1}^{n} Y_i \sim \text{Poisson}(n\theta).$$

Therefore, $T = \sum_{i=1}^{84} Y_i \overset{H_0}{\sim} \text{Poisson}(84)$, and thus

$$0.01 = P_{H_0}(T \geq k) \implies k = 106;$$

i.e., k is the 0.99 quantile (99th percentile) of a Poisson distribution with mean 84; see Figure 10.4 (next page).

```
> qpois(0.99,84)
[1] 106
```

Observation: Using $k = 106$ does not provide a Type I Error probability of **exactly** $\alpha = 0.01$; note that

$$\begin{aligned} P_{H_0}(T \geq 106) &\approx 0.0115 \\ P_{H_0}(T \geq 107) &\approx 0.0088. \end{aligned}$$

Of course, the reason this happens is because the sampling distribution of $T = \sum_{i=1}^{84} Y_i$ is discrete. Therefore, if we use $k = 106$ as a critical value, then our significance level is $\alpha \approx 0.0115$. If we use $k = 107$, then $\alpha \approx 0.0088$.

Q: What should we do?
A: There is no "right" answer, but we can think about different options.

- Using $k = 107$ would confer a slightly smaller Type I Error probability (more conservative), while using $k = 106$ would confer a slightly larger one (more anti-conservative). An analyst can make his/her decision between these two options in the light of how serious a Type I Error is.

- An alternative solution would be to create a **randomized test**. In discrete data problems (like the Poisson), these tests can be used to "hit" $\alpha = 0.01$ exactly. In this example, a randomized test would be performed as follows:

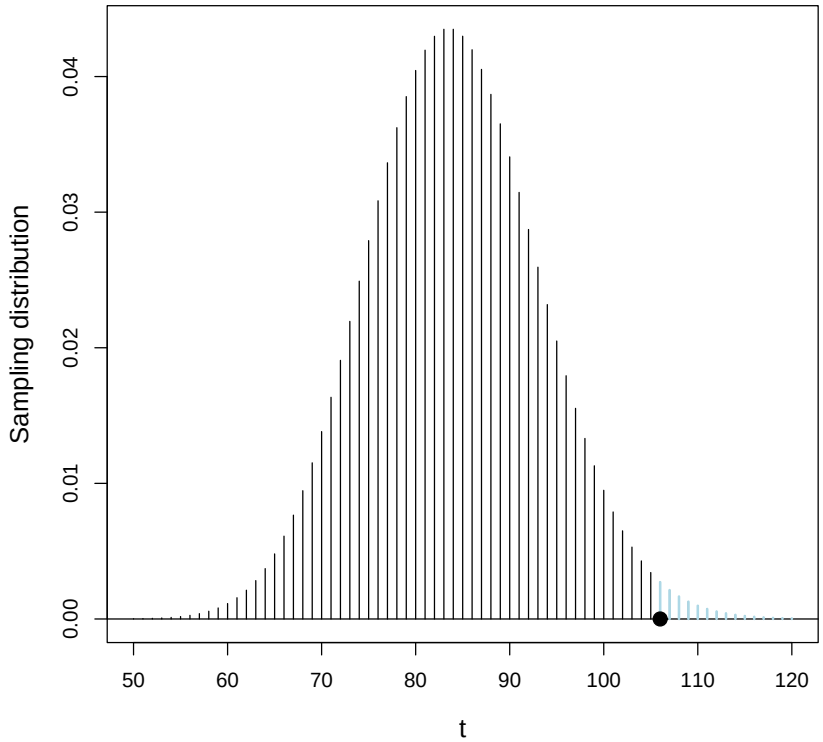

Figure 10.4: Example 10.3. Sampling distribution of $T = \sum_{i=1}^{84} Y_i$ when H_0 is true. The rejection region RR $= \{t = \sum_{i=1}^{84} y_i \geq 106\}$ is shown shaded.

- If $t \leq 105$, then do not reject H_0.
- If $t \geq 107$, then reject H_0.
- If $t = 106$, then we reject H_0 with probability p (see description below).

Description: Suppose U has a Bernoulli distribution with probability of success

$$p = P(U = 1) \approx \frac{0.01 - 0.0088}{0.0115 - 0.0088} \approx 0.44$$

and suppose U is obtained **independently** of the sample $Y_1, Y_2, ..., Y_{84}$. In this instance, the probability of Type I Error is

$$\begin{aligned} \alpha &= P_{H_0}(T \geq 107) + P_{H_0}(\{T = 106\} \cap \{U = 1\}) \\ &= 0.0088 + (0.0115 - 0.0088)\left(\frac{0.01 - 0.0088}{0.0115 - 0.0088}\right) = 0.01. \end{aligned}$$

The last step is true because U is independent of the sample and hence $U \perp\!\!\!\perp T$.

Remark: Using randomization in this way provides a beautiful mathematical solution when the test statistic T is discrete (to guarantee an exact level α test). Unfortunately, the problem with randomized tests is that no one actually uses them. The notion of potentially basing one's entire decision on a coin flip (with probability p) is just too unpalatable.

Analysis: The observed data for the number of accidents (in tabular form) are given below:

Number of accidents	Number of policies
0	32
1	26
2	12
3	7
4	4
5	2
6	1

For these data, the value of the test statistic is

$$t = \sum_{i=1}^{84} y_i = 103.$$

Therefore, we would not reject $H_0 : \theta = 1$ at the $\alpha \approx 0.0115$ level. $\square$

Q: Can we avoid the discreteness issue in Example 10.3 by using the CLT for $T = \sum_{i=1}^{84} Y_i$?
A: Yes, but the CLT only offers an approximation to the true sampling distribution of T. From STAT 512 (Chapter 7), we know

$$T = \sum_{i=1}^{84} Y_i \sim \mathcal{AN}\left(84\theta, 84\theta\right) \implies Z = \frac{T - 84}{\sqrt{84}} \overset{H_0}{\sim} \mathcal{AN}(0,1).$$

Therefore, to perform an **approximate** level α test for

$$\begin{gathered} H_0 : \theta = 1 \\ \text{versus} \\ H_a : \theta > 1, \end{gathered}$$

one could simply reject H_0 by using

$$\text{RR} = \{z > z_\alpha\},$$

where z_α is the upper α quantile of a $\mathcal{N}(0,1)$ distribution; see Figure 10.5 (next page). This is a **large-sample hypothesis test** because one is using a rejection region from a large-sample (or asymptotic) sampling distribution result (the CLT). Because the reference distribution (i.e., standard normal) is continuous, randomization is not needed. However, using $\text{RR} = \{z > z_\alpha\}$ does not provide an exact level α test; it is only an approximation.

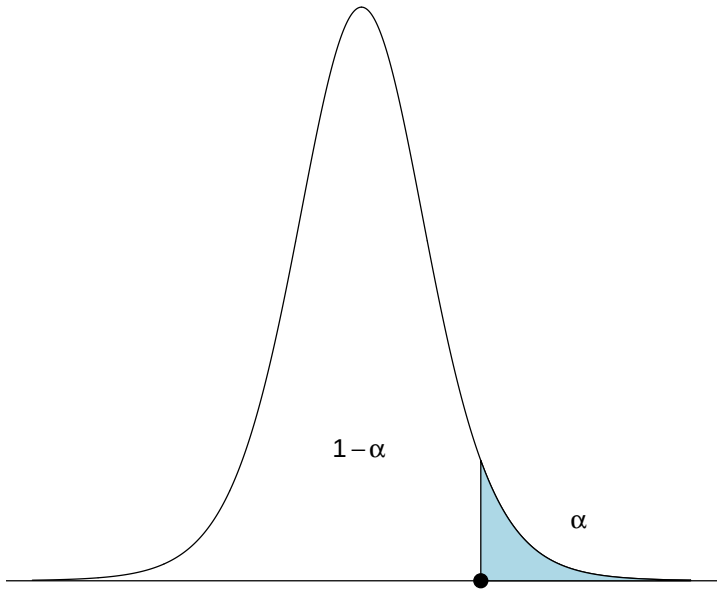

Figure 10.5: $\mathcal{N}(0,1)$ pdf. The upper α quantile z_α is shown by using a dark circle.

Example 10.4. 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

$$f_Y(y|\theta) = \begin{cases} \dfrac{2y}{\theta} e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Suppose we want to test

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta \neq \theta_0, \end{gathered}$$

where θ_0 is a specified value (e.g., $\theta_0 = 20$, etc.). We will use the test statistic $T = T(Y_1, Y_2, ..., Y_n) = \sum_{i=1}^n Y_i^2$ and a rejection region of the form

$$\text{RR} = \{t < k_1 \ \text{ or } \ t > k_2\},$$

where $t = \sum_{i=1}^n y_i^2$ and $k_2 > k_1$. Determine values of k_1 and k_2 to ensure a level α test.

Solution: First of all, it is important to note that this is an example of a **two-sided test**. This alternative hypothesis does not specify a specific direction of departure from θ_0 (e.g., greater than or less than). Therefore, it makes sense that we will reject H_0 when T is too

large *or* too small. In addition, we have not provided numerical values for α, n, and θ_0 in this example, so our answers for k_1 and k_2 will depend on them. We begin by recalling

$$Y \sim \text{Rayleigh}(\theta) \implies Y^2 \sim \text{exponential}(\theta).$$

Therefore,

$$Y_1^2, Y_2^2, ..., Y_n^2 \sim \text{iid exponential}(\theta) \implies T = \sum_{i=1}^n Y_i^2 \sim \text{gamma}(n, \theta) \implies \frac{2T}{\theta} \sim \chi^2(2n).$$

We want the probability of Type I Error to equal α. This means

$$\begin{aligned} \alpha = P(\text{Type I Error}) = P_{H_0}(\text{RR}) &= P_{H_0}(\{T < k_1\} \cup \{T > k_2\}) \\ &= P_{H_0}(T < k_1) + P_{H_0}(T > k_2) \\ &= \underbrace{P_{H_0}\left(\frac{2T}{\theta_0} < \frac{2k_1}{\theta_0}\right)}_{\stackrel{\text{set}}{=} \alpha/2} + \underbrace{P_{H_0}\left(\frac{2T}{\theta_0} > \frac{2k_2}{\theta_0}\right)}_{\stackrel{\text{set}}{=} \alpha/2}. \end{aligned}$$

We know $2T/\theta_0 \stackrel{H_0}{\sim} \chi^2(2n)$. Therefore, we can choose

$$\begin{aligned} \frac{2k_1}{\theta_0} = \chi^2_{2n,1-\alpha/2} &\implies k_1 = \frac{\theta_0 \chi^2_{2n,1-\alpha/2}}{2} \\ \frac{2k_2}{\theta_0} = \chi^2_{2n,\alpha/2} &\implies k_2 = \frac{\theta_0 \chi^2_{2n,\alpha/2}}{2}, \end{aligned}$$

where

$$\begin{aligned} \chi^2_{2n,1-\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } \chi^2(2n) \\ \chi^2_{2n,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } \chi^2(2n); \end{aligned}$$

see Figure 10.6 (next page). Therefore, a level α rejection region is

$$\text{RR} = \left\{ t < \frac{\theta_0 \chi^2_{2n,1-\alpha/2}}{2} \quad \text{or} \quad t > \frac{\theta_0 \chi^2_{2n,\alpha/2}}{2} \right\},$$

where $t = \sum_{i=1}^n y_i^2$.

Analysis: In STAT 512 (Example 9.5, pp 120), we used a Rayleigh distribution to model the time to failure for a sample of $n = 30$ light bulb filaments under "intense use" conditions. 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

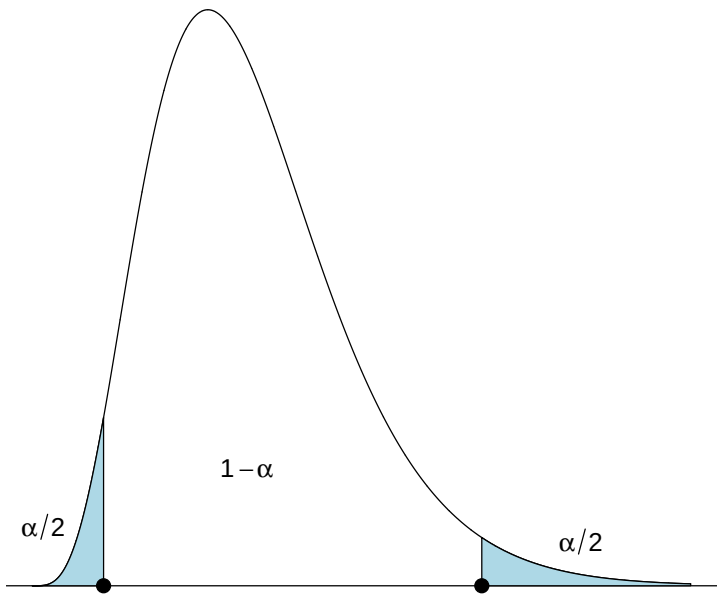

Figure 10.6: $\chi^2(2n)$ pdf. The lower $\alpha/2$ quantile $\chi^2_{2n,1-\alpha/2}$ and the upper $\alpha/2$ quantile $\chi^2_{2n,\alpha/2}$ are shown by using dark circles.

Suppose we want to test

$$\begin{gathered} H_0 : \theta = 20 \\ \text{versus} \\ H_a : \theta \neq 20 \end{gathered}$$

using $\alpha = 0.10$. With $n = 30$ and $\theta_0 = 20$, we have

$$\begin{aligned} k_1 &= \frac{20(43.2)}{2} \approx 431.9 \\ k_2 &= \frac{20(79.1)}{2} \approx 790.8. \end{aligned}$$

Therefore, a level $\alpha = 0.10$ rejection region is

$$\text{RR} = \{t < 431.9 \ \text{ or } \ t > 790.8\}.$$

The observed test statistic is $t = \sum_{i=1}^{30} y_i^2 = 645.0$, so we would not reject $H_0 : \theta = 20$ at the $\alpha = 0.10$ level. $\square$

```
> qchisq(0.05,60)
[1] 43.18796
> qchisq(0.95,60)
[1] 79.08194
```

Example 10.5. Suppose we have two independent random samples:

- $X_1, X_2, ..., X_n$ is an iid sample from a $\mathcal{N}(0, \sigma_X^2)$ population distribution
- $Y_1, Y_2, ..., Y_m$ is an iid sample from a $\mathcal{N}(0, \sigma_Y^2)$ population distribution.

The goal is to formulate a test for the equality of variances; i.e.,

$$\begin{array}{c} H_0 : \sigma_X^2 = \sigma_Y^2 \\ \text{versus} \\ H_a : \sigma_X^2 \neq \sigma_Y^2. \end{array}$$

Create a test statistic that (under H_0) has a known sampling distribution and specify a level α rejection region.

Solution. We start with what we know from STAT 512:

$$X \sim \mathcal{N}(0, \sigma_X^2) \implies \frac{X}{\sigma_X} \sim \mathcal{N}(0,1) \implies \frac{X^2}{\sigma_X^2} \sim \chi^2(1) \implies \frac{1}{\sigma_X^2}\sum_{i=1}^n X_i^2 \sim \chi^2(n).$$

Similarly,

$$\frac{1}{\sigma_Y^2}\sum_{i=1}^m Y_i^2 \sim \chi^2(m).$$

Because the two samples are independent, the statistics $\sum_{i=1}^n X_i^2$ and $\sum_{i=1}^m Y_i^2$ are also independent; hence,

$$\frac{\frac{\sum_{i=1}^n X_i^2}{\sigma_X^2}\Big/ n}{\frac{\sum_{i=1}^m Y_i^2}{\sigma_Y^2}\Big/ m} \sim F(n,m).$$

When $H_0 : \sigma_X^2 = \sigma_Y^2$ is true, we have

$$T = \frac{\frac{1}{n}\sum_{i=1}^n X_i^2}{\frac{1}{m}\sum_{i=1}^n Y_i^2} \overset{H_0}{\sim} F(n,m).$$

Therefore, we can use T (the ratio of the second sample moments) as a test statistic to test H_0 versus H_a. Define

$$\begin{aligned} F_{n,m,1-\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } F(n,m) \\ F_{n,m,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } F(n,m) \end{aligned}$$

and refer to Figure 10.7 (next page). A level α rejection region is

$$\text{RR} = \left\{ t < F_{n,m,1-\alpha/2} \ \text{ or } \ t > F_{n,m,\alpha/2} \right\},$$

where $t = \frac{1}{n}\sum_{i=1}^n x_i^2 / \frac{1}{m}\sum_{i=1}^m y_i^2$.

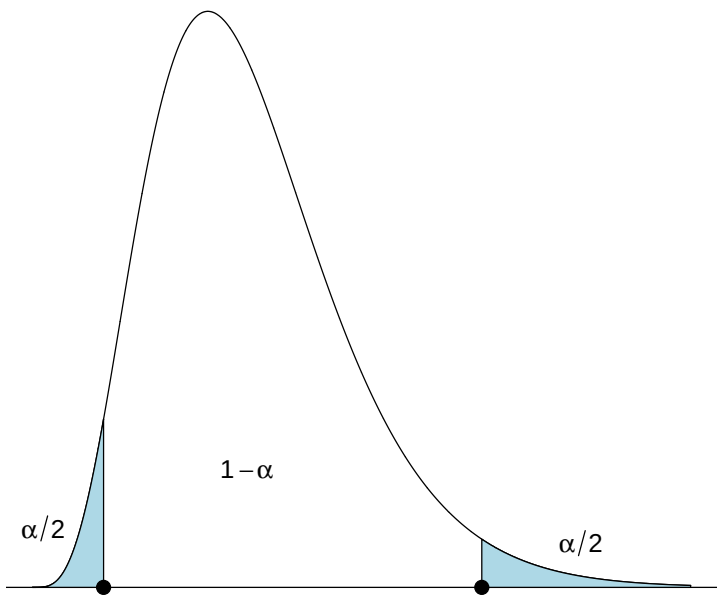

Figure 10.7: $F(n, m)$ pdf. The lower $\alpha/2$ quantile $F_{n,m,1-\alpha/2}$ and the upper $\alpha/2$ quantile $F_{n,m,\alpha/2}$ are shown by using dark circles.

Application: A study was performed to compare the serum alkaline phosphatase (ALP) levels in children with seizures who received anticonvulsant therapy (ACT) to the levels in a control group of children who did not receive ACT and had no history of having seizures. Investigators were interested in how the variation of the ALP levels compare for the two groups. We regard the (centered) ALP levels to be normally distributed with zero mean and variances σ_X^2 (control group, $n = 20$) and σ_Y^2 (treatment group, $m = 25$). Side-by-side boxplots of the centered data are shown in Figure 10.8 (next page). With $n = 20$ and $m = 25$, a level $\alpha = 0.05$ rejection region is

$$\text{RR} = \{t < 0.42 \ \text{ or } \ t > 2.30\}.$$

The value of the test statistic is

$$t = \frac{\frac{1}{20}\sum_{i=1}^{20} x_i^2}{\frac{1}{25}\sum_{i=1}^{25} y_i^2} \approx \frac{629.92}{1392.65} \approx 0.45.$$

Therefore, we would not reject $H_0 : \sigma_X^2 = \sigma_Y^2$ at the $\alpha = 0.05$ level. $\square$

```
> qf(0.025,20,25)
[1] 0.4173726
> qf(0.975,20,25)
[1] 2.300455
```

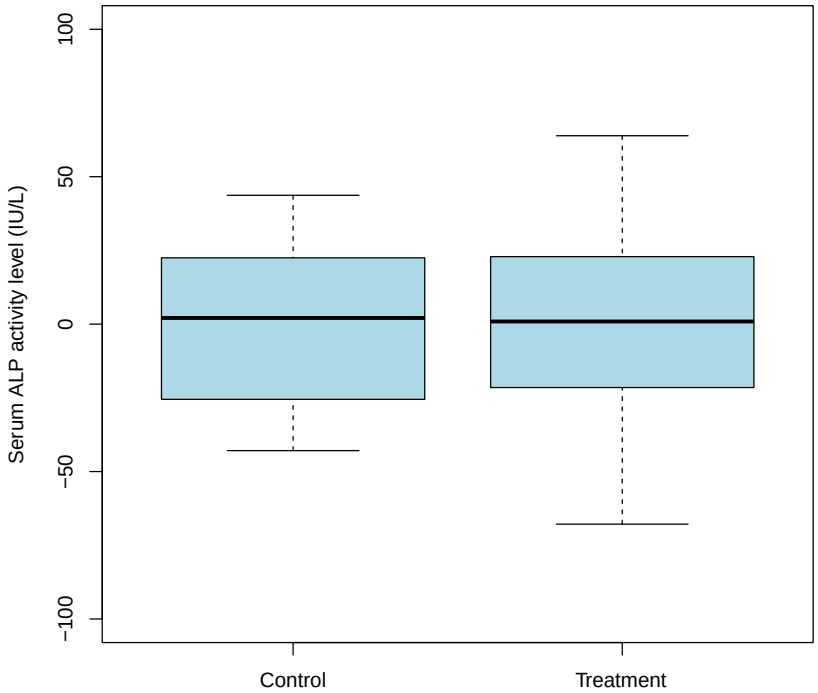

Figure 10.8: Seizure data. ALP levels for control ($n = 20$) and treatment ($m = 25$) groups.

Example 10.6. 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 are interested in testing

$$H_0 : \theta = 1$$
$$\text{versus}$$
$$H_a : \theta > 1.$$

We will use the test statistic $T = T(Y_1, Y_2, ..., Y_n) = Y_{(n)}$, the maximum order statistic, and a rejection region of the form

$$\text{RR} = \left\{ t = y_{(n)} > k \right\}.$$

(a) Find the constant k which ensures $\alpha = 0.05$.
(b) For the value of k in part (a), determine the smallest sample size n to ensure the probability of Type II Error $\beta \leq 0.10$ when $\theta = 1.25$.

Solution. Recall the $\mathcal{U}(0, \theta)$ pdf and cdf are

$$f_Y(y) = \begin{cases} \dfrac{1}{\theta}, & 0 < y < \theta \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad F_Y(y) = \begin{cases} 0, & y \leq 0 \\ \dfrac{y}{\theta}, & 0 < y < \theta \\ 1, & y \geq \theta, \end{cases}$$

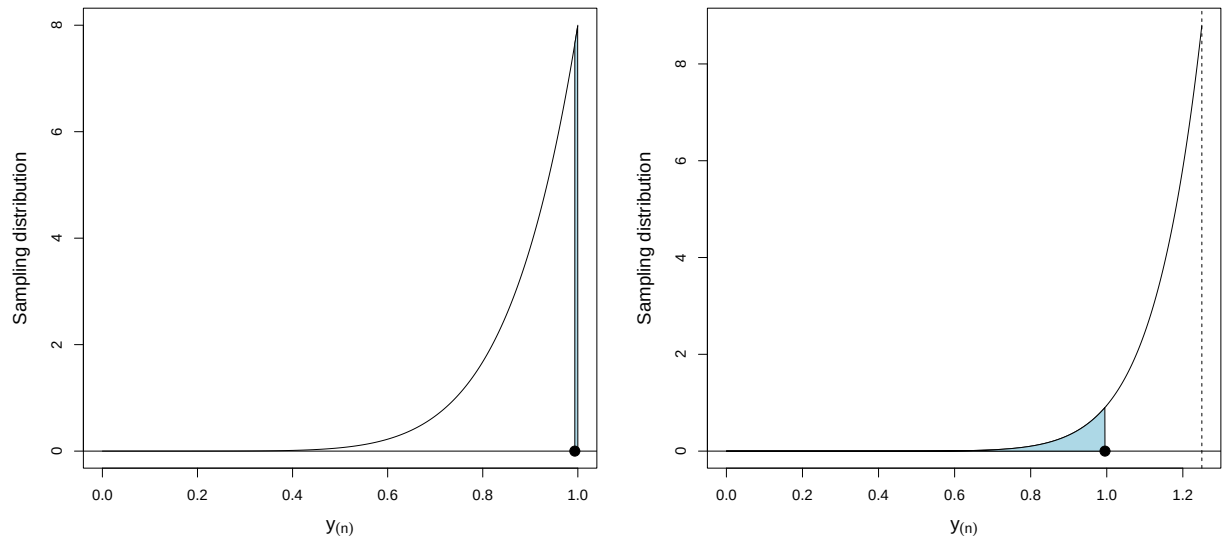

Figure 10.9: Example 10.6. Left: Sampling distribution of $T = Y_{(n)}$ when H_0 is true. The rejection region RR $= \left\{t = y_{(n)} > k\right\}$ is shown shaded. Right: Sampling distribution of $T = Y_{(n)}$ when $n = 11$ and $\theta = 1.25$; i.e., when H_a is true. The probability of Type II Error (shaded) satisfies $\beta \leq 0.10$.

respectively. The sampling distribution of the maximum order statistic $T = Y_{(n)}$ is

$$f_{Y_{(n)}}(y) = nf_Y(y)[F_Y(y)]^{n-1} = \begin{cases} \dfrac{ny^{n-1}}{\theta^n}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

In part (a), we want to determine k so that

$$0.05 = P_{H_0}(\text{RR}) = P_{H_0}\left(Y_{(n)} > k\right) = \int_k^1 ny^{n-1}dy = y^n\Big|_{y=k}^1 = 1 - k^n \implies k = 0.95^{1/n};$$

see Figure 10.9 (above, left). In part (b), we want to find the value of n which solves

$$\begin{aligned} 0.10 = P(\text{Type II Error}) &= P(\text{Do not reject } H_0|\theta = 1.25) \\ &= P\left(Y_{(n)} < 0.95^{1/n}|\theta = 1.25\right) \\ &= \int_0^{0.95^{1/n}} \frac{ny^{n-1}}{(1.25)^n}dy = \frac{1}{(1.25)^n}\ y^n\Big|_0^{0.95^{1/n}} = \frac{0.95}{(1.25)^n}. \end{aligned}$$

Solving this equation for n, we have

$$(1.25)^n = 9.5 \implies n\ln(1.25) = \ln(9.5) \implies n = \frac{\ln(9.5)}{\ln(1.25)} \approx 10.08.$$

Because $(0.95)/(1.25)^n$ is monotone decreasing in n, the solution $n = 11$ is the smallest sample size that guarantees $\beta \leq 0.10$ when $\theta = 1.25$; see Figure 10.9 (above, right). $\square$

10.3 Common large-sample hypothesis tests

Note: We revisit the four settings described in Section 8.3 (STAT 512); i.e., inference for population means and population proportions for one and two populations. Our goal now is to formulate large-sample hypothesis tests 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 (see Section 8.3), 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. These estimators and their standard errors are summarized below:

Parameter θ	Estimator $\widehat{\theta}$	Standard error $\sigma_{\widehat{\theta}}$	Estimated standard error $\widehat{\sigma}_{\widehat{\theta}}$
μ	$\overline{Y}$	$\dfrac{\sigma}{\sqrt{n}}$	$\dfrac{S}{\sqrt{n}}$
p	$\widehat{p}$	$\sqrt{\dfrac{p(1-p)}{n}}$	$\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$
$\mu_1 - \mu_2$	$\overline{Y}_{1+} - \overline{Y}_{2+}$	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$	$\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}$
$p_1 - p_2$	$\widehat{p}_1 - \widehat{p}_2$	$\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$	$\sqrt{\dfrac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \dfrac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$

10.3.1 Theoretical development

Derivation: Suppose the goal is to perform a level α test for

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta \neq \theta_0. \end{gathered}$$

We know

$$\widehat{\theta} \sim \mathcal{AN}(\theta, \sigma^2_{\widehat{\theta}}) \implies Z = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} \overset{H_0}{\sim} \mathcal{AN}(0,1),$$

when the sample size(s) is (are) large. The last statement regarding Z can be written more mathematically; i.e., when $H_0 : \theta = \theta_0$ is true,

$$Z = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} \xrightarrow{d} \mathcal{N}(0,1), \quad \text{as } n \to \infty \quad (\text{or as } \min\{n_1, n_2\} \to \infty).$$

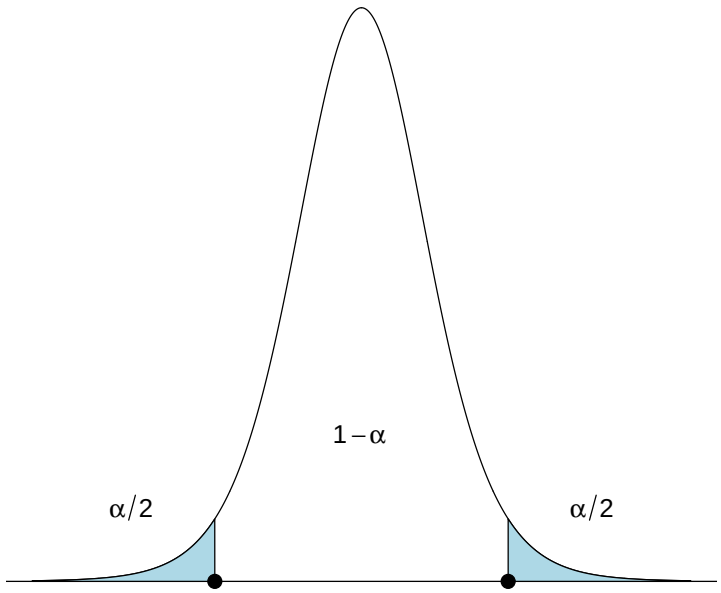

Figure 10.10: $\mathcal{N}(0,1)$ pdf. The lower $\alpha/2$ quantile $-z_{\alpha/2}$ and the upper $\alpha/2$ quantile $z_{\alpha/2}$ are shown by using dark circles.

With this convergence result, one might hope to use Z as a large-sample test statistic to test H_0 versus H_a. The problem is that the standard error $\sigma_{\widehat{\theta}}$ in all four scenarios depends on unknown population-level parameters (e.g., σ^2, p, etc.). Thus, we cannot use Z as a test statistic because it isn't even a statistic. The "work-around" involves noting that

$$\frac{\sigma_{\widehat{\theta}}}{\widehat{\sigma}_{\widehat{\theta}}} \xrightarrow{p} 1, \quad \text{as } n \to \infty \quad (\text{or as } \min\{n_1, n_2\} \to \infty).$$

This is true in each scenario because $\widehat{\theta} \xrightarrow{p} \theta$ by the WLLN and the ratio $\sigma_{\widehat{\theta}}/\widehat{\sigma}_{\widehat{\theta}}$ is a continuous function. Therefore, when $H_0 : \theta = \theta_0$ is true,

$$Z^* = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}} = \underbrace{\frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}}}_{\xrightarrow{d}\mathcal{N}(0,1)} \times \underbrace{\left(\frac{\sigma_{\widehat{\theta}}}{\widehat{\sigma}_{\widehat{\theta}}}\right)}_{\xrightarrow{p}1} \xrightarrow{d} \mathcal{N}(0,1),$$

by Slutsky's Theorem. An **approximate** level α test uses Z^* as a test statistic with rejection region

$$\text{RR} = \{z^* < -z_{\alpha/2} \ \text{ or } \ z^* > z_{\alpha/2}\} = \{|z^*| > z_{\alpha/2}\},$$

where

$$\begin{aligned} -z_{\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } \mathcal{N}(0,1) \\ z_{\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } \mathcal{N}(0,1); \end{aligned}$$

see Figure 10.10 (above).

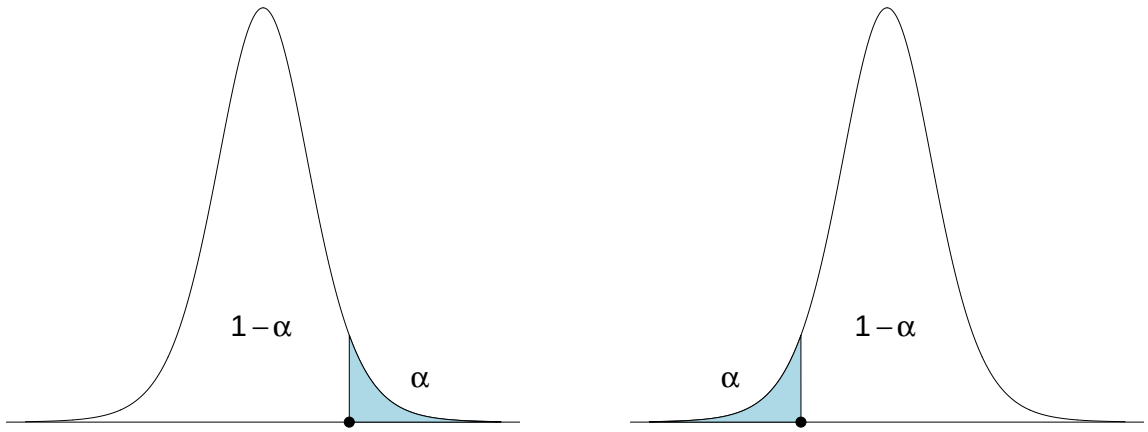

Figure 10.11: $\mathcal{N}(0,1)$ pdf. Left: The upper α quantile z_α is shown by using a dark circle. Right: The lower α quantile $-z_\alpha$ is shown by using a dark circle.

One-sided tests: The previous derivation assumes a two-sided $H_a : \theta \neq \theta_0$. Making adjustments for one-sided alternatives is easy. If we test $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$, the rejection region is

$$\text{RR} = \{z^* > z_\alpha\}.$$

If we test $H_0 : \theta = \theta_0$ versus $H_a : \theta < \theta_0$, the rejection region is

$$\text{RR} = \{z^* < -z_\alpha\};$$

see Figure 10.11 (above). Both of these (one-sided) rejection regions satisfy $\alpha \approx P_{H_0}(\text{RR})$.

10.3.2 Relationship with (large-sample) interval estimators

Interesting: By now, you are probably realizing that performing a level α hypothesis test and writing a $1-\alpha$ interval estimator are similar procedures. We have shown

$$Z^* = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}} \overset{H_0}{\sim} \mathcal{AN}(0,1),$$

when the sample size(s) is (are) large; i.e., Z^* is a large-sample pivotal quantity. This means we can write

$$\begin{aligned} 1-\alpha \approx P_{H_0}(-z_{\alpha/2} < Z^* < z_{\alpha/2}) &= P_{H_0}\left(-z_{\alpha/2} < \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}} < z_{\alpha/2}\right) \\ &= P_{H_0}\left(\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}} < \theta_0 < \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}\right). \end{aligned}$$

Of course, we recognize

$$\left(\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}},\ \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}\right)$$

as an approximate $1-\alpha$ interval estimator for θ. But, when H_0 is true, what does the event

$$\{-z_{\alpha/2} < Z^* < z_{\alpha/2}\}$$

represent? It is the complement of the rejection region (i.e., the "acceptance region") for a two-sided test of $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$. Therefore, we have discovered the following equivalence:

$$\theta_0 \text{ resides in the } 1-\alpha \text{ interval estimator} \iff H_0 : \theta = \theta_0 \text{ is not rejected at level } \alpha.$$

A similar equivalence exists for one-sided tests and one-sided interval estimators.

Discussion: Given the equivalence above, this prompts the obvious question: *Do we really need hypothesis tests?* After all, we can perform the test by writing an interval estimator for θ and then noting where θ_0 falls. In addition, interval estimators give us more information than simply saying "Reject H_0" or "Do not reject H_0." In many instances, writing interval estimators is the preferred method of statistical inference. However, the notion of testing extends far beyond the (relatively simple) problems we consider in this chapter. For example, a hypothesis test might be used to determine if two categorical variables are independent, if errors in a linear regression model follow a normal distribution, or if two distributions are (stochastically) ordered in some way. In these and other scenarios, there may not even be a single population-level parameter that describes the hypothesis under consideration.

10.3.3 Sample size determination

Setting: In the large-sample scenarios considered in this section, suppose we would test

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta > \theta_0 \end{gathered}$$

using a rejection region of the form

$$\text{RR} = \{\widehat{\theta} > k\},$$

where k is chosen to ensure the test is (approximately) level α; i.e., $\alpha \approx P_{H_0}(\widehat{\theta} > k)$. Our goal is to determine the sample size n that confers a Type II Error probability equal to β for a pre-specified value $\theta_a > \theta_0$, that is,

$$\theta_a = \theta_0 + \Delta,$$

where $\Delta > 0$ is the **practically important difference** we wish to detect. The Type I Error probability request implies

$$\alpha \approx P_{H_0}(\widehat{\theta} > k) = P_{H_0}\left(Z > \frac{k-\theta_0}{\sigma_{\widehat{\theta}}}\right) \Longrightarrow \frac{k-\theta_0}{\sigma_{\widehat{\theta}}} = z_\alpha.$$

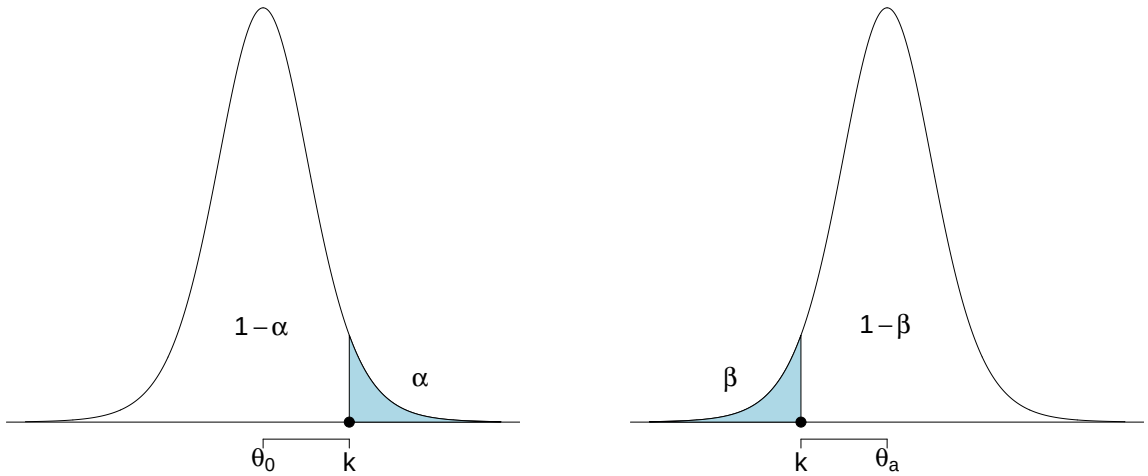

Figure 10.12: Left: Sampling distribution $\widehat{\theta} \sim \mathcal{AN}(\theta_0, \sigma^2_{\widehat{\theta}})$ under H_0. The rejection region $\text{RR} = \{\widehat{\theta} > k\}$ (shaded) has probability α. Right: Sampling distribution $\widehat{\theta} \sim \mathcal{AN}(\theta_a, \sigma^2_{\widehat{\theta}})$ under H_a. The complement set $\overline{\text{RR}} = \{\widehat{\theta} < k\}$ (shaded) has probability β.

The Type II Error probability request implies

$$\beta \approx P_{H_a}(\widehat{\theta} < k) = P_{H_a}\left(Z < \frac{k - \theta_a}{\sigma_{\widehat{\theta}}}\right) \Longrightarrow \frac{k - \theta_a}{\sigma_{\widehat{\theta}}} = -z_\beta.$$

Our goal is to select the sample size n that satisfies both of these two equations. Note that the standard error $\sigma_{\widehat{\theta}}$ is a function of sample size, so n is "hiding" in this quantity. Figure 10.12 (above) is helpful in understanding where these equations come from.

Population mean: Suppose the goal is to test

$$\begin{array}{c} H_0 : \mu = \mu_0 \\ \text{versus} \\ H_a : \mu > \mu_0 \end{array}$$

using a rejection region of the form

$$\text{RR} = \{\overline{Y} > k\},$$

where $\overline{Y}$ is the sample mean. The two equations (previously derived) are

$$\begin{aligned} \frac{k - \mu_0}{\sigma/\sqrt{n}} &= z_\alpha \quad \text{(Type I Error requirement)} \\ \frac{k - \mu_a}{\sigma/\sqrt{n}} &= -z_\beta \quad \text{(Type II Error requirement).} \end{aligned}$$

Solving for k in each equation (and then equating the solutions), we get

$$z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) + \mu_0 = -z_\beta\left(\frac{\sigma}{\sqrt{n}}\right) + \mu_a.$$

Solving the last equation for n yields

$$n = \left[\frac{(z_\alpha + z_\beta)\sigma}{\Delta}\right]^2,$$

where $\Delta = \mu_a - \mu_0$. This solution for n will ensure

- the test is approximately level α
- when $\mu = \mu_a$ (i.e., the alternative H_a is true), we will not reject H_0 with probability approximately equal to β;
  - i.e., we *will* reject H_0 with probability approximately equal to $1 - \beta$.

Remark: For this formula to be useful, we must elicit a value for the population variance σ^2 (in this instance, a "nuisance parameter"). In practice, this value can be chosen by using related studies, preliminary data, or expert opinion. In the absence of available information, one could use a conservative upper bound for σ^2 (this will make the sample size n larger).

Example 10.7. A researcher who specializes in childhood obesity is examining school-provided lunches at public elementary schools in Augusta, GA. In this population, suppose she wants to test

$$\begin{gathered} H_0 : \mu = 30 \\ \text{versus} \\ H_a : \mu > 30, \end{gathered}$$

where μ is the population mean BMI. She wants to perform a level $\alpha = 0.05$ test. In addition, if the population mean BMI is 32, she would like to reject H_0 with probability at least 0.80. How many children should she sample to perform the test with these requirements? The population standard deviation σ is assumed to be approximately 7.5.

Solution. The probability of Type I Error is $\alpha = 0.05$. The probability of Type II Error (when $\mu_a = 32$) is $\beta = 0.20$. The corresponding standard normal quantiles are $z_{0.05} \approx 1.65$ and $z_{0.20} \approx 0.84$, respectively. The practically important difference is $\Delta = \mu_a - \mu_0 = 32 - 30 = 2$. We have

$$n = \left[\frac{(z_\alpha + z_\beta)\sigma}{\Delta}\right]^2 \approx \left[\frac{(1.65 + 0.84)(7.5)}{2}\right]^2 \approx 87.19.$$

Therefore, she would have to sample $n = 88$ students.

```
> qnorm(0.95,0,1)
[1] 1.644854
> qnorm(0.80,0,1)
[1] 0.8416212
```

Q: How would the sample size change if she required $\beta = 0.01$ instead? That is, if $\mu_a = 32$, then she would like to reject H_0 with probability $1 - \beta = 0.99$.
A: Now, she would use $z_\beta = z_{0.01} \approx 2.33$ to reflect the change in the Type II Error probability requirement. We have

$$n = \left[\frac{(z_\alpha + z_\beta)\sigma}{\Delta}\right]^2 \approx \left[\frac{(1.65 + 2.33)(7.5)}{2}\right]^2 \approx 222.76.$$

She would have to sample $n = 223$ students. $\square$

```
> qnorm(0.99,0,1)
[1] 2.326348
```

Population proportion: Suppose the goal is to test

$$\begin{array}{c} H_0 : p = p_0 \\ \text{versus} \\ H_a : p > p_0 \end{array}$$

using a rejection region of the form

$$\text{RR} = \{\widehat{p} > k\},$$

where $\widehat{p}$ is the sample proportion. The two equations (previously derived) are

$$\begin{aligned} \frac{k - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} &= z_\alpha \quad \text{(Type I Error requirement)} \\ \frac{k - p_a}{\sqrt{\dfrac{p_a(1-p_a)}{n}}} &= -z_\beta \quad \text{(Type II Error requirement).} \end{aligned}$$

The value of n that satisfies both equations is

$$n = \left[\frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_a(1-p_a)}}{\Delta}\right]^2,$$

where $\Delta = p_a - p_0$. This solution for n will ensure

- the test is approximately level α
- when $p = p_a$ (i.e., the alternative H_a is true), we will not reject H_0 with probability approximately equal to β;
  - i.e., we *will* reject H_0 with probability approximately equal to $1 - \beta$.

Example 10.8. A Phase II clinical trial evaluating the effects of a new drug for metastatic prostate cancer shows positive results, so investigators would like to plan a larger Phase III trial. The goal is to test

$$\begin{gathered} H_0 : p = 0.35 \\ \text{versus} \\ H_a : p > 0.35, \end{gathered}$$

where p is the (population-level) probability of response among all eligible patients. The protocol calls for a level $\alpha = 0.05$ test. In addition, if the probability of response is $p_a = 0.40$, the investigators would like to reject H_0 with probability 0.80. How many patients should be recruited for the Phase III trial?

Solution. We have $\alpha = 0.05$ and $\beta = 0.20$. From Example 10.7, we know to use $z_\alpha = z_{0.05} \approx 1.65$ and $z_\beta = z_{0.20} \approx 0.84$. The practically important difference is $\Delta = p_a - p_0 = 0.40 - 0.35 = 0.05$. We have

$$n = \left[\frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_a(1-p_a)}}{\Delta}\right]^2 \approx \left[\frac{1.65\sqrt{0.35(0.65)} + 0.84\sqrt{0.40(0.60)}}{0.05}\right]^2 \approx 574.57.$$

Therefore, the Phase III trial will require 575 patients. □

10.4 Hypothesis tests arising from normal populations

Preview: We now derive hypothesis tests for means and variances when the population distribution is $\mathcal{N}(\mu, \sigma^2)$. We consider one and two populations. Given the duality between hypothesis tests and interval estimators, one should compare this section with Section 8.7 (notes) in STAT 512. All hypothesis tests in this section are **exact**; i.e., we are not appealing to large-sample arguments like we did in the last section. Rejection regions come from the t, χ^2, and F distributions, and they have Type I Error probability equal to α exactly.

Remark: Test statistics for all scenarios considered this section are given without proof. This is because we have derived all relevant sampling distributions in STAT 512 (see Chapters 7-8). All tests presented will assume a two-sided alternative (so that the rejection region is two-sided). Rejection regions for one-sided alternatives are formed in the obvious way.

10.4.1 Population mean μ

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. The goal is to construct a level α test for

$$\begin{gathered} H_0 : \mu = \mu_0 \\ \text{versus} \\ H_a : \mu \neq \mu_0. \end{gathered}$$

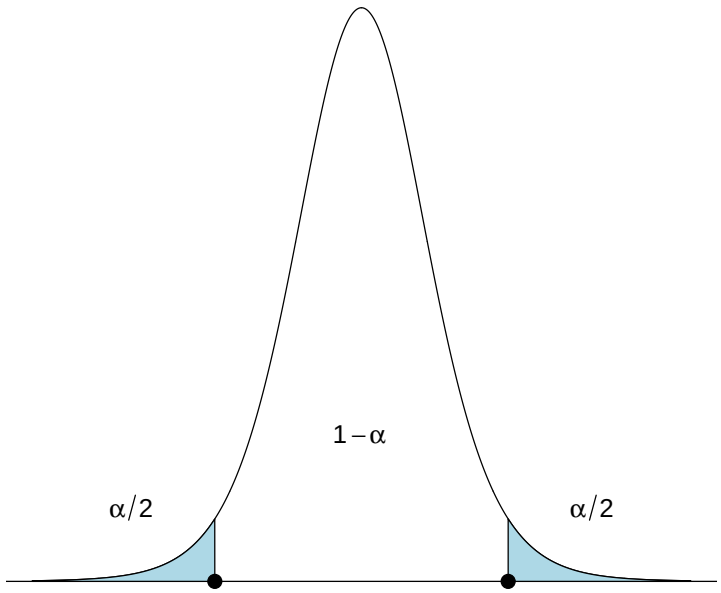

Figure 10.13: $t(n-1)$ pdf. The lower $\alpha/2$ quantile $-t_{n-1,\alpha/2}$ and the upper $\alpha/2$ quantile $t_{n-1,\alpha/2}$ are shown by using dark circles.

When H_0 is true, we know

$$T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t(n-1).$$

Therefore, a level α test uses the rejection region

$$\text{RR} = \{t < -t_{n-1,\alpha/2} \ \text{ or } \ t > t_{n-1,\alpha/2}\} = \{|t| > t_{n-1,\alpha/2}\},$$

where

$$\begin{aligned} -t_{n-1,\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } t(n-1) \\ t_{n-1,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } t(n-1); \end{aligned}$$

see Figure 10.13 (above). This inference procedure is called a **one-sample t test**.

10.4.2 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. The goal is to construct a level α test for

$$\begin{gathered} H_0 : \sigma^2 = \sigma_0^2 \\ \text{versus} \\ H_a : \sigma^2 \neq \sigma_0^2. \end{gathered}$$

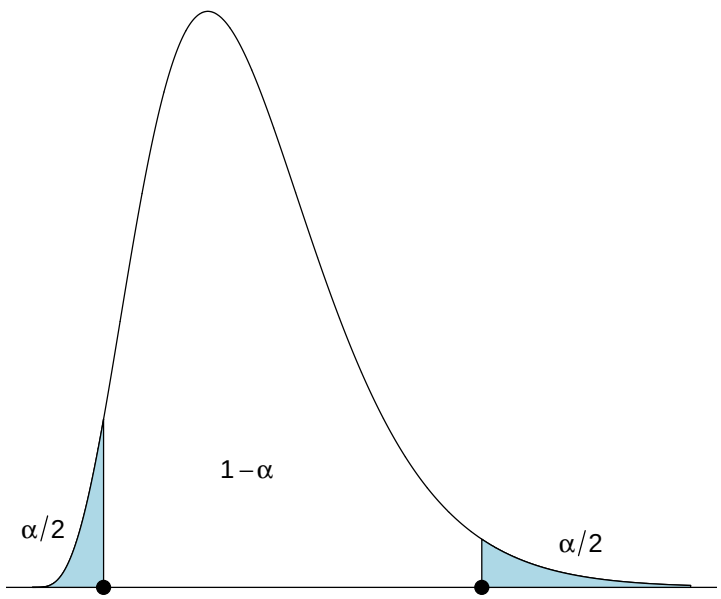

Figure 10.14: $\chi^2(n-1)$ pdf. The lower $\alpha/2$ quantile $\chi^2_{n-1,1-\alpha/2}$ and the upper $\alpha/2$ quantile $\chi^2_{n-1,\alpha/2}$ are shown by using dark circles.

When H_0 is true, we know

$$T = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1).$$

Therefore, a level α test uses the rejection region

$$\text{RR} = \{t < \chi^2_{n-1,1-\alpha/2} \ \text{ or } \ t > \chi^2_{n-1,\alpha/2}\},$$

where

$$\begin{aligned} \chi^2_{n-1,1-\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } \chi^2(n-1) \\ \chi^2_{n-1,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } \chi^2(n-1); \end{aligned}$$

see Figure 10.14 (above).

10.4.3 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. The goal is to construct a level α test for

$$H_0 : \mu_1 - \mu_2 = d_0$$
$$\text{versus}$$
$$H_a : \mu_1 - \mu_2 \neq d_0,$$

where d_0 is a pre-specified difference of the population means. In practice, one often takes $d_0 = 0$ because the goal is to test whether or not the population means are equal. This inference procedure is called a **two-sample t test**.

Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$; i.e., the population variances are **equal**. When H_0 is true, we know

$$T = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - d_0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

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. Therefore, a level α test uses the rejection region

$$\text{RR} = \{t < -t_{n_1+n_2-2,\alpha/2} \ \text{ or } \ t > t_{n_1+n_2-2,\alpha/2}\} = \{|t| > t_{n_1+n_2-2,\alpha/2}\}.$$

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 test statistic available for which the exact sampling distribution is known under H_0. When H_0 is true, it can be shown that

$$T = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - d_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

follows an **approximate** $t(\nu)$ distribution, where

$$\nu = \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{S_1^4}{n_1^2(n_1 - 1)} + \dfrac{S_2^4}{n_2^2(n_2 - 1)}}.$$

Therefore, an approximate level α test uses the same RR above with the degrees of freedom $n_1 + n_2 - 2$ replaced by ν.

Note: This equation for ν above, which is commonly referred to as **Satterthwaite's formula**, is obtained by using a MOM argument to estimate degrees of freedom for general linear combinations of χ^2 random variables. Under equal variances, recall that

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2);$$

i.e., this specific linear combination of S_1^2 and S_2^2 follows a χ^2 distribution exactly and the common population variance σ^2 "cancels out" when the two-sample t statistic is formed. When the population variances are unequal ($\sigma_1^2 \neq \sigma_2^2$), the best we can do is work with

$$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}.$$

This quantity does not have a χ^2 distribution. However, it can be regarded as a linear combination, which, in distribution, is of the form $a_1\chi^2(\nu_1) + a_2\chi^2(\nu_2)$, where a_1 and a_2 are constants (depending on sample sizes and population variances). A rigorous derivation of Satterthwaite's formula is challenging, but this is the central issue.

10.4.4 Equality of two population variances (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. The goal is to construct a level α test for

$$\begin{array}{c} H_0 : \sigma_1^2 = \sigma_2^2 \\ \text{versus} \\ H_a : \sigma_1^2 \neq \sigma_2^2. \end{array}$$

From STAT 512, we know

$$\frac{\dfrac{(n_1-1)S_1^2}{\sigma_1^2}\Big/(n_1-1)}{\dfrac{(n_2-1)S_2^2}{\sigma_2^2}\Big/(n_2-1)} = \left(\frac{S_1^2}{S_2^2}\right)\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1-1, n_2-1).$$

Note that when H_0 is true, the test statistic

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

Therefore, a level α test uses the rejection region

$$\text{RR} = \{t < F_{n_1-1,n_2-1,1-\alpha/2} \ \text{ or } \ t > F_{n_1-1,n_2-1,\alpha/2}\},$$

where

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile of } F(n_1-1, n_2-1) \\ F_{n_1-1,n_2-1,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile of } F(n_1-1, n_2-1) \end{aligned}$$

see Figure 10.15 (next page).

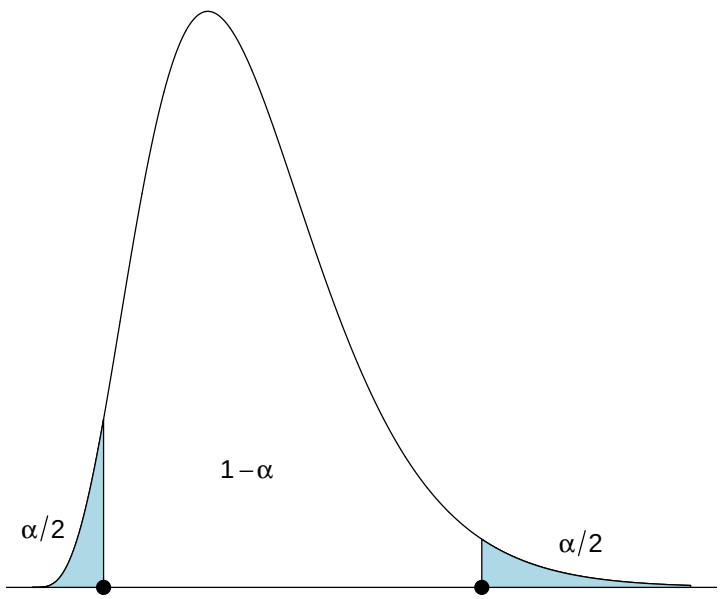

Figure 10.15: $F(n_1 - 1, n_2 - 1)$ pdf. The lower $\alpha/2$ quantile $F_{n_1-1,n_2-1,1-\alpha/2}$ and the upper $\alpha/2$ quantile $F_{n_1-1,n_2-1,\alpha/2}$ are shown by using dark circles.

10.5 Probability values

Example 10.9. *Does exercise delay the age at menarche?* A study examining the effects of exercise on the menstrual cycle compared two groups of swimmers: females who began training prior to menarche (the beginning of menstruation) and females who began training after they had reached menarche. Two independent random samples of swimmers were obtained. Our goal is to test

$$\begin{gathered} H_0 : \mu_1 - \mu_2 = 0 \\ \text{versus} \\ H_a : \mu_1 - \mu_2 > 0, \end{gathered}$$

where μ_1 is the population mean age at menarche for those who began training before they had reached menarche and μ_2 is the population mean age at menarche for those who began training after. Side-by-side boxplots of the data are shown in Figure 10.16 (next page, left).

Analysis: I used the `t.test` function in R to perform a two-sample t test while assuming normality for both populations and equal population variances (see Section 10.4.3, notes). Here is the output:

```
> t.test(pre_men,post_men,conf.level=0.95,var.equal=TRUE,alternative="greater")
t = 7.0583, df = 150, p-value = 2.914e-11
```

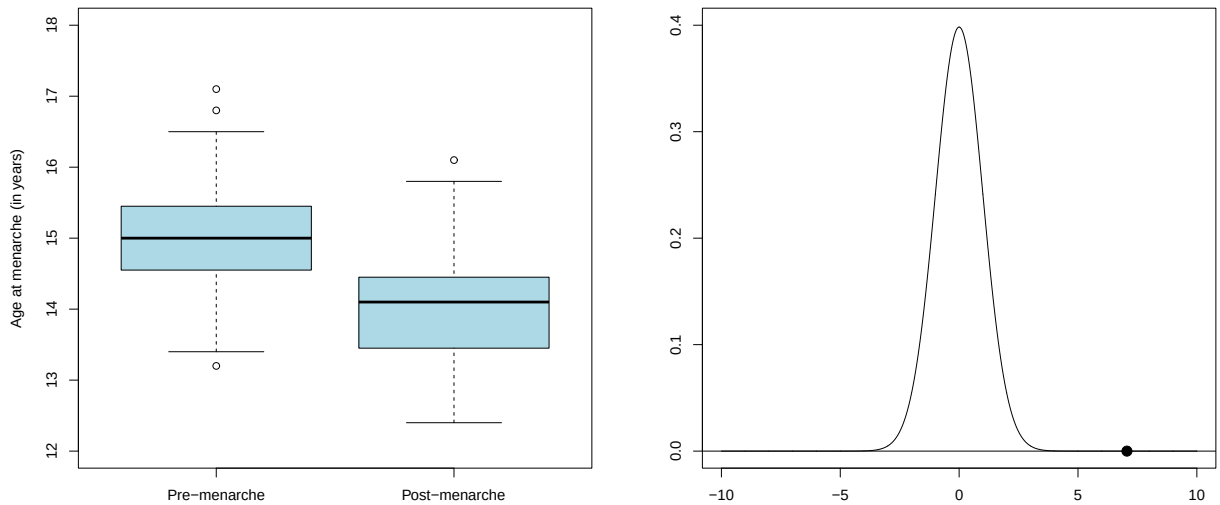

Figure 10.16: Swimmer data. Left: Age at menarche for two groups: training before menarche ($n_1 = 96$) and training after menarche ($n_2 = 56$). Right: $t(150)$ pdf; the test statistic $t \approx 7.06$ is shown by using a dark circle.

The value of the test statistic is $t = 7.0583$, which is so far out in the tail of the $t(150)$ reference distribution that H_0 would be rejected at any reasonable level. The probability value

$$\text{p-value} = 2.491 \times 10^{-11}$$

is the area to the right of $t = 7.0583$ under the $t(150)$ pdf. Note that if H_0 was true, then the test statistic t would be regarded as a realization from this pdf. The very small p-value suggests that indeed this is highly unlikely. $\square$

Terminology: The **probability value** (**p-value**) is the smallest significance level α for which H_0 would be rejected. Therefore,

- if p-value $\leq \alpha$, then we reject H_0
- if p-value $> \alpha$, then we do not reject H_0.

Remarks: The p-value is one of the most fundamentally misunderstood statistics in all of statistics. The reason for this is that applied researchers are generally not quite sure what it means.

- Because a p-value is always calculated assuming H_0 is true, it is tempting to say that it is "the probability H_0 is true." Unfortunately, this is not correct. In Example 10.9, this interpretation would literally mean

$$P(\mu_1 - \mu_2 > 0),$$

which does not even make sense mathematically because μ_1 and μ_2 are fixed (i.e., they are not random).

- Other texts might use the mysterious "more extreme" analogy in an attempt to interpret the p-value, that is,

  > "A p-value is the probability that if we replicated the study, we would observe a test statistic as extreme or more extreme than the one we observed, assuming H_0 was true."

  Although this interpretation is correct, it is built on the somewhat fanciful notion that a study would ever be replicated just for the benefit of interpreting what researchers just observed.

- I have found the easiest explanation (for non-statisticians) is that "the p-value is a measure of evidence against H_0; the smaller it is, the more evidence we have."

Example 10.3 (continued). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population, where $\theta > 0$ is unknown. Suppose we want to test

$$\begin{gathered} H_0 : \theta = 1 \\ \text{versus} \\ H_a : \theta > 1. \end{gathered}$$

For the accident data in Example 10.3 (pp 8-11, notes), our observed test statistic based on a sample of $n = 84$ policies was

$$t = \sum_{i=1}^{84} y_i = 103.$$

The rejection region in Example 10.3 involved the sample sum $T = \sum_{i=1}^{84} Y_i$ and specified to reject H_0 when T was large. Using the "as/more extreme" analogy above, the p-value is

$$p = P_{H_0}(T \geq 103) \approx 0.025.$$

Therefore, H_0 would be rejected for any significance level $\alpha \geq 0.025$. $\square$

```
> 1-ppois(102,84)
[1] 0.02452888
```

Interesting: Suppose a hypothesis test is performed using a **continuous** test statistic T. When viewed as a random variable, the p-value P is uniformly distributed over $(0, 1)$ when H_0 is true; i.e.,

$$P \overset{H_0}{\sim} \mathcal{U}(0, 1).$$

This is not true when the test statistic T is discrete.

Simulation: I prove the result above in a more advanced course, but for our purposes a simulation exercise should be sufficient to demonstrate the result. Suppose $Y_1, Y_2, ...Y_{10}$ is an iid sample from a $\mathcal{N}(0, 1)$ distribution. I used R to simulate $B = 10000$ such samples. For each one, I performed a one-sample t test for

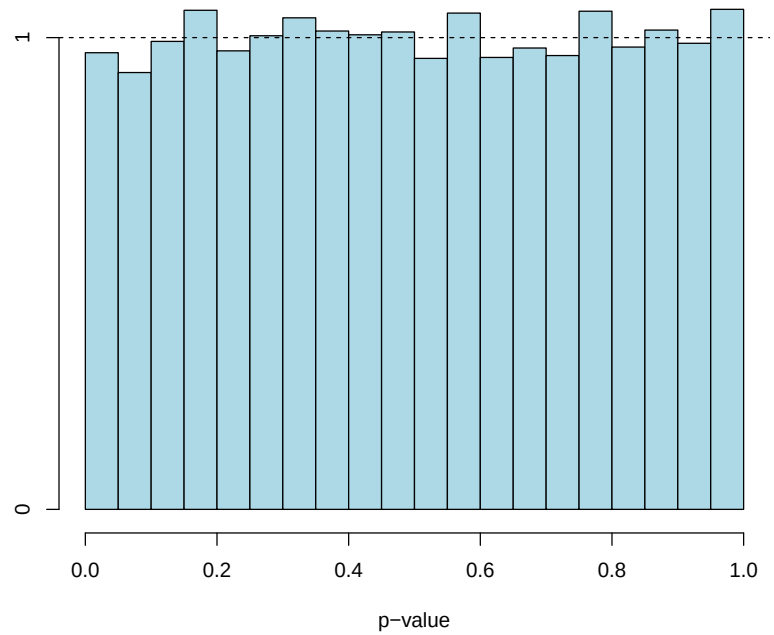

Figure 10.17: Monte Carlo simulation. Histogram of $B = 10000$ probability values.

$$H_0 : \mu = 0$$
$$\text{versus}$$
$$H_a : \mu \neq 0,$$

and I recorded the p-value for each test; note that H_0 is true under the $\mathcal{N}(0, 1)$ population model. Therefore, this simulation produced 10,000 p-values, which I plotted in a histogram; see Figure 10.17 (above). The p-values line up with a $\mathcal{U}(0, 1)$ pdf almost exactly! The discrepancies we see are attributable to the variation arising from Monte Carlo sampling.

Revelation: In the light of the previous result, it should be clear that using a p-value to decide between H_0 and H_a is equivalent to using the rejection region approach that we have espoused all along. Suppose a level α hypothesis test uses the continuous test statistic T with rejection region RR, that is,

$$\alpha = P_{H_0}(T \in \text{RR}).$$

The p-value P calculated from the sampling distribution of T satisfies $P \overset{H_0}{\sim} \mathcal{U}(0, 1)$ and hence

$$\alpha = P_{H_0}(P \leq \alpha).$$

The events $\{T \in \text{RR}\}$ and $\{P \leq \alpha\}$ are the same event, and hence "$T \in \text{RR}$" and "$P \leq \alpha$" are equivalent decision rules.

10.6 Power functions

Remark: The power function is an important part of any hypothesis test, and it emerges as relevant in formalizing definitions of test optimality (as we will see momentarily). In previous examples, we have become comfortable with specifying beforehand the significance level of the test; i.e.,

$$\alpha = P_{H_0}(\text{RR}) = P(\text{Reject } H_0 | H_0 \text{ true}).$$

Then, for a specified value of the population parameter θ which satisfies H_a, say $\theta_a \in H_a$, we might calculate the Type II Error probability β or possibly a sample size n necessary to attain a targeted value of β. These types of calculations can be made by using the power function.

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where the population parameter θ is unknown. Suppose our goal is to perform a level α test of

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta \neq \theta_0, \end{gathered}$$

where θ_0 is a fixed value. In the following, the alternative hypothesis H_a can be one sided; in addition, all definitions henceforth apply regardless of whether the test is "exact" or based on large-sample arguments. The **power function** of the test, denoted by $K(\theta)$, is given by

$$K(\theta) = P_\theta(\text{RR}) = P(\text{Reject } H_0 | \theta).$$

The power function gives the probability of rejecting H_0 when viewed as a function of θ.

- It should be clear that

  $$K(\theta_0) = P_{\theta_0}(\text{RR}) = P_{\theta_0}(\text{Reject } H_0) = P(\text{Reject } H_0 | H_0 \text{ true}) = \alpha.$$

  Therefore, the point (θ_0, α) is one point on the power function.

- For values of θ that are "close to" θ_0, one would expect the power to be smaller than when θ is "far away from" θ_0. This makes sense intuitively. It is more difficult to detect small departures from H_0 than it is to detect large ones.

- The shape of the power function always depends on the alternative hypothesis H_a. Figure 10.18 (next page) shows the typical shape of a power function for the two-sided test above.

Observation: Suppose $\theta_a \in H_a$, that is, θ_a is a value of θ which makes H_a true. Then

$$K(\theta_a) = 1 - P_{\theta_a}(\text{Type II Error}).$$

Proof. This is a straightforward application of the complement rule; i.e.,

$$K(\theta_a) = P_{\theta_a}(\text{Reject } H_0) = 1 - P_{\theta_a}(\text{Do not reject } H_0) = 1 - P_{\theta_a}(\text{Type II Error}). \square$$

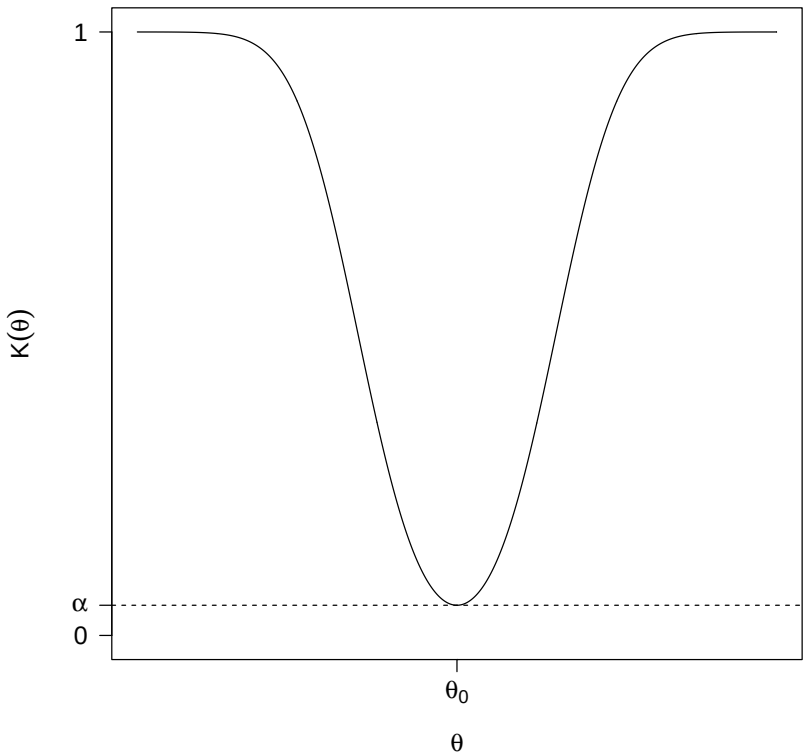

Figure 10.18: Power function for $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$. The significance level satisfies $K(\theta_0) = \alpha$.

Example 10.10. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma_0^2)$ population, where μ is unknown and σ_0^2 is known. We would like to test

$$\begin{gathered} H_0 : \mu = \mu_0 \\ \text{versus} \\ H_a : \mu > \mu_0. \end{gathered}$$

To perform the test, we will use the "one-sample z statistic"

$$Z^* = \frac{\overline{Y} - \mu_0}{\sigma_0/\sqrt{n}} \overset{H_0}{\sim} \mathcal{N}(0, 1)$$

and RR $= \{z^* > z_\alpha\}$ as a level α rejection region. Derive the power function $K(\mu)$ for this test.

Solution. From the definition of the power function, we have

$$K(\mu) = P_\mu(\text{RR}) = P_\mu(Z^* > z_\alpha) = P_\mu\left(\frac{\overline{Y} - \mu_0}{\sigma_0/\sqrt{n}} > z_\alpha\right) = P_\mu\left(\overline{Y} > z_\alpha(\sigma_0/\sqrt{n}) + \mu_0\right).$$

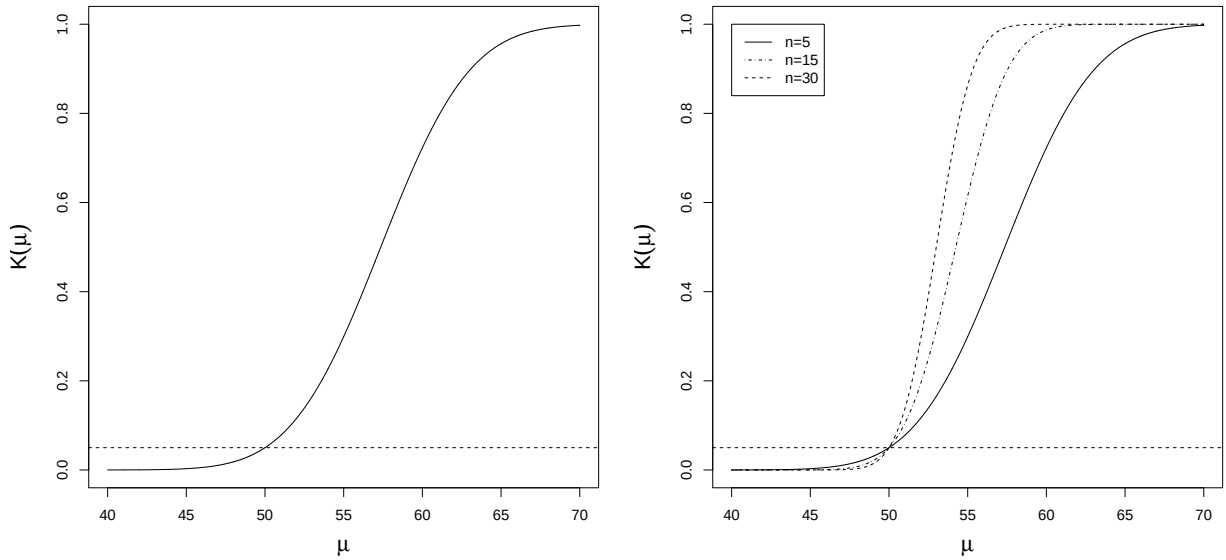

Figure 10.19: Example 10.10. Power function $K(\mu)$ for $H_0 : \mu = 50$ versus $H_a : \mu > 50$. Left: $n = 5$. Right: $n \in \{5, 15, 30\}$. A horizontal line at $\alpha = 0.05$ has been added.

Now, simply re-standardize the random variable $\overline{Y}$ as a function of μ; i.e.,

$$\begin{aligned} P_\mu\left(\overline{Y} > z_\alpha(\sigma_0/\sqrt{n}) + \mu_0\right) &= P\left(\frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} > \frac{z_\alpha(\sigma_0/\sqrt{n}) + \mu_0 - \mu}{\sigma_0/\sqrt{n}}\right) \\ &= P\left(Z > \frac{z_\alpha(\sigma_0/\sqrt{n}) + \mu_0 - \mu}{\sigma_0/\sqrt{n}}\right) \\ &= 1 - F_Z\left(\frac{z_\alpha(\sigma_0/\sqrt{n}) + \mu_0 - \mu}{\sigma_0/\sqrt{n}}\right), \end{aligned}$$

where F_Z is the $\mathcal{N}(0,1)$ cdf. This cdf can be calculated in R using the `pnorm` function.

Discussion: First, it is of interest to note that

$$K(\mu_0) = 1 - F_Z\left(\frac{z_\alpha(\sigma_0/\sqrt{n}) + \mu_0 - \mu_0}{\sigma_0/\sqrt{n}}\right) = 1 - F_Z(z_\alpha) = 1 - (1-\alpha) = \alpha;$$

i.e., the power of the test when H_0 is true is the significance level. Second, it is easy to show $K(\mu)$ is an increasing function of μ; note that

$$\begin{aligned} \frac{\partial}{\partial\mu}K(\mu) &= \frac{\partial}{\partial\mu}\left[1 - F_Z\left(\frac{z_\alpha(\sigma_0/\sqrt{n}) + \mu_0 - \mu}{\sigma_0/\sqrt{n}}\right)\right] \\ &= -f_Z\left(\frac{z_\alpha(\sigma_0/\sqrt{n}) + \mu_0 - \mu}{\sigma_0/\sqrt{n}}\right) \times \left(-\frac{1}{\sigma_0/\sqrt{n}}\right) > 0, \end{aligned}$$

because $f_Z(\cdot) > 0$; i.e., f_Z is a pdf. Figure 10.19 (above) shows examples of what $K(\mu)$ looks like when $\mu_0 = 50$, $\sigma_0^2 = 100$, and $\alpha = 0.05$. $\square$

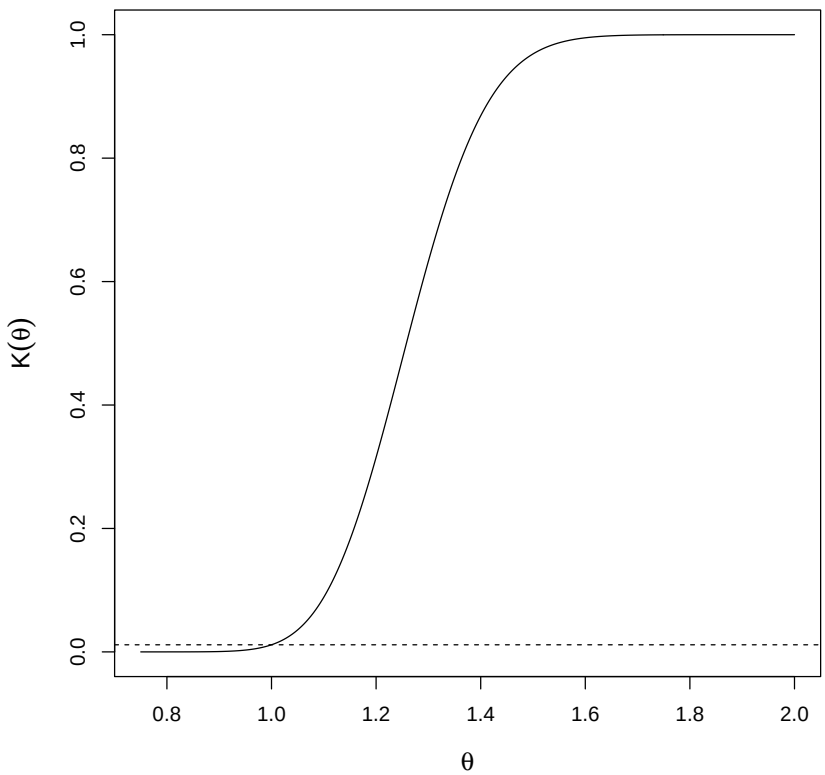

Figure 10.20: Example 10.11. Power function for $H_0 : \theta = 1$ versus $H_a : \theta > 1$. A horizontal line at $\alpha \approx 0.0115$ has been added.

Example 10.11. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population, where $\theta > 0$ is unknown. In Example 10.3 (see pp 8-11, notes), we wanted to test

$$\begin{gathered} H_0 : \theta = 1 \\ \text{versus} \\ H_a : \theta > 1 \end{gathered}$$

with a random sample of $n = 84$ policies. An $\alpha \approx 0.0115$ rejection region was

$$\text{RR} = \left\{ t = \sum_{i=1}^{84} y_i \geq 106 \right\}.$$

Derive the power function $K(\theta)$ for this test.

Solution. Recall that $T = \sum_{i=1}^{84} Y_i \sim \text{Poisson}(84\theta)$. We have

$$K(\theta) = P_\theta(\text{RR}) = P_\theta(T \geq 106) = 1 - P_\theta\left(T \leq 105\right) = 1 - \sum_{t=0}^{105} \frac{(84\theta)^t e^{-84\theta}}{t!}.$$

The probability $P_\theta\left(T \leq 105\right)$ can be calculated in R using the `ppois` function. Figure 10.20 (above) shows the graph of $K(\theta)$. Note that $K(1) = \alpha \approx 0.0115$. $\square$

10.7 Most powerful tests

Recall: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where θ is an unknown population-level parameter. In STAT 512 (Chapter 9), we posed the question,

"What is the best possible (point) estimator for θ?"

We called such an estimator $\widehat{\theta}$ the (uniformly) minimum variance unbiased estimator (MVUE). That is, among all unbiased estimators for θ, the MVUE $\widehat{\theta}$ is the one with the smallest possible variance. I sometimes add the adverb "uniformly" to reinforce the notion that $\widehat{\theta}$ is best regardless of what the true value of θ is. We learned the critical role that **sufficient statistics** play in answering this question formally.

Preview: We are now ready to embark on the same journey but for hypothesis testing instead. That is, we would like to answer the question,

"What is the best possible hypothesis test for θ?"

Just as we did in the point estimation problem, we will first have to define what we mean by "best." Then, we will need the appropriate theory to help us answer the question. Not surprisingly, sufficient statistics will again resurface when deriving best tests.

Strategy: We will attack this question in two stages. In the first stage, we will consider "simple-versus-simple tests;" i.e., tests of the form

$$\begin{array}{c} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta = \theta_a, \end{array}$$

where θ_0 and θ_a are both fixed values of θ. Recall that a hypothesis is called **simple** (or **sharp**) when it identifies exactly one probability distribution. In the second stage, we will then generalize our simple-versus-simple approach to find best rejection regions for one-sided tests like

$$\begin{array}{ccc} H_0 : \theta = \theta_0 & & H_0 : \theta = \theta_0 \\ \text{versus} & \text{and} & \text{versus} \\ H_a : \theta > \theta_0 & & H_a : \theta < \theta_0. \end{array}$$

We will learn that deriving best rejection regions for two-sided tests with $H_a : \theta \neq \theta_0$ is generally not possible (unless we alter our definition of "best").

Recall: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $p_Y(y|\theta)$ or $f_Y(y|\theta)$. 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,

$$L(\theta|\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}|\theta) = f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) = \prod_{i=1}^{n} f_Y(y_i|\theta).$$

The likelihood function plays an important role in deriving best tests, as we now see.

Neyman-Pearson Lemma: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where θ is an unknown population-level parameter, and consider testing

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta = \theta_a. \end{gathered}$$

The level α test that maximizes the power when H_a is true uses the rejection region

$$\text{RR} = \left\{ \frac{L(\theta_0|\mathbf{y})}{L(\theta_a|\mathbf{y})} < k \right\},$$

where k satisfies

$$\alpha = P_{H_0}(\text{RR}) = P(\text{Reject } H_0|H_0 \text{ true}).$$

The test which uses the rejection region above is called the **most powerful level α test**. That is, among all level α tests for H_0 versus H_a, the most powerful test maximizes the probability of rejecting H_0 when H_a is true. This is what we mean by "best."

Example 10.12. Suppose Y is a single observation from a beta$(1, \theta)$ population distribution, where $\theta > 0$. Derive the most powerful level $\alpha = 0.05$ test for

$$\begin{gathered} H_0 : \theta = 2 \\ \text{versus} \\ H_a : \theta = 4. \end{gathered}$$

Solution. Recall the beta$(1, \theta)$ pdf is given by

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

Therefore, because both H_0 and H_a are simple hypotheses, we are really testing

$$H_0 : Y \sim f_Y(y|2) = \begin{cases} 2(1-y), & 0 < y < 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{versus}$$

$$H_a : Y \sim f_Y(y|4) = \begin{cases} 4(1-y)^3, & 0 < y < 1 \\ 0, & \text{otherwise;} \end{cases}$$

see Figure 10.21 (next page). In this problem, there is only one observation; i.e., $n = 1$. Therefore, the likelihood function

$$L(\theta|y) = f_Y(y|\theta) = \theta(1-y)^{\theta-1}.$$

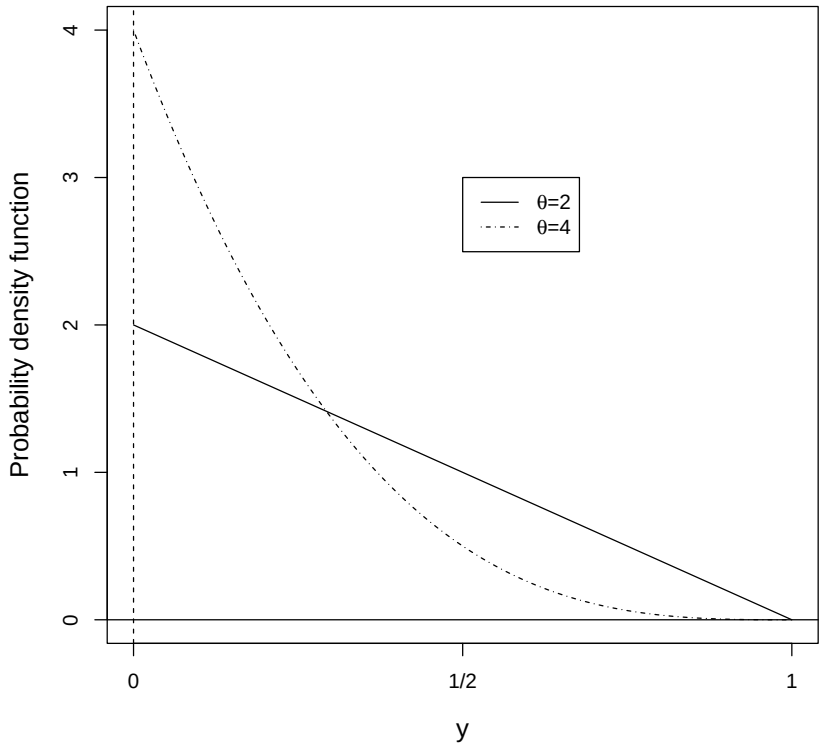

Figure 10.21: Example 10.12. Population pdfs for $H_0 : \theta = 2$ and $H_a : \theta = 4$.

To find the most powerful test/rejection region, we first calculate the ratio

$$\frac{L(\theta_0|y)}{L(\theta_a|y)} = \frac{L(2|y)}{L(4|y)} = \frac{2(1-y)}{4(1-y)^3} = \frac{1}{2(1-y)^2}.$$

The Neyman-Pearson Lemma says the most powerful level $\alpha = 0.05$ test uses

$$\text{RR} = \left\{ \frac{L(2|y)}{L(4|y)} < k \right\} = \left\{ \frac{1}{2(1-y)^2} < k \right\},$$

where k satisfies

$$0.05 = P_{H_0}(\text{RR}) = P_{H_0}\left(\frac{1}{2(1-Y)^2} < k \right).$$

On first glance, the last equation might suggest we need to find the distribution of

$$U = h(Y) = \frac{1}{2(1-Y)^2}, \qquad \text{when } Y \sim \text{beta}(1,2),$$

and then choose k to be the 0.05 quantile from this distribution. Although you could do

this, it is ultimately unnecessary. Note that the event

$$\begin{aligned}\left\{\frac{1}{2(1-Y)^2} < k\right\} &= \left\{\frac{1}{(1-Y)^2} < 2k\right\} \\ &= \left\{(1-Y)^2 > \frac{1}{2k}\right\} \\ &= \left\{1-Y > \sqrt{\frac{1}{2k}}\right\} = \left\{Y < 1-\sqrt{\frac{1}{2k}}\right\} = \{Y < k^*\},\end{aligned}$$

where $k^* = 1 - \sqrt{1/2k}$. Because all of the events above are the same event, choosing k to satisfy

$$0.05 = P_{H_0}\left(\frac{1}{2(1-Y)^2} < k\right)$$

is the same as choosing k^* to satisfy

$$0.05 = P_{H_0}(Y < k^*) \implies k^* \approx 0.0253;$$

i.e., k^* is the 0.05 quantile of a beta$(1, 2)$ distribution. Therefore, the Neyman-Pearson Lemma says

$$\text{RR} = \{y < 0.0253\}$$

is the most powerful level $\alpha = 0.05$ rejection region.

```
qbeta(0.05,1,2)
[1] 0.02532057
```

Discussion: In Example 10.12, among all level $\alpha = 0.05$ tests which exist, the one which uses the rejection region above maximizes the power when H_a is true; i.e., when $\theta = 4$. In fact, we can easily calculate what the power is; note that

$$K(4) = P_{H_a}(\text{RR}) = P_{H_a}(Y < 0.0253) \approx 0.0974.$$

Therefore, although the rejection region above is the most powerful, it is not all that powerful. This should not be surprising; after all, we are making a decision on the basis of a single observation Y and the pdfs under H_0 and H_a are not that different to begin with. □

```
> pbeta(0.0253,1,4)
[1] 0.09742383
```

Exercise: Redo Example 10.12 to find the most powerful level $\alpha = 0.05$ test of

$$\begin{gathered} H_0 : \theta = 2 \\ \text{versus} \\ H_a : \theta = 5. \end{gathered}$$

Does the rejection region change? What about if $H_a : \theta = 6$? $H_a : \theta = 100$?

Example 10.13. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(θ) population, where $\theta > 0$ is unknown. Derive the most powerful level α test for

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta = \theta_a, \end{gathered}$$

where $\theta_a < \theta_0$.

Solution. Recall the exponential(θ) pdf is given by

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

The likelihood function is given by

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i|\theta) = \frac{1}{\theta} e^{-y_1/\theta} \times \frac{1}{\theta} e^{-y_2/\theta} \times \cdots \times \frac{1}{\theta} e^{-y_n/\theta} = \frac{1}{\theta^n} e^{-\sum_{i=1}^n y_i/\theta}.$$

To find the most powerful test/rejection region, we first calculate the ratio

$$\frac{L(\theta_0|\mathbf{y})}{L(\theta_a|\mathbf{y})} = \frac{\dfrac{1}{\theta_0^n} e^{-\sum_{i=1}^n y_i/\theta_0}}{\dfrac{1}{\theta_a^n} e^{-\sum_{i=1}^n y_i/\theta_a}} = \left(\frac{\theta_a}{\theta_0}\right)^n e^{-\sum_{i=1}^n y_i \left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)} = \left(\frac{\theta_a}{\theta_0}\right)^n e^{-t\left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)},$$

where the sufficient statistic $t = \sum_{i=1}^n y_i$. The Neyman-Pearson Lemma says the most powerful level α test uses

$$\text{RR} = \left\{ \left(\frac{\theta_a}{\theta_0}\right)^n e^{-t\left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)} < k \right\},$$

where k satisfies

$$\alpha = P_{H_0}(\text{RR}) = P_{H_0}\left(\left(\frac{\theta_a}{\theta_0}\right)^n e^{-T\left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)} < k \right).$$

Working with the last equation looks terrifying. However, note that the event

$$\begin{aligned} \left\{ \left(\frac{\theta_a}{\theta_0}\right)^n e^{-T\left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)} < k \right\} &= \left\{ e^{-T\left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)} < k \left(\frac{\theta_0}{\theta_a}\right)^n \right\} \\ &= \left\{ -T\left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right) < \ln\left(k\left(\frac{\theta_0}{\theta_a}\right)^n\right) \right\} \\ &= \left\{ T < -\frac{\ln\left(k\left(\dfrac{\theta_0}{\theta_a}\right)^n\right)}{\dfrac{1}{\theta_0} - \dfrac{1}{\theta_a}} \right\} = \{T < k^*\}, \end{aligned}$$

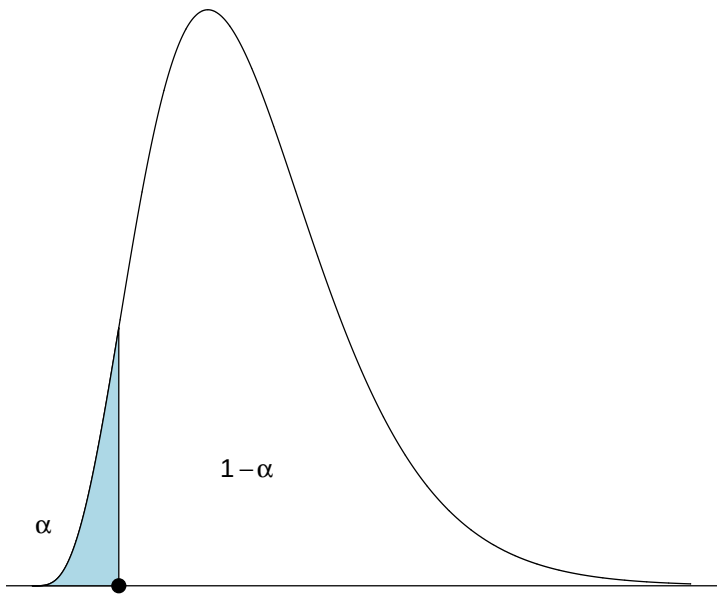

Figure 10.22: Gamma(n, θ_0) pdf. The lower α quantile $g_{n,\theta_0,1-\alpha}$ is shown using a dark circle.

where the constant

$$k^* = -\frac{\ln\left(k\left(\frac{\theta_0}{\theta_a}\right)^n\right)}{\frac{1}{\theta_0} - \frac{1}{\theta_a}}.$$

Therefore, choosing k to satisfy

$$\alpha = P_{H_0}\left(\left(\frac{\theta_a}{\theta_0}\right)^n e^{-T\left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)} < k\right)$$

is the same as choosing k^* to satisfy

$$\alpha = P_{H_0}(T < k^*).$$

This is easy! When $H_0 : \theta = \theta_0$ is true, we know

$$T = \sum_{i=1}^{n} Y_i \overset{H_0}{\sim} \text{gamma}(n, \theta_0).$$

Therefore, $k^* = g_{n,\theta_0,1-\alpha}$, the lower α quantile of a gamma distribution with shape n and scale θ_0; see Figure 10.22 (above). The Neyman-Pearson Lemma says

$$\text{RR} = \left\{ t = \sum_{i=1}^{n} y_i < g_{n,\theta_0,1-\alpha} \right\}$$

is the most powerful level α rejection region.

Discussion: Two remarks are in order.

- In Example 10.13, we see the most powerful level α test depends on the observations $Y_1, Y_2, ..., Y_n$ through a sufficient statistic $T = \sum_{i=1}^n Y_i$. This is not a coincidence. In fact, it is easy to show that if a most powerful test exists, the rejection region must depend on a sufficient statistic.

- The most powerful level α rejection region in Example 10.13,

$$\text{RR} = \left\{ t = \sum_{i=1}^n y_i < g_{n,\theta_0,1-\alpha} \right\},$$

  does not depend on the value of θ under H_a (recall $H_a : \theta = \theta_a$). This suggests that the most powerful rejection region above will be the same regardless of what θ_a is. This observation will be important later when we discuss *uniformly* most powerful tests of $H_0 : \theta = \theta_0$ versus $H_a : \theta < \theta_0$.

Application: In STAT 512 (Example 6.19, pp 35), we used the exponential(θ) distribution to model the time to treatment failure (TTF) for $n = 14$ Japanese patients with non-small cell lung cancer who were treated with two cancer drugs. Here were the times (TTF, in months):

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 most powerful level $\alpha = 0.05$ test for

$$\begin{gathered} H_0 : \theta = 24 \\ \text{versus} \\ H_a : \theta = 12 \end{gathered}$$

uses the rejection region

$$\text{RR} = \left\{ t = \sum_{i=1}^{14} y_i < 203.1 \right\}.$$

The value of the (sufficient) test statistic

$$t = \sum_{i=1}^{14} y_i = 178.7.$$

Therefore, we would reject $H_0 : \theta = 24$ in favor of $H_a : \theta = 12$ when using this most powerful level $\alpha = 0.05$ decision rule. $\square$

```
> qgamma(0.05,14,1/24)
[1] 203.1345
> 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)
> sum(ttf)
[1] 178.7
```

Result: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where θ is an unknown population-level parameter, and let $T = T(Y_1, Y_2, ..., Y_n)$ be a sufficient statistic. The rejection region for the most powerful level α test of

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta = \theta_a \end{gathered}$$

must depend on $Y_1, Y_2, ..., Y_n$ through a sufficient statistic T.

Proof. Because T is sufficient, we know we can write the likelihood function

$$L(\theta|\mathbf{y}) = L(\theta|y_1, y_2, ..., y_n) = g(t, \theta)h(y_1, y_2, ..., y_n),$$

by the Factorization Theorem. Therefore,

$$\frac{L(\theta_0|\mathbf{y})}{L(\theta_a|\mathbf{y})} = \frac{g(t, \theta_0)h(y_1, y_2, ..., y_n)}{g(t, \theta_a)h(y_1, y_2, ..., y_n)} = \frac{g(t, \theta_0)}{g(t, \theta_a)}.$$

The most powerful level α rejection region is

$$\text{RR} = \left\{ \frac{L(\theta_0|\mathbf{y})}{L(\theta_a|\mathbf{y})} < k \right\} = \left\{ \frac{g(t, \theta_0)}{g(t, \theta_a)} < k \right\},$$

which clearly depends on the value of the sufficient statistic $T = t$. $\square$

Implication: In our quest to determine best tests, we know we can immediately restrict attention to those tests whose rejection regions depend on sufficient statistics. In other words, if a level α test's rejection region does *not* involve a sufficient statistic, we know it cannot be most powerful.

Summary: The Neyman-Pearson Lemma shows us how to derive the most powerful level α test for problems which involve two simple hypotheses; i.e.,

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta = \theta_a. \end{gathered}$$

We now move on to the more practical situation where composite alternative hypotheses are allowed; i.e., tests of the form

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta > \theta_0 \end{gathered} \qquad \text{or} \qquad \begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta < \theta_0. \end{gathered}$$

Our goal is the same, namely, we would like to determine the most powerful test when H_a is composite. However, now we would like our test/rejection region to be "most powerful for *all* values of θ which satisfy H_a." We call these *uniformly* most powerful tests.

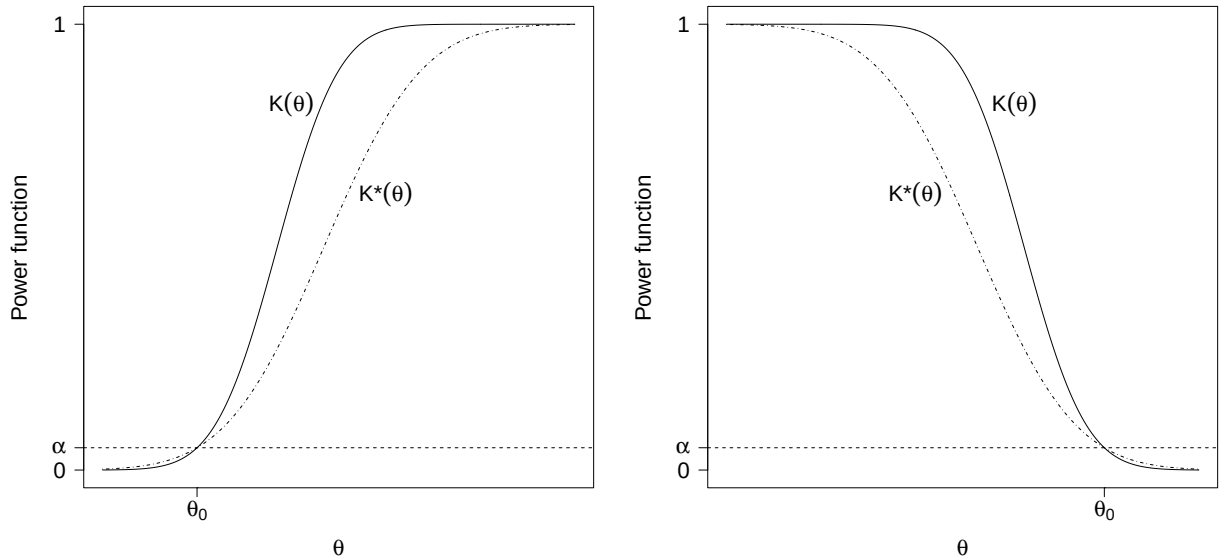

Figure 10.23: Left: UMP level α power function $K(\theta)$ for $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$. Right: UMP level α power function $K(\theta)$ for $H_0 : \theta = \theta_0$ versus $H_a : \theta < \theta_0$. In both figures $K^*(\theta)$ is the power function of another level α test.

Terminology: Suppose we are interested in testing

$$\begin{array}{ccc} H_0 : \theta = \theta_0 & & H_0 : \theta = \theta_0 \\ \text{versus} & \text{or} & \text{versus} \\ H_a : \theta > \theta_0 & & H_a : \theta < \theta_0. \end{array}$$

The **uniformly most powerful (UMP) level α test** has a power function $K(\theta)$ that satisfies

$$K(\theta) \geq K^*(\theta), \quad \text{for all } \theta \in H_a,$$

where $K^*(\theta)$ is the power function of any other level α test; see Figure 10.23 (above).

Remark: When we used the Neyman-Pearson Lemma to find the most powerful level α test for

$$\begin{array}{c} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta = \theta_a, \end{array}$$

we were actually finding the *uniformly* most powerful level α test. It's just that in this situation, there is only one value of θ allowed in H_a (i.e., H_a is a simple hypothesis). Therefore, saying "uniformly" with a simple H_a is not necessary. However, when the alternative is composite, the phrase "uniformly" is needed because we need to guarantee $K(\theta) \geq K^*(\theta)$ for *all* values of θ which satisfy H_a.

Q: How do we find UMP level α tests for composite alternatives?
A: Fortunately, we have already done most of the work. Suppose we would like to derive the UMP level α test for

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta > \theta_0. \end{gathered}$$

We first "pretend" as if we are performing the simple-versus-simple test

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a' : \theta = \theta_a, \end{gathered}$$

where θ_a is an arbitrary value which satisfies $\theta_a > \theta_0$. We then derive the most powerful level α test for H_0 versus H_a' as we have done previously (i.e., by using the Neyman-Pearson Lemma). If the rejection region for this test does not depend on θ_a, then this same rejection region must be the UMP level α rejection region for the test of $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$.

Note: This approach would be analogous to derive the UMP level α test for

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta < \theta_0. \end{gathered}$$

In this case, the simple-versus-simple test would use $H_a' : \theta = \theta_a$, where $\theta_a < \theta_0$.

Example 10.14. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(p) population, where $0 < p < 1$ is unknown. Derive the UMP level α test for

$$\begin{gathered} H_0 : p = p_0 \\ \text{versus} \\ H_a : p > p_0. \end{gathered}$$

Solution. We first use the Neyman-Pearson Lemma to find the most powerful level α test for

$$\begin{gathered} H_0 : p = p_0 \\ \text{versus} \\ H_a' : p = p_a, \end{gathered}$$

where $p_a > p_0$. Recall the Bernoulli(p) pmf is

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

The likelihood function is given by

$$\begin{aligned} L(p|\mathbf{y}) = \prod_{i=1}^n p_Y(y_i|p) &= p^{y_1}(1-p)^{1-y_1} \times p^{y_2}(1-p)^{1-y_2} \times \cdots \times p^{y_n}(1-p)^{1-y_n} \\ &= p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i}. \end{aligned}$$

To find the most powerful test/rejection region (for H_0 versus H_a'), we first calculate the ratio

$$\frac{L(p_0|\mathbf{y})}{L(p_a|\mathbf{y})} = \frac{p_0^{\sum_{i=1}^n y_i}(1-p_0)^{n-\sum_{i=1}^n y_i}}{p_a^{\sum_{i=1}^n y_i}(1-p_a)^{n-\sum_{i=1}^n y_i}} = \left[\frac{p_0(1-p_a)}{p_a(1-p_0)}\right]^{\sum_{i=1}^n y_i}\left(\frac{1-p_0}{1-p_a}\right)^n$$
$$= \left[\frac{p_0(1-p_a)}{p_a(1-p_0)}\right]^{t}\left(\frac{1-p_0}{1-p_a}\right)^n,$$

where the sufficient statistic $t = \sum_{i=1}^n y_i$. The Neyman-Pearson Lemma says the most powerful level α test for H_0 versus H_a' uses

$$\text{RR} = \left\{\left[\frac{p_0(1-p_a)}{p_a(1-p_0)}\right]^{t}\left(\frac{1-p_0}{1-p_a}\right)^n < k\right\},$$

where k satisfies

$$\alpha = P_{H_0}(\text{RR}) = P_{H_0}\left(\left[\frac{p_0(1-p_a)}{p_a(1-p_0)}\right]^{T}\left(\frac{1-p_0}{1-p_a}\right)^n < k\right).$$

Let's simplify the event above using algebra; note that

$$\begin{aligned}
\left\{\left[\frac{p_0(1-p_a)}{p_a(1-p_0)}\right]^{T}\left(\frac{1-p_0}{1-p_a}\right)^n < k\right\} &= \left\{\left[\frac{p_0(1-p_a)}{p_a(1-p_0)}\right]^{T} < k\left(\frac{1-p_a}{1-p_0}\right)^n\right\} \\
&= \left\{T\ln\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right) < \ln\left(k\left(\frac{1-p_a}{1-p_0}\right)^n\right)\right\} \\
&= \left\{T > \frac{\ln\left(k\left(\frac{1-p_a}{1-p_0}\right)^n\right)}{\ln\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right)}\right\} = \{T > k^*\},
\end{aligned}$$

where the constant

$$k^* = \frac{\ln\left(k\left(\frac{1-p_a}{1-p_0}\right)^n\right)}{\ln\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right)}.$$

Therefore, choosing k to satisfy

$$\alpha = P_{H_0}\left(\left[\frac{p_0(1-p_a)}{p_a(1-p_0)}\right]^{T}\left(\frac{1-p_0}{1-p_a}\right)^n < k\right)$$

is the same as choosing k^* to satisfy

$$\alpha = P_{H_0}(T > k^*).$$

This is easy! When $H_0 : p = p_0$ is true, we know

$$T = \sum_{i=1}^{n} Y_i \overset{H_0}{\sim} b(n, p_0).$$

Therefore, we can choose k^* to be the (upper) quantile of the $b(n, p_0)$ distribution which provides a level α test. To summarize, we have shown the most powerful level α test for

$$\begin{gathered} H_0 : p = p_0 \\ \text{versus} \\ H_a' : p = p_a \end{gathered}$$

uses the rejection region

$$\text{RR} = \left\{ t = \sum_{i=1}^{n} y_i > k^* \right\},$$

where k^* is a quantile from the $b(n, p_0)$ distribution. Now, we simply note that this rejection region does not depend on the value of p_a under H_a' (which we specified arbitrarily). Therefore, this same rejection region must be most powerful level α for all $p_a > p_0$; i.e., *uniformly* most powerful (UMP) level α for

$$\begin{gathered} H_0 : p = p_0 \\ \text{versus} \\ H_a : p > p_0. \end{gathered}$$

Application: On August 17, 2021, Dr. Pastides reinstated the mask mandate for all faculty, staff, and students at the University of South Carolina. Let p denote the population proportion of UofSC students who support this decision. A random sample of $n = 100$ students is obtained and each student is asked if s/he supports the mandate. The goal is to test

$$\begin{gathered} H_0 : p = 0.5 \\ \text{versus} \\ H_a : p > 0.5. \end{gathered}$$

The UMP level $\alpha \approx 0.0443$ test uses

$$\text{RR} = \left\{ t = \sum_{i=1}^{100} y_i \geq 59 \right\},$$

that is, among all $\alpha \approx 0.0443$ tests, this one provides the largest power for all $p > 0.5$. Because $T \sim b(100, p)$, the corresponding power function is

$$K(p) = P_p(\text{RR}) = P_p(T \geq 59) = 1 - P_p\left(T \leq 58\right) = 1 - \sum_{t=0}^{58} \binom{100}{t} p^t (1-p)^{n-t}.$$

The probability $P_p\left(T \leq 58\right)$ can be calculated in R using the `pbinom` function. Figure 10.24 (next page) shows the graph of $K(p)$. Note that $K(0.5) = \alpha \approx 0.0443$. $\square$

```
1-pbinom(58,100,0.5)
[1] 0.04431304
```

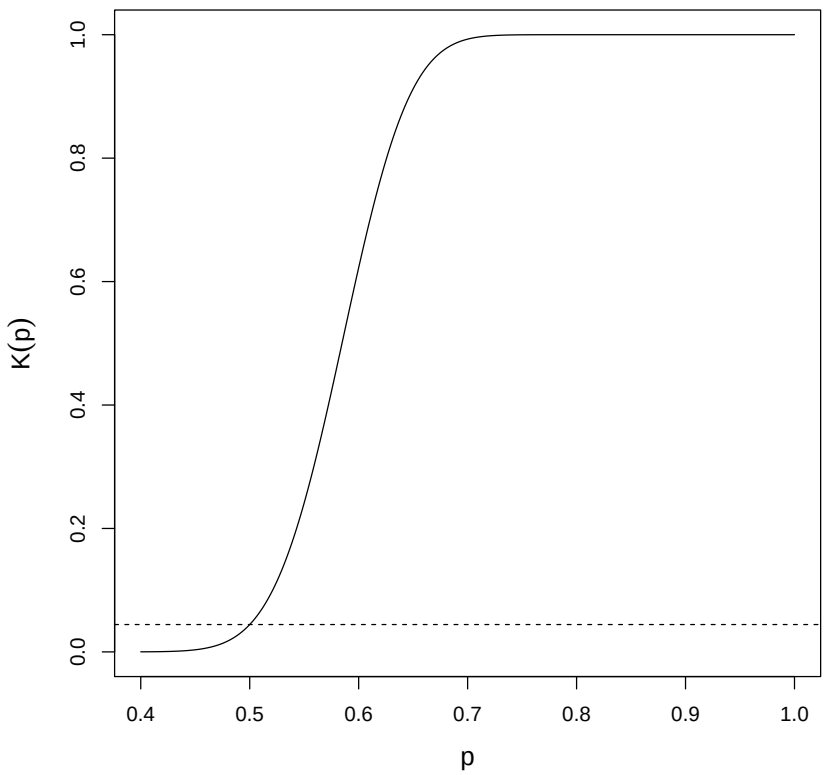

Figure 10.24: Example 10.14. Power function for $H_0 : p = 0.5$ versus $H_a : p > 0.5$. A horizontal line at $\alpha \approx 0.0443$ has been added.

Example 10.15. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(0, \sigma^2)$ population, where $\sigma^2 > 0$ is unknown. Derive the UMP level α test for

$$\begin{gathered} H_0 : \sigma^2 = \sigma_0^2 \\ \text{versus} \\ H_a : \sigma^2 < \sigma_0^2. \end{gathered}$$

Solution. We first use the Neyman-Pearson Lemma to find the most powerful level α test for

$$\begin{gathered} H_0 : \sigma^2 = \sigma_0^2 \\ \text{versus} \\ H_a' : \sigma^2 = \sigma_a^2, \end{gathered}$$

where $\sigma_a^2 < \sigma_0^2$. Recall the $\mathcal{N}(0, \sigma^2)$ pdf is

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

The likelihood function is given by

$$L(\sigma^2|\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-y_1^2/2\sigma^2} \times \frac{1}{\sqrt{2\pi\sigma^2}}e^{-y_2^2/2\sigma^2} \times \cdots \times \frac{1}{\sqrt{2\pi\sigma^2}}e^{-y_n^2/2\sigma^2}$$
$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\sum_{i=1}^n y_i^2/2\sigma^2}.$$

To find the most powerful test/rejection region (for H_0 versus H_a'), we first calculate the ratio

$$\frac{L(\sigma_0^2|\mathbf{y})}{L(\sigma_a^2|\mathbf{y})} = \frac{\left(\frac{1}{2\pi\sigma_0^2}\right)^{n/2} e^{-\sum_{i=1}^n y_i^2/2\sigma_0^2}}{\left(\frac{1}{2\pi\sigma_a^2}\right)^{n/2} e^{-\sum_{i=1}^n y_i^2/2\sigma_a^2}} = \left(\frac{\sigma_a^2}{\sigma_0^2}\right)^{n/2} e^{-\sum_{i=1}^n y_i^2 \Big/ \left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}\right)} = \left(\frac{\sigma_a^2}{\sigma_0^2}\right)^{n/2} e^{-t \Big/ \left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}\right)},$$

where the sufficient statistic $t = \sum_{i=1}^n y_i^2$. The Neyman-Pearson Lemma says the most powerful level α test for H_0 versus H_a' uses

$$\text{RR} = \left\{ \left(\frac{\sigma_a^2}{\sigma_0^2}\right)^{n/2} e^{-T \Big/ \left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}\right)} < k \right\},$$

where k satisfies

$$\alpha = P_{H_0}(\text{RR}) = P_{H_0}\left(\left(\frac{\sigma_a^2}{\sigma_0^2}\right)^{n/2} e^{-T \Big/ \left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}\right)} < k \right).$$

Let's simplify the event above using algebra; note that

$$\begin{aligned}
\left\{ \left(\frac{\sigma_a^2}{\sigma_0^2}\right)^{n/2} e^{-T \Big/ \left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}\right)} < k \right\} &= \left\{ e^{-T \Big/ \left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}\right)} < k\left(\frac{\sigma_0^2}{\sigma_a^2}\right)^{n/2} \right\} \\
&= \left\{ -\frac{T}{\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}} < \ln\left(k\left(\frac{\sigma_0^2}{\sigma_a^2}\right)^{n/2} \right) \right\} \\
&= \left\{ T < -\frac{\ln\left(k\left(\frac{\sigma_0^2}{\sigma_a^2}\right)^{n/2} \right)}{\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}} \right\} = \{T < k^*\},
\end{aligned}$$

where the constant

$$k^* = -\frac{\ln\left(k\left(\frac{\sigma_0^2}{\sigma_a^2}\right)^{n/2} \right)}{\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}}.$$

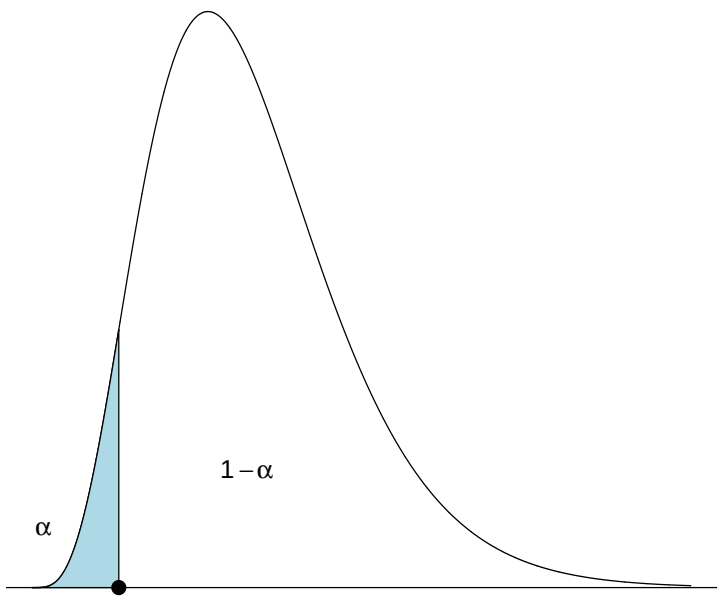

Figure 10.25: $\chi^2(n)$ pdf. The lower α quantile $\chi^2_{n,1-\alpha}$ is shown by using a dark circle.

Therefore, choosing k to satisfy

$$\alpha = P_{H_0}\left(\left(\frac{\sigma_a^2}{\sigma_0^2}\right)^{n/2} e^{-T\Big/\left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_a^2}\right)} < k\right)$$

is the same as choosing k^* to satisfy

$$\alpha = P_{H_0}(T < k^*) = P_{H_0}\left(\frac{T}{\sigma_0^2} < \frac{k^*}{\sigma_0^2}\right) \implies \frac{k^*}{\sigma_0^2} = \chi^2_{n,1-\alpha} \implies k^* = \sigma_0^2\chi^2_{n,1-\alpha},$$

where $\chi^2_{n,1-\alpha}$ is the lower α quantile of the $\chi^2(n)$ distribution; see Figure 10.25 (above).

Aside: From STAT 512, we remember that

$$\frac{T}{\sigma^2} \sim \chi^2(n) \implies \frac{T}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(n).$$

Therefore, we have shown the most powerful level α test for

$$\begin{gathered} H_0 : \sigma^2 = \sigma_0^2 \\ \text{versus} \\ H_a' : \sigma^2 = \sigma_a^2 \end{gathered}$$

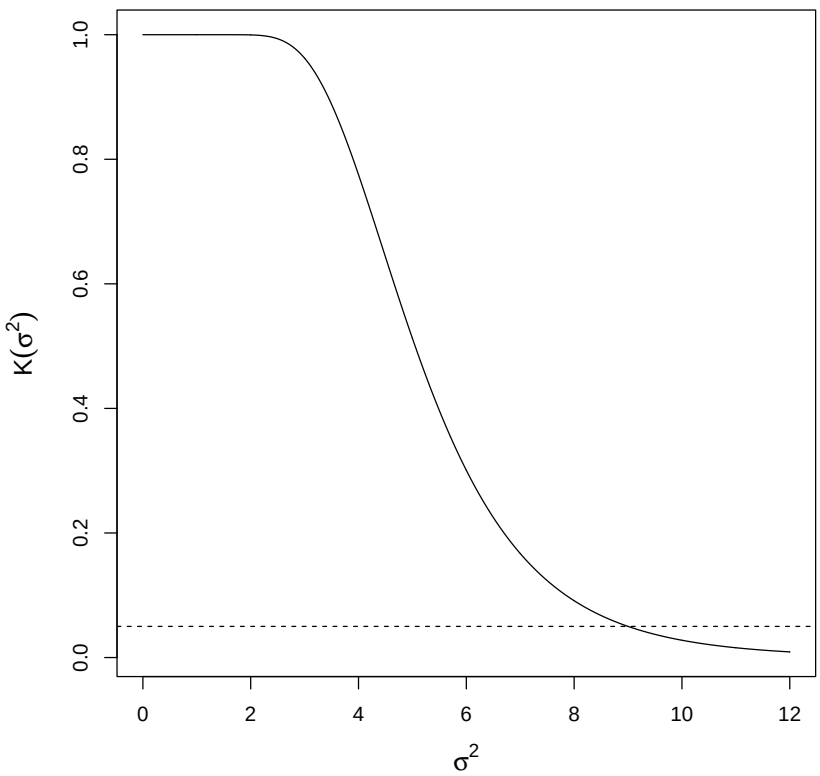

Figure 10.26: Example 10.15. Power function for $\sigma^2 = 9$ versus $H_a : \sigma^2 < 9$. A horizontal line at $\alpha = 0.05$ has been added.

uses the rejection region

$$\text{RR} = \left\{ t = \sum_{i=1}^{n} y_i^2 < \sigma_0^2 \chi^2_{n,1-\alpha} \right\}.$$

Now, we simply note that this rejection region does not depend on the value of σ_a^2 under H_a' (which we specified arbitrarily). Therefore, this same rejection region must be most powerful level α for all $\sigma_a^2 < \sigma_0^2$; i.e., *uniformly* most powerful (UMP) level α for

$$\begin{gathered} H_0 : \sigma^2 = \sigma_0^2 \\ \text{versus} \\ H_a : \sigma^2 < \sigma_0^2. \end{gathered}$$

The corresponding power function is

$$K(\sigma^2) = P_{\sigma^2}(\text{RR}) = P_{\sigma^2}(T < \sigma_0^2 \chi^2_{n,1-\alpha}) = P\left(\frac{T}{\sigma^2} < \frac{\sigma_0^2 \chi^2_{n,1-\alpha}}{\sigma^2}\right) = F_{\chi^2(n)}\left(\frac{\sigma_0^2 \chi^2_{n,1-\alpha}}{\sigma^2}\right),$$

where $F_{\chi^2(n)}$ is the $\chi^2(n)$ cdf. This cdf can be calculated in R using the `pchisq` function. Figure 10.26 (above) shows the graph of $K(\sigma^2)$ when $n = 20$, $\sigma_0^2 = 9$, and $\alpha = 0.05$. $\square$

10.8 Likelihood ratio tests

Remark: We have just learned that UMP tests are "optimal" in the sense that they maximize power when the alternative hypothesis is true. Unfortunately, UMP tests rarely exist once you get outside the simple testing problems considered in this class. This does not void their value. In fact, I would argue that in any statistical analysis, it is important to think about what the "optimal" approach is, and UMP tests help to serve that purpose.

Preview: Given the limited practical utility of UMP tests, it is important to discuss more general methods. Arguably the most common method is the likelihood ratio test (LRT). As its name suggests, the LRT is based on the ratio of two likelihood functions, one corresponding to H_0 and the other which does not put any restrictions on the population-level parameter θ. We now state new terminology needed to describe the LRT approach.

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where θ is an unknown population-level parameter. The set of all allowable values of θ is called the **parameter space**, denoted by Θ. For example,

- $Y \sim \text{Bernoulli}(\theta) \Longrightarrow \Theta = \{0 < \theta < 1\} = (0, 1)$
- $Y \sim \text{exponential}(\theta) \Longrightarrow \Theta = \{0 < \theta < \infty\} = \mathbb{R}^+$
- $Y \sim \mathcal{N}(\theta, 1) \Longrightarrow \Theta = \{-\infty < \theta < \infty\} = \mathbb{R}$
- $Y \sim \text{gamma}(\alpha, \beta) \Longrightarrow \Theta = \{\boldsymbol{\theta} = (\alpha, \beta) : \alpha > 0,\ \beta > 0\} = \mathbb{R}^+ \times \mathbb{R}^+$
- $Y \sim \mathcal{N}(\mu, \sigma^2) \Longrightarrow \Theta = \{\boldsymbol{\theta} = (\mu, \sigma^2) : -\infty < \mu < \infty,\ \sigma^2 > 0\} = \mathbb{R} \times \mathbb{R}^+$.

Terminology: For any population-level model, suppose we write the parameter space Θ as

$$\Theta = \Theta_0 \cup \Theta_a,$$

the union of two mutually exclusive sets. In our LRT formulation, we will call Θ_0 the **null parameter space**; i.e., it is the set of allowable values of θ under H_0. We will call Θ_a the **alternative parameter space**; i.e., it is the set of allowable values of θ under H_a.

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $p_Y(y|\theta)$ or $f_Y(y|\theta)$, where θ is an unknown population-level parameter (possibly vector-valued). Let $L(\theta|\mathbf{y})$ denote the likelihood function of θ. A level α **likelihood ratio test (LRT)** for

$$\begin{gathered} H_0 : \theta \in \Theta_0 \\ \text{versus} \\ H_a : \theta \in \Theta_a \end{gathered}$$

uses the test statistic

$$\lambda = \frac{\max\limits_{\theta \in \Theta_0} L(\theta|\mathbf{y})}{\max\limits_{\theta \in \Theta} L(\theta|\mathbf{y})}$$

and the rejection region

$$\text{RR} = \{\lambda < k\},$$

where k satisfies

$$\alpha = P_{H_0}(\text{RR}) = P(\text{Reject } H_0|H_0 \text{ true}).$$

Discussion: Several remarks are in order. The LRT statistic λ is the ratio of two maximized likelihood functions.

- In the denominator, we maximize $L(\theta|\mathbf{y})$ over the entire parameter space Θ. Another way to write this is

$$\max_{\theta\in\Theta} L(\theta|\mathbf{y}) = L(\widehat{\theta}|\mathbf{y}),$$

  where $\widehat{\theta}$ is the maximum likelihood estimator (MLE) of θ. In the LRT formulation, we call $\widehat{\theta}$ the "unrestricted MLE" because we are maximizing $L(\theta|\mathbf{y})$ over the entire parameter space (with no restriction).

- In the numerator, we maximize $L(\theta|\mathbf{y})$ over the null parameter space Θ_0, where $\Theta_0 \subset \Theta$. Therefore, it must be true that

$$0 \leq \frac{\max_{\theta\in\Theta_0} L(\theta|\mathbf{y})}{\max_{\theta\in\Theta} L(\theta|\mathbf{y})} \leq 1 \iff 0 \leq \lambda \leq 1$$

  because in the numerator we are maximizing $L(\theta|\mathbf{y})$ over a "smaller set" than we are in the denominator. Let $\widehat{\theta}_0$ denote the MLE of θ over Θ_0. Another way to write the numerator is

$$\max_{\theta\in\Theta_0} L(\theta|\mathbf{y}) = L(\widehat{\theta}_0|\mathbf{y}).$$

  We call $\widehat{\theta}_0$ the "restricted MLE" because we are maximizing $L(\theta|\mathbf{y})$ over the null parameter space; i.e., the parameter space "restricted" under H_0.

- Summarizing, the LRT statistic is

$$\lambda = \frac{L(\widehat{\theta}_0|\mathbf{y})}{L(\widehat{\theta}|\mathbf{y})},$$

  where $\widehat{\theta}_0$ and $\widehat{\theta}$ are the restricted and unrestricted MLEs, respectively. Small values of λ are evidence against H_0.

- A technical issue arises in how we define the level α when performing a LRT.
  - If $H_0 : \theta = \theta_0$ (i.e., a simple H_0), then the null parameter space is $\Theta_0 = \{\theta_0\}$, a singleton. In this situation,

$$\max_{\theta\in\Theta_0} L(\theta|\mathbf{y}) = L(\widehat{\theta}_0|\mathbf{y}) = L(\theta_0|\mathbf{y}),$$

    that is, we are maximizing $L(\theta|\mathbf{y})$ over a single point. There is only one value of θ which makes H_0 true so

$$\alpha = P(\text{Reject } H_0|H_0 \text{ true}) = P_{\theta_0}(\text{RR}) = P_{\theta_0}(\lambda < k).$$

- If $H_0 : \theta \leq \theta_0$ (i.e., a composite H_0), the null parameter space is $\Theta_0 = \{\theta \leq \theta_0\}$. Now, there are infinitely many values of θ which make H_0 true so

$$\alpha = P_{H_0}(\text{RR}) = P(\text{Reject } H_0 | H_0 \text{ true})$$

becomes ambiguous. For $H_0 : \theta \leq \theta_0$, we redefine the probability of Type I Error as

$$\alpha = \max_{\theta \leq \theta_0} P_\theta(\text{RR}) = \max_{\theta \leq \theta_0} K(\theta),$$

where $K(\theta)$ is the power function. In general, we define

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(\text{RR}) = \max_{\theta \in \Theta_0} K(\theta).$$

Example 10.16. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Rayleigh(θ) population, where $\theta > 0$ is unknown. Recall the Rayleigh pdf is given by

$$f_Y(y|\theta) = \begin{cases} \dfrac{2y}{\theta} e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Derive a level α LRT for

$$\begin{gathered} H_0 : \theta = \theta_0 \\ \text{versus} \\ H_a : \theta \neq \theta_0, \end{gathered}$$

where θ_0 is a specified value.
(b) Derive the power function $K(\theta)$ for the test.

Solution. (a) We start by finding the likelihood function, which is

$$\begin{aligned} L(\theta|\mathbf{y}) &= f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) \\ &= \frac{2y_1}{\theta} e^{-y_1^2/\theta} \times \frac{2y_2}{\theta} e^{-y_2^2/\theta} \times \cdots \times \frac{2y_n}{\theta} e^{-y_n^2/\theta} = \left(\frac{2}{\theta}\right)^n \left(\prod_{i=1}^n y_i\right) e^{-\sum_{i=1}^n y_i^2/\theta}. \end{aligned}$$

Now, we maximize $L(\theta|\mathbf{y})$ over the null parameter space and the entire parameter space separately.

- The null parameter space is $\Theta_0 = \{\theta_0\}$, a singleton. Clearly,

$$\max_{\theta \in \Theta_0} L(\theta|\mathbf{y}) = L(\theta_0|\mathbf{y}).$$

- The entire parameter space is $\Theta = \{0 < \theta < \infty\} = \mathbb{R}^+$. In STAT 512 (Example 9.18, pp 144-145), we showed the (unrestricted) MLE of θ is

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^n Y_i^2,$$

which is a function of $T = \sum_{i=1}^n Y_i^2$, a sufficient statistic.

The LRT statistic is

$$\lambda = \frac{L(\theta_0|\mathbf{y})}{L(\widehat{\theta}|\mathbf{y})} = \frac{\left(\frac{2}{\theta_0}\right)^n \left(\prod_{i=1}^n y_i\right) e^{-\sum_{i=1}^n y_i^2/\theta_0}}{\left(\frac{2}{\widehat{\theta}}\right)^n \left(\prod_{i=1}^n y_i\right) e^{-\sum_{i=1}^n y_i^2/\widehat{\theta}}} = \left(\frac{\widehat{\theta}}{\theta_0}\right)^n \frac{e^{-\sum_{i=1}^n y_i^2/\theta_0}}{e^{-\sum_{i=1}^n y_i^2/\widehat{\theta}}} = \left(\frac{\widehat{\theta}}{\theta_0}\right)^n \frac{e^{-t/\theta_0}}{e^{-t/\widehat{\theta}}},$$

where $t = \sum_{i=1}^n y_i^2$. We can use algebra to simplify this expression. Note the maximum likelihood estimate $\widehat{\theta}$ satsifies

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n y_i^2 \iff n\widehat{\theta} = \sum_{i=1}^n y_i^2 \iff n\widehat{\theta} = t.$$

Therefore,

$$\lambda = \left(\frac{\widehat{\theta}}{\theta_0}\right)^n \frac{e^{-t/\theta_0}}{e^{-t/\widehat{\theta}}} = \left(\frac{t}{n\theta_0}\right)^n \frac{e^{-t/\theta_0}}{e^{-n\widehat{\theta}/\widehat{\theta}}} = \underbrace{\left(\frac{e}{n\theta_0}\right)^n t^n e^{-t/\theta_0}}_{\text{think of this as a function of } t} = g(t), \quad \text{say.}$$

We have written the LRT statistic λ as a function of t; i.e., $\lambda = g(t)$. The LRT rejection region

$$\text{RR} = \{\lambda < k\}$$

says to reject H_0 when λ is small. For what values of t is $\lambda = g(t)$ small? Careful inspection of $g(t)$ reveals

$$g(t) \propto t^n e^{-t/\theta_0},$$

which is the kernel of a gamma density with shape parameter $n+1$ and scale parameter θ_0. This means the shape of $g(t)$ is similar to that of a gamma pdf; see Figure 10.27 (next page). An important observation from Figure 10.27 is that

$$\lambda = g(t) < k \iff t < k_1 \text{ or } t > k_2,$$

that is, $\lambda = g(t)$ is small whenever t is large or small. This means the rejection region can be written as

$$\text{RR} = \{\lambda < k\} = \{t < k_1 \text{ or } t > k_2\},$$

where $k_2 > k_1$. The LRT procedure says to choose k so that

$$\alpha = P_{H_0}(\text{RR}) = P_{\theta_0}(\lambda < k) = P_{\theta_0}\left(\left(\frac{e}{n\theta_0}\right)^n T^n e^{-T/\theta_0} < k\right).$$

However, given the set equivalence above, we can now instead choose k_1 and k_2 so that

$$\alpha = P_{H_0}(\text{RR}) = P_{\theta_0}(\{T < k_1\} \cup \{T > k_2\}) = P_{\theta_0}(T < k_1) + P_{\theta_0}(T > k_2).$$

This is easy! In Example 10.4 (pp 12-14, notes), we used the fact that

$$\frac{2T}{\theta_0} \overset{H_0}{\sim} \chi^2(2n)$$

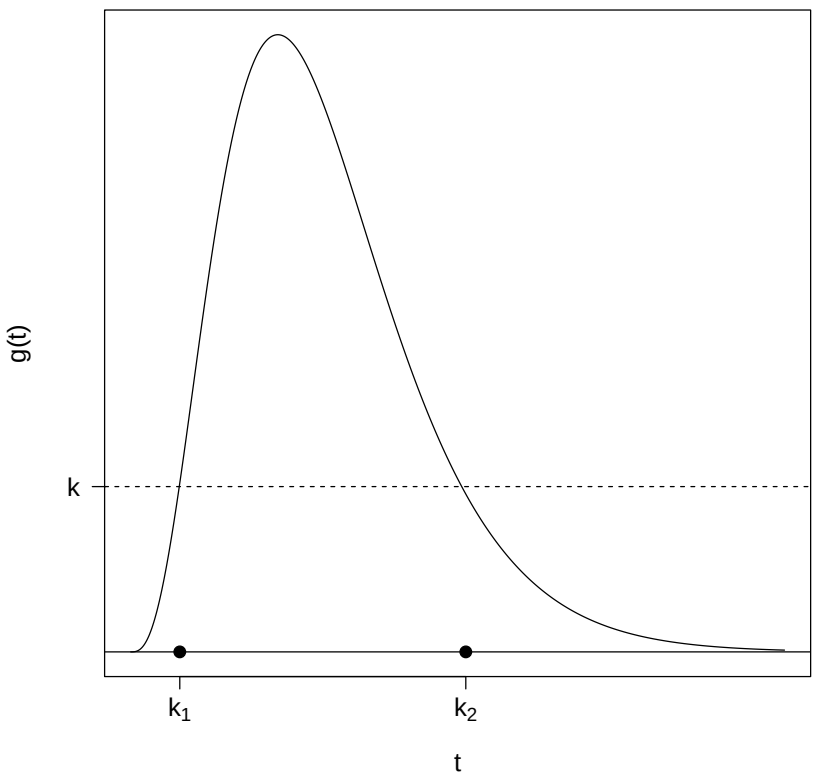

Figure 10.27: Example 10.16. Plot of $\lambda = g(t)$ versus t. Note that $\lambda = g(t) < k \iff t < k_1$ or $t > k_2$.

to calculate

$$k_1 = \frac{\theta_0 \chi^2_{2n,1-\alpha/2}}{2} \quad \text{and} \quad k_2 = \frac{\theta_0 \chi^2_{2n,\alpha/2}}{2}.$$

Therefore, a level α LRT uses the rejection region

$$\text{RR} = \left\{ t < \frac{\theta_0 \chi^2_{2n,1-\alpha/2}}{2} \ \text{ or } \ t > \frac{\theta_0 \chi^2_{2n,\alpha/2}}{2} \right\},$$

where $t = \sum_{i=1}^n y_i^2$. This was the same rejection region we used in Example 10.4.

(b) The power function is given by

$$\begin{aligned} K(\theta) = P_\theta(\text{RR}) &= P_\theta\left(T < \frac{\theta_0 \chi^2_{2n,1-\alpha/2}}{2}\right) + P_\theta\left(T > \frac{\theta_0 \chi^2_{2n,\alpha/2}}{2}\right) \\ &= P_\theta\left(\frac{2T}{\theta} < \frac{\theta_0 \chi^2_{2n,1-\alpha/2}}{\theta}\right) + P_\theta\left(\frac{2T}{\theta} > \frac{\theta_0 \chi^2_{2n,\alpha/2}}{\theta}\right) \\ &= F_{\chi^2(2n)}\left(\frac{\theta_0 \chi^2_{2n,1-\alpha/2}}{\theta}\right) + 1 - F_{\chi^2(2n)}\left(\frac{\theta_0 \chi^2_{2n,\alpha/2}}{\theta}\right), \end{aligned}$$

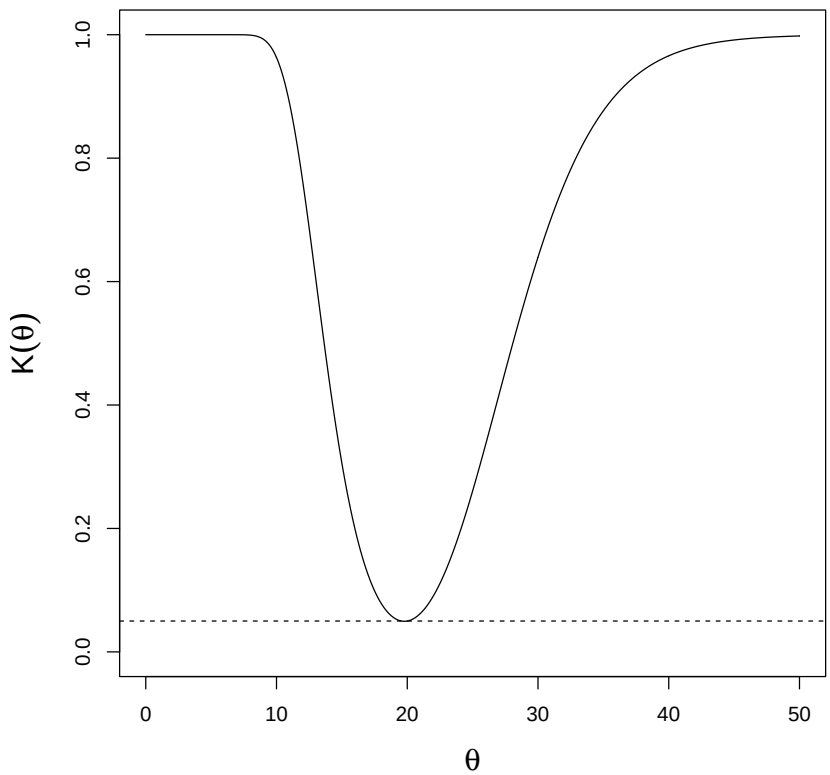

Figure 10.28: Power function $K(\theta)$ in Example 10.16 when $n = 30$, $\theta_0 = 20$, and $\alpha = 0.05$. A horizontal line at $\alpha = 0.05$ has been added.

where $F_{\chi^2(2n)}$ is the $\chi^2(2n)$ cdf. This cdf can be calculated in R using the `pchisq` function. Figure 10.28 (above) shows the graph of $K(\theta)$ when $n = 30$, $\theta_0 = 20$, and $\alpha = 0.05$. $\square$

Example 10.17. *One-sample t test.* Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population, where both parameters are unknown. In Section 10.4.1 (notes), we presented the one-sample t test for

$$\begin{gathered} H_0 : \mu = \mu_0 \\ \text{versus} \\ H_a : \mu \neq \mu_0. \end{gathered}$$

When H_0 is true, we know

$$T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t(n-1).$$

Therefore, a level α test uses the rejection region

$$\text{RR} = \{t < -t_{n-1,\alpha/2} \ \text{ or } \ t > t_{n-1,\alpha/2}\} = \{|t| > t_{n-1,\alpha/2}\}.$$

We now sketch the details to show this is a LRT.

Solution. The null hypothesis $H_0 : \mu = \mu_0$ looks like a simple hypothesis, but it is not because $\sigma^2 > 0$ is unknown (i.e., it is a **nuisance parameter**). The relevant parameter spaces are

$$\begin{aligned} \Theta_0 &= \{\boldsymbol{\theta} = (\mu, \sigma^2) : \ \mu = \mu_0, \ \sigma^2 > 0\} \\ \Theta &= \{\boldsymbol{\theta} = (\mu, \sigma^2) : -\infty < \mu < \infty, \ \sigma^2 > 0\}. \end{aligned}$$

The likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{y}) = L(\mu, \sigma^2|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-(y_i-\mu)^2/2\sigma^2} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}.$$

Here are the relevant points:

1. Over the null parameter space Θ_0, the (restricted) MLE is

$$\widehat{\boldsymbol{\theta}}_0 = \begin{pmatrix} \mu_0 \\ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_0)^2 \end{pmatrix}.$$

2. Over the entire parameter space Θ, the (unrestricted) MLE is

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \overline{Y} \\ S_b^2 \end{pmatrix} = \begin{pmatrix} \overline{Y} \\ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \end{pmatrix}.$$

   We showed this in STAT 512 (pp 148-150, notes).

3. The ratio of the two maximized likelihoods

$$\lambda = \frac{L(\widehat{\boldsymbol{\theta}}_0|\mathbf{y})}{L(\widehat{\boldsymbol{\theta}}|\mathbf{y})} = \left[\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \mu_0)^2}\right]^{n/2}.$$

4. One can show algebraically that

$$\lambda < k \iff \left|\frac{\overline{y} - \mu_0}{s/\sqrt{n}}\right| \geq k^*.$$

   Of course, we know

$$T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t(n-1),$$

   so we can choose $k^* = t_{n-1,\alpha/2}$ to ensure a level α test. This demonstrates the one-sample t test is a LRT under normality. $\square$

Remark: Deriving exact tests using the LRT method is not always possible. Even in the relatively easy problems we have discussed, it can be challenging. In this light, the following large-sample result can be useful. Of course, this is an asymptotic result, so it is only applicable when the sample size(s) is (are) large.

Large-sample result: Let $L(\theta|\mathbf{y})$ denote the likelihood function of θ (possibly vector-valued) formed after observing $Y_1, Y_2, ..., Y_n$. Consider testing

$$\begin{gathered} H_0 : \theta \in \Theta_0 \\ \text{versus} \\ H_a : \theta \in \Theta_a \end{gathered}$$

using the LRT statistic

$$\lambda = \frac{\max\limits_{\theta \in \Theta_0} L(\theta|\mathbf{y})}{\max\limits_{\theta \in \Theta} L(\theta|\mathbf{y})}.$$

Under certain "regularity conditions," it follows that

$$-2\ln\lambda \xrightarrow{d} \chi^2(\nu), \quad \text{under } H_0,$$

as $n \to \infty$, where the degrees of freedom

$$\nu = \dim(\Theta) - \dim(\Theta_0),$$

the difference between the number of free parameters specified by $\theta \in \Theta$ and the number of free parameters specified by $\theta \in \Theta_0$. The meaning of the term "free parameters" will become clear in the next example.

Remark: The difficult part about implementing an exact LRT (e.g., see Examples 10.16 and 10.17, etc.) is working out analytically what it means for the LRT statistic λ to be "small." In Example 10.16, we showed that λ was small whenever the sufficient statistic $T = \sum_{i=1}^n Y_i^2$ was large or small. In Example 10.17, we "showed" λ was small whenever the one-sample t statistic was large or small. The asymptotic result above allows us to bypass this step completely. Note that

$$\lambda < k \iff -2\ln\lambda > -2\ln k = k^*, \quad \text{say.}$$

Therefore, an **approximate** level α LRT uses the rejection region

$$\text{RR} = \{-2\ln\lambda > \chi^2_{\nu,\alpha}\},$$

where $\chi^2_{\nu,\alpha}$ is the upper α quantile of the $\chi^2(\nu)$ distribution.

Remark: The large-sample result above allows us to greatly expand the class of problems for which we can now perform LRTs (albeit large-sample versions of them). This includes problems involving multiple populations as the next example illustrates.

Example 10.18. McCann and Tebbs (2009) summarize a study examining perceived unmet need for dental health care for people with HIV infection. Baseline in-person interviews were conducted with 2,864 HIV infected individuals, aged 18 years and older, as part of the HIV Cost and Services Utilization Study. All respondents were asked,

> *"In the last six months, was there a time when you needed dental treatment but could not get it?"*

Here is the table that cross-classifies all subjects by denial of care response (yes/no) and insurance type:

	Private ins.	Medicare w/ins.	No insurance	Medicare/no ins.	Total
Denied care	49	142	181	175	547
Not denied care	609	697	630	381	2317
Total	658	839	811	556	2864

Is insurance type associated with the denial of dental health care for HIV patients? Perform a large-sample level $\alpha = 0.05$ LRT for

$$H_0 : p_1 = p_2 = p_3 = p_4$$
$$\text{versus}$$
$$H_a : H_0 \text{ not true,}$$

where p_i is the population-level probability of denial for the ith insurance group, $i = 1, 2, 3, 4$.

Solution. Let Y_i denote the number of individuals denied care in the ith insurance group and assume Y_1, Y_2, Y_3, Y_4 are mutually independent random variables satisfying

$$Y_i \sim b(n_i, p_i), \quad i = 1, 2, 3, 4,$$

where n_i is the column total (which we regard to be fixed). Set $\boldsymbol{\theta} = (p_1, p_2, p_3, p_4)$. The likelihood function of $\boldsymbol{\theta}$ is the product of the four binomial pmfs; i.e.,

$$L(\boldsymbol{\theta}) = L(p_1, p_2, p_3, p_4) = \prod_{i=1}^{4} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}.$$

The relevant parameter spaces in this problem are

$$\begin{aligned} \Theta_0 &= \{\boldsymbol{\theta} : 0 < p_1 < 1,\ 0 < p_2 < 1,\ 0 < p_3 < 1,\ 0 < p_4 < 1,\ p_1 = p_2 = p_3 = p_4\} \\ \Theta &= \{\boldsymbol{\theta} : 0 < p_1 < 1,\ 0 < p_2 < 1,\ 0 < p_3 < 1,\ 0 < p_4 < 1\}. \end{aligned}$$

Note that

- over the null parameter space Θ_0, the population-level parameters p_i are the same. Therefore, only 1 is allowed to vary freely (i.e., once we know 1, the other 3 are determined).
- over the entire parameter space Θ, all 4 parameters are allowed to vary freely.
- the difference in the number of free parameters between the two parameter spaces is $\nu = 4 - 1 = 3$.

We now maximize $L(\boldsymbol{\theta})$ over the null parameter space Θ_0 and the entire parameter space Θ separately.

MLE over Θ_0:
When H_0 is true, that is, when

$$p_1 = p_2 = p_3 = p_4 = p, \quad \text{say},$$

the likelihood function can be written as

$$L^*(p) = \prod_{i=1}^{4} \binom{n_i}{y_i} p^{y_i}(1-p)^{n_i - y_i} = \prod_{i=1}^{4} \binom{n_i}{y_i} p^{\sum_{i=1}^4 y_i}(1-p)^{\sum_{i=1}^4 (n_i - y_i)}.$$

The log-likelihood function is

$$\ln L^*(p) = \ln c + \sum_{i=1}^{4} y_i \ln p + \sum_{i=1}^{4} (n_i - y_i)\ln(1-p),$$

where the constant $c = \prod_{i=1}^4 \binom{n_i}{y_i}$ is free of p. Taking derivatives with respect to p yields

$$\frac{\partial}{\partial p} \ln L^*(p) = \frac{\sum_{i=1}^4 y_i}{p} - \frac{\sum_{i=1}^4 (n_i - y_i)}{1-p}.$$

To find the MLE over Θ_0, we set this partial derivative equal to zero and solve for p. That is,

$$\begin{aligned}
\frac{\partial}{\partial p} \ln L^*(p) \stackrel{\text{set}}{=} 0 \quad &\Longrightarrow \quad (1-p)\sum_{i=1}^{4} y_i - p\sum_{i=1}^{4}(n_i - y_i) = 0 \\
&\Longrightarrow \quad \sum_{i=1}^{4} y_i - p\sum_{i=1}^{4} y_i - p\sum_{i=1}^{4} n_i + p\sum_{i=1}^{4} y_i = 0 \quad \Longrightarrow \quad \widehat{p} = \frac{\sum_{i=1}^4 y_i}{\sum_{i=1}^4 n_i}.
\end{aligned}$$

It is straightforward to show $\partial^2/\partial p^2 \ln L^*(\widehat{p}) < 0$ so that $\widehat{p}$ maximizes $\ln L^*(p)$ by the Second Derivative Test. We have shown the maximized likelihood function over Θ_0 is

$$L(\widehat{\boldsymbol{\theta}}_0) = L(\widehat{p}, \widehat{p}, \widehat{p}, \widehat{p}) = \prod_{i=1}^{4} \binom{n_i}{y_i} \widehat{p}^{y_i}(1-\widehat{p})^{n_i - y_i},$$

where $\widehat{p} = \sum_{i=1}^4 y_i / \sum_{i=1}^4 n_i$.

MLE over Θ:
Maximizing $L(\boldsymbol{\theta}) = L(p_1, p_2, p_3, p_4)$ over Θ is a four-variable maximization problem. The likelihood function is

$$L(\boldsymbol{\theta}) = L(p_1, p_2, p_3, p_4) = \prod_{i=1}^{4} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}.$$

The log-likelihood function is

$$\ln L(\boldsymbol{\theta}) = \ln L(p_1, p_2, p_3, p_4) = \ln c + \sum_{i=1}^{4} y_i \ln p_i + \sum_{i=1}^{4} (n_i - y_i) \ln(1 - p_i).$$

The (unrestricted) MLE of $\boldsymbol{\theta} = (p_1, p_2, p_3, p_4)$ is obtained by solving

$$\begin{aligned}
\frac{\partial \ln L(p_1, p_2, p_3, p_4)}{\partial p_1} &\stackrel{\text{set}}{=} 0 \implies \frac{y_1}{p_1} - \frac{n_1 - y_1}{1 - p_1} = 0 \\
\frac{\partial \ln L(p_1, p_2, p_3, p_4)}{\partial p_2} &\stackrel{\text{set}}{=} 0 \implies \frac{y_2}{p_2} - \frac{n_2 - y_2}{1 - p_2} = 0 \\
\frac{\partial \ln L(p_1, p_2, p_3, p_4)}{\partial p_3} &\stackrel{\text{set}}{=} 0 \implies \frac{y_3}{p_3} - \frac{n_3 - y_3}{1 - p_3} = 0 \\
\frac{\partial \ln L(p_1, p_2, p_3, p_4)}{\partial p_4} &\stackrel{\text{set}}{=} 0 \implies \frac{y_4}{p_4} - \frac{n_4 - y_4}{1 - p_4} = 0
\end{aligned}$$

simultaneously. Solving this system for p_1, p_2, p_3, and p_4 gives

$$\widehat{p}_1 = \frac{y_1}{n_1}, \quad \widehat{p}_2 = \frac{y_2}{n_2}, \quad \widehat{p}_3 = \frac{y_3}{n_3}, \quad \widehat{p}_4 = \frac{y_4}{n_4},$$

the usual sample proportions. The maximized likelihood function over Θ is

$$L(\widehat{\boldsymbol{\theta}}) = L(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \widehat{p}_4) = \prod_{i=1}^{4} \binom{n_i}{y_i} \widehat{p}_i^{y_i} (1 - \widehat{p}_i)^{n_i - y_i},$$

where $\widehat{p}_i = y_i/n_i$, for $i = 1, 2, 3, 4$.

LRT: The LRT statistic is the ratio of the two maximized likelihood functions; i.e.,

$$\begin{aligned}
\lambda = \frac{L(\widehat{\boldsymbol{\theta}}_0)}{L(\widehat{\boldsymbol{\theta}})} &= \frac{\displaystyle\prod_{i=1}^{4} \binom{n_i}{y_i} \widehat{p}^{y_i} (1 - \widehat{p})^{n_i - y_i}}{\displaystyle\prod_{i=1}^{4} \binom{n_i}{y_i} \widehat{p}_i^{y_i} (1 - \widehat{p}_i)^{n_i - y_i}} \\
&= \frac{\displaystyle\prod_{i=1}^{4} \binom{n_i}{y_i} \widehat{p}^{\sum_{i=1}^4 y_i} (1 - \widehat{p})^{\sum_{i=1}^4 (n_i - y_i)}}{\displaystyle\prod_{i=1}^{4} \binom{n_i}{y_i} \prod_{i=1}^{4} \widehat{p}_i^{y_i} \prod_{i=1}^{4} (1 - \widehat{p}_i)^{n_i - y_i}} \\
&= \frac{\left(\dfrac{\sum_{i=1}^4 y_i}{\sum_{i=1}^4 n_i}\right)^{\sum_{i=1}^4 y_i} \left[1 - \left(\dfrac{\sum_{i=1}^4 y_i}{\sum_{i=1}^4 n_i}\right)\right]^{\sum_{i=1}^4 (n_i - y_i)}}{\displaystyle\prod_{i=1}^{4} \left(\frac{y_i}{n_i}\right)^{y_i} \prod_{i=1}^{4} \left[1 - \left(\frac{y_i}{n_i}\right)\right]^{n_i - y_i}}.
\end{aligned}$$

It is worth noting that to find an exact level $\alpha = 0.05$ LRT, one would have to specify the value of k that satisfies

$$P_{H_0}\left(\frac{\left(\frac{\sum_{i=1}^4 Y_i}{\sum_{i=1}^4 n_i}\right)^{\sum_{i=1}^4 Y_i}\left[1-\left(\frac{\sum_{i=1}^4 Y_i}{\sum_{i=1}^4 n_i}\right)\right]^{\sum_{i=1}^4 (n_i - Y_i)}}{\prod_{i=1}^4\left(\frac{Y_i}{n_i}\right)^{Y_i}\prod_{i=1}^4\left[1-\left(\frac{Y_i}{n_i}\right)\right]^{n_i - Y_i}} < k\right) = 0.05$$

and then reject H_0 when $\lambda < k$. Because this is completely intractable, it is much easier to use the large-sample version of the LRT. We have

$$-2\ln\lambda = -2\ln\left(\frac{\left(\frac{\sum_{i=1}^4 y_i}{\sum_{i=1}^4 n_i}\right)^{\sum_{i=1}^4 y_i}\left[1-\left(\frac{\sum_{i=1}^4 y_i}{\sum_{i=1}^4 n_i}\right)\right]^{\sum_{i=1}^4 (n_i - y_i)}}{\prod_{i=1}^4\left(\frac{y_i}{n_i}\right)^{y_i}\prod_{i=1}^4\left[1-\left(\frac{y_i}{n_i}\right)\right]^{n_i - y_i}}\right)$$

and the (large-sample) rejection region

$$\text{RR} = \{-2\ln\lambda > \chi^2_{3,0.05} \approx 7.81\}.$$

```
> qchisq(0.95,3)
[1] 7.814728
```

Analysis: For the dental care data, the binomial counts are $y_1 = 49$, $y_2 = 142$, $y_3 = 181$, and $y_4 = 175$. The sample sizes are $n_1 = 658$, $n_2 = 839$, $n_3 = 811$, and $n_4 = 556$. It is straightforward to calculate

$$\begin{aligned}&-2\ln\lambda\\ &= -2\ln\left(\frac{\left(\frac{547}{2864}\right)^{547}\left(\frac{2317}{2864}\right)^{2317}}{\left(\frac{49}{658}\right)^{49}\left(\frac{142}{839}\right)^{142}\left(\frac{181}{811}\right)^{181}\left(\frac{175}{556}\right)^{175}\left(\frac{609}{658}\right)^{609}\left(\frac{697}{839}\right)^{697}\left(\frac{630}{811}\right)^{630}\left(\frac{381}{556}\right)^{381}}\right) \approx 127.79.\end{aligned}$$

Clearly, $H_0 : p_1 = p_2 = p_3 = p_4$ is rejected at the $\alpha = 0.05$ level. The (approximate) probability value for the test is

$$\text{p-value} = P_{H_0}(\chi^2(3) > 127.79) < 1 \times 10^{-16}$$

indicating the evidence against H_0 is overwhelming. Based on these data, there is clear evidence the probability of dental health care denial is different across the four insurance groups. $\square$

11 Bayesian Inference

11.1 Introduction

Discussion: Statistical inference deals with using information in a sample of data to make statements about quantities that are not observed; e.g., population-level parameters, future observations, etc. For example, suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population, where both parameters are unknown. In this instance, we might want to perform a hypothesis test regarding μ or write an interval estimator for σ^2. In a prediction problem, we might want to predict the value of a future random variable, say Y_{n+1}.

Remark: In your exposure to statistics up until now, you have likely been taught exclusively the **classical** (or **frequentist**) approach to inference. That is, you have been taught to regard model parameters like μ and σ^2 as fixed quantities that are unknown. By "fixed," we mean they are not random. One then uses the observed data $y_1, y_2, ..., y_n$ to learn about (or "infer") their values. The classical approach to inference can be summarized generally as follows:

1. Posit a population-level probability model for Y, say $Y \sim p_Y(y|\theta)$ or $Y \sim f_Y(y|\theta)$, where θ is a fixed (and unknown) population-level parameter.
2. Observe a sample $Y_1, Y_2, ..., Y_n$ from $p_Y(y|\theta)$ or $f_Y(y|\theta)$.
3. Use the observed values $y_1, y_2, ..., y_n$ to draw statistical inference for θ.

Bayesians take a different perspective. The key difference is they regard the population-level parameter θ to be random with its own probability distribution. This distribution can be used to incorporate (or model) "prior information" about θ. For example, suppose θ is the population proportion of covid-19 positive individuals in Richland County. In this setting, we could use known information on case counts and population sizes to elicit a prior model for θ, say a beta$(1, \beta)$ distribution where β is large (e.g., $\beta = 9$, $\beta = 99$, $\beta = 999$, etc.). With a sample of test outcomes $Y_1, Y_2, ..., Y_n$ modeled as iid Bernoulli(θ) random variables (i.e., positive/negative), the Bayesian would use the observed values *and* his/her prior belief about θ to make a statement about the population proportion. The **Bayesian** approach can be summarized generally as follows:

1. Posit a population-level probability model for Y, say $Y \sim p_Y(y|\theta)$ or $Y \sim f_Y(y|\theta)$.
2. Treat θ as a random variable itself with prior probability distribution $g(\theta)$.
3. Observe a sample $Y_1, Y_2, ..., Y_n$ from $p_Y(y|\theta)$ or $f_Y(y|\theta)$.
4. Use the observed values $y_1, y_2, ..., y_n$ and one's prior belief about θ (through the prior model) to draw statistical inference for θ.

Therefore, Bayesian's put more structure into the modeling process. In addition to modeling the random variables $Y_1, Y_2, ..., Y_n$, they model the population parameters as well.

11.2 Finding posterior distributions

Preview: For the Bayesian, statistical inference proceeds by deriving (or sampling from) the **posterior distribution** of θ. This is a conditional probability distribution of the parameter θ which has been "updated" to include the information from the observed values $y_1, y_2, ..., y_n$. Schematically, Bayesians think of inference in this way:

$$\text{Model } \theta \sim g(\theta) \quad \longrightarrow \quad \text{Observe } \mathbf{Y}|\theta \sim f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta) \quad \longrightarrow \quad \text{Update with } g(\theta|\mathbf{y}).$$

The model for θ on the front end is the **prior distribution**. The posterior distribution is the model on the back end. The posterior distribution combines prior information (supplied through the prior model) and the observed data $\mathbf{y}$. All statistical inference about θ is performed by using the posterior distribution.

Five-Step Algorithm: We now present a general algorithm to find the posterior distribution in any given problem. We will learn later that some steps below can be streamlined or skipped altogether.

1. Choose a **prior distribution** for θ, say $\theta \sim g(\theta)$. This distribution reflects our *a priori* knowledge regarding θ. We will discuss approaches to choose $g(\theta)$ in due course.

2. Construct the **conditional distribution** $f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)$. This is simply the multivariate distribution of $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, but now viewed conditionally on θ.

   - For example, if $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y|\theta)$, then the conditional distribution is
   $$f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta) = \prod_{i=1}^{n} f_Y(y_i|\theta).$$

   Mathematically, $f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)$ is same as the likelihood function $L(\theta|\mathbf{y})$ except the interpretation is different.

3. Form the **joint distribution** $f_{\mathbf{Y},\theta}(\mathbf{y}, \theta)$. This distribution describes how $\mathbf{Y}$ and θ vary jointly (remembering that θ is now regarded as random). From the definition of a conditional distribution,
$$f_{\mathbf{Y},\theta}(\mathbf{y}, \theta) = f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta) \ = \text{"Likelihood} \times \text{prior."}$$

4. Calculate the **marginal distribution** $m_{\mathbf{Y}}(\mathbf{y})$. This describes how $\mathbf{Y}$ is distributed "marginally." From the definition of a marginal distribution,
$$m_{\mathbf{Y}}(\mathbf{y}) = \int_{\Theta} f_{\mathbf{Y},\theta}(\mathbf{y}, \theta)d\theta = \int_{\Theta} f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta)d\theta,$$
where Θ is the "support" of θ (remember, we are now viewing θ as a random variable).

5. The **posterior distribution** is the conditional distribution of θ given $\mathbf{Y} = \mathbf{y}$. From the definition of a conditional distribution,

$$g(\theta|\mathbf{y}) = \frac{f_{\mathbf{Y},\theta}(\mathbf{y},\theta)}{m_{\mathbf{Y}}(\mathbf{y})} = \frac{f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta)}{\displaystyle\int_{\Theta} f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta)d\theta}.$$

This is the Bayesian's "updated" distribution of θ, given the data $\mathbf{Y} = \mathbf{y}$. Under the Bayesian framework, all inference regarding θ (e.g., point estimation, interval estimation, etc.) is conducted by using the posterior distribution $g(\theta|\mathbf{y})$.

Example 11.1 (*Binomial-beta*). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(θ) population, where $0 < \theta < 1$. In turn, suppose θ is best regarded as a beta random variable with parameters $\alpha > 0$ and $\beta > 0$; i.e., $\theta \sim \text{beta}(\alpha, \beta)$. Using the five-step algorithm above, derive the posterior distribution of θ.

1. Prior distribution. This is given in the problem. If $\theta \sim \text{beta}(\alpha, \beta)$, then the prior pdf is

$$g(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{for } 0 < \theta < 1.$$

2. Conditional distribution. If $Y_1, Y_2, ..., Y_n$ are iid from a Bernoulli(θ) population, then

$$f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}.$$

Mathematically, $f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)$ is the same as the likelihood function.

3. Joint distribution. This is the conditional distribution times the prior; i.e.,

$$\begin{aligned} f_{\mathbf{Y},\theta}(\mathbf{y},\theta) &= f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta) \\ &= \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\sum_{i=1}^n y_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n y_i+\beta-1}. \end{aligned}$$

4. Marginal distribution. We get this by taking the joint distribution and integrating over θ; i.e.,

$$\begin{aligned} m_{\mathbf{Y}}(\mathbf{y}) &= \int_{\Theta} f_{\mathbf{Y},\theta}(\mathbf{y},\theta)d\theta \\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\sum_{i=1}^n y_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n y_i+\beta-1}d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{\sum_{i=1}^n y_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n y_i+\beta-1}d\theta. \end{aligned}$$

The integrand in the last integral; i.e.,

$$\theta^{\sum_{i=1}^n y_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n y_i+\beta-1}$$

is a beta kernel with parameters $\sum_{i=1}^n y_i + \alpha$ and $n - \sum_{i=1}^n y_i + \beta$. Therefore,

$$\int_0^1 \theta^{\sum_{i=1}^n y_i + \alpha - 1}(1-\theta)^{n - \sum_{i=1}^n y_i + \beta - 1} d\theta = \frac{\Gamma(\sum_{i=1}^n y_i + \alpha)\Gamma(n - \sum_{i=1}^n y_i + \beta)}{\Gamma(n+\alpha+\beta)}$$

and thus the marginal distribution is

$$m_{\mathbf{Y}}(\mathbf{y}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n y_i + \alpha)\Gamma(n - \sum_{i=1}^n y_i + \beta)}{\Gamma(n+\alpha+\beta)}.$$

5. Posterior distribution. This is the joint distribution divided by the marginal distribution of $\mathbf{Y}$; i.e.,

$$\begin{aligned} g(\theta|\mathbf{y}) = \frac{f_{\mathbf{Y},\theta}(\mathbf{y},\theta)}{m_{\mathbf{Y}}(\mathbf{y})} &= \frac{\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\sum_{i=1}^n y_i + \alpha - 1}(1-\theta)^{n - \sum_{i=1}^n y_i + \beta - 1}}{\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \dfrac{\Gamma(\sum_{i=1}^n y_i + \alpha)\Gamma(n - \sum_{i=1}^n y_i + \beta)}{\Gamma(n+\alpha+\beta)}} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\sum_{i=1}^n y_i + \alpha)\Gamma(n - \sum_{i=1}^n y_i + \beta)}\, \theta^{\sum_{i=1}^n y_i + \alpha - 1}(1-\theta)^{n - \sum_{i=1}^n y_i + \beta - 1}. \end{aligned}$$

We recognize $g(\theta|\mathbf{y})$ as a beta pdf with parameters

$$\begin{aligned} \alpha^* &= \sum_{i=1}^n y_i + \alpha \\ \beta^* &= n - \sum_{i=1}^n y_i + \beta. \end{aligned}$$

Therefore, when we start with a beta(α, β) prior, the posterior distribution is also beta but with these "updated" parameters. Note that these updated parameter values are the same ones identified in the beta kernel

$$\theta^{\sum_{i=1}^n y_i + \alpha - 1}(1-\theta)^{n - \sum_{i=1}^n y_i + \beta - 1}$$

in the joint distribution $f_{\mathbf{Y},\theta}(\mathbf{y},\theta)$ in Step 3.

Discussion: Several remarks are in order.

- Instead of getting distracted by the math for the moment, it is helpful to see the "big picture" here; i.e.,

$$\begin{aligned} \text{Start with } \theta \sim \text{beta}(\alpha,\beta) &\longrightarrow \text{Observe data } \mathbf{y} \\ &\longrightarrow \text{Update with } \theta|\mathbf{y} \sim \text{beta}\left(\sum_{i=1}^n y_i + \alpha,\ n - \sum_{i=1}^n y_i + \beta\right). \end{aligned}$$

  The posterior distribution combines the information from the prior (through α and β) with the information in the data (through the sufficient statistic $\sum_{i=1}^n y_i$ and the sample size n).

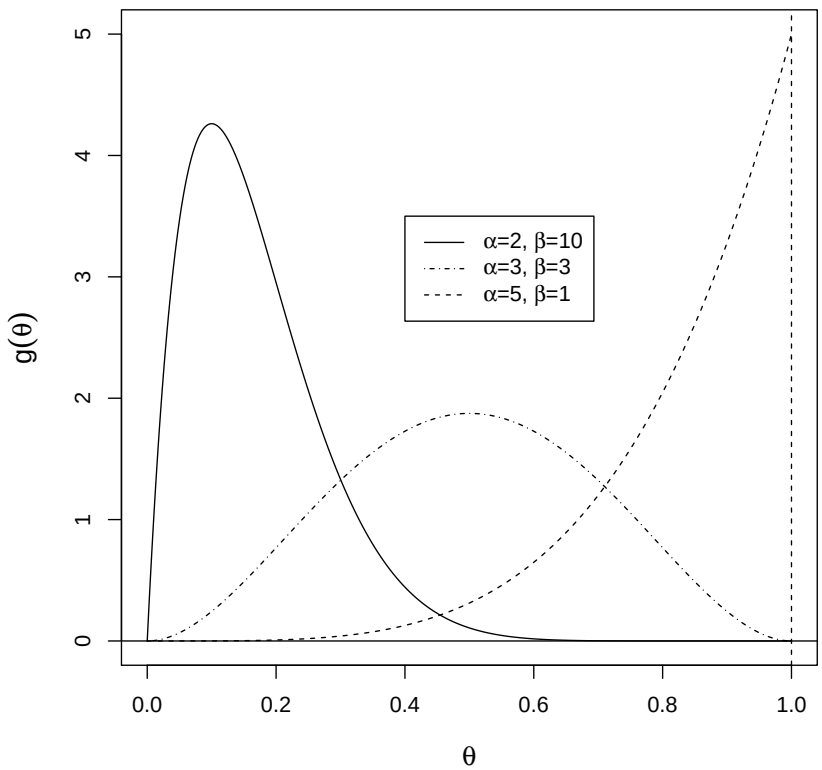

Figure 11.1: Different beta prior distributions.

- *Why did we start with a beta prior distribution?*
  - There is nothing which mandates us to use a beta prior distribution. However, it is certainly reasonable because

    $$Y \sim \text{Bernoulli}(\theta) \Longrightarrow \Theta = \{0 < \theta < 1\} = (0, 1).$$

    In turn, the support of a beta random variable is also $(0, 1)$. Therefore, the parameter space for θ in the Bernoulli(θ) distribution matches the support for θ in the prior distribution.
  - Prior information regarding θ can be incorporated by selecting a prior model that accurately reflects that information; refer to Figure 11.1 (above). In the beta prior, values of $\alpha < \beta$ are consistent with smaller values of θ; values of $\alpha > \beta$ are consistent with larger values of θ. Recall that the beta pdf is symmetric about $1/2$ when $\alpha = \beta$. This type of prior model might be selected when θ is thought to be near $1/2$.
- *Is there a reason the prior and posterior distributions are both beta?* Yes, in this problem, the beta distribution is a **conjugate prior**. We will discuss this more later.

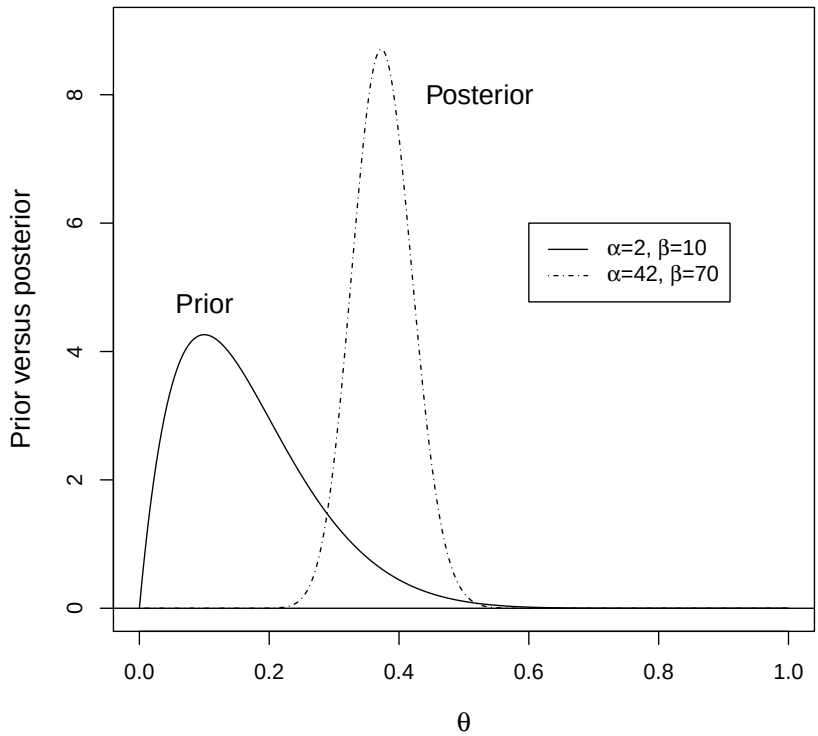

Figure 11.2: Prior and posterior distributions for the population proportion of UofSC students supporting the mask mandate. Prior: $\theta \sim \text{beta}(2, 10)$. Posterior: $\theta|\mathbf{y} \sim \text{beta}(42, 70)$.

Application: On August 17, 2021, Dr. Pastides reinstated the mask mandate for all faculty, staff, and students at the University of South Carolina. Let θ denote the population proportion of UofSC students who support this decision. A random sample of $n = 100$ students is obtained and each student is asked if s/he supports the mandate. Suppose the yes/no responses $Y_1, Y_2, ..., Y_{100}$ are modeled as iid Bernoulli(θ), where $\theta \sim \text{beta}(\alpha, \beta)$ is used to incorporate prior information. Suppose 40 students answered "yes" so that

$$\sum_{i=1}^{100} y_i = 40.$$

For an analyst who uses a beta($\alpha = 2, \beta = 10$) prior, which incorporates the *a priori* belief that θ is "small," the posterior distribution is beta with parameters

$$\begin{aligned} \alpha^* &= 40 + 2 = 42 \\ \beta^* &= 100 - 40 + 10 = 70; \end{aligned}$$

see Figure 11.2 (above). Note how the posterior distribution shifts notably to the right when compared to the prior model; this is the due to the effect of observing 40 "successes," which is not consistent with a beta$(2, 10)$ prior. In addition, note how the posterior distribution is

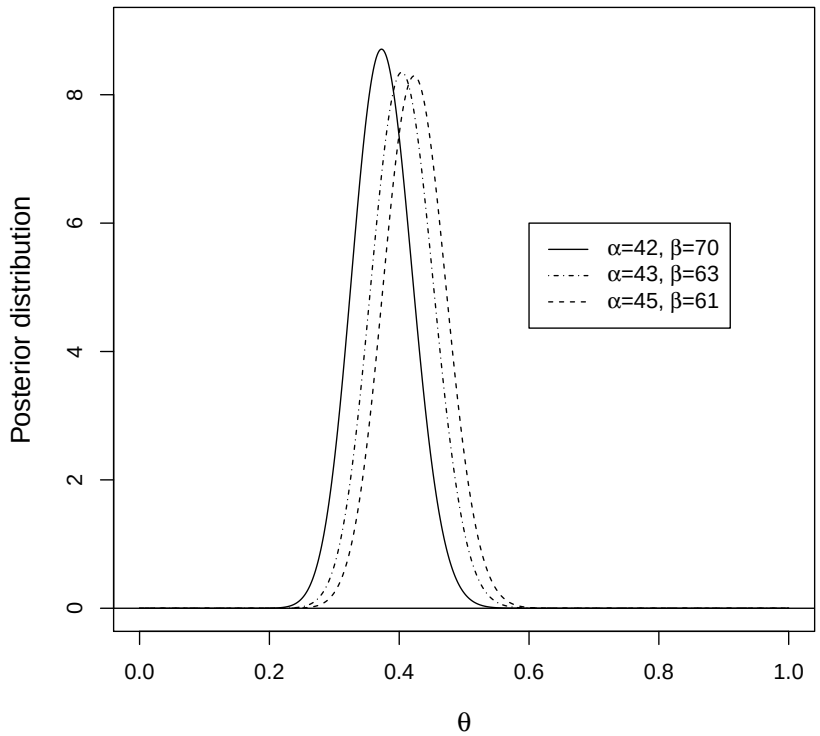

Figure 11.3: Posterior distributions $g(\theta|\mathbf{y})$ for the different prior models in Figure 11.1.

much less variable than the prior distribution. The prior model is selected before seeing the data $\mathbf{y}$. The posterior distribution incorporates the data so it uses more information.

Observation: Out of curiosity, I went ahead and constructed the posterior distributions for each of the prior model choices in Figure 11.1; i.e.,

- Prior: $\theta \sim \text{beta}(2, 10) \implies$ Posterior: $\theta|\mathbf{y} \sim \text{beta}(42, 70)$
- Prior: $\theta \sim \text{beta}(3, 3) \implies$ Posterior: $\theta|\mathbf{y} \sim \text{beta}(43, 63)$
- Prior: $\theta \sim \text{beta}(5, 1) \implies$ Posterior: $\theta|\mathbf{y} \sim \text{beta}(45, 61)$,

and I plotted each one in Figure 11.3 (above). This figure shows the posterior distributions are very similar! This begs the question, *"Does the prior model choice even matter?"* The answer is "Yes, it can matter," although there is a good reason why the posterior distributions are similar here. In general, a posterior distribution is formed by combining the prior model with the data (i.e., the information contained in the likelihood function). With a large sample size like $n = 100$ students, the likelihood function is contributing so much information that it is "down-weighting" the influence of the prior. Had the sample size been something like $n = 10$ students, then the prior model choices would be much more influential in determining where the posteriors reside. $\square$

Example 11.2 (*Poisson-gamma*). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population, where $\theta > 0$. In turn, suppose θ is best regarded as a gamma random variable with parameters $\alpha > 0$ and $\beta > 0$; i.e., $\theta \sim$ gamma(α, β). Using the five-step algorithm described previously, derive the posterior distribution of θ.

1. Prior distribution. This is given in the problem. If $\theta \sim$ gamma(α, β), then the prior pdf is

$$g(\theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\,\theta^{\alpha-1}e^{-\theta/\beta}, \quad \text{for } \theta > 0.$$

2. Conditional distribution. If $Y_1, Y_2, ..., Y_n$ are iid from a Poisson(θ) population, then

$$f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta) = \prod_{i=1}^n \frac{\theta^{y_i}e^{-\theta}}{y_i!} = \frac{\theta^{\sum_{i=1}^n y_i}e^{-n\theta}}{\prod_{i=1}^n y_i!}.$$

Mathematically, $f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)$ is the same as the likelihood function.

3. Joint distribution. This is the conditional distribution times the prior; i.e.,

$$\begin{aligned} f_{\mathbf{Y},\theta}(\mathbf{y},\theta) &= f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta) \\ &= \frac{\theta^{\sum_{i=1}^n y_i}e^{-n\theta}}{\prod_{i=1}^n y_i!} \times \frac{1}{\Gamma(\alpha)\beta^\alpha}\,\theta^{\alpha-1}e^{-\theta/\beta} = \frac{1}{\Gamma(\alpha)\beta^\alpha\prod_{i=1}^n y_i!}\,\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}. \end{aligned}$$

4. Marginal distribution. We get this by taking the joint distribution and integrating over θ; i.e.,

$$\begin{aligned} m_{\mathbf{Y}}(\mathbf{y}) &= \int_\Theta f_{\mathbf{Y},\theta}(\mathbf{y},\theta)d\theta \\ &= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha\prod_{i=1}^n y_i!}\,\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}d\theta \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha\prod_{i=1}^n y_i!}\int_0^\infty \theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}d\theta. \end{aligned}$$

The integrand in the last integral; i.e.,

$$\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}$$

is a gamma kernel with parameters $\sum_{i=1}^n y_i + \alpha$ and $(n + 1/\beta)^{-1}$. Therefore,

$$\int_0^\infty \theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}d\theta = \Gamma\left(\sum_{i=1}^n y_i + \alpha\right)\left[\left(n+\frac{1}{\beta}\right)^{-1}\right]^{\sum_{i=1}^n y_i+\alpha}$$

and thus the marginal distribution is

$$m_{\mathbf{Y}}(\mathbf{y}) = \frac{\Gamma\left(\sum_{i=1}^n y_i + \alpha\right)\left[\left(n+\frac{1}{\beta}\right)^{-1}\right]^{\sum_{i=1}^n y_i+\alpha}}{\Gamma(\alpha)\beta^\alpha\prod_{i=1}^n y_i!}.$$

5. Posterior distribution. This is the joint distribution divided by the marginal distribution of $\mathbf{Y}$; i.e.,

$$\begin{aligned} g(\theta|\mathbf{y}) = \frac{f_{\mathbf{Y},\theta}(\mathbf{y},\theta)}{m_{\mathbf{Y}}(\mathbf{y})} &= \frac{\dfrac{1}{\Gamma(\alpha)\beta^\alpha \prod_{i=1}^n y_i!}\,\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}}{\dfrac{\Gamma\left(\sum_{i=1}^n y_i+\alpha\right)\left[\left(n+\frac{1}{\beta}\right)^{-1}\right]^{\sum_{i=1}^n y_i+\alpha}}{\Gamma(\alpha)\beta^\alpha \prod_{i=1}^n y_i!}} \\ &= \frac{1}{\Gamma\left(\sum_{i=1}^n y_i+\alpha\right)\left[\left(n+\frac{1}{\beta}\right)^{-1}\right]^{\sum_{i=1}^n y_i+\alpha}}\,\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}. \end{aligned}$$

We recognize $g(\theta|\mathbf{y})$ as a gamma pdf with parameters

$$\begin{aligned} \alpha^* &= \sum_{i=1}^n y_i + \alpha \\ \beta^* &= \left(n+\frac{1}{\beta}\right)^{-1}. \end{aligned}$$

Therefore, when we start with a gamma(α,β) prior, the posterior distribution is also gamma but with these "updated" parameters. Note that these updated parameter values are the same ones identified in the gamma kernel

$$\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}$$

in the joint distribution $f_{\mathbf{Y},\theta}(\mathbf{y},\theta)$ in Step 3.

Discussion: Several remarks are in order.

- As in Example 11.1 (binomial-beta), it is helpful to see the "big picture" here; i.e.,

  $$\begin{aligned} &\text{Start with } \theta \sim \text{gamma}(\alpha,\beta) \longrightarrow \text{Observe data } \mathbf{y} \\ &\qquad\qquad \longrightarrow \text{Update with } \theta|\mathbf{y} \sim \text{gamma}\left(\sum_{i=1}^n y_i+\alpha, \left(n+\frac{1}{\beta}\right)^{-1}\right). \end{aligned}$$

  The posterior distribution combines the information from the prior (through α and β) with the information in the data (through the sufficient statistic $\sum_{i=1}^n y_i$ and the sample size n).

- *Why did we start with a gamma prior distribution?*
  - There is nothing which mandates us to use a gamma prior distribution. However, it is certainly reasonable because

    $$Y \sim \text{Poisson}(\theta) \Longrightarrow \Theta = \{0 < \theta < \infty\} = (0,\infty).$$

    In turn, the support of a gamma random variable is also $(0,\infty)$. Therefore, the parameter space for θ in the Poisson(θ) distribution matches the support for θ in the prior distribution.

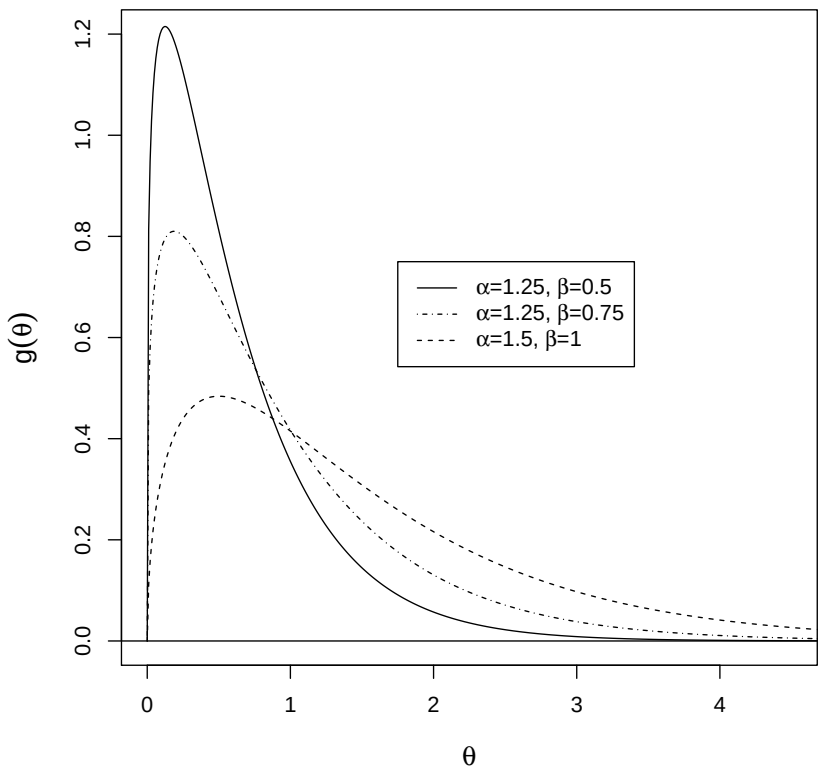

Figure 11.4: Different gamma prior distributions.

- Prior information regarding θ can be incorporated by selecting a prior model that accurately reflects that information; refer to Figure 11.4 (above). The gamma prior can assume a variety of shapes depending on the values of α and β (e.g., when $\alpha = 1$, the gamma distribution reduces to an exponential distribution with mean β). Recall the mean of a gamma(α, β) random variable is $\alpha\beta$.

- *Is there a reason the prior and posterior distributions are both gamma?* Yes, in this problem, the gamma distribution is a **conjugate prior**. We will discuss this more later.

Application: The number of events (e.g., claims per day, accidents per year, etc.) in property/casualty insurance is often assumed to follow a Poisson distribution. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population, where $\theta \sim$ gamma(α, β) is used to incorporate prior information. In STAT 512 (Example 9.4, pp 118), we used a Poisson distribution to model the number of accidents per year for a sample of $n = 84$ policies. The total number of accidents in the sample was

$$\sum_{i=1}^{84} y_i = 103.$$

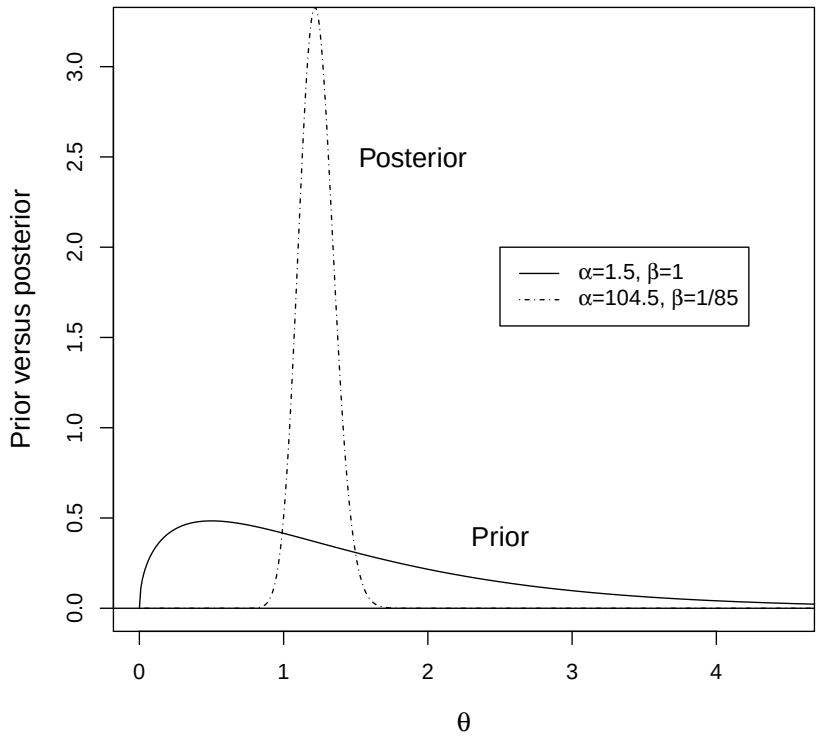

Figure 11.5: Prior and posterior distributions for the mean number of accidents per year. Prior: $\theta \sim \text{gamma}(1.5, 1)$. Posterior: $\theta|\mathbf{y} \sim \text{gamma}(104.5, 1/85)$.

For an analyst who uses a gamma($\alpha = 1.5, \beta = 1$) prior, which incorporates the *a priori* belief that the (prior) mean is 1.5 accidents per year, the posterior distribution is gamma with parameters

$$\begin{aligned} \alpha^* &= 103 + 1.5 = 104.5 \\ \beta^* &= \left(84 + \frac{1}{1}\right)^{-1} = \frac{1}{85}; \end{aligned}$$

see Figure 11.5 (above). Again, note how the posterior distribution is much less variable than the prior distribution. The posterior distribution combines the information supplied by the prior model with the information observed in data $\mathbf{y}$. As in Example 11.1, the sample size $n = 84$ is pretty large, so the posterior distribution is more influenced by the contribution from the likelihood function than by the contribution from the prior model. $\square$

Remark: We have presented a five-step algorithm to construct the posterior distribution $g(\theta|\mathbf{y})$. It turns out that Step 4, the step which painstakingly derives the marginal distribution of $\mathbf{Y}$, isn't really needed to determine $g(\theta|\mathbf{y})$. In addition, when a sufficient statistic $T = T(\mathbf{Y}) = T(Y_1, Y_2, ..., Y_n)$ exists, there is no harm in working with the sampling distribution of T from the start (this streamlines Step 2). We now discuss these issues.

Posterior shortcut: Let's start with the joint distribution

$$f_{\mathbf{Y},\theta}(\mathbf{y},\theta) = f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta) \ = \text{"Likelihood} \times \text{prior"}$$

in Step 3. This distribution describes how the random vector $\mathbf{Y}$ and the random variable θ vary jointly. Step 4 then proceeds to find the marginal distribution $m_{\mathbf{Y}}(\mathbf{y})$ so that we can determine the posterior distribution

$$g(\theta|\mathbf{y}) = \frac{f_{\mathbf{Y},\theta}(\mathbf{y},\theta)}{m_{\mathbf{Y}}(\mathbf{y})}.$$

However, if our only goal is to determine $g(\theta|\mathbf{y})$, then is Step 4 really needed? After all, we know that $m_{\mathbf{Y}}(\mathbf{y})$ is *free of* θ, that is, as far as the posterior distribution $g(\theta|\mathbf{y})$ is concerned, the marginal pdf/pmf $m_{\mathbf{Y}}(\mathbf{y})$ is nothing more than a proportionality constant. This means we can write

$$g(\theta|\mathbf{y}) \propto f_{\mathbf{Y},\theta}(\mathbf{y},\theta) = f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta) \ = \text{"Likelihood} \times \text{prior"}.$$

The posterior distribution in Step 5 is proportional to the joint distribution in Step 3; i.e., dividing by $m_{\mathbf{Y}}(\mathbf{y})$ in Step 4 simply takes the joint distribution and makes it a bona fide (updated) density function for θ. For example,

- In Example 11.1 (binomial-beta), we wrote in Step 3

$$f_{\mathbf{Y},\theta}(\mathbf{y},\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\sum_{i=1}^n y_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n y_i+\beta-1}.$$

  Dropping the constant which is free of θ, we know

$$g(\theta|\mathbf{y}) \propto \theta^{\sum_{i=1}^n y_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n y_i+\beta-1},$$

  the kernel of a beta density with parameters

$$\begin{aligned}\alpha^* &= \sum_{i=1}^n y_i + \alpha \\ \beta^* &= n - \sum_{i=1}^n y_i + \beta.\end{aligned}$$

  Therefore, we know at this point (after Step 3) that the posterior distribution *must* be beta with these parameters. Dividing the marginal distribution $m_{\mathbf{Y}}(\mathbf{y})$ in Step 4 is now unnecessary; all this step does is determine the proportionality constant which makes $g(\theta|\mathbf{y})$ a bona fide density of θ.

- In Example 11.2 (Poisson-gamma), we wrote in Step 3

$$f_{\mathbf{Y},\theta}(\mathbf{y},\theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha\prod_{i=1}^n y_i!}\,\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}.$$

Dropping the constant which is free of θ, we know

$$g(\theta|\mathbf{y}) \propto \theta^{\sum_{i=1}^n y_i+\alpha-1} e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}},$$

the kernel of a gamma density with parameters

$$\begin{aligned}\alpha^* &= \sum_{i=1}^n y_i + \alpha \\ \beta^* &= \left(n+\frac{1}{\beta}\right)^{-1}.\end{aligned}$$

Therefore, we know at this point (after Step 3) that the posterior distribution *must* be gamma with these parameters. Again, dividing by $m_{\mathbf{Y}}(\mathbf{y})$ in Step 4 is just a normalization step.

Example 11.3. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from the population-level pdf

$$f_Y(y|\theta) = \begin{cases} \theta^2 y e^{-\theta y}, & y > 0 \\ 0, & \text{otherwise,} \end{cases}$$

a gamma distribution with shape parameter 2 and scale parameter $1/\theta$. In turn, suppose θ is best regarded as random with prior pdf

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

Determine the posterior distribution $g(\theta|\mathbf{y})$.

Solution. We use this example as an opportunity to highlight our recently discovered "streamlined approach." The conditional distribution $f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)$ in Step 2 is simply the likelihood function, that is,

$$\begin{aligned} f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta) &= f_Y(y_1|\theta) \times f_Y(y_2|\theta) \times \cdots \times f_Y(y_n|\theta) \\ &= \theta^2 y_1 e^{-\theta y_1} \times \theta^2 y_2 e^{-\theta y_2} \times \cdots \times \theta^2 y_n e^{-\theta y_n} = \theta^{2n} \left(\prod_{i=1}^n y_i\right) e^{-\theta \sum_{i=1}^n y_i}. \end{aligned}$$

The joint distribution in Step 3 is

$$f_{\mathbf{Y},\theta}(\mathbf{y},\theta) = f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta) g(\theta) = \theta^{2n} \left(\prod_{i=1}^n y_i\right) e^{-\theta \sum_{i=1}^n y_i} \times e^{-\theta} \propto \theta^{2n} e^{-\theta\left(\sum_{i=1}^n y_i + 1\right)},$$

a gamma kernel with parameters

$$\begin{aligned}\alpha^* &= 2n+1 \\ \beta^* &= \left(\sum_{i=1}^n y_i + 1\right)^{-1}.\end{aligned}$$

Therefore, the posterior distribution $g(\theta|\mathbf{y})$ must be gamma with these (updated) parameters. $\square$

Sufficient statistics: When a sufficient statistic $T = T(\mathbf{Y}) = T(Y_1, Y_2, ..., Y_n)$ exists, determining the posterior distribution $g(\theta|\mathbf{y})$ becomes even easier. When T is sufficient, we know the conditional distribution $f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)$ can be written as

$$f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta) = k_1(t, \theta)k_2(y_1, y_2, ..., y_n) = k_1(t, \theta)k_2(\mathbf{y}),$$

by the Factorization Theorem. Therefore, the posterior distribution $g(\theta|\mathbf{y})$ satisfies

$$g(\theta|\mathbf{y}) \propto f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta) = k_1(t, \theta)k_2(\mathbf{y})g(\theta) \propto k_1(t, \theta)g(\theta).$$

This should convince us of the following:

1. The posterior distribution $g(\theta|\mathbf{y})$ must depend on the data $\mathbf{y}$ through the value of the sufficient statistic $T = T(\mathbf{y}) = t$.
2. We can write $g(\theta|t)$ to denote the posterior distribution instead and there is no harm in doing so (i.e., we lose no information about the posterior).

Therefore, instead of working with the conditional distribution $f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)$ in Step 2, we can instead work with $f_{T|\theta}(t|\theta)$, the (sampling) distribution of T. The posterior distribution $g(\theta|t)$ must satisfy

$$g(\theta|t) \propto f_{T|\theta}(t|\theta)g(\theta).$$

Example 11.1 (revisited). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(θ) population, where $0 < \theta < 1$. In turn, suppose θ is best regarded as a beta random variable with parameters $\alpha > 0$ and $\beta > 0$; i.e., $\theta \sim \text{beta}(\alpha, \beta)$. A sufficient statistic is

$$T = T(Y_1, Y_2, ..., Y_n) = \sum_{i=1}^{n} Y_i$$

and $T \sim b(n, \theta)$. Therefore,

$$g(\theta|t) \propto f_{T|\theta}(t|\theta)g(\theta) = \binom{n}{t}\theta^t(1-\theta)^{n-t} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{t+\alpha-1}(1-\theta)^{n-t+\beta-1}.$$

We can immediately conclude the posterior distribution is beta with parameters $t + \alpha$ and $n - t + \beta$, where $t = \sum_{i=1}^{n} y_i$. This was the same conclusion as before. $\square$

Example 11.2 (revisited). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population, where $\theta > 0$. In turn, suppose θ is best regarded as a gamma random variable with parameters $\alpha > 0$ and $\beta > 0$; i.e., $\theta \sim \text{gamma}(\alpha, \beta)$. A sufficient statistic is

$$T = T(Y_1, Y_2, ..., Y_n) = \sum_{i=1}^{n} Y_i$$

and $T \sim \text{Poisson}(n\theta)$. Therefore,

$$g(\theta|t) \propto f_{T|\theta}(t|\theta)g(\theta) = \frac{(n\theta)^t e^{-n\theta}}{t!} \times \frac{1}{\Gamma(\alpha)\beta^\alpha}\,\theta^{\alpha-1}e^{-\theta/\beta} \propto \theta^{t+\alpha-1}e^{-\theta\big/\left(n+\frac{1}{\beta}\right)^{-1}}.$$

We can immediately conclude the posterior distribution is gamma with parameters $t+\alpha$ and $(n + 1/\beta)^{-1}$, where $t = \sum_{i=1}^{n} y_i$. This was the same conclusion as before. $\square$

11.3 Prior model selection

Recall: In Example 11.2, we considered the Poisson-gamma modeling problem, that is, $Y_1, Y_2, ..., Y_n$ are iid Poisson(θ) and the prior distribution $\theta \sim$ gamma(α, β), where α and β are known. We observed the following:

$$\text{Prior: } \theta \sim \text{gamma}(\alpha, \beta) \longrightarrow \text{Posterior: } \theta|\mathbf{y} \sim \text{gamma}\left(\sum_{i=1}^n y_i + \alpha, \left(n + \frac{1}{\beta}\right)^{-1}\right).$$

That is, the prior and the posterior both reside in the same family of distributions. Suppose instead we had selected the prior distribution

$$\theta \sim \mathcal{N}(\mu, \sigma^2),$$

where μ and σ^2 are known. Will the posterior distribution now reside in the normal family? With a $\mathcal{N}(\mu, \sigma^2)$ prior, the joint distribution in Step 3 is

$$\begin{aligned} f_{\mathbf{Y},\theta}(\mathbf{y}, \theta) &= f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)g(\theta) \\ &= \frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{\prod_{i=1}^n y_i!} \times \frac{1}{\sqrt{2\pi}\sigma} e^{-(\theta-\mu)^2/2\sigma^2} = \frac{\theta^{\sum_{i=1}^n y_i} e^{-[n\theta + (\theta-\mu)^2/2\sigma^2]}}{\sqrt{2\pi}\sigma \prod_{i=1}^n y_i!}. \end{aligned}$$

Unfortunately, there is no easily identified normal kernel from this distribution, and in fact the marginal distribution of $\mathbf{Y}$; i.e.,

$$m_{\mathbf{Y}}(\mathbf{y}) = \int_\Theta f_{\mathbf{Y},\theta}(\mathbf{y}, \theta) d\theta = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma \prod_{i=1}^n y_i!} \theta^{\sum_{i=1}^n y_i} e^{-[n\theta + (\theta-\mu)^2/2\sigma^2]} d\theta$$

involves a messy integral which appears to be intractable.

Q: Why is it when $\theta \sim$ gamma(α, β), the posterior $g(\theta|\mathbf{y})$ is a gamma pdf, but when $\theta \sim \mathcal{N}(\mu, \sigma^2)$, the posterior is not normal?
A: Because the $\mathcal{N}(\mu, \sigma^2)$ family is not conjugate in this example.

Terminology: Let $\mathcal{F} = \{f_Y(y|\theta); \theta \in \Theta\}$ denote a class of probability density (mass) functions indexed by the parameter θ. A class $\mathcal{G}$ of prior distributions is said to be a **conjugate family** for $\mathcal{F}$ if the posterior distribution $g(\theta|\mathbf{y}) \in \mathcal{G}$, for all $f_Y(y|\theta) \in \mathcal{F}$ and for all priors $g(\theta) \in \mathcal{G}$. Table 11.1 (next page) gives examples of common probability distributions and their conjugate priors.

Terminology: Parameters which index a prior distribution are called **hyperparameters**.

- In Example 11.1 (binomial-beta), the prior distribution $\theta \sim$ beta(α, β). The hyperparameters are α and β.
- In Example 11.2 (Poisson-gamma), the prior distribution $\theta \sim$ gamma(α, β). The hyperparameters are α and β.

Table 11.1: Common conjugate families and their hyperparameters.

Family	Parameter	Conjugate family	Prior hyperparameters
$b(n,p)$	p	beta(α,β)	α, β
nib(r,p)	p	beta(α,β)	α, β
Poisson(λ)	λ	gamma(α,β)	α, β
$\mathcal{N}(\mu,\sigma_0^2)$	μ	$\mathcal{N}(\eta,\delta^2)$	η, δ^2
$\mathcal{N}(\mu_0,\sigma^2)$	σ^2	IG(α,β)	α, β
exponential$(1/\theta)$	θ	gamma(α,β)	α, β
mult$(n,\mathbf{p})$	$\mathbf{p}$	Dirichlet$(\alpha_1,\alpha_2,...,\alpha_k)$	$\alpha_1,\alpha_2,...,\alpha_k$
$\mathcal{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma}_0)$	$\boldsymbol{\mu}$	$\mathcal{N}_p(\boldsymbol{\eta},\boldsymbol{\Sigma})$	$\boldsymbol{\eta}$, $\boldsymbol{\Sigma}$
$\mathcal{N}_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma})$	$\boldsymbol{\Sigma}$	inverse Wishart$(\nu,\boldsymbol{\Psi})$	ν, $\boldsymbol{\Psi}$

Remark: To carry out a **classical Bayes** analysis (as we have done so far), the researcher must specify the precise values of all hyperparameters which index the prior. In this light, it is certainly reasonable to criticize (or at least be skeptical of) the Bayesian paradigm because of this requirement. For example, in the mask mandate problem (Example 11.1), how do we know *a priori* the population proportion of UofSC students favoring the mask mandate $\theta \sim$ beta$(2,10)$? Of course, we probably don't know this. However, this is certainly a reasonable prior model if the analyst believes θ is "small." The beauty of the Bayesian paradigm is that both the prior and the data (through the likelihood function) "have their say" in determining where the posterior distribution resides. Furthermore, we know from Examples 11.1 and 11.2 that when the sample size n is large, the effect of the prior distribution is reduced anyway. Therefore, even a "bad" prior choice may not create large problems.

Alternative approaches: In addition to classical Bayes (where hyperparameters are specified beforehand), other Bayesian approaches to inference have been developed. We will not pursue these approaches in depth in this course, but it is helpful to have a passing familiarity with them.

- *Hierarchical Bayes.* In this approach, instead of eliciting values of the hyperparameters beforehand, one assigns probability distributions to them. For example, in Example 11.1 (binomial-beta), we could enrich the model as follows:

$$\begin{aligned} T|\theta &\sim b(n,\theta) \\ \theta|\alpha,\beta &\sim \text{beta}(\alpha,\beta) \\ \alpha,\beta &\sim \text{exponential}(1). \end{aligned}$$

  In this hierarchy, the exponential(1) distribution is called a **hyperprior** for α and β. Adding this extra layer increases the level of complexity in deriving the posterior distribution $g(\theta|t)$ analytically. Computational (simulation-based) Bayesian methods are usually needed to estimate models like this.

- *Empirical Bayes.* In this approach, instead of eliciting values of the hyperparameters beforehand, one uses the marginal distribution of the data (usually through a sufficient

statistic) to estimate them. For example, in Example 11.1 (binomial-beta), we assumed

$$\begin{aligned} T|\theta &\sim b(n,\theta) \\ \theta|\alpha,\beta &\sim \text{beta}(\alpha,\beta). \end{aligned}$$

In Step 4, we derived the marginal distribution of the data; the pmf corresponding to this distribution can be written as

$$m_T(t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(t+\alpha)\Gamma(n-t+\beta)}{\Gamma(n+\alpha+\beta)},$$

where $t = \sum_{i=1}^n y_i$, for $t = 0, 1, 2, ..., n$. This is called the **beta-binomial** pmf. An empirical Bayesian approach would first determine estimates $\widehat{\alpha}$ and $\widehat{\beta}$ from this distribution (using MLE, MOM, etc.) and then set the prior distribution at these estimates, beta$(\widehat{\alpha}, \widehat{\beta})$. An obvious criticism of empirical Bayes is that one is actually using the data twice: once to select the prior and then again to derive the posterior. Another criticism of empirical Bayes is that it violates the "spirit" of the Bayesian approach, where a prior distribution should be selected on the basis of *a priori* knowledge−not currently observed data.

- *Nonparametric Bayes.* This approach generally avoids making parametric assumptions altogether. Instead of using prior distributions for population-level parameters, prior distributions are assigned to the population-level distributions themselves. For example, in a density estimation problem, one might model

$$\begin{aligned} Y_1, Y_2, ..., Y_n &\sim \text{iid } F \\ F &\sim \text{DP}, \end{aligned}$$

where F is a cdf and "DP" stands for "Dirichlet process." The DP is basically a probability model for an infinite dimensional parameter, viewing F to be of "infinite" dimension. The nonparametric Bayes approach is used in a variety of statistical problems, including regression, survival analysis, and clustering.

Q: In a classical Bayes analysis, do we have to use a conjugate prior distribution?
A: No, but it does simplify things, and conjugate priors are often used primarily for this reason. Non-conjugate priors could be used, but then simulation-based methods would be needed to approximate posterior distributions numerically (through a type of Monte Carlo sampling). It should be noted that in some problems conjugate priors may not exist, especially when the population-level parameter θ is of higher dimension.

Remark: When there is a general lack of *a priori* knowledge about the parameter θ, prior models may be difficult to select. It might also be desired for the prior $g(\theta)$ to play a minimal role in determining where the posterior distribution $g(\theta|\mathbf{y})$ resides. In this situation, one can choose a **noninformative prior** for θ. The motivation for this is to "let the data speak for themselves" and to have the prior distribution contribute only minimally. These are also known as "vague," "flat," or "diffuse" priors.

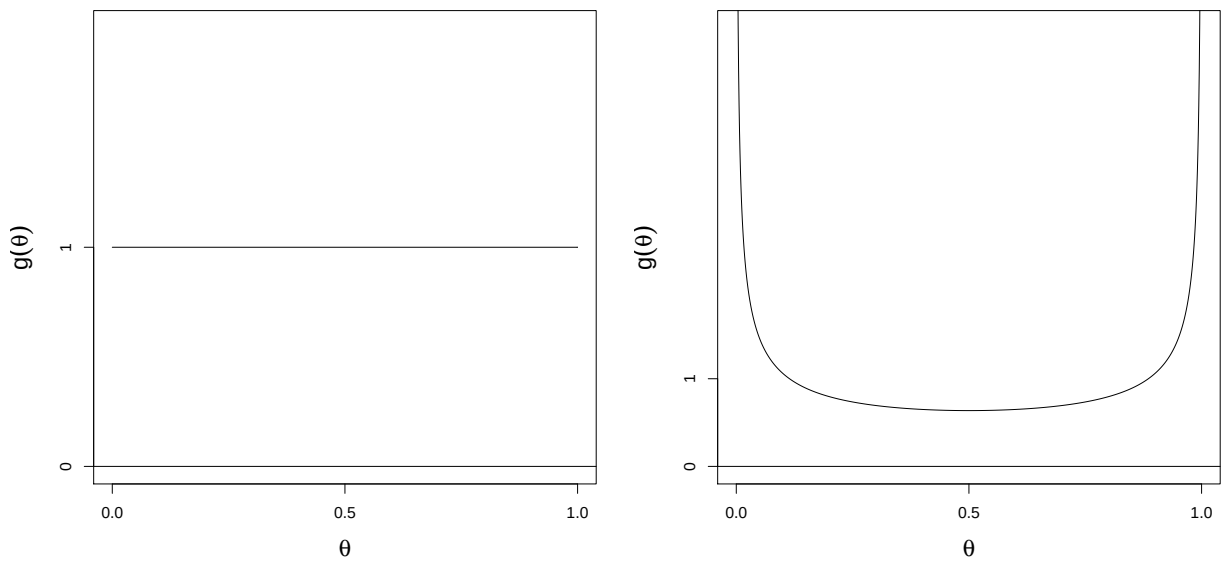

Figure 11.6: Noninformative beta(α, β) prior distributions. Left: $\alpha = \beta = 1$ (uniform prior). Right: $\alpha = \beta = 1/2$ (Jeffreys' prior). Note that the vertical axis scales are different in the figures.

Example 11.1 (revisited). In the UofSC mask mandate problem, let $Y_1, Y_2, ..., Y_{100}$ denote the "yes/no" responses from students who are asked if they favor the mask mandate (1 = in favor; 0 = not). Suppose we model these responses as iid Bernoulli(θ), where $\theta \sim$ beta(α, β) is used to incorporate prior knowledge as before.

- One example of a noninformative prior is the $\mathcal{U}(0, 1)$ distribution; i.e., a beta(α, β) distribution when $\alpha = \beta = 1$. This injects the minimal amount of information into the prior model; in fact, it simply promises that $0 < \theta < 1$ and acknowledges there is no *a priori* knowledge.
- Another noninformative prior is the beta(α, β) distribution when $\alpha = \beta = 1/2$; this is known as **Jeffreys' prior** (to be discussed momentarily). This distribution is nearly uniform (flat) over a large range of $(0, 1)$, but allows for θ to be very small or very large with larger probability. $\square$

Example 11.2 (revisited). In the actuarial claims problem, let $Y_1, Y_2, ..., Y_{84}$ denote the number of accidents observed from a random sample of $n = 84$ policies, modeled as iid counts from a Poisson(θ) distribution. As we will see later, a noninformative gamma(α, β) prior for θ would choose α to be small and β to be large. For example, Figure 11.7 (next page) shows this prior when $\alpha = 1/2$ and $\beta = 100$. Note this prior distribution is mostly flat over a large range of values of $\theta > 0$. $\square$

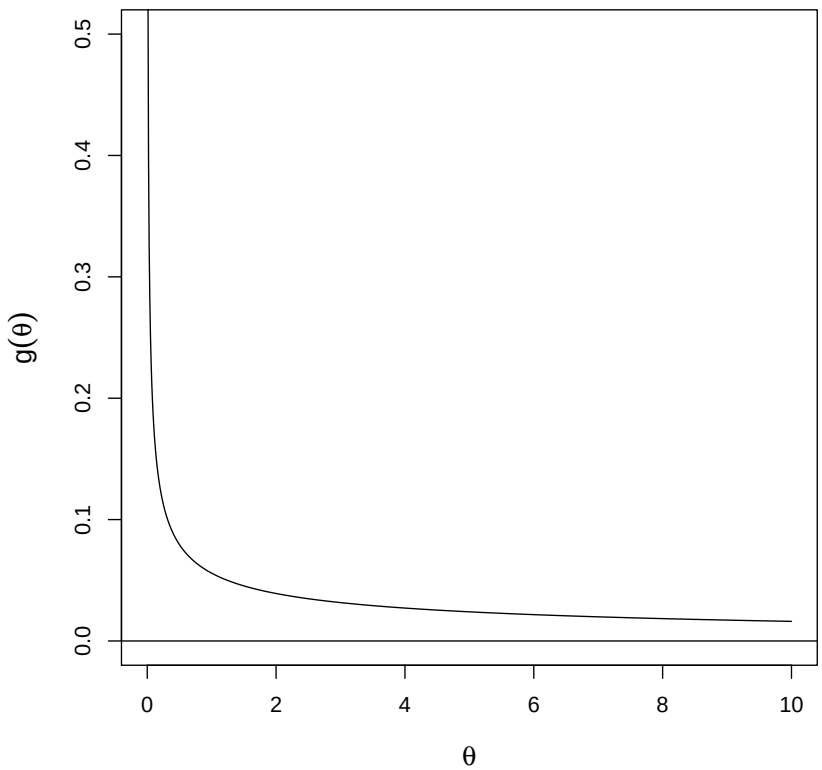

Figure 11.7: Gamma(α, β) prior distribution with $\alpha = 1/2$ and $\beta = 100$.

Jeffreys' prior: A general strategy for (noninformative) prior model selection uses "Jeffreys' principle." Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y|\theta)$, where $\theta \sim g(\theta)$ is a prior distribution. When θ is a scalar parameter (one-dimensional), Jeffreys' principle says to choose

$$g(\theta) \propto [I(\theta)]^{1/2},$$

where

$$I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} \ln f_Y(Y|\theta)\right]$$

is the **Fisher information**. The mathematical reasoning behind Jeffreys' principle (and hence this choice of prior model) is that it respects **parameter invariance** under all monotone transformations. That is, whether one models $\theta \sim g(\theta)$ or some monotone function of θ, say $h(\theta)$, the prior model should be chosen to respect both scales; i.e., if $\theta \sim g(\theta)$, then the prior for $\xi = h(\theta)$ should satisfy

$$g(h^{-1}(\xi)) \left|\frac{d}{d\xi} h^{-1}(\xi)\right|.$$

Jeffreys showed that choosing $g(\theta) \propto [I(\theta)]^{1/2}$ leads to this type of parameter invariance property.

Recall: We have seen the Fisher information before. In STAT 512 (Chapter 9), we learned the Fisher information appears when studying the large-sample properties of maximum likelihood estimators. Back in our classical (i.e., non-Bayesian) paradigm where θ is best regarded as fixed, 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,

$$\frac{\widehat{\theta} - \theta}{\sqrt{\dfrac{v(\theta)}{n}}} \xrightarrow{d} \mathcal{N}(0,1) \quad \Longleftrightarrow \quad \sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

as $n \to \infty$, where

$$v(\theta) = \frac{1}{E\left[-\dfrac{\partial^2}{\partial\theta^2} \ln p_Y(Y|\theta)\right]} \quad \text{(discrete case)}$$

$$v(\theta) = \frac{1}{E\left[-\dfrac{\partial^2}{\partial\theta^2} \ln f_Y(Y|\theta)\right]} \quad \text{(continuous case).}$$

In other words, the "large-sample variance" $v(\theta)$ in asymptotic distributions for MLEs is the reciprocal of the Fisher information.

Example 11.4. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(p) population distribution, where $p \sim g(p)$. Show that applying Jeffreys' principle leads to a beta prior distribution with $\alpha = \beta = 1/2$.

Solution. We first calculate the Fisher information $I(p)$. The pmf of $Y \sim$ Bernoulli(p), where nonzero, is

$$f_Y(y|p) = p^y(1-p)^{1-y} \implies \ln f_Y(y|p) = y \ln p + (1-y)\ln(1-p).$$

The first derivative of $\ln f_Y(y|p)$ is

$$\frac{\partial}{\partial p} \ln f_Y(y|p) = \frac{y}{p} - \frac{1-y}{1-p}.$$

The second derivative of $\ln f_Y(y|p)$ is

$$\frac{\partial^2}{\partial p^2} \ln f_Y(y|p) = -\frac{y}{p^2} - \frac{1-y}{(1-p)^2}.$$

Recalling $E(Y) = p$, the Fisher information is

$$\begin{aligned} I(p) = E\left[-\frac{\partial^2}{\partial p^2} \ln f_Y(Y|p)\right] &= E\left[\frac{Y}{p^2} + \frac{1-Y}{(1-p)^2}\right] \\ &= \frac{E(Y)}{p^2} + \frac{1-E(Y)}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}. \end{aligned}$$

Jeffreys' principle says

$$g(p) \propto [I(p)]^{1/2} = \left[\frac{1}{p(1-p)}\right]^{1/2} = p^{\frac{1}{2}-1}(1-p)^{\frac{1}{2}-1},$$

which we recognize as the kernel of a beta pdf with $\alpha = \beta = 1/2$. Therefore, Jeffreys' prior for p is beta$(1/2, 1/2)$; see Figure 11.6. $\square$

Example 11.5. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(λ) population distribution, where $\lambda \sim g(\lambda)$. Derive Jeffreys' prior distribution for λ.

Solution. We first calculate the Fisher information $I(\lambda)$. The pmf of $Y \sim$ Poisson(λ), where nonzero, is

$$f_Y(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \implies \ln f_Y(y|\lambda) = y \ln \lambda - \lambda - \ln y!.$$

The first derivative of $\ln f_Y(y|\lambda)$ is

$$\frac{\partial}{\partial \lambda} \ln f_Y(y|\lambda) = \frac{y}{\lambda} - 1.$$

The second derivative of $\ln f_Y(y|\lambda)$ is

$$\frac{\partial^2}{\partial \lambda^2} \ln f_Y(y|\lambda) = -\frac{y}{\lambda^2}.$$

Recalling $E(Y) = \lambda$, the Fisher information is

$$I(\lambda) = E\left[-\frac{\partial^2}{\partial \lambda^2} \ln f_Y(Y|\lambda)\right] = E\left(\frac{Y}{\lambda^2}\right) = \frac{E(Y)}{\lambda^2} = \frac{1}{\lambda}.$$

Jeffreys' principle says

$$g(\lambda) \propto [I(\lambda)]^{1/2} = \left(\frac{1}{\lambda}\right)^{1/2} = \frac{1}{\sqrt{\lambda}}.$$

Interesting: Using Jeffreys' principle for prior selection in this problem leads to a prior distribution $g(\lambda)$ which is not even a legitimate pdf! Letting c denote a constant which is free of λ, note that

$$\int_0^\infty g(\lambda) d\lambda = \int_0^\infty \frac{c}{\sqrt{\lambda}}\, d\lambda = 2c\sqrt{\lambda}\,\Big|_0^\infty = +\infty.$$

This is an example of an **improper prior**. What's more interesting is that even though the prior is improper here, the posterior is still proper (i.e., it is a valid pdf). To see why, note that

$$\frac{1}{\sqrt{\lambda}} = \lim_{\beta \to \infty} \lambda^{\frac{1}{2}-1} e^{-\lambda/\beta}.$$

We recognize $\lambda^{\frac{1}{2}-1} e^{-\lambda/\beta}$ as the kernel of a gamma pdf with shape $\alpha = 1/2$ and scale β. Therefore, we can think of $1/\sqrt{\lambda}$ as the kernel of the "gamma$(1/2, \infty)$" distribution. Of

course, this isn't a real distribution, but it is the noninformative prior that arises when using Jeffreys' principle. Now, recall that in the Poisson-gamma problem:

$$\text{Prior: } \lambda \sim \text{gamma}(\alpha, \beta) \longrightarrow \text{Posterior: } \lambda|\mathbf{y} \sim \text{gamma}\left(\sum_{i=1}^{n} y_i + \alpha, \left(n + \frac{1}{\beta}\right)^{-1}\right);$$

see Example 11.2. Letting $\alpha = 1/2$ and "$\beta = \infty$," the posterior becomes

$$\lambda|\mathbf{y} \sim \text{gamma}\left(\sum_{i=1}^{n} y_i + \frac{1}{2}, \frac{1}{n}\right),$$

which is a valid pdf. Therefore, even though the prior distribution is not proper, the posterior still is. □

11.4 Point estimation

Preview: Bayesians usually report point estimates for population-level parameters by using commonly known measures of central tendency from posterior distributions. Because a posterior distribution combines information from the prior distribution $g(\theta)$ and the observed data $\mathbf{y}$ (through the likelihood function), these functionals of $g(\theta|\mathbf{y})$ will also incorporate both sources of information. Bayesian interval estimators and hypothesis tests also use the posterior distribution as we will see in due course.

Background: Bayesian point estimation theory relies on advanced topics like loss functions and decision theory. A rigorous treatment of these topics is slightly beyond the scope of this course. However, it does suffice to note that determining a Bayesian point estimator, say $\delta(\mathbf{Y})$, involves making a decision in light of the penalty (loss) incurred for making a wrong choice. Therefore, it is helpful to think of point estimation as a "decision problem" in the following way:

$$\begin{aligned} \theta &\longleftarrow \text{parameter we want to estimate} \\ \delta(\mathbf{Y}) &\longleftarrow \text{point estimator we will use (choosing which one is a "decision").} \end{aligned}$$

For example, suppose we would like to penalize overestimation of θ and underestimation of θ equally. Two **loss functions** which capture this notion beautifully are

$$\begin{aligned} L_1(\theta, \delta(\mathbf{y})) &= [\theta - \delta(\mathbf{y})]^2 \\ L_2(\theta, \delta(\mathbf{y})) &= |\theta - \delta(\mathbf{y})|. \end{aligned}$$

Both loss functions increase as the distance between θ and $\delta(\mathbf{y})$ increases (and hence a larger "penalty" is incurred), and $L_1(\theta, \delta(\mathbf{y}))$ penalizes larger distances more severely than $L_2(\theta, \delta(\mathbf{y}))$. The Bayesian will report different point estimates for θ under each loss function, as we are now ready to describe.

Terminology: A **Bayesian point estimate** $\delta(\mathbf{y})$ is the "decision" which minimizes the expected loss

$$E[L(\theta, \delta(\mathbf{Y}))] = \int_{\Theta} L(\theta, \delta(\mathbf{y}))g(\theta|\mathbf{y})d\theta,$$

where $g(\theta|\mathbf{y})$ is the posterior distribution. Another way to say this is

$$\delta(\mathbf{y}) = \arg\min \int_{\Theta} L(\theta, \delta(\mathbf{y}))g(\theta|\mathbf{y})d\theta.$$

This definition shows us that Bayesians will use the point estimate $\delta(\mathbf{y})$ which makes the expected (average) loss as small as possible, and the form of the estimate will change depending on what type of loss function is used. The corresponding random version $\delta(\mathbf{Y})$ is called a **Bayesian point estimator**.

Common estimates: We now summarize common Bayesian point estimates and identify which loss functions correspond to each one.

- When squared-error loss $L_1(\theta, \delta(\mathbf{y})) = [\theta - \delta(\mathbf{y})]^2$ is used, the Bayesian point estimate of θ is the **posterior mean**

  $$\delta(\mathbf{y}) = \widehat{\theta}_B = E(\theta|\mathbf{Y} = \mathbf{y}),$$

  i.e., $\widehat{\theta}_B$ is the mean of the posterior distribution.

- When absolute-error loss $L_2(\theta, \delta(\mathbf{y})) = |\theta - \delta(\mathbf{y})|$ is used, the Bayesian point estimate of θ is the **posterior median**

  $$\delta(\mathbf{y}) = \widetilde{\theta}_B = \text{median}(\theta|\mathbf{Y} = \mathbf{y}),$$

  i.e., $\widetilde{\theta}_B$ is the median of the posterior distribution.

- When the loss function is

  $$L_3(\theta, \delta(\mathbf{y})) = \begin{cases} 1, & \theta \neq \delta(\mathbf{y}) \\ 0, & \theta = \delta(\mathbf{y}), \end{cases}$$

  which is also known as the "0-1 loss function," the Bayesian point estimate of θ is the **posterior mode**

  $$\delta(\mathbf{y}) = \widehat{\theta}_B^* = \text{mode}(\theta|\mathbf{Y} = \mathbf{y}),$$

  i.e., $\widehat{\theta}_B^*$ is the mode of the posterior distribution. This is simply the value of θ which maximizes $g(\theta|\mathbf{y})$.

Q: When will the posterior mean, median, and mode be equal?
A: When the posterior distribution $g(\theta|\mathbf{y})$ is symmetric.

Example 11.1 (revisited). Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(p) population, where the prior distribution is $p \sim \text{beta}(\alpha, \beta)$. We have learned that

$$\underbrace{p \sim \text{beta}(\alpha, \beta)}_{\text{prior distribution}} \longrightarrow \text{Observe data } \mathbf{y} \longrightarrow \underbrace{p|\mathbf{y} \sim \text{beta}\left(\sum_{i=1}^n y_i + \alpha,\ n - \sum_{i=1}^n y_i + \beta\right)}_{\text{posterior distribution}}.$$

Under squared-error loss, the Bayesian would report

$$\widehat{p}_B = E(p|\mathbf{Y} = \mathbf{y}) = \frac{\displaystyle\sum_{i=1}^n y_i + \alpha}{\displaystyle\sum_{i=1}^n y_i + \alpha + n - \sum_{i=1}^n y_i + \beta} = \frac{\displaystyle\sum_{i=1}^n y_i + \alpha}{n + \alpha + \beta},$$

which is simply the mean of the beta posterior distribution identified above. An interesting observation in this example is that the posterior mean $\widehat{p}_B$ can be written as

$$\widehat{p}_B = \frac{\sum_{i=1}^n y_i + \alpha}{n + \alpha + \beta} = \left(\frac{n}{n + \alpha + \beta}\right)\frac{\sum_{i=1}^n y_i}{n} + \left(\frac{\alpha + \beta}{n + \alpha + \beta}\right)\frac{\alpha}{\alpha + \beta},$$

a weighted average of the MLE $\overline{y} = \sum_{i=1}^n y_i/n$ and the prior mean $\alpha/(\alpha + \beta)$. Note that when the sample size n is large, the prior mean receives less weight in its contribution to posterior mean $\widehat{p}_B$. This makes sense intuitively. When the sample size is large, we should weight the MLE more and the prior mean less. The opposite is true when n is small; the prior mean will then play a larger role in determining the value of the posterior mean $\widehat{p}_B$.

Application: Among prematurely born infants at Richland Hospital, researchers would like to estimate p, the probability of developing necrotizing enterocolitis (NEC) for a "high-risk" group ($<$ 1500 g birth weight and $<$ 32 weeks gestational age). Over a 6-month period, there were $n = 37$ infants who were classified as high risk. Denote the NEC statuses (1/0) by $Y_1, Y_2, ..., Y_{37}$ and assume these are iid Bernoulli(p), where p is modeled noninformatively as $p \sim \text{beta}(1/2, 1/2)$. Among the 37 infants,

$$t = \sum_{i=1}^{37} y_i = 9$$

developed NEC during their stay in the neonatal intensive care unit. The posterior distribution based on these data is

$$p|\mathbf{y} \sim \text{beta}\left(\sum_{i=1}^{37} y_i + \frac{1}{2},\ 37 - \sum_{i=1}^{37} y_i + \frac{1}{2}\right) \implies p|\mathbf{y} \sim \text{beta}(9.5, 28.5)$$

and is shown in Figure 11.8 (next page). The posterior mean is

$$\widehat{p}_B = E(p|\mathbf{Y} = \mathbf{y}) = \frac{9.5}{9.5 + 28.5} = 0.25.$$

The posterior median is

$$\widetilde{p}_B = \text{median}(p|\mathbf{Y} = \mathbf{y}) \approx 0.246.$$

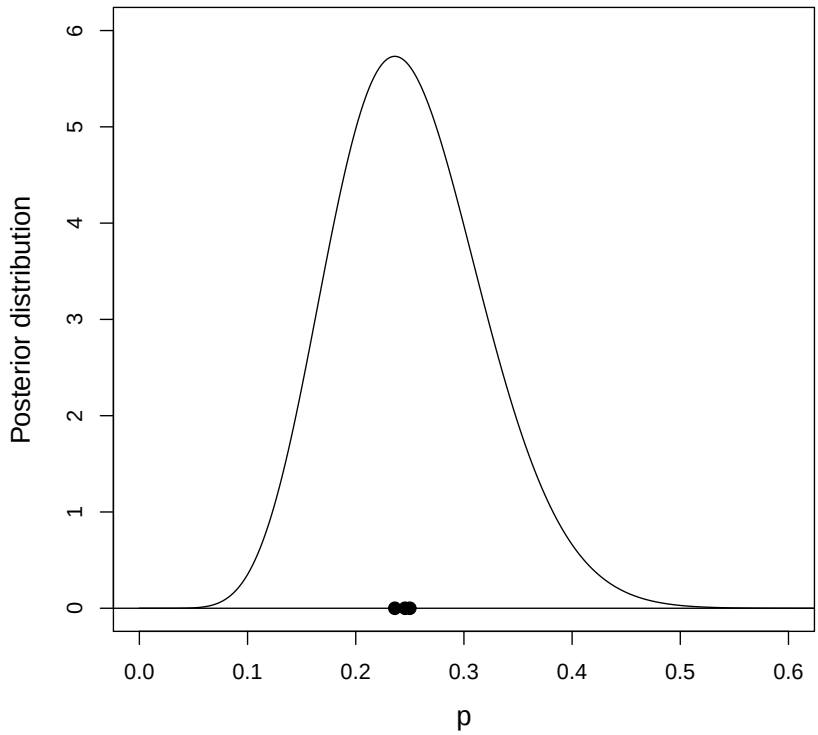

Figure 11.8: Premature infant data. Posterior distribution $p|\mathbf{y} \sim \text{beta}(9.5, 28.5)$. The posterior mean, median, and mode are shown by using dark circles.

```
> qbeta(0.5,9.5,28.5)
[1] 0.2455784
```

To find the posterior mode, we need to maximize the posterior $g(p|\mathbf{y})$ as a function of p. Note that

$$g(p|\mathbf{y}) = \frac{\Gamma(38)}{\Gamma(9.5)\Gamma(28.5)}\, p^{9.5-1}(1-p)^{28.5-1} = cp^{8.5}(1-p)^{27.5},$$

where c is a constant free of p. The value of p which maximizes $g(p|\mathbf{y})$ is the same as the value of p which maximizes $\ln g(p|\mathbf{y})$. We have

$$\ln g(p|\mathbf{y}) = \ln c + 8.5 \ln p + 27.5 \ln(1-p) \implies \frac{d}{dp} \ln g(p|\mathbf{y}) = \frac{8.5}{p} - \frac{27.5}{1-p}.$$

Setting $(d/dp) \ln g(p|\mathbf{y})$ equal to 0 and solving for p gives

$$\widehat{p}_B^* = \text{mode}(p|\mathbf{Y} = \mathbf{y}) = \frac{8.5}{36} \approx 0.236.$$

Figure 11.8 (above) shows the locations of the posterior mean, median, and mode in this example. These point estimates are similar because the posterior distribution $g(p|\mathbf{y})$ is fairly symmetrical in shape. $\square$

11.5 Interval estimation

Recall: 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.,

$$P(\theta_L < \theta < \theta_U) = 1 - \alpha.$$

In the classical (i.e., non-Bayesian) paradigm, a $1-\alpha$ interval estimator is also called a $100(1-\alpha)\%$ **confidence interval**. Because the non-Bayesian regards θ to be fixed, we remember it is the endpoints θ_L and θ_U that are random in the probability equation above−not θ.

Illustration: If $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma_0^2)$ population, where σ_0^2 is known, we can derive a confidence interval for μ by using the pivotal quantity

$$Z = \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}(0,1).$$

This leads to

$$1 - \alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P\Bigg(\underbrace{\overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}}_{\mu_L} < \mu < \underbrace{\overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}}_{\mu_U}\Bigg)$$

so that

$$\left(\overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}},\ \overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for μ.

Q: What does the term "confidence" really mean?
A: This is a term widely used in statistical inference but is widely misunderstood. The interval for μ above is a **random** interval; i.e., the endpoints depend on the random variable $\overline{Y}$ so the endpoints are random. Therefore, the statement

$$1 - \alpha = P\left(\overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} < \mu < \overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right)$$

makes perfect sense, and I prefer to use the term "interval estimator" to reinforce this point. Where things start to get murky in the interpretation is when we observe a *realization* of this random interval; e.g.,

$$\left(0.54 - 1.96 \times \frac{7.3}{\sqrt{20}},\ 0.54 + 1.96 \times \frac{7.3}{\sqrt{20}}\right) \longrightarrow (-2.66, 3.74).$$

Now, the interval $(-2.66, 3.74)$ is no longer random so the probability equation

$$0.95 = P\left(-2.66 < \mu < 3.74\right)$$

no longer makes sense. Therefore, in an attempt to explain what the realized interval $(-2.66, 3.74)$ means, we are forced to dream up the following hypothetical scenario. Suppose

we took many iid $\mathcal{N}(\mu, \sigma_0^2 = 7.3)$ samples, each one of size $n = 20$, and we calculated the 95% confidence interval estimate

$$\left(\overline{y} - 1.96 \times \frac{7.3}{\sqrt{20}},\ \overline{y} + 1.96 \times \frac{7.3}{\sqrt{20}}\right)$$

with each sample. Then, acknowledging the variation in the sampling distribution of $\overline{Y}$, we would expect 95% of the intervals in this collection to contain the population parameter μ. The one we calculated, $(-2.66, 3.74)$, is one of these possible intervals. We say "95% confident," but the confidence coefficient 95% refers to the *long-run percentage* of the intervals that would contain μ, noting that in different iid samples we would get different values of $\overline{y}$. Interestingly, the coefficient 95% has little or nothing to do with the interval we calculated, and, of course, we never get to know if our interval $(-2.66, 3.74)$ contains μ or not.

Note: We now describe how Bayesians do interval estimation. Their approach is far simpler, and it leads to easy interpretation as we will now see.

Terminology: If $g(\theta|\mathbf{y})$ is the posterior pdf of θ (a scalar parameter), the **credible probability** of the interval $A = (\theta_L, \theta_U)$ is

$$P(\theta_L < \theta < \theta_U|\mathbf{y}) = \int_{\theta_L}^{\theta_U} g(\theta|\mathbf{y})d\theta.$$

If $g(\theta|\mathbf{y})$ is a discrete pmf, then the integral above is replaced with a sum. If the interval $A = (\theta_L, \theta_U)$ has credible probability equal to $1 - \alpha$, that is,

$$1 - \alpha = P(\theta_L < \theta < \theta_U|\mathbf{y}),$$

we call A a $100(1 - \alpha)\%$ **credible interval** for θ. Another name for "credible interval" is "posterior probability interval."

Interpretation: Because $g(\theta|\mathbf{y})$ is a valid pdf (pmf) of θ, the interpretation of a $100(1-\alpha)\%$ credible interval is strikingly simple:

"The probability θ is between θ_L and θ_U is $1 - \alpha$."

This is far simpler than the interpretation of a classical confidence interval. Of course, the Bayesian requires an elicitation of the prior distribution $\theta \sim g(\theta)$, whereas the classical statistician does not.

Construction: A $100(1 - \alpha)\%$ credible interval for θ is *any* interval $A = (\theta_L, \theta_U)$ with credible probability equal to $1 - \alpha$. Here are two common ways to construct them:

- *Equal-tail.* Simply take the endpoints θ_L and θ_U to be the lower and upper $\alpha/2$ quantiles of $g(\theta|\mathbf{y})$, respectively. Clearly, these choices satisfy

$$1 - \alpha = P(\theta_L < \theta < \theta_U|\mathbf{y}) = \int_{\theta_L}^{\theta_U} g(\theta|\mathbf{y})d\theta.$$

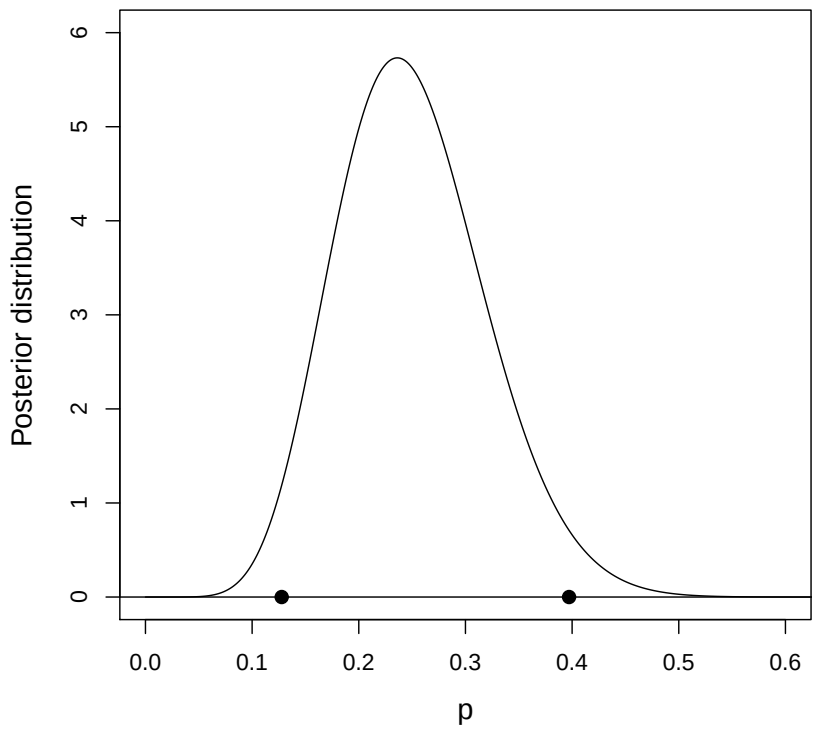

Figure 11.9: Premature infant data. Posterior distribution $p|\mathbf{y} \sim \text{beta}(9.5, 28.5)$. The lower and upper 0.025 quantiles are shown by using dark circles.

- *Highest posterior density.* Select A to be the interval so that the posterior pdf $g(\theta|\mathbf{y})$ within the region A is never lower than outside A. This method will produce an interval consisting of the "most likely" values of θ (as determined by the posterior distribution).

Remark: These two approaches will provide the same interval when the posterior distribution $g(\theta|\mathbf{y})$ is **symmetric**. Even though highest posterior density intervals seem more sophisticated, equal-tail intervals are often preferred in practice because they are easier to compute; e.g., a 95% equal-tail credible interval for θ is formed simply by taking the 0.025 and 0.975 quantiles of $g(\theta|\mathbf{y})$.

Illustration: In the premature infant study, recall the probability p of developing NEC among high-risk infants was modeled as

$$\text{Prior: } p \sim \text{beta}(1/2, 1/2) \;\longrightarrow\; \text{Posterior: } p|\mathbf{y} \sim \text{beta}(9.5, 28.5).$$

The posterior distribution $g(p|\mathbf{y})$ is shown in Figure 11.9 (above). A 95% equal-tail credible interval for p is $(0.128, 0.397)$, which is formed by identifying the 0.025 and 0.975 quantiles of the beta$(9.5, 28.5)$ distribution. Therefore, the population proportion p of high-risk infants developing NEC is between 0.128 and 0.397 with probability 0.95.

```
> qbeta(0.025,9.5,28.5)
[1] 0.1277291
> qbeta(0.975,9.5,28.5)
[1] 0.39715
```

HPD interval: I used the R package `HDInterval` to construct a 95% highest posterior density interval for p:

```
> post = rbeta(1e6,9.5,28.5)
> hdi(post,credMass=0.95)
    lower     upper
0.1211636 0.3882725
```

It is not surprising both methods produce very similar answers as the posterior distribution $g(p|\mathbf{y})$ is fairly symmetric; see Figure 11.9. $\square$

Example 11.6. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma_0^2)$ population, where σ_0^2 is known. In turn, suppose the population mean μ is modeled *a priori* using a conjugate $\mathcal{N}(\eta, \delta^2)$ distribution, where the hyperparameters η and δ^2 are known. Derive an equal-tail $100(1-\alpha)\%$ credible interval for μ. Note this will be the Bayesian analogue of the $100(1-\alpha)\%$ confidence interval for μ we discussed at the beginning of this subsection.

Solution. The first step is to derive the posterior distribution. It is easy to show

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

is a sufficient statistic in the $\mathcal{N}(\mu, \sigma_0^2)$ family (when σ_0^2 is known). Therefore, we know immediately the posterior distribution

$$g(\mu|t) \propto f_{T|\mu}(t|\mu)g(\mu),$$

where $f_{T|\mu}(t|\mu)$ is the pdf corresponding to the (sampling) distribution of $T = \overline{Y}$, $t = \overline{y}$, and $g(\mu)$ is the $\mathcal{N}(\eta, \delta^2)$ prior pdf. Recall that

$$T \sim \mathcal{N}\left(\mu, \frac{\sigma_0^2}{n}\right) \implies f_{T|\mu}(t|\mu) = \frac{1}{\sqrt{2\pi(\sigma_0^2/n)}} e^{-(t-\mu)^2/2(\sigma_0^2/n)},$$

and the $\mathcal{N}(\eta, \delta^2)$ prior pdf is

$$g(\mu) = \frac{1}{\sqrt{2\pi\delta^2}} e^{-(\mu-\eta)^2/2\delta^2}.$$

Therefore,

$$g(\mu|t) \propto \frac{1}{\sqrt{2\pi(\sigma_0^2/n)}} e^{-(t-\mu)^2/2(\sigma_0^2/n)} \times \frac{1}{\sqrt{2\pi\delta^2}} e^{-(\mu-\eta)^2/2\delta^2} \propto \exp\left\{-\left[\frac{(t-\mu)^2}{2(\sigma_0^2/n)} + \frac{(\mu-\eta)^2}{2\delta^2}\right]\right\}.$$

The tedious part is now working algebraically with

$$\exp\left\{-\left[\frac{(t-\mu)^2}{2(\sigma_0^2/n)} + \frac{(\mu-\eta)^2}{2\delta^2}\right]\right\}$$

which, when viewed as a function of μ, I claim is proportional to

$$\exp\left\{-\frac{(\mu - \text{something})^2}{2 \times \text{something}}\right\}.$$

Therefore, all we have to do is figure out what these "something's" are. If we can, then we know the posterior distribution is proportional to a normal kernel with these quantities as the (updated) mean and variance, and hence we have the posterior distribution. Expanding the squares and getting a common denominator, we have

$$\begin{aligned}
\frac{(t-\mu)^2}{2(\sigma_0^2/n)} + \frac{(\mu-\eta)^2}{2\delta^2} &= \frac{(t^2 - 2\mu t + \mu^2)\delta^2 + (\mu^2 - 2\mu\eta + \eta^2)(\sigma_0^2/n)}{2(\sigma_0^2/n)\delta^2} \\
&= \frac{[\delta^2 + (\sigma_0^2/n)]\mu^2 - 2[\delta^2 t + (\sigma_0^2/n)\eta]\mu + [\delta^2 t^2 + (\sigma_0^2/n)\eta^2]}{2(\sigma_0^2/n)\delta^2}.
\end{aligned}$$

Dividing the numerator and denominator by the leading coefficient $[\delta^2 + (\sigma_0^2/n)]$, the last expression equals

$$\frac{\mu^2 - 2\left[\dfrac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]\mu + \left[\dfrac{\delta^2 t^2 + (\sigma_0^2/n)\eta^2}{\delta^2 + (\sigma_0^2/n)}\right]}{\dfrac{2(\sigma_0^2/n)\delta^2}{\delta^2 + (\sigma_0^2/n)}}$$

$$= \frac{\mu^2 - 2\left[\dfrac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]\mu}{\dfrac{2(\sigma_0^2/n)\delta^2}{\delta^2 + (\sigma_0^2/n)}} + \underbrace{\frac{\dfrac{\delta^2 t^2 + (\sigma_0^2/n)\eta^2}{\delta^2 + (\sigma_0^2/n)}}{\dfrac{2(\sigma_0^2/n)\delta^2}{\delta^2 + (\sigma_0^2/n)}}}_{\text{constant free of } \mu}.$$

Completing the square in the numerator of the first term, we have

$$\begin{aligned}
\mu^2 - 2\left[\frac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]\mu &= \underbrace{\mu^2 - 2\left[\frac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]\mu + \left[\frac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]^2}_{\text{perfect square}} - \left[\frac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]^2 \\
&= \left(\mu - \left[\frac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]\right)^2 - \left[\frac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]^2.
\end{aligned}$$

Therefore, we have shown

$$\frac{(t-\mu)^2}{2(\sigma_0^2/n)} + \frac{(\mu-\eta)^2}{2\delta^2} = \frac{\left(\mu - \left[\dfrac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]\right)^2}{\dfrac{2(\sigma_0^2/n)\delta^2}{\delta^2 + (\sigma_0^2/n)}} \underbrace{- \frac{\left[\dfrac{\delta^2 t + (\sigma_0^2/n)\eta}{\delta^2 + (\sigma_0^2/n)}\right]^2}{\dfrac{2(\sigma_0^2/n)\delta^2}{\delta^2 + (\sigma_0^2/n)}} + \frac{\dfrac{\delta^2 t^2 + (\sigma_0^2/n)\eta^2}{\delta^2 + (\sigma_0^2/n)}}{\dfrac{2(\sigma_0^2/n)\delta^2}{\delta^2 + (\sigma_0^2/n)}}}_{\text{constant free of } \mu\text{; call this } c}$$

and hence

$$\exp\left\{-\left[\frac{(t-\mu)^2}{2(\sigma_0^2/n)}+\frac{(\mu-\eta)^2}{2\delta^2}\right]\right\} = \exp\left\{-\frac{\left(\mu-\left[\frac{\delta^2 t+(\sigma_0^2/n)\eta}{\delta^2+(\sigma_0^2/n)}\right]\right)^2}{\frac{2(\sigma_0^2/n)\delta^2}{\delta^2+(\sigma_0^2/n)}}\right\}\exp(-c)$$

$$\propto \exp\left\{-\frac{\left(\mu-\left[\frac{\delta^2 t+(\sigma_0^2/n)\eta}{\delta^2+(\sigma_0^2/n)}\right]\right)^2}{\frac{2(\sigma_0^2/n)\delta^2}{\delta^2+(\sigma_0^2/n)}}\right\}.$$

We have shown the posterior distribution $g(\mu|t)$ is proportional to a normal kernel with mean

$$\frac{\delta^2 t+(\sigma_0^2/n)\eta}{\delta^2+(\sigma_0^2/n)}$$

and variance

$$\frac{(\sigma_0^2/n)\delta^2}{\delta^2+(\sigma_0^2/n)}.$$

Therefore, we have shown

$$\text{Prior: } \mu \sim \mathcal{N}(\eta,\delta^2) \;\longrightarrow\; \text{Posterior: } \mu|t \sim \mathcal{N}\left(\frac{\delta^2 t+(\sigma_0^2/n)\eta}{\delta^2+(\sigma_0^2/n)},\ \frac{(\sigma_0^2/n)\delta^2}{\delta^2+(\sigma_0^2/n)}\right),$$

where $t=\overline{y}$. Having just derived the posterior distribution, it is interesting to note the posterior mean (i.e., the Bayesian point estimate under squared-error loss) can be written as

$$\frac{\delta^2 t+(\sigma_0^2/n)\eta}{\delta^2+(\sigma_0^2/n)} = \left[\frac{\delta^2}{\delta^2+(\sigma_0^2/n)}\right]t + \left[\frac{\sigma_0^2/n}{\delta^2+(\sigma_0^2/n)}\right]\eta,$$

a weighted average of the MLE $t=\overline{y}$ and the prior mean η. An $100(1-\alpha)\%$ equal-tail credible interval for μ is formed by selecting the lower and upper $\alpha/2$ quantiles of the posterior distribution $g(\mu|t)$. Because $g(\mu|t)$ is a normal pdf, this interval is

$$\left(\frac{\delta^2 t+(\sigma_0^2/n)\eta}{\delta^2+(\sigma_0^2/n)} - z_{\alpha/2}\sqrt{\frac{(\sigma_0^2/n)\delta^2}{\delta^2+(\sigma_0^2/n)}},\ \frac{\delta^2 t+(\sigma_0^2/n)\eta}{\delta^2+(\sigma_0^2/n)} + z_{\alpha/2}\sqrt{\frac{(\sigma_0^2/n)\delta^2}{\delta^2+(\sigma_0^2/n)}}\right).$$

On the basis of the observed data (through the sufficient statistic $t=\overline{y}$) and the $\mathcal{N}(\eta,\delta^2)$ prior model, this interval contains μ with probability $1-\alpha$. $\square$

11.6 Hypothesis testing

Remark: Our treatment of hypothesis testing under the Bayesian paradigm will be far less formal than it was in Chapter 10 under the classical one. There is a good reason for this. Most

Bayesians do not endorse the role hypothesis testing plays in statistical inference. Instead, on a practical level, they are much happier simply summarizing posterior distributions with (Bayesian) point estimates and interval estimates. More generally, Bayesians are usually not interested in population-level parameters themselves but rather the statistical models that contain them. Therefore, Bayesians are more focused on comparing one or more statistical models between or among each other, and they use different criteria to do this.

Bayesian tests: There is certainly nothing to prevent us from framing a hypothesis testing problem within the Bayesian paradigm. Suppose we are interested in testing

$$\begin{gathered} H_0 : \theta \in \Theta_0 \\ \text{versus} \\ H_a : \theta \in \Theta_a, \end{gathered}$$

where $\Theta = \Theta_0 \cup \Theta_a$. As we have already learned, all inference for the Bayesian flows from using the posterior distribution $g(\theta|\mathbf{y})$. This is a valid probability distribution, so the probabilities

$$P(H_0 \text{ is true}|\mathbf{y}) = P(\theta \in \Theta_0|\mathbf{y}) = \int_{\Theta_0} g(\theta|\mathbf{y})d\theta$$

and

$$P(H_a \text{ is true}|\mathbf{y}) = P(\theta \in \Theta_a|\mathbf{y}) = \int_{\Theta_a} g(\theta|\mathbf{y})d\theta$$

make perfect sense and can be computed directly. As for a decision rule, one can simply choose to reject H_0 when

$$P(\theta \in \Theta_0|\mathbf{y}) < P(\theta \in \Theta_a|\mathbf{y}),$$

that is, when the probability of H_0 is less than that of H_a. One could use a more stringent criterion; e.g., rejecting H_0 when the posterior probability $P(\theta \in \Theta_0|\mathbf{y})$ is small, say less than 0.01 or 0.05.

Remark: It is worth emphasizing that statements like

$$P(H_0 \text{ is true}|\mathbf{y}) \;\text{ and }\; P(H_a \text{ is true}|\mathbf{y})$$

make absolutely no sense in the classical hypothesis testing framework described in the last chapter. To the classical statistician, the population-level parameter θ is best regarded as fixed, so $\{H_0 \text{ is true}\}$ and $\{H_a \text{ is true}\}$ are not even random events to which probability can be assigned. In a casual sense, these probabilities are either 0 or 1; e.g.,

$$\begin{aligned} &\text{if } \theta \in \Theta_0 \implies P(H_0 \text{ is true}|\mathbf{y}) = 1 \;\text{ and }\; P(H_a \text{ is true}|\mathbf{y}) = 0 \\ &\text{if } \theta \in \Theta_a \implies P(H_0 \text{ is true}|\mathbf{y}) = 0 \;\text{ and }\; P(H_a \text{ is true}|\mathbf{y}) = 1. \end{aligned}$$

The problem is one never gets to know which probability is which. This is why the classical framework introduces concepts like "Type I Error" and "Type II Error" so that one can quantify the chance that certain errors will be made. However, these probabilities are calculated from the sampling distribution of a test statistic T. The test statistic is a random variable because it depends on the sample $Y_1, Y_2, ..., Y_n$.

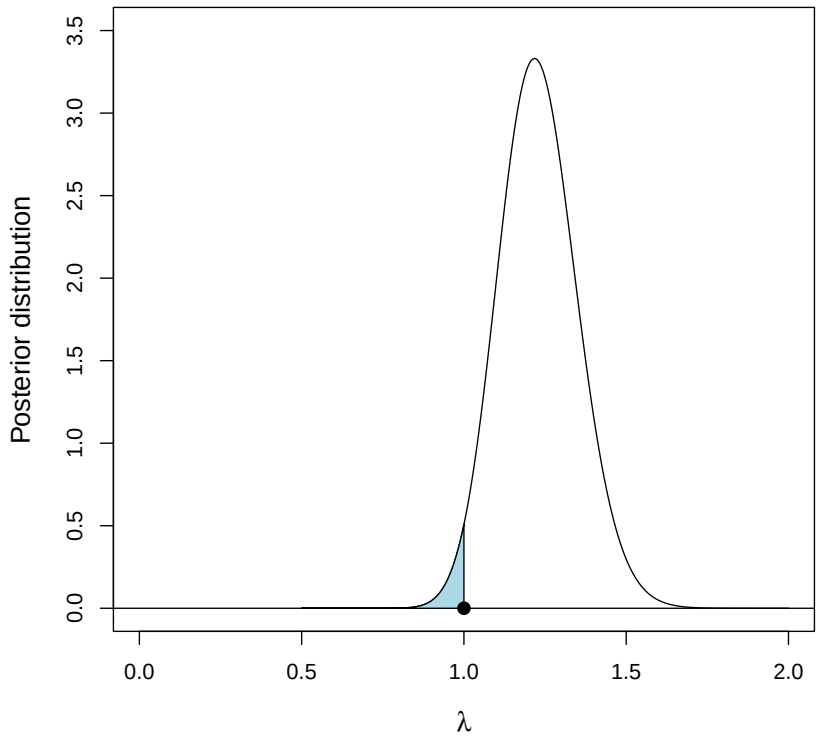

Figure 11.10: Accident data. Posterior distribution $g(\lambda|\mathbf{y})$ for the mean number of accidents per year. The posterior probability $P(\lambda \leq 1|\mathbf{y})$ is shown shaded.

Illustration: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(λ) distribution, where $\lambda \sim$ gamma(α, β). In Example 11.2 (notes, pp 76-77), we used a Poisson distribution to model the number of accidents per year for a sample of $n = 84$ policies, and we regarded the population mean λ to be random with a gamma(1.5, 1) prior distribution. On the basis of the observed data (103 total accidents), we determined

$$\text{Prior: } \lambda \sim \text{gamma}(1.5, 1) \longrightarrow \text{Posterior: } \lambda|\mathbf{y} \sim \text{gamma}(104.5, 1/85).$$

Suppose we would like to perform a Bayesian test for

$$\begin{array}{c} H_0 : \lambda \leq 1 \\ \text{versus} \\ H_a : \lambda > 1. \end{array}$$

From the gamma(104.5, 1/85) posterior distribution (see Figure 11.10 above), we calculate

$$P(H_0 \text{ is true}|\mathbf{y}) = P(\lambda \leq 1|\mathbf{y}) \approx 0.022.$$

Therefore, it is unlikely H_0 is true. The posterior evidence highly favors H_a.

```
> pgamma(1,104.5,85)
[1] 0.02232586
```

12 Linear Models

12.1 Introduction

Discussion: A problem that often arises in economics, engineering, medicine, and other areas is that of investigating the mathematical relationship between two (or more) variables. The goal is often to model a continuous random variable Y as a function of one or more independent variables, say, $x_1, x_2, ..., x_k$. One can express this model as

$$Y = g(x_1, x_2, ..., x_k) + \epsilon,$$

where $g : \mathbb{R}^k \to \mathbb{R}$ and ϵ is a random error term that satisfies certain conditions. This is called a **regression model**.

- The presence of the error term ϵ conveys that the relationship between Y and the independent variables through $g(x_1, x_2, ..., x_k)$ is likely not perfect (if it was perfect, this would be a deterministic model).
- The independent variables $x_1, x_2, ..., x_k$ are assumed to be fixed (not random), and they are measured without error.

There are different types of regression models. A **nonparametric model** would leave the form of g unspecified, essentially regarding the relationship between Y and the independent variables $x_1, x_2, ..., x_k$ to be characterized by some function. A **parametric model** would dictate the specific form of g, for example,

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}_{g(x_1, x_2, ..., x_k)} + \epsilon,$$

where $\beta_0, \beta_1, ..., \beta_k$ are unknown regression parameters. This is called a **linear regression model**. The adjective "linear" does not refer to the shape of $g(x_1, x_2, ..., x_k)$. Instead, it refers to the manner in which the regression parameters β_0, β_1, ..., β_k appear in the g function. With the g function above, note that

$$\begin{aligned}
\frac{\partial g(x_1, x_2, ..., x_k)}{\partial \beta_0} &= 1 \\
\frac{\partial g(x_1, x_2, ..., x_k)}{\partial \beta_1} &= x_1 \\
&\vdots \\
\frac{\partial g(x_1, x_2, ..., x_k)}{\partial \beta_k} &= x_k.
\end{aligned}$$

All of these partial derivatives are free of $\beta_0, \beta_1, ..., \beta_k$, meaning that g is a linear function of the regression parameters. With this definition in mind, we see that all of the following

models are linear in the regression parameters:

$$\begin{aligned}
Y &= \underbrace{\beta_0 + \beta_1 x}_{g(x)} + \epsilon \\
Y &= \underbrace{\beta_0 + \beta_1 x + \beta_2 x^2}_{g(x)} + \epsilon \\
Y &= \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2}_{g(x_1,x_2)} + \epsilon \\
Y &= \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}_{g(x_1,x_2)} + \epsilon.
\end{aligned}$$

These are all examples of linear regression models. This is true even though only two of these models are linear functions of the independent variables (the first one and the third one). An example of a **nonlinear model** is

$$Y = \underbrace{\frac{\beta_0}{1 + \beta_1 e^{\beta_2 x}}}_{g(x)} + \epsilon.$$

This model is not linear in its parameters. For example,

$$\frac{\partial g(x)}{\partial \beta_0} = \frac{1}{1 + \beta_1 e^{\beta_2 x}},$$

which is not free of β_1 and β_2.

Preview: This chapter is about **linear models**, which includes linear regression models and other types of linear models (e.g., ANOVA models, etc.). We will start by studying the underlying theory of simple linear regression (one independent variable) and then move to multiple linear regression (multiple independent variables). Multiple linear regression models are best presented by using vector and matrix notation.

12.2 Simple linear regression

Terminology: A **simple linear regression model** is of the form

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

If $E(\epsilon) = 0$, then

$$E(Y) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x.$$

Therefore, the (population-level) regression parameter β_1 quantifies the change in $E(Y)$ brought about by a one-unit change in x. The (population-level) regression parameter β_0 represents the mean of Y when the independent variable $x = 0$.

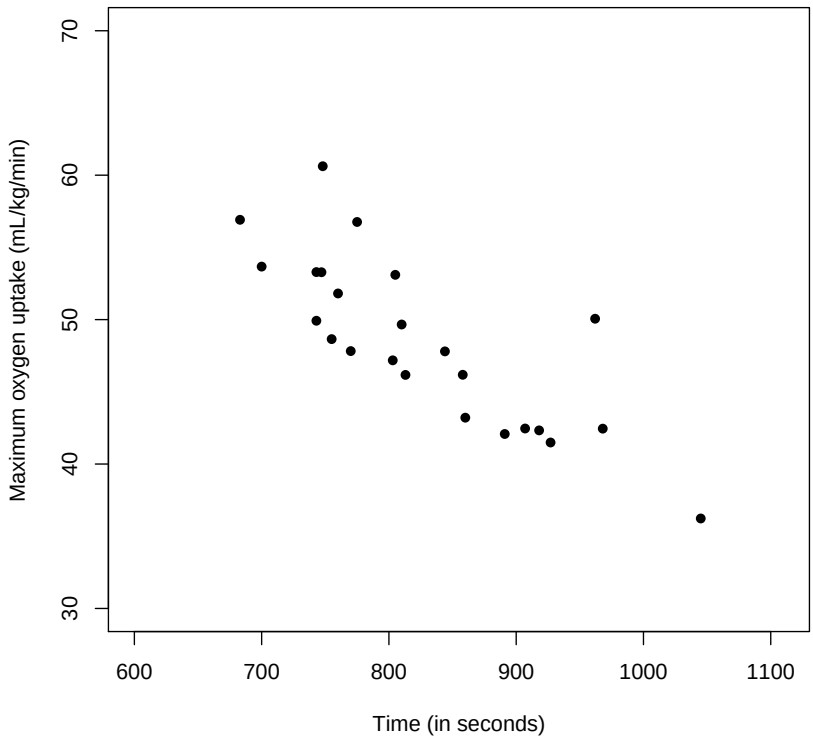

Figure 12.1: Exercise data. Scatterplot of $n = 24$ observations from middle-aged men. The response variable is Y = maximum O_2 uptake. The independent variable is x = time.

Example 12.1. A study was conducted involving a random sample of $n = 24$ middle-aged men to determine the relationship between maximum oxygen uptake (Y, measured in mL/kg/min) and the time required to complete a two-mile run (x, measured in seconds). Maximum oxygen uptake was measured with standard laboratory methods as the subjects performed on a treadmill. Here are the data from the study:

Max O_2	Time	Max O_2	Time	Max O_2	Time
42.33	918	36.23	1045	53.29	743
53.10	805	49.66	810	47.18	803
42.08	891	41.49	927	56.91	683
50.06	962	46.17	813	47.80	844
42.45	968	46.18	858	48.65	755
42.46	907	43.21	860	53.67	700
47.82	770	51.81	760	60.62	748
49.92	743	53.28	747	56.76	775

A scatterplot of the observations is shown in Figure 12.1 (above). Based on the empirical evidence in this figure, a simple linear regression model (for the population of "middle-aged" men) seems appropriate.

12.2.1 Estimation and sampling distributions

Terminology: When we say "fit a model" or "estimate a model," we mean we would like to estimate the population-level model parameters (e.g., β_0 and β_1) with the observed data. Suppose we observe a random sample of individuals from a larger population and the pairs

$$(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$$

are obtained which follow

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

for $i = 1, 2, ..., n$. Our first goal is to estimate β_0 and β_1. Formal assumptions for the error terms ϵ_i will be needed when we investigate sampling distributions of the estimators.

Side note: Later in this chapter we will learn that linear models can be written more generally as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

The simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, ..., n,$$

is a special case in this class of models with

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta}_{2\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\epsilon}_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Estimation: A widely accepted method of estimating the population parameters β_0 and β_1 is to use least squares, which says to choose the values of β_0 and β_1 that minimize the objective function

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Denote the least squares estimators by $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively. These are the values of β_0 and β_1 that minimize $Q(\beta_0, \beta_1)$. Taking partial derivatives of $Q(\beta_0, \beta_1)$, we obtain

$$\begin{aligned} \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) \stackrel{\text{set}}{=} 0 \\ \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)x_i \stackrel{\text{set}}{=} 0. \end{aligned}$$

Solving for β_0 and β_1 gives

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}$$

and

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

These are the **least-squares estimators** of β_0 and β_1, respectively.

Assumptions: We wish to investigate the sampling properties of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as estimators of β_0 and β_1 in the simple linear regression model

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

for $i = 1, 2, ..., n$. To do this thoroughly, we will assume $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. This means

- $E(\epsilon_i) = 0$
- $V(\epsilon_i) = \sigma^2$, that is, the variance is constant
- the random variables ϵ_i are independent
- the random variables ϵ_i are normally distributed.

Remark: Under the assumption that $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, it is easy to see

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

This means

- $E(Y_i) = \beta_0 + \beta_1 x_i$, for $i = 1, 2, ..., n$
- $V(Y_i) = \sigma^2$, that is, the variance is constant
- the random variables Y_i are independent (because they are functions of ϵ_i)
- the random variables Y_i are normally distributed.

Fact 1. The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively, that is,

$$\begin{aligned} E(\widehat{\beta}_0) &= \beta_0 \\ E(\widehat{\beta}_1) &= \beta_1. \end{aligned}$$

Proof. Algebraically,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

because

$$\begin{aligned} \sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y}) &= \sum_{i=1}^n (x_i - \overline{x})Y_i - \sum_{i=1}^n (x_i - \overline{x})\overline{Y} \\ &= \sum_{i=1}^n (x_i - \overline{x})Y_i - \overline{Y}\sum_{i=1}^n (x_i - \overline{x}) \end{aligned}$$

and $\sum_{i=1}^n (x_i - \overline{x}) = 0$. Therefore, if we let

$$c_i = \frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

for $i = 1, 2, ..., n$, we see the least-squares slope estimator $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = \sum_{i=1}^n c_i Y_i,$$

a **linear combination** of $Y_1, Y_2, ..., Y_n$. Taking expectations, we have

$$E(\widehat{\beta}_1) = E\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n c_i E(Y_i) = \sum_{i=1}^n c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i.$$

However, note that

$$\sum_{i=1}^n c_i = \sum_{i=1}^n \left[\frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] = \frac{\sum_{i=1}^n (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = 0$$

and

$$\sum_{i=1}^n c_i x_i = \sum_{i=1}^n \left[\frac{(x_i - \overline{x})x_i}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = 1.$$

Therefore, $E(\widehat{\beta}_1) = \beta_1$ as claimed. To show $\widehat{\beta}_0$ is unbiased, we first note

$$E(\widehat{\beta}_0) = E(\overline{Y} - \widehat{\beta}_1 \overline{x}) = E(\overline{Y}) - \overline{x}E(\widehat{\beta}_1).$$

However, $E(\widehat{\beta}_1) = \beta_1$ and

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \frac{1}{n}\sum_{i=1}^n E(Y_i) = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i) = \frac{1}{n}\sum_{i=1}^n \beta_0 + \frac{1}{n}\sum_{i=1}^n \beta_1 x_i = \beta_0 + \beta_1 \overline{x}.$$

Therefore,

$$E(\widehat{\beta}_0) = E(\overline{Y}) - \overline{x}E(\widehat{\beta}_1) = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0,$$

as claimed. $\square$

Important: The only assumption we used in the preceding argument (to establish unbiasedness) was that $E(\epsilon_i) = 0$, for $i = 1, 2, ..., n$. This is a sufficient condition. The other three assumptions (constant variance, independence, normality) are not needed to establish unbiasedness.

Fact 2. The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have the following characteristics:

$$\begin{aligned}
V(\widehat{\beta}_0) &= \sigma^2 \left[\frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2}\right] \\
V(\widehat{\beta}_1) &= \sigma^2 \left[\frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \\
\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) &= \sigma^2 \left[\frac{-\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}\right].
\end{aligned}$$

Important: To derive these expressions, we will use the second and third assumptions, that is, $V(\epsilon_i) = \sigma^2$ and ϵ_i independent. The fourth assumption (normality) is not needed.

Proof. Recall that $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i,$$

where the constant

$$c_i = \frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

for $i = 1, 2, ..., n$. Therefore,

$$\begin{aligned} V(\widehat{\beta}_1) = V\left(\sum_{i=1}^{n} c_i Y_i\right) &= \sum_{i=1}^{n} c_i^2 V(Y_i) \\ &= \sigma^2 \sum_{i=1}^{n} \left[\frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]^2 \\ &= \frac{\sigma^2}{[\sum_{i=1}^{n}(x_i - \overline{x})^2]^2}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = \sigma^2\left[\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right], \end{aligned}$$

as claimed. The variance of $\widehat{\beta}_0$ is

$$V(\widehat{\beta}_0) = V(\overline{Y} - \widehat{\beta}_1\overline{x}) = V(\overline{Y}) + \overline{x}^2 V(\widehat{\beta}_1) - 2\overline{x}\text{Cov}(\overline{Y}, \widehat{\beta}_1).$$

Note that

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

Also,

$$\begin{aligned} \text{Cov}(\overline{Y}, \widehat{\beta}_1) &= \text{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} c_i Y_i\right) \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}\text{Cov}(Y_i, c_i Y_i) + \sum_{i \neq j}\text{Cov}(Y_i, c_j Y_j)\right] = \frac{1}{n}\sum_{i=1}^{n} c_i V(Y_i) = \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0. \end{aligned}$$

Therefore,

$$\begin{aligned} V(\widehat{\beta}_0) = \frac{\sigma^2}{n} + \sigma^2\left[\frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] &= \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \\ &= \sigma^2\left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 + n\overline{x}^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \\ &= \sigma^2\left[\frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 + n\overline{x}^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] = \sigma^2\left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\right], \end{aligned}$$

as claimed. Finally, the covariance between $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is

$$\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \text{Cov}(\overline{Y} - \widehat{\beta}_1\overline{x}, \widehat{\beta}_1) = \text{Cov}(\overline{Y}, \widehat{\beta}_1) - \overline{x}V(\widehat{\beta}_1).$$

We have already shown that $\text{Cov}(\overline{Y}, \widehat{\beta}_1) = 0$. Therefore,

$$\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\overline{x}V(\widehat{\beta}_1) = \sigma^2 \left[\frac{-\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right],$$

as claimed. □

Fact 3. The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are normally distributed.

Proof. Recall $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = \sum_{i=1}^n c_i Y_i,$$

where

$$c_i = \frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

for $i = 1, 2, ..., n$. Under our model assumptions,

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

Therefore, $\widehat{\beta}_1$ is normally distributed because it is a linear combination of $Y_1, Y_2, ..., Y_n$. That $\widehat{\beta}_0$ is also normally distributed follows because

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x},$$

a linear combination of $\overline{Y}$ and $\widehat{\beta}_1$, both of which are normally distributed. □

Note: In this argument (to establish normality), we have used the final assumption that the errors ϵ_i are normally distributed.

Summary: In the simple linear regression model

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

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, so far we have shown the least squares estimators satisfy

$$\widehat{\beta}_0 \sim \mathcal{N}(\beta_0, c_{00}\sigma^2) \quad \text{and} \quad \widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2),$$

where

$$c_{00} = \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2} \quad \text{and} \quad c_{11} = \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

We have also shown $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are negatively correlated and have derived the covariance between them. This covariance will be needed when we derive confidence intervals and prediction intervals later.

Terminology: In 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 $\mathcal{N}(0, \sigma^2)$, we have just derived the sampling distributions of the least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$. We now turn our attention to estimating σ^2, the **error variance**. Define the ith **fitted value** by

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least squares estimators. Define the ith **residual** by

$$e_i = Y_i - \widehat{Y}_i.$$

The **error (residual) sum of squares** by

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

We now state the following distributional results which hold for the simple linear regression model under our model assumptions. Unfortunately, the proofs of some of these results are beyond the scope of this course (at least at this point).

Fact 4. The **mean-squared error**

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

is an unbiased estimator of the error variance σ^2, that is,

$$E(\widehat{\sigma}^2) = E\left(\frac{\text{SSE}}{n-2}\right) = \sigma^2.$$

Remark: We could actually prove this now, but it is rather messy. To see how we would, note that

$$E(\text{SSE}) = E\left[\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2\right] = \sum_{i=1}^{n} E[(Y_i - \widehat{Y}_i)^2] = \sum_{i=1}^{n} V(Y_i - \widehat{Y}_i),$$

because $E(Y_i - \widehat{Y}_i) = 0$. Therefore, all we need to do is work with $V(Y_i - \widehat{Y}_i)$. We will prove this result later under the more general linear model setting.

Fact 5. The pivotal quantity

$$\frac{\text{SSE}}{\sigma^2} = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Fact 6. The mean-squared error $\widehat{\sigma}^2$ is **independent** of both $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Remark: The last two facts will be needed when we pursue statistical inference for β_0, β_1, and other relevant quantities.

12.2.2 Statistical inference

Relevance: In 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 $\mathcal{N}(0, \sigma^2)$, the population-level regression parameters β_0 and β_1 and the error variance σ^2 are unknown. It is therefore of interest to perform statistical inference for these parameters (i.e., write an interval estimator or perform a hypothesis test). In most settings, statistical inference for the slope parameter β_1 is of primary interest because of its connection to the independent variable x in the model. Inference for β_0 is often less relevant, unless one is explicitly interested in the mean of Y when $x = 0$.

Inference for β_1: Under our model assumptions, recall the least squares estimator

$$\widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2) \implies Z = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\sigma^2}} \sim \mathcal{N}(0, 1),$$

where $c_{11} = 1/\sum_{i=1}^n (x_i - \overline{x})^2$. Recall Fact 5, which says

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_1$ (Fact 6), it follows that Z and W are also independent. Therefore,

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\widehat{\sigma}^2}} = \frac{\dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\sigma^2}}}{\sqrt{\dfrac{(n-2)\widehat{\sigma}^2}{\sigma^2} \Big/ (n-2)}} \sim t(n-2).$$

Because T is pivotal, we can write

$$P(-t_{n-2,\alpha/2} < T < t_{n-2,\alpha/2}) = P\left(-t_{n-2,\alpha/2} < \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\widehat{\sigma}^2}} < t_{n-2,\alpha/2}\right) = 1 - \alpha,$$

where $t_{n-2,\alpha/2}$ is the upper $\alpha/2$ quantile of the $t(n-2)$ distribution. Rewriting the event above using algebra leads to

$$P\left(\widehat{\beta}_1 - t_{n-2,\alpha/2}\sqrt{c_{11}\widehat{\sigma}^2} < \beta_1 < \widehat{\beta}_1 + t_{n-2,\alpha/2}\sqrt{c_{11}\widehat{\sigma}^2}\right) = 1 - \alpha,$$

showing that

$$\widehat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{c_{11}\widehat{\sigma}^2}$$

is a $100(1-\alpha)\%$ confidence interval for β_1. To perform a hypothesis test for

$$\begin{gathered} H_0 : \beta_1 = \beta_{1,0} \\ \text{versus} \\ H_a : \beta_1 \neq \beta_{1,0}, \end{gathered}$$

where $\beta_{1,0}$ is a specified value (often, $\beta_{1,0} = 0$), we would use

$$T = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{c_{11}\widehat{\sigma}^2}}$$

as a test statistic and

$$\text{RR} = \{t : |t| > t_{n-2,\alpha/2}\}$$

as a level α rejection region. One sided tests would use a suitably adjusted rejection region. Probability values are computed as areas under the $t(n-2)$ distribution.

Inference for β_0: Confidence intervals and hypothesis tests for β_0 would be constructed similarly by using the pivotal quantity

$$T = \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{c_{00}\widehat{\sigma}^2}} \sim t(n-2).$$

The derivations are analogous, and the forms of the confidence interval and rejection region are analogous.

Confidence interval for $E(Y)$ when $x = x^*$: In the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, the quantity

$$\theta = E(Y|x^*) = \beta_0 + \beta_1 x^*$$

is the population mean of Y when $x = x^*$. In other words, among all individuals in the population whose independent variable is $x = x^*$, the parameter $\theta = E(Y|x^*) = \beta_0 + \beta_1 x^*$ is the mean corresponding to this group of individuals. We now describe inference for this population mean. An obvious point estimator for θ is

$$\widehat{\theta} = \widehat{E(Y|x^*)} = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

This is an unbiased estimator of θ because both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators; i.e.,

$$E(\widehat{\theta}) = E(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) = E(\widehat{\beta}_0) + E(\widehat{\beta}_1 x^*) = \beta_0 + \beta_1 x^* = \theta.$$

The variance of $\widehat{\theta}$ is

$$V(\widehat{\theta}) = V(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right].$$

Because $\widehat{\theta} = \widehat{\beta}_0 + \widehat{\beta}_1 x^*$ is a linear combination of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, both of which are normally distributed (Fact 3), we have

$$\widehat{\theta} \sim \mathcal{N}\left(\theta,\ \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]\right) \implies Z = \frac{\widehat{\theta} - \theta}{\sqrt{\sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} \sim \mathcal{N}(0,1).$$

Recall Fact 5, which says

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (Fact 6), it follows that Z and W are also independent. Therefore,

$$T = \frac{\widehat{\theta} - \theta}{\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} = \frac{\dfrac{\widehat{\theta} - \theta}{\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}}}{\sqrt{\frac{(n-2)\widehat{\sigma}^2}{\sigma^2}\Big/(n-2)}} \sim t(n-2).$$

Because T is pivotal, we can write

$$P\left(-t_{n-2,\alpha/2} < \frac{\widehat{\theta} - \theta}{\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} < t_{n-2,\alpha/2}\right) = 1 - \alpha,$$

where $t_{n-2,\alpha/2}$ is the upper $\alpha/2$ quantile of the $t(n-2)$ distribution. Rewriting the event above using algebra leads to

$$P\left(\widehat{\theta} - t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]} < \theta < \right.$$
$$\left. \widehat{\theta} + t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}\right) = 1 - \alpha,$$

showing that

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}$$

is a $100(1-\alpha)\%$ confidence interval for the population mean $\theta = \beta_0 + \beta_1 x^*$.

Remark: The confidence interval for $\theta = \beta_0 + \beta_1 x^*$ will be different for different values of x^*. The length of the confidence interval will be smallest when $x^* = \overline{x}$ and will increase as the distance between x^* and $\overline{x}$ increases. Therefore, more precise inference for $\theta = \beta_0 + \beta_1 x^*$ will result when x^* is close to $\overline{x}$. It is sometimes desired to estimate the population mean $\theta = \beta_0 + \beta_1 x^*$ for a value of x^* outside the range of x values in the observed data. This is called **extrapolation**. In order for this inference to be valid, one must believe the simple linear regression model is reasonable even for values of x^* outside the range where we have observed data. In some situations, this may be reasonable. In others, it may not.

Discussion: In the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

suppose we want to predict a future value of Y when $x = x^*$. On the surface, this sounds like the last problem where we were estimating $E(Y|x^*)$. However, they are very different problems. For example, suppose Y = 1st year final course percentage in MATH 141 and x = SAT MATH score. Consider these (very different) questions:

- What is an estimate of the mean MATH 141 course percentage for all students who made a SAT math score of $x = 700$?
- What MATH 141 course percentage would you predict for your friend Joe, who made a SAT math score of $x = 700$?

The first question deals with *estimating* $E(Y|x^* = 700)$, a population mean. The second question deals with *predicting* the value of a random variable Y that comes from a population distribution with mean $E(Y|x^* = 700)$. Estimating the mean of a population distribution is much easier (to do precisely) than predicting where one value from the distribution will be.

Prediction interval for Y when $x = x^*$: Our goal is to construct a $100(1-\alpha)\%$ prediction interval for Y^*, a new value of Y when $x = x^*$. An obvious point predictor is

$$\widehat{Y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

This is the same as $\widehat{E(Y|x^*)}$, the point estimator we used to estimate $E(Y|x^*) = \beta_0 + \beta_1 x^*$. However, we use a different symbol in this context to remind ourselves that we are predicting the random variable Y^*, not estimating $E(Y|x^*)$. Define the random variable

$$U = Y^* - \widehat{Y}^*.$$

We call U the **prediction error**. Note that

$$E(U) = E(Y^*) - E(\widehat{Y}^*) = (\beta_0 + \beta_1 x^*) - E(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) = (\beta_0 + \beta_1 x^*) - (\beta_0 + \beta_1 x^*) = 0.$$

The variance of U is

$$V(U) = V(Y^* - \widehat{Y}^*) = V(Y^*) + V(\widehat{Y}^*) - 2\text{Cov}(Y^*, \widehat{Y}^*).$$

Under our simple linear regression model assumptions, we know $V(Y^*) = \sigma^2$. In addition,

$$V(\widehat{Y}^*) = V(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right],$$

which is the same as the variance of $\widehat{E(Y|x^*)}$. Finally, $\text{Cov}(Y^*, \widehat{Y}^*) = 0$ because of the independence assumption. More specifically, $\widehat{Y}^*$ is a function of $Y_1, Y_2, ..., Y_n$, the observed

data. The random variable Y^* is a new value of Y, and hence is independent of $Y_1, Y_2, ..., Y_n$. Therefore,

$$V(U) = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right].$$

Finally, the prediction error $U = Y^* - \widehat{Y}^*$ is normally distributed because it is a linear combination of Y^* and $\widehat{Y}^*$, both of which are normally distributed. Therefore,

$$Y^* - \widehat{Y}^* \sim \mathcal{N}\left(0,\ \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]\right)$$

$$\Longrightarrow Z = \frac{Y^* - \widehat{Y}^*}{\sqrt{\sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} \sim \mathcal{N}(0, 1).$$

Recall Fact 5, which says

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Because $\widehat{\sigma}^2$ is independent of Y^* (why?) and $\widehat{Y}^*$ (Fact 6), it follows that Z and W are also independent. Therefore,

$$T = \frac{Y^* - \widehat{Y}^*}{\sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} = \frac{\dfrac{Y^* - \widehat{Y}^*}{\sqrt{\sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}}}{\sqrt{\dfrac{(n-2)\widehat{\sigma}^2}{\sigma^2} \Big/ (n-2)}} \sim t(n-2).$$

Because T is pivotal, we can write

$$P\left(-t_{n-2,\alpha/2} < \frac{Y^* - \widehat{Y}^*}{\sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} < t_{n-2,\alpha/2}\right) = 1 - \alpha,$$

where $t_{n-2,\alpha/2}$ is the upper $\alpha/2$ quantile of the $t(n-2)$ distribution. Rewriting the event above using algebra leads to

$$P\left(\widehat{Y}^* - t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]} < Y^* < \right.$$

$$\left. \widehat{Y}^* + t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}\right) = 1 - \alpha,$$

showing that

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}$$

is a $100(1-\alpha)\%$ **prediction interval** for the random variable Y^*.

Remark: It is of interest to compare the confidence interval for $E(Y|x^*)$,

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]},$$

to the prediction interval for Y^*,

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}.$$

The prediction interval when $x = x^*$ will always be wider than the corresponding confidence interval for $E(Y|x^*)$. This is a result of the additional uncertainty which arises from having to predict the value of a new random variable.

Example 12.1 (continued). We use R to estimate the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon \iff \text{max.O2} = \beta_0 + \beta_1(\text{time}) + \epsilon$$

for the exercise data in Example 12.1 (notes, pp 102) under the assumptions for ϵ stated in this section. Here is the output:

```
> fit = lm(max.O2 ~ time)
> summary(fit)

Call:
lm(formula = max.O2 ~ time)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5425 -2.5733 -0.8386  0.8226  8.5555

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 90.897200   6.542737  13.893 2.27e-12 ***
time        -0.051344   0.007869  -6.525 1.46e-06 ***

Residual standard error: 3.497 on 22 degrees of freedom
Multiple R-squared: 0.6593,  Adjusted R-squared: 0.6438
F-statistic: 42.57 on 1 and 22 DF, p-value: 1.458e-06
```

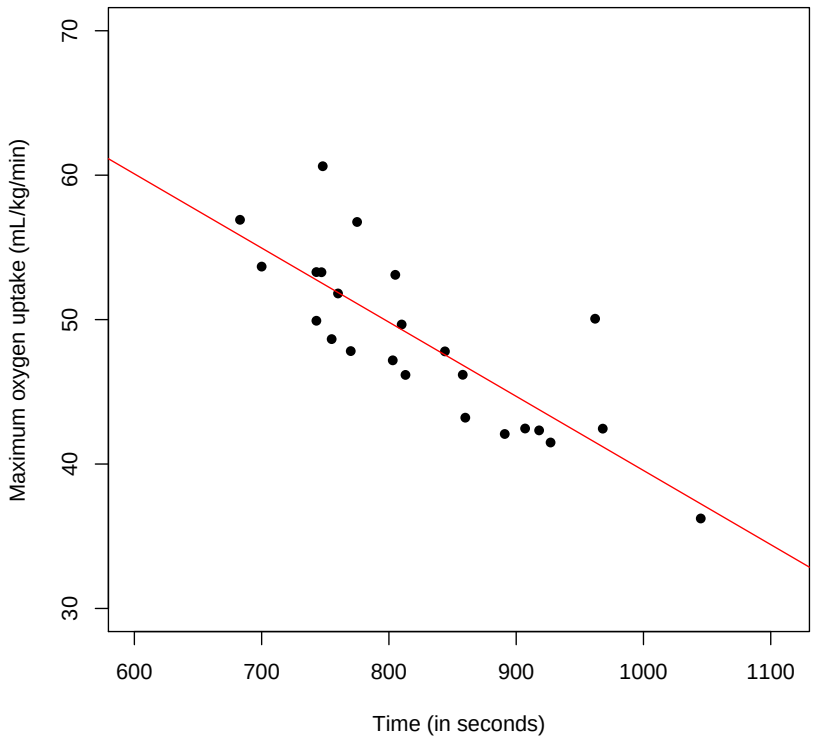

Figure 12.2: Exercise data. Scatterplot of $n = 24$ observations from middle-aged men. The least-squares regression line is superimposed.

R output: The `Estimate` output gives the least squares estimates

$$\begin{aligned} \widehat{\beta}_0 &\approx 90.897 \\ \widehat{\beta}_1 &\approx -0.051. \end{aligned}$$

The estimated model is

$$\widehat{Y} = 90.897 - 0.051x \quad \Longleftrightarrow \quad \widehat{\text{max.O2}} = 90.897 - 0.051(\text{time}).$$

This line is shown superimposed over the exercise data in Figure 12.2 (above). The `Std.Error` output gives

$$\begin{aligned} \widehat{\text{se}}(\widehat{\beta}_0) = \sqrt{c_{00}\widehat{\sigma}^2} &= 6.542737 \\ \widehat{\text{se}}(\widehat{\beta}_1) = \sqrt{c_{11}\widehat{\sigma}^2} &= 0.007869. \end{aligned}$$

These are the *estimated* standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively. These are point estimates of the true standard errors

$$\text{se}(\widehat{\beta}_0) = \sqrt{c_{00}\sigma^2} \quad \text{and} \quad \text{se}(\widehat{\beta}_1) = \sqrt{c_{11}\sigma^2}.$$

The output

```
Residual standard error: 3.497 on 22 degrees of freedom
```

gives the square root of the mean-squared error $\widehat{\sigma}^2$; i.e.,

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{24-2} = (3.497)^2 \approx 12.229.$$

The output

```
            t value Pr(>|t|)
(Intercept) 13.893 2.27e-12 ***
time        -6.525 1.46e-06 ***
```

gives the test statistics

$$\begin{aligned} t = \frac{\widehat{\beta}_0 - 0}{\sqrt{c_{00}\widehat{\sigma}^2}} &= 13.893 \\ t = \frac{\widehat{\beta}_1 - 0}{\sqrt{c_{11}\widehat{\sigma}^2}} &= -6.525. \end{aligned}$$

The first test statistic is used to test

$$\begin{gathered} H_0 : \beta_0 = 0 \\ \text{versus} \\ H_a : \beta_0 \neq 0, \end{gathered}$$

which is nonsensical. Recall that β_0 equals $E(Y)$ when $x = 0$. This is the population mean maximum O_2 uptake for all middle-aged men who run two miles in 0 seconds. Therefore, the population-level intercept term β_0 has no practical meaning here. The second test statistic is used to test

$$\begin{gathered} H_0 : \beta_1 = 0 \\ \text{versus} \\ H_a : \beta_1 \neq 0, \end{gathered}$$

which can be used to assess whether maximum O_2 uptake (Y) and time (x) are linearly related *in the population* of all middle-aged men. Two-sided probability values are in `Pr(>|t|)`. The probability value for this test is

$$\text{p-value} = P_{H_0}(|T| > 6.525) < 1.5 \times 10^{-6},$$

indicating the evidence against H_0 is overwhelming. The random variable T above satisfies $T \overset{H_0}{\sim} t(22)$; see Figure 12.3 (next page).

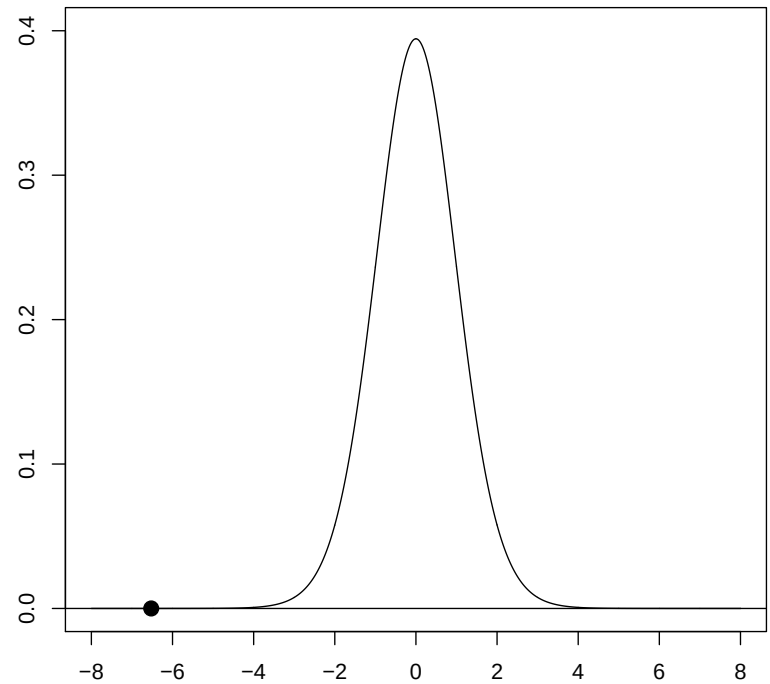

Figure 12.3: $t(22)$ pdf; the test statistic $t = -6.525$ is shown by using a dark circle. This pdf represents the sampling distribution of T when $H_0 : \beta_1 = 0$ is true.

Confidence intervals for the population-level regression parameters β_0 and β_1 are not provided in the `fit` summary output, but they can be calculated easily. Because the intercept parameter β_0 has no practical relevance in this problem, we focus only on the slope parameter β_1. A 95% confidence interval for β_1 is calculated as

$$\widehat{\beta}_1 \pm t_{22,0.025}\widehat{\text{se}}(\widehat{\beta}_1) \;\longrightarrow\; -0.051 \pm 2.074(0.008) \;\longrightarrow\; (-0.068, -0.035).$$

Interpretation: For the population of middle-aged men, we would expect the maximum O_2 uptake to decrease between 0.035 and 0.068 mL/kg/min for each one second increase in the time it takes to run two miles.

Estimating $E(Y|x^*)$ and predicting Y^*: Suppose we are interested in the population of middle-aged men who run a two-minute mile in 900 seconds (i.e., $x^* = 900$). It is first interesting to observe that no one from this population is in the sample of $n = 24$ men observed. However, we can still make inferential statements about this population by making use of the assumed relationship between maximum O_2 uptake and time across all times. In R, calculating a confidence interval for $E(Y|x^* = 900)$ and a prediction interval for Y^* is done as follows:

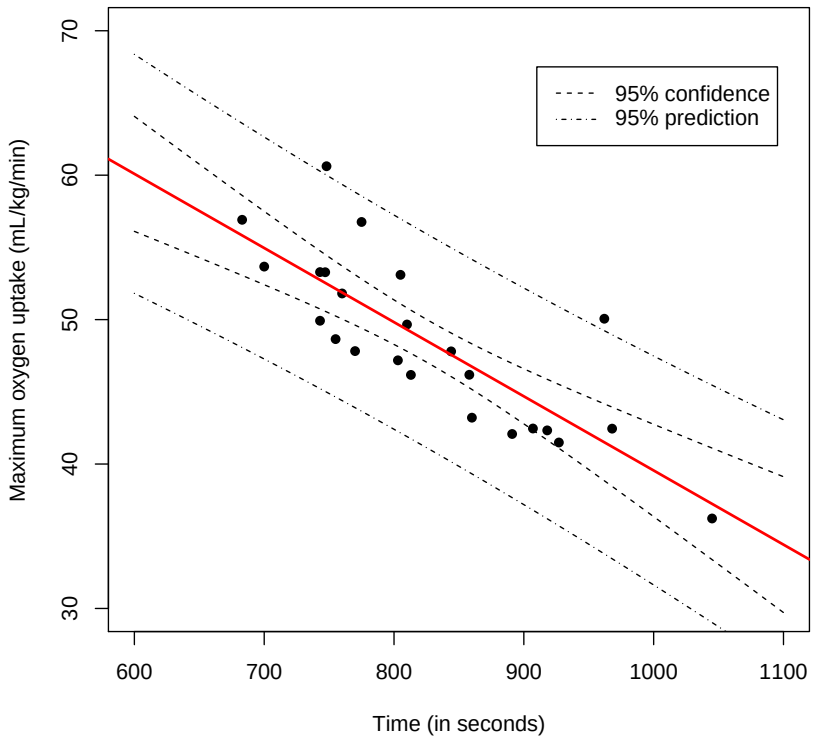

Figure 12.4: Exercise data. Scatterplot of $n = 24$ observations from middle-aged men. The least-squares regression line is superimposed; 95% confidence intervals for $E(Y|x^*)$ and 95% prediction intervals for Y^* are shown for each $x^* \in \{600, 601, ..., 1100\}$.

```
> predict(fit,data.frame(time=900),level=0.95,interval="confidence")
       fit      lwr      upr
  44.68785 42.78192 46.59377
> predict(fit,data.frame(time=900),level=0.95,interval="prediction")
       fit      lwr      upr
  44.68785 37.18834 52.18735
```

The output `fit` gives the point estimate/point prediction

$$\widehat{\beta}_0 + \widehat{\beta}_1 x^* \approx 90.897 - 0.051(900) \approx 44.688.$$

We now interpret the intervals shown above:

- Among all middle-aged men who run two miles in 900 seconds, we are 95% confident the population mean maximum O_2 uptake is between 42.782 and 46.584 mL/kg/min.
- For an individual middle-aged man who runs two miles in 900 seconds, his maximum O_2 uptake will fall between 37.188 and 52.187 mL/kg/min with probability 0.95.

12.3 Random vectors, quadratic forms, and the multivariate normal distribution

Recall: A general linear model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is a random vector; i.e.,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},$$

$\mathbf{X}$ is a design matrix (fixed), $\boldsymbol{\beta}$ is a vector of parameters (fixed and unknown; to be estimated), and $\boldsymbol{\epsilon}$ is a random vector; i.e.,

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

We have already seen the simple linear regression model

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

for $i = 1, 2, ..., n$, is a special case of the general linear model with

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\times 2} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\times 1}.$$

The class of general linear models includes simple linear regression, multiple linear regression, analysis of variance (ANOVA), and other well known statistical models.

Importance: We have just discussed simple linear regression at length, focusing on

- estimation via least squares
- sampling distributions of least squares estimators
- statistical inference for regression parameters, confidence intervals, and prediction intervals.

We now pursue these same goals but more generally. To accomplish this, we first need to discuss random vectors and present results for expectations, variance and covariance, and their multivariate probability distributions (especially the multivariate normal distribution). Much of this is a review of STAT 511 concepts, but it is presented more generally to facilitate our discussion with linear models.

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ are random variables. We call

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n \times 1}$$

a **random vector**. The multivariate probability density function (pdf) of $\mathbf{Y}$ is denoted by $f_{\mathbf{Y}}(\mathbf{y})$. The function $f_{\mathbf{Y}}(\mathbf{y})$ describes probabilistically how the random variables $Y_1, Y_2, ..., Y_n$ are jointly distributed.

- If $Y_1, Y_2, ..., Y_n$ are independent variables, then

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i),$$

  where $f_{Y_i}(y_i)$ is the marginal pdf of Y_i.

- If $Y_1, Y_2, ..., Y_n$ are iid from a common marginal pdf, say $f_Y(y)$, then

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i).$$

Terminology: Suppose $Y_1, Y_2, ..., Y_n$ are random variables with means $E(Y_i) = \mu_i$ and variances $V(Y_i) = \sigma_i^2$, for $i = 1, 2, ..., n$, and covariances $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$ for $i \neq j$. The **mean** of a random vector $\mathbf{Y}$ is

$$E(\mathbf{Y}) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}.$$

That is, the mean of a random vector $\mathbf{Y}$ is the vector of the marginal means $E(Y_i)$. The **variance-covariance matrix** of $\mathbf{Y}$ is

$$\text{Cov}(\mathbf{Y}) = E[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'] = \begin{pmatrix} V(Y_1) & \text{Cov}(Y_1, Y_2) & \cdots & \text{Cov}(Y_1, Y_n) \\ \text{Cov}(Y_2, Y_1) & V(Y_2) & \cdots & \text{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_n, Y_1) & \text{Cov}(Y_n, Y_2) & \cdots & V(Y_n) \end{pmatrix}_{n \times n}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}_{n \times n} = \mathbf{V}.$$

- The variance-covariance matrix $\mathbf{V}$ contains the variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ on the diagonal of the matrix and the $2\binom{n}{2}$ covariance terms $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$, for $i \neq j$, on the off-diagonal.
- Because $\text{Cov}(Y_i, Y_j) = \text{Cov}(Y_j, Y_i)$, the variance-covariance matrix $\mathbf{V}$ is **symmetric**; i.e., $\mathbf{V}' = \mathbf{V}$.
- If $Y_1, Y_2, ..., Y_n$ are pairwise independent; i.e., $Y_i \perp\!\!\!\perp Y_j$, $i \neq j$, then all the covariances are zero and $\mathbf{V}$ is a **diagonal** matrix. That is,

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}_{n \times n}.$$

Terminology: Suppose

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n \times 1} \quad \text{and} \quad \mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}_{m \times 1}$$

are random vectors. The **covariance** of $\mathbf{Y}$ and $\mathbf{Z}$ is

$$\text{Cov}(\mathbf{Y}, \mathbf{Z}) = \begin{pmatrix} \text{Cov}(Y_1, Z_1) & \text{Cov}(Y_1, Z_2) & \cdots & \text{Cov}(Y_1, Z_m) \\ \text{Cov}(Y_2, Z_1) & \text{Cov}(Y_2, Z_2) & \cdots & \text{Cov}(Y_2, Z_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_n, Z_1) & \text{Cov}(Y_n, Z_2) & \cdots & \text{Cov}(Y_n, Z_m) \end{pmatrix}_{n \times m}.$$

We say the random vectors $\mathbf{Y}_{n\times 1}$ and $\mathbf{Z}_{m\times 1}$ are uncorrelated if $\text{Cov}(\mathbf{Y}, \mathbf{Z}) = \mathbf{0}_{n\times m}$.

Results: Suppose $\mathbf{Y}_{n\times 1}$ is a random vector with mean $E(\mathbf{Y}) = \boldsymbol{\mu}_{n\times 1}$ and variance-covariance matrix $\text{Cov}(\mathbf{Y}) = \mathbf{V}_{n\times n}$. Suppose $\mathbf{c}_{m\times 1}$ is a non-random (constant) vector and $\mathbf{A}_{m\times n}$ is a non-random matrix. Then

$$\begin{aligned} E(\mathbf{c} + \mathbf{A}\mathbf{Y}) &= \mathbf{c} + \mathbf{A}E(\mathbf{Y}) = \mathbf{c} + \mathbf{A}\boldsymbol{\mu} \\ \text{Cov}(\mathbf{c} + \mathbf{A}\mathbf{Y}) &= \mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}' = \mathbf{A}\mathbf{V}\mathbf{A}'. \end{aligned}$$

Result: Suppose $\mathbf{Y}_{n\times 1}$ and $\mathbf{Z}_{m\times 1}$ are random vectors and $\mathbf{A}_{q\times n}$ and $\mathbf{B}_{p\times m}$ are non-random matrices. Then

$$\text{Cov}(\mathbf{A}\mathbf{Y}, \mathbf{B}\mathbf{Z}) = \mathbf{A}\text{Cov}(\mathbf{Y}, \mathbf{Z})\mathbf{B}'.$$

Terminology: Suppose $\mathbf{Y}_{n\times 1}$ is a random vector with mean $E(\mathbf{Y}) = \boldsymbol{\mu}_{n\times 1}$ and variance-covariance matrix $\text{Cov}(\mathbf{Y}) = \mathbf{V}_{n\times n}$. Suppose $\mathbf{A}_{n\times n}$ is a non-random matrix. We call $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ a **quadratic form**. The mean of a quadratic form is

$$E(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{A}\mathbf{V}),$$

where $\text{tr}(\cdot)$ means "trace," that is, $\text{tr}(\mathbf{A}\mathbf{V})$ is the sum of the diagonal elements of $\mathbf{A}\mathbf{V}$.

Remark: It is important to note that a quadratic form $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ is a univariate random variable. Therefore, its mean $E(\mathbf{Y}'\mathbf{A}\mathbf{Y})$ is a scalar constant. Quadratic forms are important in the theory of linear models. We will see later that sums of squares from regression and analysis of variance can be written as quadratic forms like $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ and degrees of freedom are determined by the matrix $\mathbf{A}$.

Remark: To derive the expression for $E(\mathbf{Y}'\mathbf{A}\mathbf{Y})$, we need to recall properties about the trace operator from linear algebra. For matrices $\mathbf{A}$ and $\mathbf{B}$ of conformable dimensions,

1. $\text{tr}(c\mathbf{A}) = c\,\text{tr}(\mathbf{A})$, for any constant c
2. $\text{tr}(\mathbf{A}+\mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$
3. $\text{tr}(\mathbf{A}\mathbf{B}) = \text{tr}(\mathbf{B}\mathbf{A})$.

Properties (1) and (2) above identify that the trace operator $\text{tr}(\cdot)$ is linear (just as mathematical expectation is a linear operator). We have

$$\begin{aligned}
E(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = E[\text{tr}(\mathbf{Y}'\mathbf{A}\mathbf{Y})] &= E[\text{tr}(\mathbf{A}\mathbf{Y}\mathbf{Y}')] \\
&= \text{tr}[E(\mathbf{A}\mathbf{Y}\mathbf{Y}')] \\
&= \text{tr}[\mathbf{A}E(\mathbf{Y}\mathbf{Y}')] \\
&= \text{tr}[\mathbf{A}(\mathbf{V}+\boldsymbol{\mu}\boldsymbol{\mu}')] \\
&= \text{tr}(\mathbf{A}\mathbf{V}) + \text{tr}(\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}') \\
&= \text{tr}(\mathbf{A}\mathbf{V}) + \text{tr}(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}) = \text{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}. \ \square
\end{aligned}$$

Terminology: Suppose $Z_1, Z_2, ..., Z_n$ are iid $\mathcal{N}(0,1)$ random variables. The joint pdf of

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}$$

is given by

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} f_Z(z_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n z_i^2/2} = (2\pi)^{-n/2}\exp(-\mathbf{z}'\mathbf{z}/2),$$

for all $\mathbf{z} \in \mathbb{R}^n$. We say $\mathbf{Z}$ has a **standard multivariate normal distribution** and write $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$. In this distribution, note that

$$E(\mathbf{Z}) = \begin{pmatrix} E(Z_1) \\ E(Z_2) \\ \vdots \\ E(Z_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}_{n\times 1}$$

and

$$\text{Cov}(\mathbf{Z}) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{n\times n} = \mathbf{I}_{n\times n}.$$

This is the multivariate analogue of a standard normal distribution; i.e., $Z \sim \mathcal{N}(0,1)$. Note that

$$Z_1, Z_2, ..., Z_n \sim \text{iid } \mathcal{N}(0,1) \iff \mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}).$$

Terminology: The random vector

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

has a **multivariate normal distribution** with mean $\boldsymbol{\mu}_{n\times 1}$ and variance-covariance matrix $\mathbf{V}_{n\times n}$ if its pdf is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-n/2}|\mathbf{V}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\},$$

for all $\mathbf{y} \in \mathbb{R}^n$. We write $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$.

Notes: A few remarks are in order about this distribution and the associated pdf.

- The notation $|\cdot|$ above means "determinant."
- Writing $\mathbf{V}^{-1}$ assumes $\mathbf{V}$ is a full rank matrix; i.e., $\text{rank}(\mathbf{V}) = n$. Recall if a square matrix is of full rank, then its inverse exists.
- In the univariate normal case (i.e., $n = 1$), recall the following relationship:

  $$Z \sim \mathcal{N}(0,1) \implies Y = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2).$$

  A similar relationship holds in the multivariate case; i.e.,

  $$\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}) \implies \mathbf{Y} = \mathbf{V}^{1/2}\mathbf{Z} + \boldsymbol{\mu} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V}),$$

  where the (symmetric square root) matrix $\mathbf{V}^{1/2}$ satisfies $\mathbf{V}^{1/2}\mathbf{V}^{1/2} = \mathbf{V}$.
- Note that

  $$\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V}) \implies Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad \text{for } i = 1, 2, ..., n.$$

  That is, joint normality implies marginal normality. The relationship does not necessarily go the other way.

Result: Suppose $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$. Let $\mathbf{c}_{m\times 1}$ denote a non-random (constant) vector and $\mathbf{A}_{m\times n}$ denote a non-random matrix. Then

$$\mathbf{U} = \mathbf{c} + \mathbf{A}\mathbf{Y} \sim \mathcal{N}_m(\mathbf{c} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\mathbf{V}\mathbf{A}').$$

12.4 Multiple linear regression

Preview: We have already considered the simple linear regression model

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

for $i = 1, 2, ..., n$. Our interest now is to extend this model to include multiple independent variables $x_1, x_2, ..., x_k$. Specifically, we consider models of the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$. We call this a **multiple linear regression model**.

- There are now $p = k+1$ population-level regression parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$. These are unknown (and fixed) and are to be estimated with the observed data.
- Schematically, we can envision the observed data as follows:

Individual	Y	x_1	x_2	$\cdots$	x_k
1	Y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	Y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	Y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

  That is, each of the n individuals in the sample contributes a response Y and a value of each of the independent variables $x_1, x_2, ..., x_k$.

Remark: To estimate the model, we will continue to use least squares. After we estimate the model, we then will pursue similar topics as we did in the simple linear regression case; e.g., determining the sampling distributions of the least squares estimators, writing confidence intervals and performing hypothesis tests for (population-level) regression parameters, and writing confidence intervals and prediction intervals for $E(Y|x_1, x_2, ..., x_k)$ and Y^*, respectively. Not surprisingly, these topics are best presented by making use of notation for random vectors we have described previously.

Matrix representation: Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$. Define

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}_{n\times p}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}_{p\times 1}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}.$$

With these definitions, the multiple linear regression model on the previous page can be expressed equivalently as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

In this representation,

- $\mathbf{Y}$ is an $n \times 1$ (random) vector of responses
- $\mathbf{X}$ is an $n \times p$ (fixed) matrix of independent variable measurements ($p = k + 1$)
- $\boldsymbol{\beta}$ is a $p \times 1$ (fixed) vector of unknown population-level regression parameters (to be estimated)
- $\boldsymbol{\epsilon}$ is an $n \times 1$ (random) vector of unobserved errors. Formal assumptions on $\boldsymbol{\epsilon}$ will be stated later when needed.

Illustration: Consider the four linear models presented at the beginning of this chapter (see pp 101, notes). For $i = 1, 2, ..., n$, consider

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\
Y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \\
Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \\
Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i.
\end{aligned}$$

Each of these models can be written as a general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}.$$

The differences among the models are in how $\mathbf{X}$ and $\boldsymbol{\beta}$ are defined. Of course, for the first model (simple linear regression), we have

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\times 2} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\times 1}.$$

For the second model (quadratic regression), we have

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \\ 1 & x_n & x_n^2 \end{pmatrix}_{n\times 3} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}_{3\times 1}.$$

For the third model (regression plane in $\mathbb{R}^3$), we have

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} \end{pmatrix}_{n\times 3} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}_{3\times 1}.$$

For the fourth model (curvilinear surface in $\mathbb{R}^3$), we have

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}x_{22} \\ \vdots & \vdots & & \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} \end{pmatrix}_{n\times 4} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}_{4\times 1}.$$

Observations: It is interesting to note that the first model above

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

is a special case of the second model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

with $\beta_2 = 0$. Also, the third model above

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

is a special case of the fourth model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

with $\beta_3 = 0$. The third and fourth models are both special cases of the fifth model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i3} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \beta_7 x_{i1} x_{i2} x_{i3} + \epsilon_i,$$

(a curvilinear surface in $\mathbb{R}^4$). The third model is a special case with $\beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$. The fourth model is a special case with $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$. The fifth model above has

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} & x_{13} & x_{11}x_{13} & x_{12}x_{13} & x_{11}x_{12}x_{13} \\ 1 & x_{21} & x_{22} & x_{21}x_{22} & x_{23} & x_{21}x_{23} & x_{22}x_{23} & x_{21}x_{22}x_{23} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} & x_{n3} & x_{n1}x_{n3} & x_{n2}x_{n3} & x_{n1}x_{n2}x_{n3} \end{pmatrix}_{n\times 8} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{pmatrix}_{8\times 1}.$$

Preview: We will learn later how to perform statistical inference for population-level regression parameters in multiple linear regression. For example, testing

$$H_0 : \beta_2 = 0$$

allows us to "test" the first model versus the second. Testing

$$H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

allows us to "test" the third model versus the fifth.

Example 12.2. A recent article in the *Journal of Air and Waste Management Association* described an observational study in Kaohsiung City, Taiwan. The goal was to develop a multiple linear regression model to explain how the response variable

$$Y = \text{ energy content of solid waste specimen}$$

was related to four independent variables

$$\begin{aligned} x_1 &= \text{plastic by weight (measured as \% of total weight)} \\ x_2 &= \text{paper by weight (measured as \% of total weight)} \\ x_3 &= \text{garbage by weight (measured as \% of total weight)} \\ x_4 &= \text{moisture percentage.} \end{aligned}$$

The authors describe how a random sample of $n = 30$ solid waste specimens were available to estimate the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. The energy content Y (kcal/kg) was measured on each waste specimen after it was incinerated. This model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{30} \end{pmatrix}_{30\times 1}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{30,1} & x_{30,2} & x_{30,3} & x_{30,4} \end{pmatrix}_{30\times 5},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}_{5\times 1}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{30} \end{pmatrix}_{30\times 1}.$$

The data from this study are on the course web site; I used R to construct the design matrix $\mathbf{X}_{30\times 5}$ (shown on the next page). □

```
> X = cbind(intercept,plastic,paper,garbage,moisture)
> X
      intercept plastic paper garbage moisture
 [1,]         1   18.69 15.65   45.01    58.21
 [2,]         1   19.43 23.51   39.69    46.31
 [3,]         1   19.24 24.23   43.16    46.63
 [4,]         1   22.64 22.20   35.76    45.85
 [5,]         1   16.54 23.56   41.20    55.14
 [6,]         1   21.44 23.65   35.56    54.24
 [7,]         1   19.53 24.45   40.18    47.20
 [8,]         1   23.97 19.39   44.11    43.82
 [9,]         1   21.45 23.84   35.41    51.01
[10,]         1   20.34 26.50   34.21    49.06
[11,]         1   17.03 23.46   32.45    53.23
[12,]         1   21.03 26.99   38.19    51.78
[13,]         1   20.49 19.87   41.35    46.69
[14,]         1   20.45 23.03   43.59    53.57
[15,]         1   18.81 22.62   42.20    52.98
[16,]         1   18.28 21.87   41.50    47.44
[17,]         1   21.41 20.47   41.20    54.68
[18,]         1   25.11 22.59   37.02    48.74
[19,]         1   21.04 26.27   38.66    53.22
[20,]         1   17.99 28.22   44.18    53.37
[21,]         1   18.73 29.39   34.77    51.06
[22,]         1   18.49 26.58   37.55    50.66
[23,]         1   22.08 24.88   37.07    50.72
[24,]         1   14.28 26.27   35.80    48.24
[25,]         1   17.74 23.61   37.36    49.92
[26,]         1   20.54 26.58   35.40    53.58
[27,]         1   18.25 13.77   51.32    51.38
[28,]         1   19.09 25.62   39.54    50.13
[29,]         1   21.25 20.63   40.72    48.67
[30,]         1   21.62 22.71   36.22    48.19
```

12.4.1 Estimation and sampling distributions

Estimation: Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$. Least squares estimation involves finding the values of $\beta_0, \beta_1, \beta_2, ..., \beta_k$ that minimize the objective function

$$Q(\beta_0, \beta_1, \beta_2, ..., \beta_k) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})]^2.$$

It is much easier to write this objective function using our notation for random vectors and matrices. To see how, one needs only note that the sum of squares above is the inner product

of $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ with itself; i.e.,

$$Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Because $Q(\boldsymbol{\beta})$ is a scalar function of the $p = k+1$ elements of $\boldsymbol{\beta}$, it is possible to use calculus to determine the values of the p elements that minimize it. Formally, we can take the p partial derivatives with respect to each of $\beta_0, \beta_1, \beta_2, ..., \beta_k$ and set these equal to zero; i.e.,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_0} \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_k} \end{pmatrix} \stackrel{\text{set}}{=} \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{p \times 1}.$$

These are called the **normal equations**. Solving the normal equations for $\beta_0, \beta_1, \beta_2, ..., \beta_k$ gives the least squares estimators, which we denote by $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_k$.

Results: Let $\mathbf{a}$ and $\mathbf{b}$ be $p \times 1$ vectors and $\mathbf{A}$ be a $p \times p$ matrix of constants. Then

$$\frac{\partial \mathbf{a}'\mathbf{b}}{\partial \mathbf{b}} = \mathbf{a} \quad \text{and} \quad \frac{\partial \mathbf{b}'\mathbf{A}\mathbf{b}}{\partial \mathbf{b}} = (\mathbf{A} + \mathbf{A}')\mathbf{b}.$$

These two results from matrix calculus allow us to derive the least squares estimator $\widehat{\boldsymbol{\beta}}$ in closed form. Recalling $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ for conformable matrices $\mathbf{A}$ and $\mathbf{B}$, we have

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{Y}' - \boldsymbol{\beta}'\mathbf{X}')(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

because $\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{X}\boldsymbol{\beta}$ (why?). Using the results above, we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta},$$

because $\mathbf{X}'\mathbf{X}$ is symmetric. Setting this expression equal to $\mathbf{0}$ and rearranging gives

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}.$$

In other words, we have shown the normal equations

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_0} \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_k} \end{pmatrix} \stackrel{\text{set}}{=} \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \iff \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}.$$

If $\mathbf{X}$ is $n \times p$, then $\mathbf{X}'\mathbf{X}$ is a $p \times p$ (square) matrix. If $\mathbf{X}'\mathbf{X}$ is nonsingular, then it has a unique inverse $(\mathbf{X}'\mathbf{X})^{-1}$. Pre-multiplying each side of the normal equations by $(\mathbf{X}'\mathbf{X})^{-1}$ gives

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \implies \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

This is the formula for the least squares estimator of $\boldsymbol{\beta}$.

Q: What happens if $\mathbf{X}'\mathbf{X}$ is singular?
A: If $\mathbf{X}'\mathbf{X}$ is singular, then its inverse $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist and the least squares estimator of $\boldsymbol{\beta}$ is not unique. When might this happen? Mathematically, it turns out that

$$\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}'\mathbf{X}).$$

Therefore, if there are linear dependencies among the columns of $\mathbf{X}$, then

$$\text{rank}(\mathbf{X}'\mathbf{X}) < p \iff (\mathbf{X}'\mathbf{X})^{-1} \text{ does not exist}$$

and the least squares estimator $\widehat{\boldsymbol{\beta}}$ cannot be calculated uniquely. This can happen in ANOVA models depending how they are parameterized. It can also happen in regression models if there is perfect collinearity among two or more of the independent variables.

Example 12.2 (continued). We calculate the least squares estimate

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

for the waste data using matrix calculations in R. The $\mathbf{X}'\mathbf{X}$ matrix is

```
> t(X)%*%X
          intercept  plastic    paper  garbage moisture
intercept     30.00   596.98   702.41  1180.38  1515.72
plastic      596.98 12023.37 13944.86 23465.33 30106.63
paper        702.41 13944.86 16776.81 27387.00 35486.85
garbage     1180.38 23465.33 27387.00 46918.77 59665.18
moisture    1515.72 30106.63 35486.85 59665.18 76896.85
```

We can calculate $(\mathbf{X}'\mathbf{X})^{-1}$ in R using the `solve` function; i.e.,

```
> round(solve(t(X)%*%X),6)
          intercept   plastic     paper   garbage  moisture
intercept 31.930929 -0.314886 -0.279648 -0.233927 -0.195548
plastic   -0.314886  0.008044  0.001730  0.001230  0.001304
paper     -0.279648  0.001730  0.005402  0.002921  0.000076
garbage   -0.233927  0.001230  0.002921  0.003705 -0.000093
moisture  -0.195548  0.001304  0.000076 -0.000093  0.003394
```

I used the `round` function to control the number of decimal places in the output.

The vector $\mathbf{X}'\mathbf{Y}$ is

```
> t(X)%*%Y
               [,1]
intercept   38438.0
plastic    770763.3
paper      900549.2
garbage   1510802.3
moisture  1928724.2
```

Finally, the least squares estimate $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ for the waste data is

```
> solve(t(X)%*%X)%*%t(X)%*%Y
                 [,1]
intercept 2244.922664
plastic     28.925002
paper        7.643614
garbage      4.296642
moisture   -37.353831
```

That is,

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \\ \widehat{\beta}_4 \end{pmatrix} = \begin{pmatrix} 2244.923 \\ 28.925 \\ 7.644 \\ 4.297 \\ -37.354 \end{pmatrix}$$

(to 3 dp) and the estimated model is

$$\widehat{Y} = 2244.923 + 28.925x_1 + 7.644x_2 + 4.297x_3 - 37.354x_4,$$

or, in other words,

$$\widehat{\text{energy}} = 2244.923 + 28.925(\text{plastic}) + 7.644(\text{paper}) + 4.297(\text{garbage}) - 37.354(\text{moisture}).$$

Remark: In practice, it is not necessary to code in $\mathbf{X}$ and $\mathbf{Y}$ and perform matrix operations yourself. The `lm` function in R does all this in the background. For example, with the waste data,

```
> fit = lm(energy~plastic+paper+garbage+moisture)
> fit

Coefficients:
(Intercept)      plastic        paper      garbage     moisture
   2244.923       28.925        7.644        4.297      -37.354
```

These numerical values of the least squares estimates match the ones we obtained by doing the matrix calculations directly (to 3 dp). □

Assumptions: Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$. As in simple linear regression, we will continue to assume $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Recall this means

- $E(\epsilon_i) = 0$
- $V(\epsilon_i) = \sigma^2$, that is, the variance is constant
- the random variables ϵ_i are independent
- the random variables ϵ_i are normally distributed.

In our matrix formulation, the model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}),$$

that is, $\boldsymbol{\epsilon}$ is a multivariate normal random vector with mean

$$E(\boldsymbol{\epsilon}) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}_{n\times 1}$$

and variance-covariance matrix

$$\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{n\times n} = \sigma^2 \mathbf{I}_{n\times n}.$$

This model assumption for $\boldsymbol{\epsilon}$ is equivalent to the four conditions above. Under these assumptions, it is easy to show

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

Note that

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{0} = \mathbf{X}\boldsymbol{\beta}$$

and

$$\text{Cov}(\mathbf{Y}) = \text{Cov}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}.$$

Finally, because $\mathbf{Y}$ is a linear (vector-valued) function of $\boldsymbol{\epsilon}$, it is also normally distributed.

Discussion: Because $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, this means the $n \times 1$ vector $E(\mathbf{Y})$ is a linear combination of the $(n \times 1)$ columns of $\mathbf{X}$. To see why, write

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \end{pmatrix},$$

where

$$\mathbf{1}_{n\times 1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

an $n \times 1$ vector of 1's, and the k remaining $(n \times 1)$ columns of $\mathbf{X}$ are denoted by $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$. For $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \cdots + \beta_k\mathbf{x}_k,$$

a linear combination of the columns of $\mathbf{X}$. The collection of $n \times 1$ vectors above is called the **column space** of $\mathbf{X}$ and is denoted by $\mathcal{C}(\mathbf{X})$. This interpretation allows us to characterize the least squares estimator $\widehat{\boldsymbol{\beta}}$ in a new way. We say an estimate $\widehat{\boldsymbol{\beta}}$ is a least squares estimate of $\boldsymbol{\beta}$ if $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is the $(n \times 1)$ vector in $\mathcal{C}(\mathbf{X})$ that is "closest" to $\mathbf{Y}$. In other words, $\widehat{\boldsymbol{\beta}}$ is a least squares estimate of $\boldsymbol{\beta}$ if it minimizes

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

the squared distance between $\mathbf{Y}$ and $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})$. We will learn later about a special matrix that "projects" $\mathbf{Y}$ onto $\mathcal{C}(\mathbf{X})$ perpendicularly so as to minimize this squared distance.

Sampling distribution of $\widehat{\boldsymbol{\beta}}$: Consider the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. We now derive the sampling distribution of the least squares estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The mean of $\widehat{\boldsymbol{\beta}}$ is

$$E(\widehat{\boldsymbol{\beta}}) = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{=\mathbf{I}_{p\times p}}\boldsymbol{\beta} = \boldsymbol{\beta};$$

i.e., $\widehat{\boldsymbol{\beta}}$ is an **unbiased estimator** of $\boldsymbol{\beta}$. The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \text{Cov}(\widehat{\boldsymbol{\beta}}) &= \text{Cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Cov}(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

because $(\mathbf{X}'\mathbf{X})^{-1}$ is symmetric. Therefore,

$$\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\underbrace{\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}_{=\mathbf{I}_{p\times p}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Finally, because $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is a linear (vector-valued) function of $\mathbf{Y}$, which is multivariate normal, $\widehat{\boldsymbol{\beta}}$ is also multivariate normal. Therefore, we arrive at the sampling distribution

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right).$$

Remark: Because multivariate normality implies marginal normality, we know

$$\begin{aligned}
\widehat{\beta}_0 &\sim \mathcal{N}(\beta_0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{11}) \\
\widehat{\beta}_1 &\sim \mathcal{N}(\beta_1, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{22}) \\
\widehat{\beta}_2 &\sim \mathcal{N}(\beta_2, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{33}) \\
&\vdots \\
\widehat{\beta}_k &\sim \mathcal{N}(\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{pp}),
\end{aligned}$$

where recall $p = k+1$ and $(\mathbf{X}'\mathbf{X})^{-1}_{jj}$ is the jth diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix, for $j = 0, 1, 2, ..., k$. Covariances of different least squares estimators use the off-diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$; e.g.,

$$\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{23}.$$

Recall: In 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 $\mathcal{N}(0, \sigma^2)$, we proved the least squares estimators

$$\widehat{\beta}_0 \sim \mathcal{N}(\beta_0, c_{00}\sigma^2) \quad \text{and} \quad \widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2),$$

where

$$c_{00} = \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2} \quad \text{and} \quad c_{11} = \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

We also showed

$$\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \sigma^2\left[\frac{-\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}\right].$$

These results are a special case of our sampling distribution result for $\widehat{\boldsymbol{\beta}}$ above. To see why, recall that in simple linear regression

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta}_{2\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\epsilon}_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Straightforward (but tedious) calculations show

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \dfrac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2} & -\dfrac{\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ -\dfrac{\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} & \dfrac{1}{\sum_{i=1}^n (x_i - \overline{x})^2} \end{pmatrix},$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \displaystyle\sum_{i=1}^{n} Y_i \\ \displaystyle\sum_{i=1}^{n} x_i Y_i \end{pmatrix}.$$

Thus, the least squares estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} &= \begin{pmatrix} \dfrac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\overline{x})^2} & -\dfrac{\overline{x}}{\sum_{i=1}^n (x_i-\overline{x})^2} \\ -\dfrac{\overline{x}}{\sum_{i=1}^n (x_i-\overline{x})^2} & \dfrac{1}{\sum_{i=1}^n (x_i-\overline{x})^2}. \end{pmatrix} \begin{pmatrix} \displaystyle\sum_{i=1}^{n} Y_i \\ \displaystyle\sum_{i=1}^{n} x_i Y_i \end{pmatrix} \\ &= \begin{pmatrix} \overline{Y} - \widehat{\beta}_1\overline{x} \\ \dfrac{\sum_{i=1}^n (x_i-\overline{x})(Y_i-\overline{Y})}{\sum_{i=1}^n (x_i-\overline{x})^2} \end{pmatrix} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}.\end{aligned}$$

That is, the least squares estimator formula $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ produces the same answers we saw in simple linear regression. Furthermore,

$$\begin{aligned} V(\widehat{\beta}_0) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{11} = \sigma^2\left[\frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\overline{x})^2}\right] \\ V(\widehat{\beta}_1) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{22} = \sigma^2\left[\frac{1}{\sum_{i=1}^n (x_i-\overline{x})^2}\right] \\ \text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{12} = \sigma^2\left[\frac{-\overline{x}}{\sum_{i=1}^n (x_i-\overline{x})^2}\right], \end{aligned}$$

where the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix (for simple linear regression) is on the previous page. These are the same formulas we derived previously; see Fact 2 (pp 105, notes).

Terminology: In the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, we have derived the sampling distribution of the least squares estimator

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right).$$

We now turn our attention to estimating σ^2, the error variance. In vector notation, the **error (residual) sum of squares** is

$$\text{SSE} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = (\mathbf{Y} - \widehat{\mathbf{Y}})'(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{e}'\mathbf{e}.$$

The $n \times 1$ vector $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ contains the least squares **fitted values**, that is,

$$\widehat{\mathbf{Y}} = \begin{pmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \vdots \\ \widehat{Y}_n \end{pmatrix}_{n\times 1}.$$

The $n \times 1$ vector $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ contains the least squares **residuals**, that is,

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} - \begin{pmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \vdots \\ \widehat{Y}_n \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}_{n \times 1}.$$

Terminology: Note that the vector of fitted values

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where the $(n \times n)$ matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

In regression analysis, it is common to call $\mathbf{H}$ the **hat matrix** because it "puts the hat on $\mathbf{Y}$;" i.e., $\mathbf{HY} = \widehat{\mathbf{Y}}$. That hat matrix has interesting properties:

- $\mathbf{H}$ is symmetric, i.e., $\mathbf{H}' = \mathbf{H}$.
- $\mathbf{H}$ is idempotent, i.e., $\mathbf{H}^2 = \mathbf{H}$.
- $\mathbf{HX} = \mathbf{X}$, i.e., $\mathbf{H}$ projects each column of $\mathbf{X}$ onto itself.
- $\mathbf{H}$ is the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$.

The vector of residuals can also be written in terms of $\mathbf{H}$. Note that

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y},$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. The matrix $\mathbf{I} - \mathbf{H}$ also has interesting properties:

- $\mathbf{I} - \mathbf{H}$ is symmetric, i.e., $(\mathbf{I} - \mathbf{H})' = \mathbf{I} - \mathbf{H}$.
- $\mathbf{I} - \mathbf{H}$ is idempotent, i.e., $(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} - \mathbf{H}$.
- $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$, i.e., $\mathbf{I} - \mathbf{H}$ projects each column of $\mathbf{X}$ to the $\mathbf{0}$ vector.
- $\mathbf{I} - \mathbf{H}$ is the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})^{\perp}$, the orthogonal complement of $\mathcal{C}(\mathbf{X})$.

Interesting: Because

$$\widehat{\mathbf{Y}} = \mathbf{HY} \in \mathcal{C}(\mathbf{X}) \quad \text{and} \quad \mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \in \mathcal{C}(\mathbf{X})^{\perp},$$

this means the $n \times 1$ vectors $\widehat{\mathbf{Y}}$ and $\mathbf{e}$ are **orthogonal**; i.e., the inner product

$$\widehat{\mathbf{Y}}'\mathbf{e} = \sum_{i=1}^{n} \widehat{Y}_i e_i = 0.$$

This geometric discovery is interesting in its own right, and we will see its utility later when we examine analysis of variance for multiple linear regression. Another interesting characteristic of least squares is the residuals from an estimated model (that includes an intercept term) must sum to 0. This can be argued geometrically. If the model includes the intercept β_0, then the first column of the design matrix $\mathbf{X}$ is $\mathbf{1} = \mathbf{1}_{n\times 1}$ and $\mathbf{1} \in \mathcal{C}(\mathbf{X})$. Because $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \in \mathcal{C}(\mathbf{X})^{\perp}$, the vectors $\mathbf{1}$ and $\mathbf{e}$ are orthogonal and hence the inner product

$$\mathbf{1}'\mathbf{e} = \sum_{i=1}^{n} e_i = 0.$$

Fact 4 (restated). For simple linear regression ($k = 1$, $p = 2$), we saw (pp 108, notes)

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

is an unbiased estimator of the error variance σ^2, that is,

$$E(\widehat{\sigma}^2) = E\left(\frac{\text{SSE}}{n-2}\right) = \sigma^2.$$

For multiple linear regression, this result generalizes immediately, that is,

$$E(\widehat{\sigma}^2) = E\left(\frac{\text{SSE}}{n-p}\right) = \sigma^2.$$

Proof. The residual sum of squares SSE can be written as a quadratic form. Note that

$$\begin{aligned}
\text{SSE} = (\mathbf{Y} - \widehat{\mathbf{Y}})'(\mathbf{Y} - \widehat{\mathbf{Y}}) &= (\mathbf{Y} - \mathbf{HY})'(\mathbf{Y} - \mathbf{HY}) \\
&= [(\mathbf{I} - \mathbf{H})\mathbf{Y}]'(\mathbf{I} - \mathbf{H})\mathbf{Y} \\
&= \mathbf{Y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{Y} \\
&= \mathbf{Y}'(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})\mathbf{Y} \;=\; \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y},
\end{aligned}$$

because $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent. Recall $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$. Therefore,

$$E(\text{SSE}) = E[\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}] = \underbrace{(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta}}_{=\,0} + \text{tr}[(\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}].$$

The first term is 0 because $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})$ and $(\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})^{\perp}$. Hence, $\mathbf{X}\boldsymbol{\beta}$ and $(\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta}$ are orthogonal and therefore $(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta} = 0$. Another way to see this is through direct algebra, that is,

$$(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\boldsymbol{\beta} - \mathbf{HX}\boldsymbol{\beta}) = (\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}\boldsymbol{\beta})'\mathbf{0} = 0.$$

For the second term, because the $\text{tr}(\cdot)$ function is linear, we have

$$\text{tr}[(\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}] = \sigma^2\text{tr}(\mathbf{I} - \mathbf{H}) = \sigma^2[\text{tr}(\mathbf{I}) - \text{tr}(\mathbf{H})] = \sigma^2\{n - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\}.$$

Recall $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ for any conformable matrices $\mathbf{A}$ and $\mathbf{B}$. Taking $\mathbf{A} = \mathbf{X}$ and $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, we can write the last expression as

$$\sigma^2\{n - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\} = \sigma^2\{n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]\} = \sigma^2[n - \text{tr}(\mathbf{I})] = \sigma^2(n-p),$$

because $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$ is $p \times p$. We have shown $E(\text{SSE}) = \sigma^2(n-p)$. Therefore,

$$E(\widehat{\sigma}^2) = E\left(\frac{\text{SSE}}{n-p}\right) = \frac{\sigma^2(n-p)}{n-p} = \sigma^2. \ \square$$

Fact 5 (restated). The pivotal quantity

$$\frac{\text{SSE}}{\sigma^2} = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

This is a generalization of Fact 5 (pp 108, notes) to multiple linear regression. Its proof is beyond the scope of this course; among other things, it would require us to be familiar with the sampling distribution of quadratic forms like $\mathbf{Y}'\mathbf{AY}$, where $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$.

Fact 6 (restated). The mean-squared error $\widehat{\sigma}^2$ and the least squares estimator $\widehat{\boldsymbol{\beta}}$ are **independent**. The proof of this result is also beyond the scope of this course.

12.4.2 Statistical inference

Relevance: Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$. In our matrix formulation, the model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Just as we did in simple linear regression (Section 12.2.2, notes), we target three statistical inference questions:

- inference for (population-level) regression parameters β_j, for $j = 0, 1, 2, ..., k$
- confidence intervals for the population mean $E(Y)$ when $\mathbf{x} = \mathbf{x}^*$
- prediction intervals for the random variable Y^* when $\mathbf{x} = \mathbf{x}^*$.

This treatment generalizes our discussion of these same topics for simple linear regression. Later, we will present a statistical inference approach to test

$$\begin{gathered} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ \text{versus} \\ H_a : \text{ not } H_0, \end{gathered}$$

which can be used to assess the significance of the overall model.

Inference for β_j: Under our model assumptions, recall the least squares estimator

$$\widehat{\beta}_j \sim \mathcal{N}(\beta_j, c_{jj}\sigma^2) \implies Z = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\sigma^2}} \sim \mathcal{N}(0,1),$$

where c_{jj} is the corresponding diagonal entry in the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix; i.e.,

$$\begin{array}{rcll} c_{00} & = & (\mathbf{X}'\mathbf{X})^{-1}_{11}, & j=0 \\ c_{jj} & = & (\mathbf{X}'\mathbf{X})^{-1}_{j+1,j+1}, & j=1,2,...,k. \end{array}$$

Recall Fact 5, which says

$$W = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_j$ (Fact 6), it follows that Z and W are also independent. Therefore,

$$T = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\widehat{\sigma}^2}} = \frac{\dfrac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\sigma^2}}}{\sqrt{\dfrac{(n-p)\widehat{\sigma}^2}{\sigma^2}\Big/(n-p)}} \sim t(n-p).$$

Because T is pivotal, we can write

$$\begin{aligned} 1-\alpha = P(-t_{n-p,\alpha/2} < T < t_{n-p,\alpha/2}) &= P\left(-t_{n-p,\alpha/2} < \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\widehat{\sigma}^2}} < t_{n-p,\alpha/2}\right) \\ &= P\left(\widehat{\beta}_j - t_{n-p,\alpha/2}\sqrt{c_{jj}\widehat{\sigma}^2} < \beta_j < \widehat{\beta}_j + t_{n-p,\alpha/2}\sqrt{c_{jj}\widehat{\sigma}^2}\right), \end{aligned}$$

where $t_{n-p,\alpha/2}$ is the upper $\alpha/2$ quantile of the $t(n-p)$ distribution. Therefore,

$$\widehat{\beta}_j \pm t_{n-p,\alpha/2}\sqrt{c_{jj}\widehat{\sigma}^2}$$

is a $100(1-\alpha)\%$ confidence interval for β_j. To perform a hypothesis test for

$$\begin{array}{c} H_0 : \beta_j = \beta_{j,0} \\ \text{versus} \\ H_a : \beta_j \neq \beta_{j,0}, \end{array}$$

where $\beta_{j,0}$ is a specified value (often, $\beta_{j,0} = 0$), we would use

$$T = \frac{\widehat{\beta}_j - \beta_{j,0}}{\sqrt{c_{jj}\widehat{\sigma}^2}}$$

as a test statistic and

$$\text{RR} = \{t : |t| > t_{n-p,\alpha/2}\}$$

as a level α rejection region. One sided tests would use a suitably adjusted rejection region. Probability values are computed as areas under the $t(n-p)$ distribution.

Remark: Confidence intervals and hypothesis tests for β_j can help us to assess the importance of using the independent variable x_j in a (population-level) linear regression model that includes the other independent variables. If we test $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$, we are basically asking

> *Is there a linear relationship between $E(Y)$ and x_j in the population of individuals after accounting for the other variables in the model?*

That is, inference regarding β_j is always **conditional** on the other variables being included in the model. For example, consider the linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon,$$

in Example 12.2 (pp 127, notes). Testing

$$\begin{array}{c} H_0 : \beta_4 = 0 \\ \text{versus} \\ H_a : \beta_4 \neq 0 \end{array}$$

allows us to assess the linear relationship between the expected energy $E(Y)$ and x_4 (moisture) in a population-level regression model that already includes x_1, x_2, and x_3 (paper, plastic, and garbage, respectively).

Confidence interval for $E(Y)$ when $\mathbf{x} = \mathbf{x}^*$: In the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, the quantity

$$\theta = E(Y|\mathbf{x}^*) = \beta_0 + \beta_1 x^* + \beta_2 x_2^* + \cdots + \beta_k x_k^*$$

is the population mean of Y when

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_k^* \end{pmatrix} = \mathbf{x}^*.$$

To include the intercept term β_0 in our calculations, define

$$\mathbf{a} = \begin{pmatrix} 1 \\ x_1^* \\ x_2^* \\ \vdots \\ x_k^* \end{pmatrix}_{p \times 1}$$

so that the population mean

$$\theta = \mathbf{a}'\boldsymbol{\beta} = \beta_0 + \beta_1 x^* + \beta_2 x_2^* + \cdots + \beta_k x_k^*.$$

A natural point estimator of θ is

$$\widehat{\theta} = \mathbf{a}'\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x^* + \widehat{\beta}_2 x_2^* + \cdots + \widehat{\beta}_k x_k^*,$$

where $\widehat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$. Note that

$$E(\widehat{\theta}) = E(\mathbf{a}'\widehat{\boldsymbol{\beta}}) = \mathbf{a}'E(\widehat{\boldsymbol{\beta}}) = \mathbf{a}'\boldsymbol{\beta} = \theta,$$

showing that $\widehat{\theta}$ is **unbiased**. The variance of $\widehat{\theta}$ is

$$V(\widehat{\theta}) = V(\mathbf{a}'\widehat{\boldsymbol{\beta}}) = \mathbf{a}'\text{Cov}(\widehat{\boldsymbol{\beta}})\mathbf{a} = \mathbf{a}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a} = \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}.$$

Finally, we know $\widehat{\theta}$ is normally distributed because it is a linear function of $\widehat{\boldsymbol{\beta}}$, which is multivariate normal. Therefore, we have shown

$$\widehat{\theta} \sim \mathcal{N}\left(\theta, \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}\right) \implies Z = \frac{\widehat{\theta} - \theta}{\sqrt{\sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}} \sim \mathcal{N}(0,1).$$

Recall Fact 5, which says

$$W = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\boldsymbol{\beta}}$ (Fact 6), it follows that Z and W are also independent. Therefore,

$$T = \frac{\widehat{\theta} - \theta}{\sqrt{\widehat{\sigma}^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}} = \frac{\dfrac{\widehat{\theta} - \theta}{\sqrt{\sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}}{\sqrt{\dfrac{(n-p)\widehat{\sigma}^2}{\sigma^2}\Big/(n-p)}} \sim t(n-p).$$

Because T is pivotal, we can write

$$\begin{aligned}
1-\alpha &= P(-t_{n-p,\alpha/2} < T < t_{n-p,\alpha/2}) \\
&= P\left(-t_{n-p,\alpha/2} < \frac{\widehat{\theta} - \theta}{\sqrt{\widehat{\sigma}^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}} < t_{n-p,\alpha/2}\right) \\
&= P\left(\widehat{\theta} - t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}} < \theta < \widehat{\theta} + t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}\right),
\end{aligned}$$

where $t_{n-p,\alpha/2}$ is the upper $\alpha/2$ quantile of the $t(n-p)$ distribution. Therefore,

$$\widehat{\theta} \pm t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}} \implies \mathbf{a}'\widehat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$$

is a $100(1-\alpha)\%$ **confidence interval** for the population mean

$$\theta = \mathbf{a}'\boldsymbol{\beta} = E(Y|\mathbf{x}^*).$$

That is, for the population of individuals with $\mathbf{x} = \mathbf{x}^*$, we are $100(1-\alpha)\%$ confident the mean of this population is between the endpoints above.

Note: In the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

one uses

$$\mathbf{a} = \begin{pmatrix} 1 \\ x^* \end{pmatrix}, \quad \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}, \quad \text{and} \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \dfrac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2} & -\dfrac{\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ -\dfrac{\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} & \dfrac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}. \end{pmatrix}.$$

It is straightforward (but tedious) to show the preceding formula collapses to

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}$$

as we derived previously (see pp 110-111, notes).

Prediction interval for Y when $\mathbf{x} = \mathbf{x}^*$: Our goal is to construct a $100(1-\alpha)\%$ prediction interval for Y^*, a new value of Y when $\mathbf{x} = \mathbf{x}^*$. This interval derivation mimics the one for simple linear regression so only the salient points are noted. The natural point predictor is

$$\widehat{Y}^* = \mathbf{a}'\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x^* + \widehat{\beta}_2 x_2^* + \cdots + \widehat{\beta}_k x_k^*,$$

and the prediction error

$$U = Y^* - \widehat{Y}^* \sim \mathcal{N}\left(0, \sigma^2[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]\right) \implies Z = \frac{Y^* - \widehat{Y}^*}{\sqrt{\sigma^2[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]}} \sim \mathcal{N}(0,1).$$

Recall Fact 5, which says

$$W = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

Because $\widehat{\sigma}^2$ is independent of Y^* (why?) and $\widehat{Y}^*$ (Fact 6), it follows that Z and W are also independent. Therefore,

$$T = \frac{Y^* - \widehat{Y}^*}{\sqrt{\widehat{\sigma}^2[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]}} = \frac{\dfrac{Y^* - \widehat{Y}^*}{\sqrt{\sigma^2[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]}}}{\sqrt{\dfrac{(n-p)\widehat{\sigma}^2}{\sigma^2}\Big/(n-p)}} \sim t(n-p).$$

Finishing the derivation uses this sampling distribution. It follows that

$$\widehat{Y}^* \pm t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}\right]}$$

is a $100(1-\alpha)\%$ **prediction interval** for Y^*. Note the presence of the extra "1" in the estimated standard error, just as we saw in simple linear regression, when compared to the corresponding confidence interval for $E(Y|\mathbf{x}^*)$.

Example 12.2 (continued). We use R to estimate the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

for $i = 1, 2, ..., 30$, for the waste data in Example 12.2 (notes, pp 127). Recall the response variable

$$Y = \text{ energy content of solid waste specimen}$$

and the four independent variables

$$\begin{aligned} x_1 &= \text{plastic by weight (measured as \% of total weight)} \\ x_2 &= \text{paper by weight (measured as \% of total weight)} \\ x_3 &= \text{garbage by weight (measured as \% of total weight)} \\ x_4 &= \text{moisture percentage.} \end{aligned}$$

Here is the output:

```
> fit = lm(energy~plastic+paper+garbage+moisture)
> summary(fit)

Call:
lm(formula = energy ~ plastic + paper + garbage + moisture)

Residuals:
   Min     1Q Median     3Q    Max
-41.32 -24.03 -11.01  22.55  59.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2244.923    177.902  12.619 2.43e-12 ***
plastic       28.925      2.824  10.244 1.97e-10 ***
paper          7.644      2.314   3.303  0.00288 **
garbage        4.297      1.916   2.242  0.03406 *
moisture     -37.354      1.834 -20.365  < 2e-16 ***

Residual standard error: 31.48 on 25 degrees of freedom
Multiple R-squared: 0.9641,  Adjusted R-squared: 0.9583
F-statistic: 167.7 on 4 and 25 DF,  p-value: < 2.2e-16
```

R output: The `Estimate` output gives the least squares estimate

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \\ \widehat{\beta}_4 \end{pmatrix} = \begin{pmatrix} 2244.923 \\ 28.925 \\ 7.643 \\ 4.297 \\ -37.354 \end{pmatrix},$$

which we have already seen (pp 131, notes) when we calculated $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ ourselves using matrix operations directly. The estimated model is

$$\widehat{Y} = 2244.923 + 28.925x_1 + 7.643x_2 + 4.297x_3 - 37.354x_4.$$

The `Std.Error` output gives the *estimated* standard errors

$$\begin{aligned}
177.902 &= \widehat{\text{se}}(\widehat{\beta}_0) = \sqrt{c_{00}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{11}} \\
2.824 &= \widehat{\text{se}}(\widehat{\beta}_1) = \sqrt{c_{11}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{22}} \\
2.314 &= \widehat{\text{se}}(\widehat{\beta}_2) = \sqrt{c_{22}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{33}} \\
1.916 &= \widehat{\text{se}}(\widehat{\beta}_3) = \sqrt{c_{33}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{44}} \\
1.834 &= \widehat{\text{se}}(\widehat{\beta}_4) = \sqrt{c_{44}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{55}},
\end{aligned}$$

where the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix is on pp 130 (notes) and the mean-squared error

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{30-5} \approx (31.48)^2$$

is the square of the `Residual standard error` (see output). Recall

$$\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the 30×30 hat matrix. The error (residual) sums of squares SSE can be calculated directly in R as follows:

```
> X = cbind(intercept,plastic,paper,garbage,moisture) # design matrix
> Y = energy # response
> I = diag(30) # 30 by 30 identity matrix
> H = X%*%solve(t(X)%*%X)%*%t(X) # hat matrix
> SSE = t(Y)%*%(I-H)%*%Y # residual sum of squares
> SSE
         [,1]
[1,] 24779.22
```

so that

$$\widehat{\sigma}^2 = \frac{24779.22}{30-5} \approx 991.16.$$

We can calculate the fitted value vector $\widehat{\mathbf{Y}} = \mathbf{HY}$ and the residual vector $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ using the definitions in the R code above:

```
> Y.hat = H%*%Y # vector of fitted values
> e = (I-H)%*%Y # vector of residuals
```

Recall $\widehat{\mathbf{Y}}'\mathbf{e} = 0$ (i.e., the fitted value and residual vectors are orthogonal) and $\mathbf{1}'\mathbf{e} = 0$ (the residuals sum to zero). This can be "verified" by using R:

```
> t(Y.hat)%*%e
              [,1]
[1,] -1.432306e-06
> sum(e)
[1] -1.086953e-09
```

Note that modest rounding error is incurred when carrying out these matrix calculations directly (especially when inverting $\mathbf{X}'\mathbf{X}$). The output

```
           t value Pr(>|t|)
(Intercept) 12.619 2.43e-12 ***
plastic     10.244 1.97e-10 ***
paper        3.303  0.00288 **
garbage      2.242  0.03406 *
moisture   -20.365  < 2e-16 ***
```

gives values of the five test statistics

$$T = \frac{\widehat{\beta}_j - 0}{\sqrt{c_{jj}\widehat{\sigma}^2}},$$

for $j = 0, 1, 2, 3, 4$. For example, as noted earlier (pp 140, notes), testing

$$\begin{array}{c} H_0 : \beta_4 = 0 \\ \text{versus} \\ H_a : \beta_4 \neq 0 \end{array}$$

allows us to assess the linear relationship between the expected energy $E(Y)$ and x_4 (moisture) in a population-level regression model that already includes x_1, x_2, and x_3 (plastic, paper, and garbage, respectively). Two-sided probability values are in `Pr(>|t|)`. The probability value for the test above is

$$\text{p-value} = P_{H_0}(|T| > 20.365) < 2 \times 10^{-16},$$

indicating the evidence against H_0 is overwhelming. The random variable T satisfies $T \overset{H_0}{\sim} t(25)$, noting that $n-p = 30-5 = 25$. Confidence intervals for the population-level regression parameters β_j are not provided in the `fit` summary output, but they can be calculated easily. A 95% confidence interval for β_4 is calculated as

$$\widehat{\beta}_4 \pm t_{25,0.025}\widehat{\text{se}}(\widehat{\beta}_4) \;\longrightarrow\; -37.354 \pm 2.059(1.834) \;\longrightarrow\; (-41.13, -33.58).$$

Interpretation: For the population of waste specimens, we would expect the energy content to decrease between 33.58 and 41.13 kcal/kg for each one percentage increase in moisture (x_4). This statement is conditional on the values of the other variables x_1 (plastic), x_2 (paper), and x_3 (garbage) being included in the model and remaining fixed.

Estimating $E(Y|\mathbf{x}^*)$ and predicting Y^*: To illustrate estimation and prediction at a given value of $\mathbf{x} = \mathbf{x}^*$, set

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ x_2^* \\ x_3^* \\ x_4^* \end{pmatrix} = \begin{pmatrix} \overline{x}_{+1} \\ \overline{x}_{+2} \\ \overline{x}_{+3} \\ \overline{x}_{+4} \end{pmatrix} = \begin{pmatrix} 19.9 \\ 23.4 \\ 39.3 \\ 50.5 \end{pmatrix},$$

the vector of sample means of each independent variable (i.e., the "center of gravity" of the 30 measurements of $\mathbf{x}$). In R, confidence intervals and prediction intervals are calculated by using the `predict` function (see complete code online):

```
> predict(fit,...,level=0.95,interval="confidence")
       fit      lwr      upr
1 1281.267 1269.429 1293.105
> predict(fit,...,level=0.95,interval="prediction")
       fit      lwr      upr
1 1281.267 1215.355 1347.179
```

The output `fit` gives the point estimate/point prediction

$$\begin{aligned}&\widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \widehat{\beta}_2 x_2^* + \widehat{\beta}_3 x_3^* + \widehat{\beta}_4 x_4^* \\ &\qquad \approx 2244.923 + 28.925(19.9) + 7.643(23.4) + 4.297(39.3) - 37.354(50.5) \approx 1281.267.\end{aligned}$$

Interpreting the intervals shown above is done as follows:

- Among all waste specimens with independent variable measurements specified by $\mathbf{x}^*$ above, we are 95% confident the population mean energy content $E(Y|\mathbf{x}^*)$ is between 1269.429 and 1293.105 kcal/kg.
- For an individual waste specimen with independent variable measurements specified by $\mathbf{x}^*$ above, its energy content Y^* will fall between 1215.355 and 1347.179 kcal/kg with probability 0.95. □

12.5 Analysis of variance for linear regression models

Remark: The overall fit of a linear regression model (simple or multiple) can be summarized by using an **analysis of variance (ANOVA).** The results of this analysis are often presented in a table to show how variability in the response data $Y_1, Y_2, ..., Y_n$ is partitioned into different sources. This partition allows us to assess the overall fit of the model.

Recall: Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, or, in matrix notation,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Recall $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix, and $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ and $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ are the vectors of fitted values and residuals, respectively. The matrix $\mathbf{I}$ is the $n \times n$ identity matrix.

Approach: To create an analysis of variance partition, start with the simple quadratic form $\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{I}\mathbf{Y}$. Note that

$$\begin{aligned}
\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{I}\mathbf{Y} &= \mathbf{Y}'(\mathbf{H} + \mathbf{I} - \mathbf{H})\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{H}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{H}\mathbf{H}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})\mathbf{Y} = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} + \mathbf{e}'\mathbf{e},
\end{aligned}$$

because both $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are symmetric and idempotent. This equation can be expressed equivalently as

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

We use the following terminology:

$$\begin{aligned}
\text{(uncorrected) } \textbf{total} \text{ sum of squares} &\longrightarrow \mathbf{Y}'\mathbf{I}\mathbf{Y} = \mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} Y_i^2 \\
\text{(uncorrected) } \textbf{regression} \text{ sum of squares} &\longrightarrow \mathbf{Y}'\mathbf{H}\mathbf{Y} = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} = \sum_{i=1}^{n} \widehat{Y}_i^2 \\
\textbf{error (residual)} \text{ sum of squares} &\longrightarrow \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.
\end{aligned}$$

This shows how the (uncorrected) total sum of squares $\mathbf{Y}'\mathbf{I}\mathbf{Y} = \mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} Y_i^2$ can be partitioned into two parts: one part due to estimating the linear regression model and the other part which is "left over" after estimating the model.

Convention: When we estimate a linear regression model, we are usually interested in the regression coefficients that are attached to independent variables; i.e., $\beta_1, \beta_2, ..., \beta_k$. We are usually not interested in the intercept term β_0, the population mean of Y when each independent variable equals 0. Therefore, it is common to "remove the effects" of estimating this overall mean. This can be accomplished by subtracting $n\overline{Y}^2$ from both sides of the last equation, that is,

$$\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = \sum_{i=1}^{n} \widehat{Y}_i^2 - n\overline{Y}^2 + \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2,$$

which is algebraically the same as

$$\underbrace{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}_{\text{SSE}}.$$

We call

$$\begin{aligned} \text{(corrected) } \textbf{total} \text{ sum of squares} &\longrightarrow \text{SST} \\ \text{(corrected) } \textbf{regression} \text{ sum of squares} &\longrightarrow \text{SSR} \\ \textbf{error (residual)} \text{ sum of squares} &\longrightarrow \text{SSE}, \end{aligned}$$

and these quantities obey

$$\text{SST} = \text{SSR} + \text{SSE}.$$

This partition still shows how variation in the response values $Y_1, Y_2, ..., Y_n$ can be partitioned into two parts (one part due to the model and the "left over" part), but now the part due to the model disregards the intercept term. Note that

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

is simply the numerator of the sample variance

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

which we know is an overall measure of variation in the response values $Y_1, Y_2, ..., Y_n$.

Interesting: We already know the error (residual) sum of squares SSE can be written as a quadratic form, that is,

$$\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

The other sums of squares, SST and SSR, can also be written as quadratic forms. To see how, note that the correction term

$$n\overline{Y}^2 = \mathbf{Y}' n^{-1} \mathbf{J} \mathbf{Y},$$

where

$$\mathbf{J} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}_{n \times n}$$

is the $n \times n$ matrix of ones. The matrix $\mathbf{J}$ can be written as $\mathbf{J} = \mathbf{1}\mathbf{1}'$, where

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times 1}.$$

Therefore,

$$\mathbf{Y}' n^{-1} \mathbf{J} \mathbf{Y} = n^{-1} \mathbf{Y}' \mathbf{1}\mathbf{1}' \mathbf{Y} = \frac{1}{n} \left(\sum_{i=1}^{n} Y_i \right)^2 = \frac{1}{n} (n\overline{Y})^2 = n\overline{Y}^2,$$

as claimed. The corrected partition can now be written as

$$\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'n^{-1}\mathbf{J}\mathbf{Y} = \mathbf{Y}'\mathbf{H}\mathbf{Y} - \mathbf{Y}'n^{-1}\mathbf{J}\mathbf{Y} + \mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}$$

or, equivalently,

$$\underbrace{\mathbf{Y}'(\mathbf{I}-n^{-1}\mathbf{J})\mathbf{Y}}_{\text{SST}} = \underbrace{\mathbf{Y}'(\mathbf{H}-n^{-1}\mathbf{J})\mathbf{Y}}_{\text{SSR}} + \underbrace{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}}_{\text{SSE}}.$$

Result: If $\mathbf{A}$ is an idempotent matrix, then

$$\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A}).$$

This fact might be proven in a linear algebra course. Its proof relies on the fact that the trace of a matrix $\mathbf{A}$ equals the sum of the eigenvalues of $\mathbf{A}$ and all eigenvalues of an idempotent matrix must be either 0 or 1.

Revelation: The matrices $\mathbf{I}-n^{-1}\mathbf{J}$, $\mathbf{H}-n^{-1}\mathbf{J}$, and $\mathbf{I}-\mathbf{H}$ in the ANOVA partition above are all idempotent! For example, note that

$$\begin{aligned}(\mathbf{I}-n^{-1}\mathbf{J})^2 = (\mathbf{I}-n^{-1}\mathbf{J})(\mathbf{I}-n^{-1}\mathbf{J}) &= \mathbf{I}^2 - n^{-1}\mathbf{J}\mathbf{I} - \mathbf{I}n^{-1}\mathbf{J} + n^{-2}\mathbf{J}^2 \\ &= \mathbf{I} - n^{-1}\mathbf{J} - n^{-1}\mathbf{J} + n^{-2}(\mathbf{1}\mathbf{1}')^2.\end{aligned}$$

Now, write

$$n^{-2}(\mathbf{1}\mathbf{1}')^2 = n^{-2}\mathbf{1}\mathbf{1}'\mathbf{1}\mathbf{1}' = n^{-2}\mathbf{1}n\mathbf{1}' = n^{-1}\mathbf{1}\mathbf{1}' = n^{-1}\mathbf{J},$$

where we have used the fact that $\mathbf{1}'\mathbf{1} = n$. Therefore,

$$(\mathbf{I}-n^{-1}\mathbf{J})^2 = \mathbf{I} - n^{-1}\mathbf{J} - n^{-1}\mathbf{J} + n^{-1}\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{J},$$

showing that $\mathbf{I}-n^{-1}\mathbf{J}$ is idempotent. Showing $\mathbf{H}-n^{-1}\mathbf{J}$ and $\mathbf{I}-\mathbf{H}$ are idempotent is done similarly.

Interesting: The ranks of the matrices $\mathbf{I}-n^{-1}\mathbf{J}$, $\mathbf{H}-n^{-1}\mathbf{J}$, and $\mathbf{I}-\mathbf{H}$ correspond to the degrees of freedom attached to SST, SSR, and SSE, respectively. For example,

$$\text{rank}(\mathbf{I}-n^{-1}\mathbf{J}) = \text{tr}(\mathbf{I}-n^{-1}\mathbf{J}) = \text{tr}(\mathbf{I}) - n^{-1}\text{tr}(\mathbf{J}) = n - n^{-1}n = n-1.$$

Recall $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ for any conformable matrices $\mathbf{A}$ and $\mathbf{B}$. Note that

$$\text{rank}(\mathbf{H}) = \text{tr}(\mathbf{H}) = \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \text{tr}(\mathbf{I}) = p,$$

because $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$ is $p \times p$. Therefore,

$$\text{rank}(\mathbf{H}-n^{-1}\mathbf{J}) = \text{tr}(\mathbf{H}-n^{-1}\mathbf{J}) = \text{tr}(\mathbf{H}) - n^{-1}\text{tr}(\mathbf{J}) = p - n^{-1}n = p-1$$

and

$$\text{rank}(\mathbf{I}-\mathbf{H}) = \text{tr}(\mathbf{I}-\mathbf{H}) = \text{tr}(\mathbf{I}) - \text{tr}(\mathbf{H}) = n-p.$$

Terminology: It is common to organize sums of squares (SST, SSR, and SSE) and their degrees of freedom in an **ANOVA table**, which generally looks like this:

Source	df	SS	MS	F
Regression	$p-1$	$\text{SSR} = \mathbf{Y}'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{Y}$	$\text{MSR} = \dfrac{\text{SSR}}{p-1}$	$F = \dfrac{\text{MSR}}{\text{MSE}}$
Error	$n-p$	$\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$	$\text{MSE} = \dfrac{\text{SSE}}{n-p} = \widehat{\sigma}^2$	
Total	$n-1$	$\text{SST} = \mathbf{Y}'(\mathbf{I} - n^{-1}\mathbf{J})\mathbf{Y}$		

Mean squares: In the ANOVA table above, mean squares (MS) are formed by taking the sums of squares (SS) and dividing by the corresponding degrees of freedom (df). On pp 137-138 (notes), we showed MSE is an unbiased estimator of σ^2, that is,

$$E(\text{MSE}) = E\left(\frac{\text{SSE}}{n-p}\right) = \sigma^2,$$

by making use of the fact that $\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$; i.e., SSE is a quadratic form in $\mathbf{Y}$. What about $E(\text{MSR})$? *When is MSR also an unbiased estimator of σ^2?*

Investigation: Recall $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$ under our linear model assumptions. We have

$$E(\text{SSR}) = E[\mathbf{Y}'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{Y}] = (\mathbf{X}\boldsymbol{\beta})'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{X}\boldsymbol{\beta} + \text{tr}[(\mathbf{H} - n^{-1}\mathbf{J})\sigma^2\mathbf{I}].$$

The second term

$$\text{tr}[(\mathbf{H} - n^{-1}\mathbf{J})\sigma^2\mathbf{I}] = \sigma^2\text{tr}(\mathbf{H} - n^{-1}\mathbf{J}) = \sigma^2(p-1),$$

as shown on the previous page. Because

$$\text{MSR} = \frac{\text{SSR}}{p-1},$$

it follows that MSR will be an unbiased estimator of σ^2 when the first term

$$(\mathbf{X}\boldsymbol{\beta})'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{X}\boldsymbol{\beta} = 0.$$

This occurs when $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. To see why, note that

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \end{pmatrix} \begin{pmatrix} \beta_0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \beta_0\mathbf{1}.$$

Therefore,

$$\begin{aligned}
(\mathbf{X}\boldsymbol{\beta})'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{X}\boldsymbol{\beta} &= (\beta_0\mathbf{1})'(\mathbf{H} - n^{-1}\mathbf{J})\beta_0\mathbf{1} \\
&= \beta_0^2\mathbf{1}'(\mathbf{H}\mathbf{1} - n^{-1}\mathbf{J}\mathbf{1}) \\
&= \beta_0^2\mathbf{1}'(\mathbf{1} - n^{-1}\mathbf{1}\mathbf{1}'\mathbf{1}) \\
&= \beta_0^2\mathbf{1}'(\mathbf{1} - \mathbf{1}) = \beta_0^2\mathbf{1}'\mathbf{0} = 0,
\end{aligned}$$

and hence

$$E(\text{SSR}) = \sigma^2(p-1).$$

Above, we have used the fact that $\mathbf{1}'\mathbf{1} = n$ and $\mathbf{H}\mathbf{1} = \mathbf{1}$. The latter fact is true because $\mathbf{H}$ projects vectors onto $\mathcal{C}(\mathbf{X})$ and $\mathbf{1} \in \mathcal{C}(\mathbf{X})$ already.

Summary: The argument above shows two things.

1. When $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ is true, then

$$E(\text{MSR}) = E\left(\frac{\text{SSR}}{p-1}\right) = \frac{E(\text{SSR})}{p-1} = \frac{\sigma^2(p-1)}{p-1} = \sigma^2,$$

   that is, MSR is also an unbiased estimator of σ^2. When this is true,

$$F = \frac{\text{MSR}}{\text{MSE}}$$

   is the ratio of two **unbiased** estimators of σ^2. Because MSR and MSE are both estimating the same thing on average, we would expect F to be around 1.

2. When $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ is *not* true, then

$$\begin{aligned}
E(\text{MSR}) = E\left(\frac{\text{SSR}}{p-1}\right) &= \frac{E(\text{SSR})}{p-1} \\
&= \frac{(\mathbf{X}\boldsymbol{\beta})'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{X}\boldsymbol{\beta} + \sigma^2(p-1)}{p-1} \\
&= \frac{(\mathbf{X}\boldsymbol{\beta})'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{X}\boldsymbol{\beta}}{p-1} + \sigma^2 > \sigma^2
\end{aligned}$$

   because $(\mathbf{X}\boldsymbol{\beta})'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{X}\boldsymbol{\beta} > 0$. That is, MSR is estimating something *larger* than σ^2 on average (maybe a lot larger depending on where $\mathbf{X}\boldsymbol{\beta}$ is). In this situation, we would expect

$$F = \frac{\text{MSR}}{\text{MSE}}$$

   to be "'large" (i.e., larger than 1). MSE is still estimating σ^2 on average, but MSR is estimating something larger than σ^2 on average (possibly much larger). This will make F large.

Remark: I find the preceding summary provides a compelling conceptual understanding of how to perform the hypothesis test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$\text{versus}$$
$$H_a : \text{at least one of the } \beta_j \text{ is nonzero.}$$

We have shown that values of F around 1 are consistent with H_0, and "large" values of F are consistent with H_a. Therefore, H_0 should be rejected when F is large. Note that this conceptual explanation of the "F test" does not rely on the (multivariate) normal assumption for $\mathbf{Y}$. It relies only on

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \text{ and } \text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I} \iff E(\boldsymbol{\epsilon}) = \mathbf{0} \text{ and } \text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I},$$

the so-called first and second order moment assumptions for $\boldsymbol{\epsilon}$.

Remark: When we include the multivariate normal assumption for $\boldsymbol{\epsilon}$, we can derive the exact sampling distribution of F when H_0 is true. Using distribution theory for quadratic forms like $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ (under multivariate normality), a more advanced treatment of linear models would show

$$\frac{\text{SSR}}{\sigma^2} \overset{H_0}{\sim} \chi^2(p-1), \quad \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-p), \quad \text{and} \quad \text{SSR} \perp\!\!\!\perp \text{SSE}.$$

Therefore,

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} = \frac{\dfrac{\text{SSR}/\sigma^2}{p-1}}{\dfrac{\text{SSE}/\sigma^2}{n-p}} \overset{H_0}{\sim} F(p-1, n-p).$$

A level α rejection region to test H_0 versus H_a is

$$\text{RR} = \{F > F_{p-1,n-p,\alpha}\},$$

where $F_{p-1,n-p,\alpha}$ is the upper α quantile of the F distribution with $p-1$ (numerator) and $n-p$ (denominator) degrees of freedom. Probability values are computed as areas to the right of F on the $F(p-1, n-p)$ distribution.

Q: Why would we ever want to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ to begin with?
A: When H_0 is true, the (population-level) multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

reduces to

$$Y = \beta_0 + \epsilon,$$

that is, none of the independent variables $x_1, x_2, ..., x_k$ are linearly related to $E(Y)$ in the population. Therefore, the F test above provides an "overall assessment" of the regression model. Of course, if H_0 is rejected, then we do not know which β_j's are different from 0 (or how many). Recall we have already discussed how to perform individual inference for the β_j's as needed (see pp 139, notes).

Example 12.2 (continued). We use R to construct the ANOVA table with the waste data (pp 127, notes). The multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. Here, $n = 30$ and $p = 5$. We have already calculated

$$\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \approx 24779.2.$$

The sample variance of $Y_1, Y_2, ..., Y_{30}$ is

$$S^2 = \frac{1}{30-1} \underbrace{\sum_{i=1}^{30} (Y_i - \overline{Y})^2}_{\text{SST}} \implies \text{SST} = 29S^2 \approx 689709.9$$

so that

$$\text{SSR} = \text{SST} - \text{SSE} \approx 689709.9 - 24779.2 = 664930.7,$$

by subtraction.

```
> SST = 29*var(Y)
> SST
[1] 689709.9
```

The complete ANOVA table is below:

Source	df	SS	MS	F
Regression	4	664930.7	166232.7	167.7
Error	25	24779.2	991.2	
Total	29	689709.9		

Analysis: If $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is true, that is, if none of x_1 (plastic), x_2 (paper), x_3 (garbage), and x_4 (moisture) are linearly related to expected energy $E(Y)$ in the population, then we would expect F to be "close" to 1. This is clearly not the case here. A level $\alpha = 0.05$ rejection region for testing

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$\text{versus}$$
$$H_a : \text{at least one of the } \beta_j \text{ is nonzero}$$

is

$$\text{RR} = \{F > F_{4,25,0.05}\} = \{F > 2.76\}.$$

Because $F \approx 167.7$, this indicates the evidence against H_0 is overwhelming; see Figure 12.5 (next page). That is, there is overwhelming evidence that at least one of the independent variables x_1, x_2, x_3, x_4 is linearly related to expected energy $E(Y)$ in the population.

```
> qf(0.95,4,25)
[1] 2.75871
```

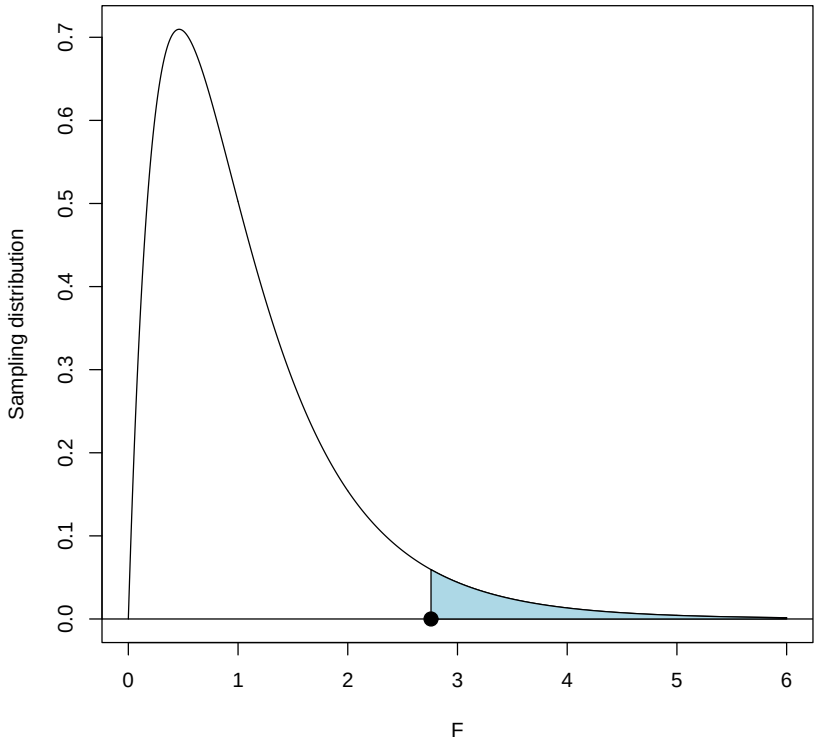

Figure 12.5: $F(4, 25)$ pdf. This pdf represents the sampling distribution of F when $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is true. The level $\alpha = 0.05$ rejection region RR $= \{F > 2.76\}$ is shown shaded.

Q: How does R summarize the ANOVA table?
A: Instead of presenting the overall partition (SST = SSR + SSE) as on the previous page, R takes the (corrected) regression sum of squares

$$664930.7 \approx \text{SSR} = \mathbf{Y}'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{Y}$$

($n = 30$) and partitions it into the components contributed by the four independent variables x_1, x_2, x_3, and x_4 separately. For the waste data, recall the design matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 \end{pmatrix}.$$

Define the hat matrices

$$\begin{aligned}
\mathbf{H}_1 &= \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' \quad \text{corresponding to} \quad \mathbf{X}_1 = \begin{pmatrix} \mathbf{1} & \mathbf{x}_1 \end{pmatrix} \\
\mathbf{H}_2 &= \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' \quad \text{corresponding to} \quad \mathbf{X}_2 = \begin{pmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 \end{pmatrix} \\
\mathbf{H}_3 &= \mathbf{X}_3(\mathbf{X}_3'\mathbf{X}_3)^{-1}\mathbf{X}_3' \quad \text{corresponding to} \quad \mathbf{X}_3 = \begin{pmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{pmatrix}
\end{aligned}$$

and, of course,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad \text{corresponding to} \quad \mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 \end{pmatrix},$$

the full design matrix. Observe that

$$\mathbf{H} - n^{-1}\mathbf{J} = (\mathbf{H}_1 - n^{-1}\mathbf{J}) + (\mathbf{H}_2 - \mathbf{H}_1) + (\mathbf{H}_3 - \mathbf{H}_2) + (\mathbf{H} - \mathbf{H}_3).$$

A stunning geometric interpretation is that the 4 matrices on the RHS are perpendicular projection matrices onto orthogonal subspaces of the rank 4 vector space $\mathcal{C}(\mathbf{H} - n^{-1}\mathbf{J})$. For us, this means we can write

$$\begin{aligned}
\text{SSR} &= \mathbf{Y}'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{Y} \\
&= \mathbf{Y}'(\mathbf{H}_1 - n^{-1}\mathbf{J})\mathbf{Y} + \mathbf{Y}'(\mathbf{H}_2 - \mathbf{H}_1)\mathbf{Y} + \mathbf{Y}'(\mathbf{H}_3 - \mathbf{H}_2)\mathbf{Y} + \mathbf{Y}'(\mathbf{H} - \mathbf{H}_3)\mathbf{Y} \\
&= \text{SSR}(\mathbf{x}_1) + \text{SSR}(\mathbf{x}_2|\mathbf{x}_1) + \text{SSR}(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2) + \text{SSR}(\mathbf{x}_4|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3).
\end{aligned}$$

These are called **sequential sum of squares** because they account for how regression sums of squares accumulate in sequence; i.e., as independent variables "are added to the model." In other words,

$$\begin{aligned}
\text{SSR}(\mathbf{x}_1) &\longrightarrow \text{SS from regressing on } \mathbf{x}_1 \text{ only (while including } \beta_0) \\
\text{SSR}(\mathbf{x}_2|\mathbf{x}_1) &\longrightarrow \text{additional SS from regressing on } \mathbf{x}_2 \text{ (including } \mathbf{x}_1 \text{ and } \beta_0) \\
\text{SSR}(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2) &\longrightarrow \text{additional SS from regressing on } \mathbf{x}_3 \text{ (including } \mathbf{x}_1, \mathbf{x}_2, \text{ and } \beta_0) \\
\text{SSR}(\mathbf{x}_4|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) &\longrightarrow \text{additional SS from regressing on } \mathbf{x}_4 \text{ (including } \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \text{ and } \beta_0).
\end{aligned}$$

R output: Here is the ANOVA partition R provides for the waste data:

```
> fit = lm(energy~plastic+paper+garbage+moisture)
> anova(fit)
Analysis of Variance Table
          Df Sum Sq Mean Sq  F value    Pr(>F)
plastic    1 239735  239735 241.8709 2.311e-14 ***
paper      1  11239   11239  11.3392  0.002458 **
garbage    1   2888    2888   2.9136  0.100231
moisture   1 411069  411069 414.7313 < 2.2e-16 ***
Residuals 25  24779     991
```

It is straightforward to verify the partition

$$664930.7 \approx \text{SSR} = 239735 + 11239 + 2888 + 411069$$

(up to rounding error). We know $\text{SSR} = \mathbf{Y}'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{Y}$ has 4 degrees of freedom in total; i.e., $\text{rank}(\mathbf{H} - n^{-1}\mathbf{J}) = 4$. One degree of freedom is allocated to each of the sequential SS above because this is the dimension of each orthogonal subspace of $\mathcal{C}(\mathbf{H} - n^{-1}\mathbf{J})$.

Sequential F tests: The F statistics in the R output above are

$$\begin{aligned}
F_1 = \frac{\text{SSR}(\mathbf{x}_1)/1}{\text{MSE}} &\approx 241.9 \\
F_2 = \frac{\text{SSR}(\mathbf{x}_2|\mathbf{x}_1)/1}{\text{MSE}} &\approx 11.3 \\
F_3 = \frac{\text{SSR}(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)/1}{\text{MSE}} &\approx 2.9 \\
F_4 = \frac{\text{SSR}(\mathbf{x}_4|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)/1}{\text{MSE}} &\approx 414.7.
\end{aligned}$$

What exactly are these statistics used to test? In words,

- F_1 allows us to test whether x_1 (plastic) "adds to the regression" when compared to a model that includes only the intercept β_0.
- F_2 allows us to test whether x_2 (paper) "adds to the regression" when compared to a model that includes x_1 (plastic) and the intercept β_0.
- F_3 allows us to test whether x_3 (garbage) "adds to the regression" when compared to a model that includes x_1 (plastic), x_2 (paper), and the intercept β_0.
- F_4 allows us to test whether x_4 (moisture) "adds to the regression" when compared to a model that includes x_1 (plastic), x_2 (paper), x_3 (garbage), and the intercept β_0.

Large values of F indicate that the additional independent variable is important after accounting for the variables that preceded it. Under multivariate normality, a level α rejection region for tests of this type in general is

$$\text{RR} = \{F > F_{1,n-p,\alpha}\},$$

where $F_{1,n-p,\alpha}$ is the upper α quantile of the $F(1, n-p)$ distribution (here $n - p = 25$). For the tests above, values of $F > F_{1,25,\alpha}$ indicate the corresponding variable does add to the regression model significantly (after accounting for the variables that preceded it). Probability values in `Pr(>F)` are areas to the right of F under the $F(1, 25)$ pdf.

Interesting: Sequential sums of squares are used to assess the relative contribution of each independent variable as it is added to the model in sequence. Therefore, if you change the ordering of the independent variables x_1, x_2, x_3, x_4, you change the partition of SSR! Before, we used

```
> fit = lm(energy~plastic+paper+garbage+moisture)
```

which adds the independent variables x_1, x_2, x_3, and x_4 in this order. Using

```
> fit = lm(energy~garbage+moisture+paper+plastic)
```

adds x_3, x_4, x_2, and x_1 in this order. Here is the ANOVA partition R provides for this ordering:

```
> fit = lm(energy~garbage+moisture+paper+plastic)
> anova(fit)
Analysis of Variance Table
          Df Sum Sq Mean Sq  F value    Pr(>F)
garbage    1   5245    5245   5.2916   0.03005 *
moisture   1 555276  555276 560.2233 < 2.2e-16 ***
paper      1    402     402   0.4060   0.52980
plastic    1 104007  104007 104.9340 1.968e-10 ***
Residuals 25  24779     991
```

It is straightforward to verify the partition

$$664930.7 \approx \text{SSR} = 5245 + 555276 + 402 + 104007$$

(up to rounding error). However, note that our assessment of the relative contributions of each independent variable changes to reflect this new ordering.

- In the first partition, x_3 (garbage) did not add significantly to a model that included x_1 (plastic), x_2 (paper), and the intercept; p-value ≈ 0.100. However, in the second partition, x_3 (garbage) does add significantly when compared to a model that includes only the intercept; p-value ≈ 0.030.
- In the first partition, x_2 (paper) added significantly to a model that included x_1 (plastic) and the intercept; p-value ≈ 0.002. However, in the second partition, x_2 (paper) does not add significantly to a model that includes x_3 (garbage), x_4 (moisture), and the intercept; p-value ≈ 0.530.

These are not contradictory findings. The differences in the conclusions are based entirely on the ordering of the independent variables; i.e., the F statistics (and probability values) for the two partitions are simply testing different things. And both partitions test different hypotheses than we saw earlier with the `summary` output (pp 143, notes):

```
> fit = lm(energy~plastic+paper+garbage+moisture)
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2244.923    177.902  12.619 2.43e-12 ***
plastic       28.925      2.824  10.244 1.97e-10 ***
paper          7.644      2.314   3.303  0.00288 **
garbage        4.297      1.916   2.242  0.03406 *
moisture     -37.354      1.834 -20.365  < 2e-16 ***
```

This output gives t statistics of the form

$$T = \frac{\widehat{\beta}_j - 0}{\sqrt{c_{jj}\widehat{\sigma}^2}},$$

which are used to test

$$H_0 : \beta_j = 0$$
$$\text{versus}$$
$$H_a : \beta_j \neq 0.$$

Inference here is conditional on *all other independent variables* being included in the model. For example, the `summary` output above shows x_2 (paper) adds significantly to a model that includes x_1 (plastic), x_3 (garbage), x_4 (moisture), and the intercept; p-value ≈ 0.003. Compare this with the second sequential SS partition, where x_2 (paper) does not add significantly to a model that includes only x_3 (garbage), x_4 (moisture), and the intercept. $\square$

13 Survival Analysis

13.1 Introduction

Remark: The statistical analysis of **lifetime data** is important in many areas, including biomedical applications (e.g., clinical trials, etc.), engineering, and actuarial science. The term "lifetime" means "time to event," where an event may refer to death, part failure, insurance claim, natural disaster, eradication of infection, etc.

- In chronic disease clinical trials; e.g., trials involving cancer, diabetes, cardiovascular disease, etc., the primary endpoint (variable) of interest may be time to death, time to relapse of disease, time to disease progression, etc. For such trials, we are usually interested in comparing the distribution of the time to event among two or more treatments.
- Typically, clinical trials occur over a finite period of time; therefore, the time to event is not measured on all patients in the study. This results in what is referred to as **censored data**. Also, because patients generally enter a clinical trial at different calendar times (staggered entry), the amount of follow-up time varies for different individuals.
- The combination of censoring and staggered entry creates challenges in the analysis of such data that do not allow basic statistical techniques to be used. This area of statistics is called **survival analysis**.

Example 13.1. A randomized clinical trial (RCT) involves 64 cancer patients with severe aplastic anemia. This occurs when an individual's bone marrow stops making enough new blood cells. This is a serious condition; patients who are left untreated usually die in less than one year. Prior to the trial, all 64 patients were treated with a high dose of cyclophosphamide (a drug designed to prepare patients for transplant by lowering the body's immune system), followed by an infusion of bone marrow from a family member. Patients were then randomly assigned to one of two treatment groups:

- Group 1: Cyclosporine and methotrexate (CSP+MTX)
- Group 2: Methotrexate only (MTX).

Cyclosporine also lowers the body's immune system (to prevent rejection of marrow from a donor). Methotrexate is designed to slow the growth of cancer cells. In this trial, the primary endpoint (variable) was

$$T = \text{time from treatment assignment until diagnosis of AGVHD.}$$

Acute graft versus host disease (AGVHD) is a condition where the donor's bone marrow cells attack the patient's organs and tissue. One goal of the trial was to compare the distribution of T for both treatment groups.

CSP+MTX				MTX only			
3*	65	324	528*	9*	25	104*	395*
8	77*	356*	547*	11	28	106*	428*
10	82*	378*	691	12	28	156	469
12*	98*	408*	769*	20*	31	218	602
16	155*	411	1111*	20	35*	230*	681*
17	189	420*	1173	22	35*	231*	690
22	199*	449*	1213*	25	46	316*	1112*
64*	247*	490	1357	25*	49*	393	1180

Table 13.1: RCT data. Time to diagnosis of AGVHD. Starred entries represent censored observations.

Data: Table 13.1 gives the times to diagnosis (in days) for the 64 patients. Note that only 30 of the 64 patents actually "reached the endpoint" (i.e., were actually diagnosed with AGVHD). The remaining 34 patients were **censored**, that is, these patients were never diagnosed with AGVHD.

- What probability distribution should we use to model the diagnosis times?
- What effects do censoring and staggered entry have on the resulting analysis?
- Figure 13.1 (next page) displays estimates of the **survivor functions**. How are these constructed? Is the difference between the two groups statistically significant? □

13.2 Describing the distribution of time to an event

Terminology: Let T denote the time to event. It is understood to mean that T is a nonnegative random variable for which there is an unambiguous start (e.g., point of infection, start of treatment, etc.) and an unambiguous end (e.g., death, diagnosis, etc.) with the time in between corresponding to T. Therefore, $P(T \geq 0) = 1$. Random variables T with positive (nonnegative) support are called **lifetime random variables**. For example,

- T = survival time (from birth to death)
- T = time from treatment of disease to death (this may be tricky if individuals die from "other causes;" more about this later)
- T = time to diagnosis of a more severe condition (e.g., Alzheimer's disease, etc.).

Remark: The time of interest may not always correspond to something deleterious such as "death." For example, we may consider the time to the eradication of an infection, measured from the initiation of an antibiotic used to treat patients. In this situation, it is preferable to *shorten* the distribution of times, whereas, in the other situations (e.g., when death is the endpoint), it is desirable to *lengthen* time.

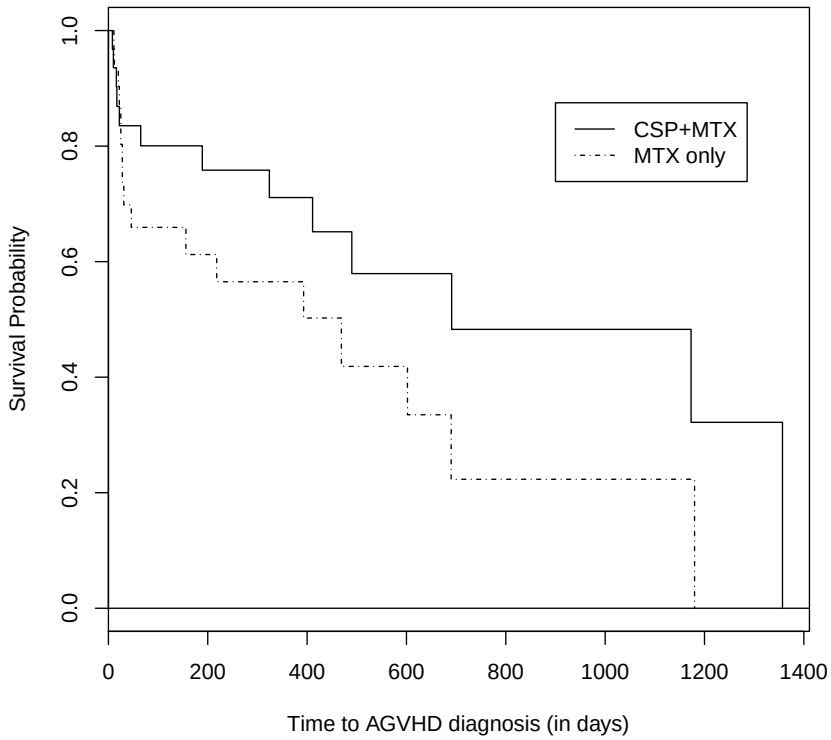

Figure 13.1: RCT data. Kaplan-Meier estimates for the time to diagnosis of AGVHD for two treatment groups.

Review: We now describe some different, but equivalent, ways of defining the distribution of a lifetime random variable T (the time to an event). In our discussion, we assume that T is continuous.

- The **cumulative distribution function** (cdf):

$$F_T(t) = P(T \leq t).$$

  This is the proportion of the population that has experienced the event at or before time t. If the event is something bad like "death" or "failure," then $F_T(t)$ is the proportion that has "failed" by time t.

- The **survivor function**:

$$S_T(t) = P(T > t) = 1 - F_T(t).$$

  This is the proportion of the population that has *not* failed at or before time t; i.e., the proportion that is "still alive" at time t.

- The **probability density function**:

$$f_T(t) = \frac{d}{dt}F_T(t) = -\frac{d}{dt}S_T(t).$$

Also, recall

$$F_T(t) = \int_0^t f_T(u)du \quad \text{and} \quad S_T(t) = \int_t^\infty f_T(u)du.$$

Note the lower limit on the integral for $F_T(t)$ is 0 because T is a lifetime random variable (it has nonnegative support).

Example 13.2. A simple parametric model for T is the exponential distribution with mean $\beta > 0$. Recall if $T \sim \text{exponential}(\beta)$, then the pdf of T is

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

The cdf of T is

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

The survivor function of T is

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

A graph of the survivor function appears in Figure 13.2 (next page, left) when $\beta = 2$. Note that

$$S_T(1) = e^{-1/2} \approx 0.607,$$

that is, approximately 60.7% of the population "is alive at" or "has survived up to" time $t = 1$. Also,

$$S_T(\phi_{0.5}) = e^{-\phi_{0.5}/2} \stackrel{\text{set}}{=} 0.5 \implies \frac{\phi_{0.5}}{2} = -\ln 0.5 \implies \phi_{0.5} = 2\ln 2 \approx 1.39;$$

i.e., the median survival is $\phi_{0.5} \approx 1.39$. $\square$

Terminology: We say the distribution of a survival time T_1 is **stochastically larger** than another survival time T_2, and write $T_1 \geq_{\text{st}} T_2$, if the survivor function of T_1 is greater than or equal to the survivor function of T_2 for all t; that is,

$$S_{T_1}(t) = P(T_1 > t) \geq P(T_2 > t) = S_{T_2}(t),$$

for all $t \geq 0$. In other words, "T_1 tends to be larger than T_2." See Figure 13.2 (next page, right).

Terminology: The **mortality rate** at time t is the proportion of the population who "fail" between times t and $t + 1$ among those individuals "alive" at time t. This is a conditional probability and is given by

$$m_T^*(t) = P(t \leq T < t + 1 | T \geq t).$$

Usually, t is an integer of some unit of time (e.g., day, month, year, etc.).

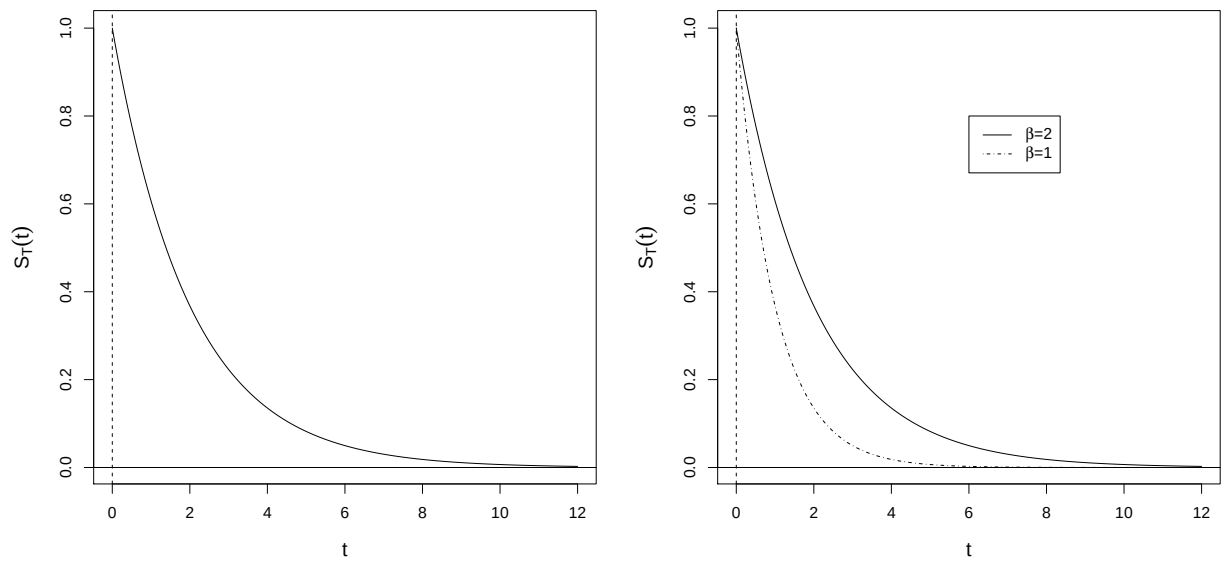

Figure 13.2: Left: Survivor function of $T \sim$ exponential($\beta = 2$). Right: Survivor functions of $T_1 \sim$ exponential($\beta = 2$) and $T_2 \sim$ exponential($\beta = 1$). Note that $T_1 \geq_{\text{st}} T_2$.

Q: Suppose

$$T = \text{survival time (from birth to death)}$$

for the human population. What does the corresponding mortality rate (mortality function) $m_T^*(t)$ look like?

Terminology: The **hazard rate** is simply a "continuous version" of a mortality rate, defined as follows:

$$\lambda_T(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h | T \geq t)}{h}.$$

In other words, the hazard rate $\lambda_T(t)$ is the limit of the mortality rate when the interval of time is taken to be arbitrarily small. The hazard rate is the instantaneous rate of failure at time t, given that the individual is alive at time t. Note the hazard rate is not a probability; rather, it is a *probability rate*. Therefore, it is possible that a hazard rate may exceed one.

Remark: The hazard rate (or hazard function) is a very important characteristic of a lifetime distribution. It indicates the way the risk of failure varies over time, and this is of interest in most applications.

- Distributions with **increasing** hazard functions are seen for individuals for whom some kind of aging or "wear out" takes place (e.g., people, car batteries, etc.).
- Certain types of devices may actually display a **decreasing** hazard function (i.e., the population of individuals "strengthens" over time).

Examples of hazard functions and their shapes are shown in Figure 13.3 (next page).

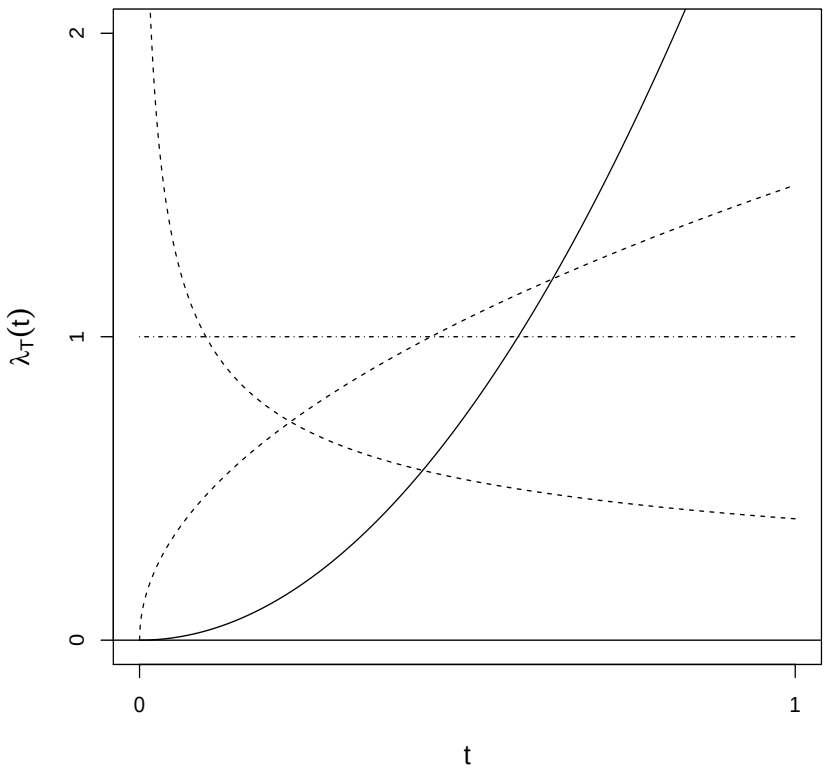

Figure 13.3: Examples of hazard functions.

Observation: Note that

$$\begin{aligned}
\lambda_T(t) &= \lim_{h\to 0} \frac{P(t \le T < t+h | T \ge t)}{h} \\
&= \lim_{h\to 0} \frac{P(t \le T < t+h)}{hP(T \ge t)} \\
&= \frac{1}{P(T \ge t)} \lim_{h\to 0} \frac{F_T(t+h) - F_T(t)}{h} = \frac{f_T(t)}{S_T(t)} = \frac{-\frac{d}{dt}S_T(t)}{S_T(t)} = -\frac{d}{dt}\ln\{S_T(t)\}.
\end{aligned}$$

Integrating both sides of the last equation, we get

$$-\ln\{S_T(t)\} = \int_0^t \lambda_T(u)du = \Lambda_T(t),$$

the so called **cumulative hazard function**. Consequently,

$$S_T(t) = \exp\left\{-\int_0^t \lambda_T(u)du\right\} = \exp\left\{-\Lambda_T(t)\right\}.$$

Because of these one-to-one relationships, we can describe the distribution of a continuous survival time T by using $f_T(t)$, $F_T(t)$, $S_T(t)$, $\lambda_T(t)$, or $\Lambda_T(t)$.

Example 13.2 (continued). Let's calculate the hazard function $\lambda_T(t)$ for $T \sim \text{exponential}(\beta)$. We have

$$\lambda_T(t) = \frac{f_T(t)}{S_T(t)} = \frac{\frac{1}{\beta}e^{-t/\beta}}{e^{-t/\beta}} = \frac{1}{\beta},$$

which is a constant function of t. Therefore, the *rate* of failure is constant over time. In other words, failures will occur, but the rate of failure does not increase or decrease over time. Does this remind you of a "special property" that (for continuous distributions) only the exponential distribution enjoys? $\square$

Example 13.3. Another parametric model which is commonly assumed in engineering is $T \sim \text{Weibull}(\beta, \eta)$, where $\beta > 0$ and $\eta > 0$. The pdf of T can be written as

$$f_T(t) = \begin{cases} \dfrac{\beta}{\eta}\left(\dfrac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^\beta}, & t > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where

$$\begin{aligned} \beta &= \text{shape parameter} \\ \eta &= \text{scale parameter.} \end{aligned}$$

This pdf has been parameterized differently than our book so as to facilitate an interpretation commonly used in engineering applications. The cdf of T is

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

The survivor function of T is

$$S_T(t) = 1 - F_T(t) = \begin{cases} 1, & t \leq 0 \\ e^{-(t/\eta)^\beta}, & t > 0. \end{cases}$$

Therefore, the hazard function, for $t > 0$, is

$$\lambda_T(t) = \frac{f_T(t)}{S_T(t)} = \frac{\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^\beta}}{e^{-(t/\eta)^\beta}} = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}.$$

Note: For $T \sim \text{Weibull}(\beta, \eta)$, it is easy to show

- $\lambda_T(t)$ is increasing if $\beta > 1$ (population gets weaker with aging)
- $\lambda_T(t)$ is constant if $\beta = 1$ (constant hazard; exponential distribution)
- $\lambda_T(t)$ is decreasing if $\beta < 1$ (population gets stronger with aging).

Therefore, under the Weibull model assumption for T, the strength of the population over time can be described exclusively through the value of the shape parameter β. $\square$

Example 13.4. The data below are taken from Xu et al. (2003, *Applied Soft Computing*), who describe a reliability study on turbochargers in diesel engines. These are failure time data for $n = 40$ turbochargers; the failure time T is measured in 1000s of hours. All turbochargers eventually failed; i.e., no times below were censored.

1.6	2.0	2.6	3.0	3.5	3.9	4.5	4.6	4.8	5.0
5.1	5.3	5.4	5.6	5.8	6.0	6.0	6.1	6.3	6.5
6.5	6.7	7.0	7.1	7.3	7.3	7.3	7.7	7.7	7.8
7.9	8.0	8.1	8.3	8.4	8.4	8.5	8.7	8.8	9.0

Under the assumption that $T \sim \text{Weibull}(\beta, \eta)$, the likelihood function of β and η (assuming independent units) is given by

$$L(\beta, \eta|\mathbf{t}) = \prod_{i=1}^{40} \frac{\beta}{\eta}\left(\frac{t_i}{\eta}\right)^{\beta-1} e^{-(t_i/\eta)^\beta} = \left(\frac{\beta}{\eta^\beta}\right)^{40}\left(\prod_{i=1}^{40} t_i\right)^{\beta-1} e^{-\sum_{i=1}^{40}(t_i/\eta)^\beta}.$$

Maximizing $L(\beta, \eta|\mathbf{t})$ or $\ln L(\beta, \eta|\mathbf{t})$ is difficult to do analytically, so numerical optimization methods are preferred. In R, maximizing $L(\beta, \eta|\mathbf{t})$ numerically can be carried out by using the `fitdist` function as follows:

```
> fitdist(turbo,"weibull")
Fitting of the distribution ‘weibull’ by maximum likelihood
Parameters:
      estimate Std. Error
shape 3.873157  0.5176799
scale 6.920191  0.2946851
```

Therefore, the maximum likelihood estimates of β and η based on these data (and under the Weibull model assumption) are

$$\begin{aligned} \widehat{\beta} &\approx 3.87 \\ \widehat{\eta} &\approx 6.92. \end{aligned}$$

The estimated standard errors are obtained numerically from the second derivative matrix (Hessian) of the log-likelihood function. Because MLEs are approximately normally distributed in large samples, an approximate 95% confidence interval for β is

$$\widehat{\beta} \pm 1.96 \times \widehat{\text{se}}(\widehat{\beta}) \quad \longrightarrow \quad 3.87 \pm 1.96(0.52) \quad \longrightarrow \quad (2.85, 4.89).$$

Therefore, we are 95% confident (under the Weibull model assumption) that the population parameter β is between 2.85 and 4.89. By the invariance property of maximum likelihood estimators, the MLE of the survivor function $S_T(t)$ is

$$\widehat{S}_T(t) = \begin{cases} 1, & t \leq 0 \\ e^{-(t/6.92)^{3.87}}, & t > 0, \end{cases}$$

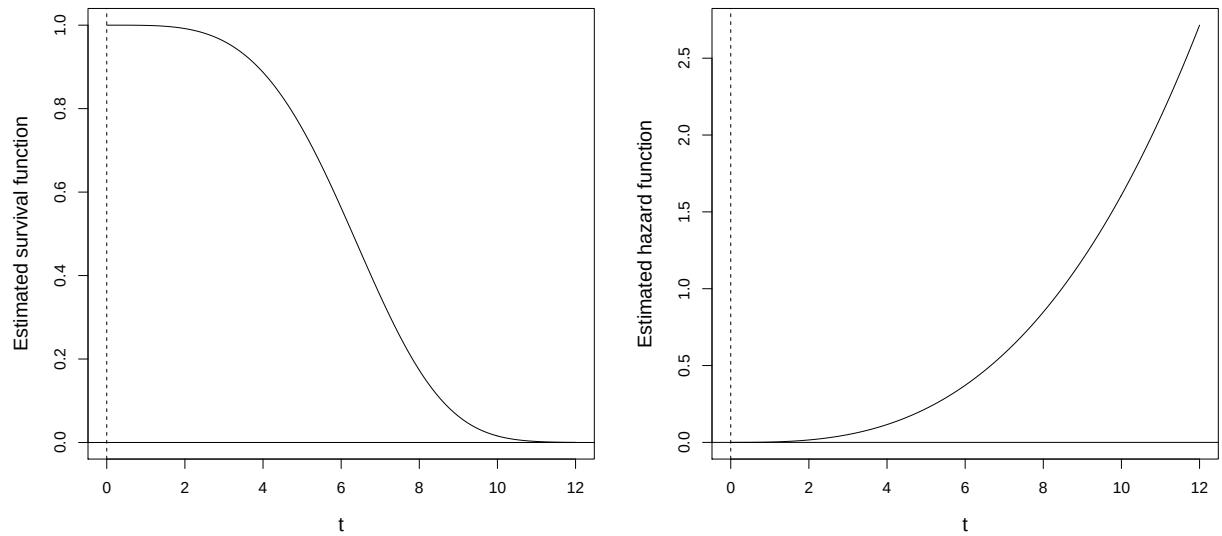

Figure 13.4: Turbocharger data. Left: Estimated survivor function $\widehat{S}_T(t)$. Right: Estimated hazard function $\widehat{\lambda}_T(t)$.

and the estimated hazard function is

$$\widehat{\lambda}_T(t) = \frac{3.87}{6.92}\left(\frac{t}{6.92}\right)^{3.87-1}.$$

The estimated survivor and hazard functions are shown in Figure 13.4 (above). □

Remark: Investigators in different areas approach the analysis of lifetime data from different perspectives. A **parametric approach** is generally espoused in engineering and actuarial science applications (see Example 13.4). That is, a parametric probability model is assigned to describe the distribution of the random variable T. Common model choices include

- exponential
- Weibull
- log-normal
- gamma
- other less well known distributions, such as log-logistic, inverse Gaussian, Pareto, log-gamma, Burr, Gompertz-Makeham, etc.

On the other hand, it is more common in biostatistics and medical applications to take a **nonparametric approach**, where the probability distribution of T is left unspecified. This is the approach we take going forward.

13.3 Censoring and life table estimates

Remark: Two important issues arise in survival analysis (in particular, in clinical trials) when time to event data are considered:

- Some individuals "are still alive" at the time of analysis (i.e., the event of interest has not yet occurred). This results in **right censored data**.
- The length of follow-up varies due to staggered entry over "calendar time." Note that "patient time" is measured from entry into the study.

In addition to censoring occurring because of insufficient follow-up (i.e., due to the study ending), it may also occur for other reasons. For example,

- loss to follow-up; e.g., the patient stops drops out of the study, moves away, etc.
- death from other causes (competing risks).

These different forms of censoring are referred to as **random right censoring**. How do we account for censoring in the analysis of survival data?

Example 13.1 (revisited). Let's once again consider the two-arm clinical trial with patients assigned to one of the following treatment groups:

- Group 1: Cyclosporine and methotrexate (CSP+MTX)
- Group 2: Methotrexate only (MTX).

For this illustration, we will consider Group 1 only (CSP+MTX; see data below):

CSP+MTX							
3*	8	10	12*	16	17	22	64*
65	77*	82*	98*	155*	189	199*	247*
324	356*	378*	408*	411	420*	449*	490
528*	547*	691	769*	1111*	1173	1213*	1357

Recall these are observations of

$$T = \text{time from treatment assignment until diagnosis of AGVHD}$$

for 32 patients. Thirteen (13) patients reached the endpoint and 19 patients were censored. Our goal is to estimate the survivor function

$$S_T(t) = P(T > t).$$

If we make a **parametric** model assumption for T, then the survivor function $S_T(t)$ will be a function of the parameters in the model. For example, suppose $T \sim \text{exponential}(\beta)$, where $\beta > 0$. Under this model assumption,

$$S_T(t) = e^{-t/\beta}, \quad t > 0.$$

Therefore, to estimate $S_T(t)$ under the exponential assumption, all we have to do is estimate β under the exponential model. Recall the sample mean

$$\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$$

is the MLE of β if the diagnosis times $T_1, T_2, ..., T_n$ are iid and are completely observed. To account for censoring (under the exponential assumption), consider the following three approaches:

1. Discard the censored observations and estimate β using ML with the complete observations only. There are 13 complete observations whose average AGVHD diagnosis time is 367.2 days. Therefore, the estimated survivor function is

$$\widehat{S}_T(t) = e^{-t/367.2}, \quad t > 0.$$

2. Ignore the censoring aspect and estimate β using ML with all the observations (treating censored observations as complete observations). There are 32 observations whose average is 371.5 days. Therefore, the estimated survivor function is

$$\widehat{S}_T(t) = e^{-t/371.5}, \quad t > 0.$$

Remark: Although the first two approaches produce similar estimates of β, both approaches are terrible. In the first, one is discarding information about the survival distribution contained in the censored observations. For example, knowing that $\{T > 1213\}$ has occurred for the 31st patient gives information about survival time. In the second, regarding censored times as observed times will also greatly distort the analysis. For the 1st patient, treating $\{T > 3\}$ as $\{T = 3\}$ is misleading as these two events mean very different things.

3. We can write out the likelihood function of β based on the complete observations *and* the censored ones. To see how, define

$$\Delta_i = \begin{cases} 1, & i\text{th time is observed} \\ 0, & i\text{th time is censored,} \end{cases}$$

for $i = 1, 2, ..., 32$. Under the iid exponential(β) assumption, the likelihood function consists of two parts: the part due to the complete observations and the part due to the censored observations, that is,

$$\begin{aligned} L(\beta) = \prod_{i=1}^{32} [f_T(t_i)]^{\Delta_i}[S_T(t_i)]^{1-\Delta_i} &= \prod_{i=1}^{32}\left(\frac{1}{\beta}e^{-t_i/\beta}\right)^{\Delta_i}\left(e^{-t_i/\beta}\right)^{1-\Delta_i} \\ &= \left(\frac{1}{\beta}\right)^{\sum_{i=1}^{32}\Delta_i} e^{-\sum_{i=1}^{32} t_i\Delta_i/\beta} \times e^{-\sum_{i=1}^{32} t_i(1-\Delta_i)/\beta} \\ &= \left(\frac{1}{\beta}\right)^{\sum_{i=1}^{32}\Delta_i} e^{-\sum_{i=1}^{32}[t_i\Delta_i + t_i(1-\Delta_i)]/\beta} \\ &= \left(\frac{1}{\beta}\right)^{r} e^{-\sum_{i=1}^{32} t_i/\beta}, \end{aligned}$$

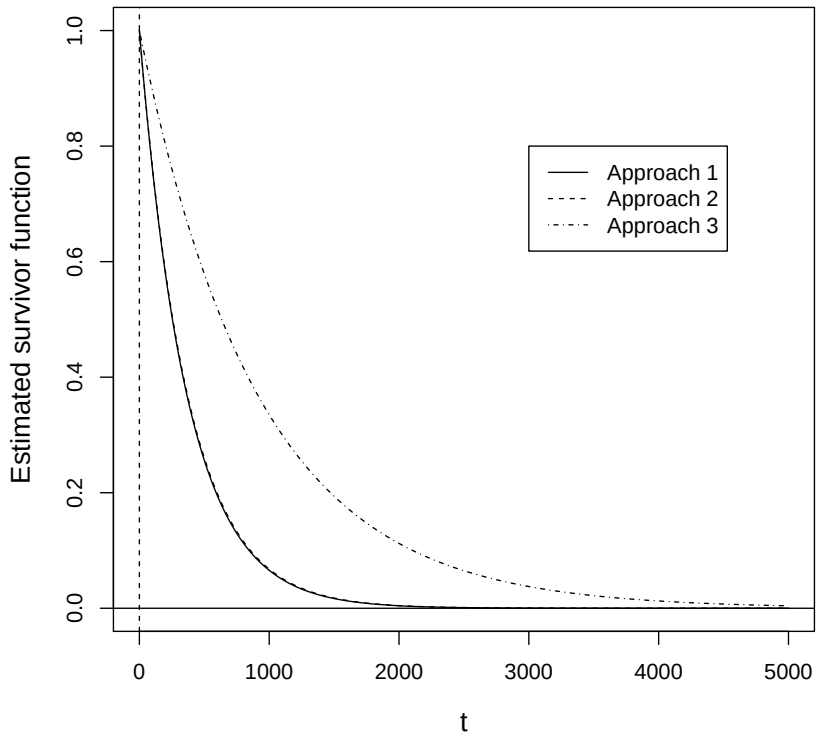

Figure 13.5: RCT data. Estimated survivor functions (under the exponential assumption) for three approaches to incorporate censoring.

where $r = \sum_{i=1}^{32} \Delta_i$ is the number of complete observations ($r = 13$). The log-likelihood function is

$$\ln L(\beta) = -r \ln \beta - \frac{\sum_{i=1}^{32} t_i}{\beta}.$$

The derivative of the log-likelihood function is

$$\frac{\partial}{\partial \beta} \ln L(\beta) = -\frac{r}{\beta} + \frac{\sum_{i=1}^{32} t_i}{\beta^2} \stackrel{\text{set}}{=} 0 \quad \Longrightarrow \quad \widehat{\beta} = \frac{1}{r} \sum_{i=1}^{32} t_i \approx 914.5.$$

Therefore, the estimated survivor function is

$$\widehat{S}_T(t) = e^{-t/914.5}, \quad t > 0.$$

Remark: Estimated survivor functions $\widehat{S}_T(t)$ from the three approaches are shown in Figure 13.5 (above). Clearly the third approach is preferred as it distinguishes between complete and censored observations and incorporates both types. However, a remaining limitation is that one is "locked in" to the exponential distribution as a population-level model for time to AGVHD diagnosis. All estimates (and resulting inference) are valid under this assumption but not necessarily otherwise. □

Remark: We now discuss estimation in survival analysis while making no parametric assumptions about the distribution of the time to event T, and we do this while accounting for censoring and staggered entry. We start with a clinical trial example where observations are grouped into one-year intervals of time.

Example 13.5. Data from $n = 146$ individuals, who previously had myocardial infarction (MI) and participated in a clinical trial for an antihypertensive drug (to lower high blood pressure), are given in Table 13.2 (below). All times are measured in terms of patient time (not calendar time). The endpoint T is time to death.

Year since entry into study	Number alive and under observation at beginning of interval	Number dying during interval	Number censored or withdrawn
$[0, 1)$	146	27	3
$[1, 2)$	116	18	10
$[2, 3)$	88	21	10
$[3, 4)$	57	9	3
$[4, 5)$	45	1	3
$[5, 6)$	41	2	11
$[6, 7)$	28	3	5
$[7, 8)$	20	1	8
$[8, 9)$	11	2	1
$[9, 10)$	8	2	6

Table 13.2: Myocardial infarction data. All data are measured in patient time.

Q: How should we estimate the five-year survival probability $S_T(5)$?
A: Two naive answers are given by

$$\frac{\text{76 deaths in 5 years}}{\text{146 individuals}} = 0.521 \implies \widehat{S}_T(5) = 0.479$$

$$\frac{\text{76 deaths in 5 years}}{\text{146}-\text{29 individuals}} = 0.650 \implies \widehat{S}_T(5) = 0.350.$$

- The first estimate would be appropriate if all 29 individuals withdrawn in the first 5 years were withdrawn (censored) exactly at the 5-year mark; i.e., at time $t = 5$. This corresponds to censoring on the **right** of the interval $[0, 5)$. This is not the case, so this estimate is overly optimistic; i.e., this overestimates $S_T(5)$.

- The second estimate would be appropriate if all 29 individuals withdrawn in the first 5 years were withdrawn (censored) immediately upon entering the study; i.e., at time $t = 0$. This corresponds to censoring on the **left** of the interval $[0, 5)$. This is not the case either, so this estimate is overly pessimistic; i.e., this underestimates $S_T(5)$.

Remark: A better estimate of $S_T(5)$ obviously is somewhere between these two extremes, one that allows for individuals to be censored while using smaller time intervals than $[0, 5)$.

Using conditional probabilities, it is easy to show $S_T(5)$ can be expressed as

$$S_T(5) = P(T \geq 5) = P(T \geq 1|T \geq 0) \times P(T \geq 2|T \geq 1) \times P(T \geq 3|T \geq 2) \times P(T \geq 4|T \geq 3) \times P(T \geq 5|T \geq 4).$$

Using the complement rule for conditional probabilities, we have

$$\begin{aligned} P(T \geq 1|T \geq 0) &= 1 - P(0 \leq T < 1|T \geq 0) = 1 - m_T^*(0) \\ P(T \geq 2|T \geq 1) &= 1 - P(1 \leq T < 2|T \geq 1) = 1 - m_T^*(1) \\ &\vdots \\ P(T \geq 5|T \geq 4) &= 1 - P(4 \leq T < 5|T \geq 4) = 1 - m_T^*(4), \end{aligned}$$

where recall $m_T^*(t)$ is the mortality rate. Therefore, we can write

$$S_T(5) = \prod_{j=1}^{5} \{1 - m_T^*(j-1)\}.$$

The mortality rate in each interval is easy to estimate nonparametrically; simply use the proportion of patients who die in each interval. Choosing which denominator to use in this proportion then depends on what we do with the censored observations.

Right censoring: Suppose any individual who is withdrawn (censored) in an interval of time is censored at the **end** of that interval (right censoring). Our table then looks like

Time	$n(t)$	$d(t)$	$w(t)$	$\widehat{m}_T^*(t) = \dfrac{d(t)}{n(t)}$	$1 - \widehat{m}_T^*(t)$	$\widehat{S}_T^R(t) = \prod\{1 - \widehat{m}_T^*(t)\}$
$[0,1)$	146	27	3	0.185	0.815	0.815
$[1,2)$	116	18	10	0.155	0.845	0.689
$[2,3)$	88	21	10	0.239	0.761	0.524
$[3,4)$	57	9	3	0.158	0.842	0.441
$[4,5)$	45	1	3	0.022	0.972	0.432

Thus, if right censoring was used, our estimate of the five-year survival probability, based on the life-table, would be $\widehat{S}_T^R(5) = 0.432$.

Left censoring: Suppose any individual who is withdrawn (censored) in an interval of time is censored at the **beginning** of that interval (left censoring). Our table then looks like

Time	$n(t)$	$d(t)$	$w(t)$	$\widehat{m}_T^*(t) = \dfrac{d(t)}{n(t) - w(t)}$	$1 - \widehat{m}_T^*(t)$	$\widehat{S}_T^L(t) = \prod\{1 - \widehat{m}_T^*(t)\}$
$[0,1)$	146	27	3	0.189	0.811	0.811
$[1,2)$	116	18	10	0.170	0.830	0.673
$[2,3)$	88	21	10	0.269	0.731	0.492
$[3,4)$	57	9	3	0.167	0.833	0.410
$[4,5)$	45	1	3	0.024	0.976	0.400

Thus, if left censoring was used, our estimate of the five-year survival probability, based on the life-table, would be $\widehat{S}_T^L(5) = 0.400$.

Observation: Naive estimates of $S_T(5)$ ranged from 0.350 to 0.479, which assumed censoring occurred at the beginning or the end of the interval $[0, 5)$, respectively. When we restricted attention to patient time on a year-by-year basis, our estimates of $S_T(5)$ ranged from 0.400 to 0.432, a tremendous improvement in precision! Therefore, sharper (nonparametric) estimates of the survivor function result when we consider smaller intervals of time.

Remark: It is likely censoring occurs at a time *inside* of each interval (i.e., not always on the endpoints). Therefore, our improved estimates $\widehat{S}_T^L(5) = 0.400$ and $\widehat{S}_T^R(5) = 0.432$ based on one-year intervals are still too pessimistic and optimistic, respectively. A compromise is to use the following table:

Time	$n(t)$	$d(t)$	$w(t)$	$\widehat{m}_T^*(t) = \dfrac{d(t)}{n(t) - w(t)/2}$	$1 - \widehat{m}_T^*(t)$	$\widehat{S}_T(t) = \prod\{1 - \widehat{m}_T^*(t)\}$
$[0, 1)$	146	27	3	0.187	0.813	0.813
$[1, 2)$	116	18	10	0.162	0.838	0.681
$[2, 3)$	88	21	10	0.253	0.747	0.509
$[3, 4)$	57	9	3	0.162	0.838	0.426
$[4, 5)$	45	1	3	0.023	0.977	0.417

This table forms the basis for the **life-table estimate** of $S_T(5)$; i.e., $\widehat{S}_T(5) = 0.417$. The denominator $n(t) - w(t)/2$ is called the **effective sample size**. A plot of the estimated survivor function is shown in Figure 13.6 (next page). This plot is based on the effective sample size life table (above) and the completed table below:

Time	$n(t)$	$d(t)$	$w(t)$	$\widehat{m}_T^*(t) = \dfrac{d(t)}{n(t) - w(t)/2}$	$1 - \widehat{m}_T^*(t)$	$\widehat{S}_T(t) = \prod\{1 - \widehat{m}_T^*(t)\}$
$[5, 6)$	41	2	11	0.056	0.944	0.393
$[6, 7)$	28	3	5	0.118	0.882	0.347
$[7, 8)$	20	1	8	0.063	0.937	0.325
$[8, 9)$	11	2	1	0.190	0.810	0.264
$[9, 10)$	8	2	6	0.400	0.600	0.158

Inference: For life-table estimators to provide unbiased results, we must assume individuals who are censored are at the same risk of failure as those who are "still alive" and uncensored. The "risk set" $n(t)$, that is, those who are still alive and uncensored, should be representative of the entire population alive at the same time. Under these assumptions, for fixed t, theoretical arguments show the life-table estimator $\widehat{S}_T(t)$ is approximately normal with mean $S_T(t)$ and variance which is consistently estimated by

$$\widehat{\sigma}^2_{\widehat{S}_T(t)} = \{\widehat{S}_T(t)\}^2 \sum_{j=1}^{t} \frac{d_j}{(n_j - w_j/2)(n_j - d_j - w_j/2)},$$

where $n_j = n(j)$, $d_j = d(j)$, and $w_j = w(j)$. The formula for $\widehat{\sigma}^2_{\widehat{S}_T(t)}$ is called **Greenwood's formula**. An approximate $100(1 - \alpha)\%$ confidence interval for $S_T(t)$ is therefore given by

$$\widehat{S}_T(t) \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{S}_T(t)},$$

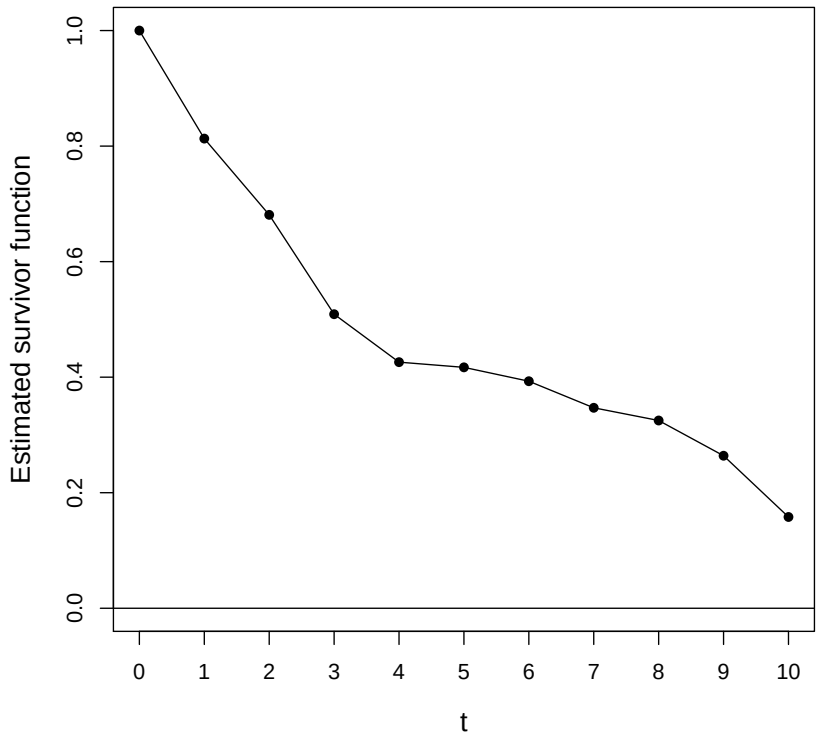

Figure 13.6: Myocardial infarction data. Life-table estimate of $S_T(t)$ in Example 13.5.

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution and $\widehat{\sigma}_{\widehat{S}_T(t)}$ is the estimated standard error.

MI data: The following table calculates estimated standard errors for the myocardial infarction data in Example 13.5:

Time	$n(t)$	$d(t)$	$w(t)$	$\widehat{S}_T(t)$	$\sum_j \frac{d_j}{(n_j - w_j/2)(n_j - d_j - w_j/2)}$	$\widehat{\sigma}_{\widehat{S}_T(t)}$
$[0,1)$	146	27	3	0.813	0.00159	0.032
$[1,2)$	116	18	10	0.681	0.00327	0.039
$[2,3)$	88	21	10	0.509	0.00735	0.044
$[3,4)$	57	9	3	0.426	0.01084	0.044
$[4,5)$	45	1	3	0.417	0.01138	0.044

Therefore, an approximate 95% confidence interval for $S_T(5)$ using the life table estimate is

$$0.417 \pm 1.96(0.044) \longrightarrow (0.331, 0.503).$$

That is, we are 95% confident the five-year survival probability $S_T(5)$ after an MI episode is between 0.331 and 0.503. $\square$

13.4 Kaplan-Meier estimator

Motivation: In Example 13.5, we saw that the bias when estimating the survivor function $S_T(t)$ incorrectly (i.e., assuming that censoring occurs at the left or the right of each time interval) decreases when the length of the interval was reduced (e.g., from five years to one year). Therefore, if the data are not grouped, that is, we know the *exact* failure and censoring times, we could apply the life-table estimator using intervals with very small lengths. The **Kaplan-Meier estimator** is the "limit" of the life-table estimator when intervals of time are so small that at most one observation occurs in a unit of time.

Example 13.6. Consider the small (fictitious) data set below. The endpoint T is time to death. We have the following death and censoring times for $n = 10$ individuals:

Time (t)	4.5	7.5	8.5	11.5	13.5	15.5	16.5	17.5	19.5	21.5
Censoring indicator	1	1	0	1	0	1	1	0	1	0

Here, "1" means the observation was a death and "0" means the observation was censored. We have 6 deaths and 4 censored observations (out of the 10 individuals). Let

$$\widehat{m}_T^*(t) = \frac{d(t)}{n(t)} = \frac{\text{number of deaths in an interval}}{\text{number at risk at beginning of the interval}}$$

be an estimate of the mortality rate at time t. Taking the interval of time to be one unit, consider the following calculations:

Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$\widehat{m}_T^*(t)$	0	0	0	0	$\frac{1}{10}$	0	0	$\frac{1}{9}$	0	0	0	$\frac{1}{7}$	0	0	0	$\frac{1}{5}$	$\frac{1}{4}$	0	0	$\frac{1}{2}$
$1 - \widehat{m}_T^*(t)$	1	1	1	1	$\frac{9}{10}$	1	1	$\frac{8}{9}$	1	1	1	$\frac{6}{7}$	1	1	1	$\frac{4}{5}$	$\frac{3}{4}$	1	1	$\frac{1}{2}$
$\widehat{S}_T(t)$	1	.	.	.	$\frac{9}{10}$	.	.	$\frac{8}{10}$	.	.	.	$\frac{48}{70}$	.	.	.	$\frac{192}{350}$	$\frac{144}{350}$	.	.	$\frac{144}{700}$

The Kaplan-Meier estimate $\widehat{S}_T(t)$ is a **step function** taking jumps precisely at those times where an event (death) occurs; see Figure 13.7 (left, next page). By convention, the Kaplan-Meier estimator is assumed to be right continuous. Pointwise confidence bands, which acknowledge uncertainty in the point estimate $\widehat{S}_T(t)$, can be obtained by using Greenwood's formula; see Figure 13.7 (right, next page). □

Objective: We now embark on a more general discussion about survival data; this discussion introduces notation that will be useful in developing the Kaplan-Meier estimator of $S_T(t)$ in the one-sample problem and the log-rank test to compare two or more survival functions (see, e.g., Example 13.1).

Discission: When describing censored survival data, it is useful to conceptualize the existence of two "latent" random variables for each individual corresponding to the failure time and the censoring time. The term "latent" means "missing" or "not observed."

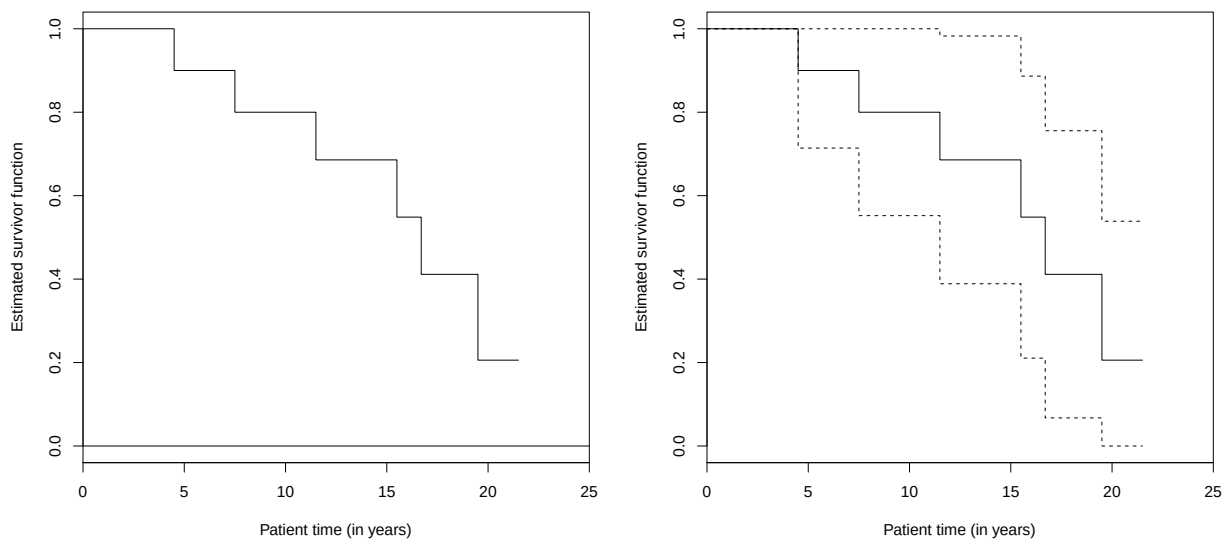

Figure 13.7: Left: Kaplan-Meier estimate of $S_T(t)$ in Example 13.6. Right: Kaplan-Meier estimate with 95% pointwise confidence bands.

- For the ith individual, denote the **failure time** by T_i and the **censoring time** by C_i. Only one of these variables is observed for the ith individual (the other is not).
- The random variable T_i corresponds to the ith individual's survival time if that individual was observed until death. The random variable C_i corresponds to the time that the ith individual is censored provided that death does not intervene first.
- For example, C_i may be the time from entry into the study until the time of analysis. Of course, censoring could occur for other reasons; e.g., loss to follow up, death from other causes, etc.

In a survival study, for the ith individual, we get to observe the **minimum** of T_i and C_i, which we denote by the random variable

$$X_i = \min\{T_i, C_i\}.$$

We also get to observe whether the individual failed (died) or was censored; i.e., we get to observe the binary random variable

$$\Delta_i = I(T_i \leq C_i) = \begin{cases} 1, & \text{if } T_i \leq C_i \\ 0, & \text{if } T_i > C_i, \end{cases}$$

where $I(\cdot)$ is the indicator function. Therefore, $\{(X_i, \Delta_i), i = 1, 2, ..., n\}$ are the observed data in a survival study, whereas T_i and C_i are latent variables which are useful in conceptualizing the problem.

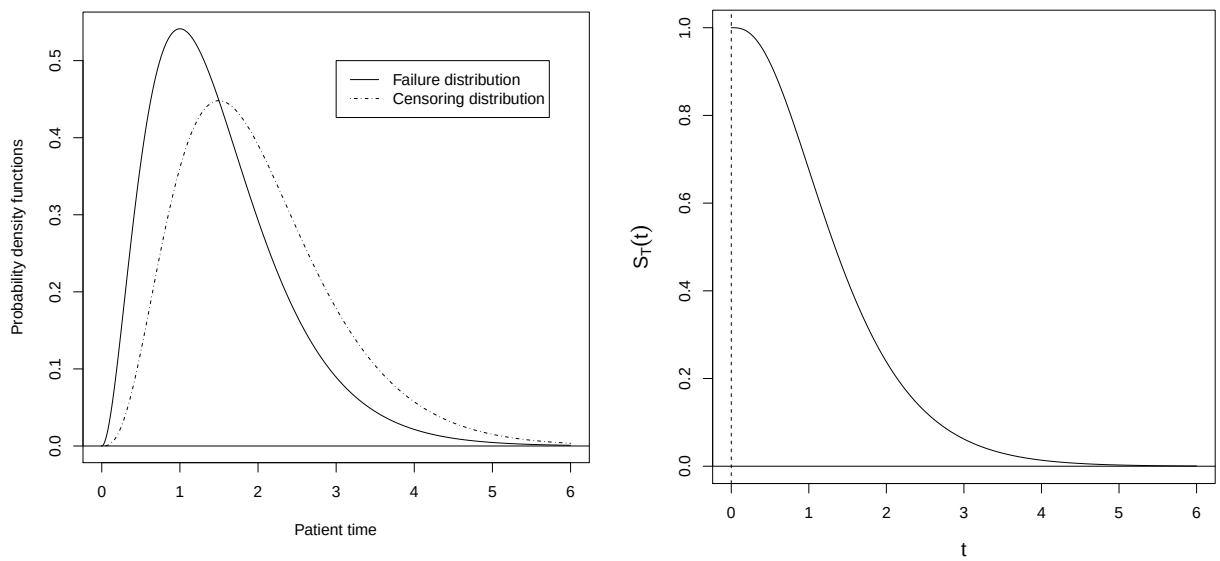

Figure 13.8: Left: Probability density functions of $T \sim \text{gamma}(3, 0.5)$ and $C \sim \text{gamma}(4, 0.5)$. Right: Survivor function of T.

Example 13.7. To gain insight on the notation previously defined, and to get a handle on the challenges that arise when estimating $S_T(t)$ with censored survival data, let's perform a small simulation study. In this study, we will assume we know the true failure and censoring distributions. Specifically, suppose

$$\begin{aligned} T &\sim \text{gamma}(\alpha = 3, \beta = 0.5) \\ C &\sim \text{gamma}(\alpha = 4, \beta = 0.5). \end{aligned}$$

Probability density functions of T and C are given in Figure 13.8 (above, left). The survivor function of T, $S_T(t)$, is shown in Figure 13.8 (above, right).

Challenge: In a censored survival study, we are attempting to estimate $S_T(t)$ with observations from both $f_T(t)$ and $f_C(c)$. Of course, both of these distributions are unknown in practice; the purpose of a simulation study like this is investigate what happens in a controlled setting where we know the underlying distributions.

Simulation study: I generated $n = 100$ observations from both distributions (independently, to emulate non-informative censoring). For each pair of observations (T_i, C_i), $i = 1, 2, ..., 100$, I calculated

$$X_i = \min\{T_i, C_i\}$$

and

$$\Delta_i = I(T_i \leq C_i) = \begin{cases} 1, & \text{if } T_i \leq C_i \\ 0, & \text{if } T_i > C_i. \end{cases}$$

Recall that $\{(X_i, \Delta_i),\ i = 1, 2, ..., 100\}$ would be the observed data in a censored survival study, not $\{(T_i, C_i),\ i = 1, 2, ..., 100\}$.

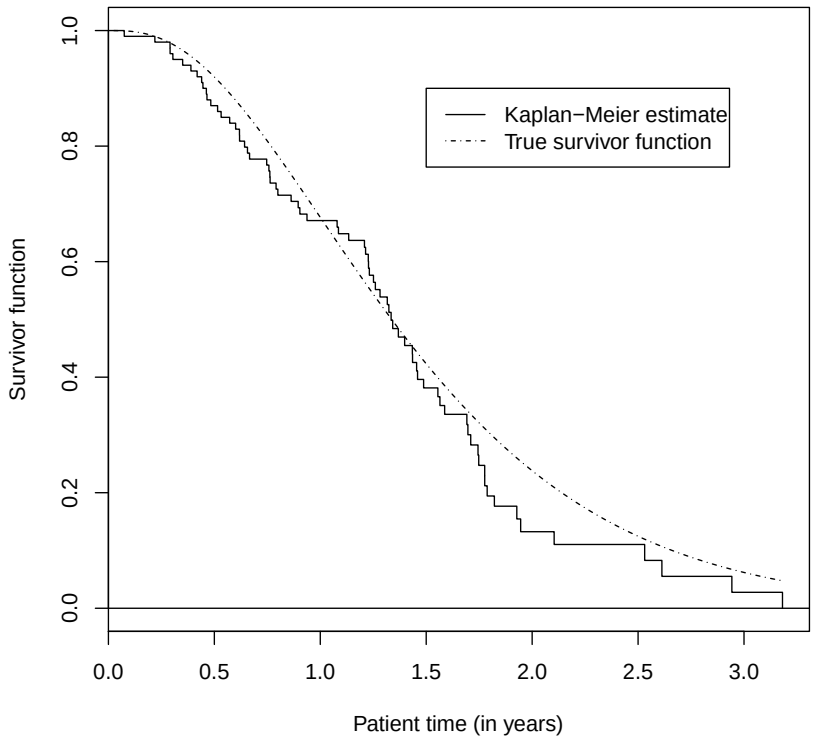

Figure 13.9: Simulation study. Kaplan-Meier estimate $\widehat{S}_T(t)$ (solid, step function) and the true survivor function $S_T(t)$, where $T \sim \text{gamma}(3, 0.5)$.

Results: The R code I used to perform this simulation is on the course web site. Below are the values of T, C, X, and Δ for the first five (simulated) patients:

```
> sim # first 5 patients only
  failure.time cens.time obs.time delta
1         2.53      1.12     1.12     0
2         1.56      2.94     1.56     1
3         1.84      1.37     1.37     0
4         1.33      2.70     1.33     1
5         1.82      1.90     1.82     1
```

The Kaplan-Meier estimate $\widehat{S}_T(t)$ based on all $n = 100$ observations and the true survivor function $S_T(t)$ are shown in Figure 13.9 (above). For the most part, we see the estimate $\widehat{S}_T(t)$ is in general agreement with the true $S_T(t)$ despite being a nonparametric estimate calculated from the observed data $\{(X_i, \Delta_i),\ i = 1, 2, ..., 100\}$. $\square$

Terminology: The main goal in survival analysis is to make inference about the probability distribution of the latent random variable T. For example, in the one-sample problem, we are usually interested in estimating the survivor function $S_T(t) = P(T > t)$ with the available

data

$$\{(X_i, \Delta_i);\ i = 1, 2, ..., n\}.$$

If we define the number of individuals **at risk** at time t in the sample by

$$n(t) = \sum_{i=1}^{n} I(X_i \geq t);$$

i.e., $n(t)$ is the number of individuals in the sample who have neither died nor have been censored by time t, then the Kaplan-Meier estimator of the survivor function $S_T(t)$ can be written as

$$\widehat{S}_T(t) = \prod_{\{i:X_i \leq t\}} \left\{1 - \frac{1}{n(X_i)}\right\}^{\Delta_i}.$$

This is the definition of the Kaplan-Meier estimator when there are no tied survival times in the sample.

- Note that $\widehat{S}_T(t)$ is simply the product of "one minus the estimated mortalities" across all observed death times up to an including t.
- Downward jumps will occur at those times where a death (event) occurs; i.e., when $\Delta_i = 1$. At all other times, the $\widehat{S}_T(t)$ remains constant.

Dealing with "ties:" Let $d(t)$ denote the number of observed deaths in the sample at time t, that is,

$$d(t) = \sum_{i=1}^{n} I(X_i = t, \Delta_i = 1).$$

Note that $d(t)$ is equal to 0 or 1 with continuous survival data when there are no ties. More generally, $d(t)$ may be greater than 1 when ties are allowed. In this situation, we can write the Kaplan-Meier estimator as

$$\widehat{S}_T(t) = \prod_{A(u)} \left\{1 - \frac{d(u)}{n(u)}\right\},$$

where $A(u) = \{\text{all death times } u \leq t\}$, $n(u)$ is the number of individuals at risk (i.e., "still alive" and not censored) at time u, and

$$d(u) = \sum_{i=1}^{n} I(X_i = u, \Delta_i = 1)$$

is the total number of deaths at time u. This is the most general version of the Kaplan-Meier estimator under (random) right censoring.

Remarks: For the Kaplan-Meier estimator to give unbiased results, there is an implicit assumption that individuals who are censored are at the same risk of failure as those who are still alive and are uncensored. This is called the **non-informative censoring** assumption.

- In our latent variable conceptualization, this means that $T_i \perp\!\!\!\perp C_i$, for $i = 1, 2, ..., n$. Those at risk, at any time t, should be representative of the entire population alive at the same time so estimated mortality rates in the Kaplan-Meier estimator reflect the true population mortality rates.

- If censoring occurs only because of staggered entry, then the assumption of non-informative censoring is probably plausible. However, when censoring results from loss to follow-up or death from a competing risk, then this assumption may be suspect because the censoring process depends on the survival time.

Inference: Under the non-informative censoring assumption, for fixed t, theoretical arguments show the Kaplan-Meier estimator $\widehat{S}_T(t)$ is approximately normal with mean $S_T(t)$ and variance which is consistently estimated by

$$\widehat{\sigma}^2_{\widehat{S}_T(t)} = \{\widehat{S}_T(t)\}^2 \sum_{A(u)} \frac{d(u)}{n(u)[n(u) - d(u)]},$$

which is the limit of Greenwood's formula given earlier for life table estimates. An approximate $100(1-\alpha)\%$ confidence interval for $S_T(t)$ is therefore given by

$$\widehat{S}_T(t) \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{S}_T(t)},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution and $\widehat{\sigma}_{\widehat{S}_T(t)}$ is the estimated standard error. R calculates these confidence intervals upon request.

Example 13.8. The data below in Table 13.3 are from a survival study with $n = 80$ males subjects with advanced tongue cancer. There were two types of cancerous tumors in this study, but we will not distinguish between them in this analysis. The endpoint was

$$T = \text{time to death (measured in weeks)}.$$

Among the 80 subjects, there were 52 death times and 28 censored times. Note that some of the death times are "ties."

1	3	3	4	10	13	13	16	16	24
26	27	28	30	30	32	41	51	65	67
70	72	73	77	91	93	96	100	104	157
167	61*	74*	79*	80*	81*	87*	87*	88*	89*
93*	97*	101*	104*	108*	109*	120*	131*	150*	231*
240*	400*	1	3	4	5	5	8	12	13
18	23	26	27	30	42	56	62	69	104
104	112	129	181	8*	67*	76*	104*	176*	231*

Table 13.3: Tongue cancer data. Times to death or censoring for $n = 80$ patients. Censored observations are starred.

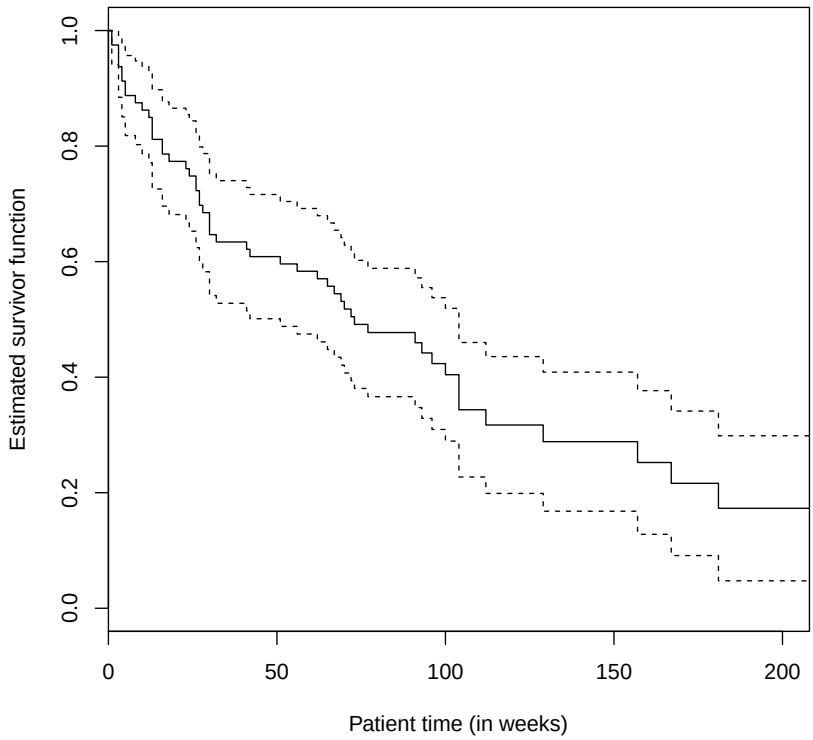

Figure 13.10: Tongue cancer data. Kaplan-Meier estimate of the survivor function $S_T(t)$. Pointwise confidence bands are included.

R output: The `survfit` function in R records the at risk process `n.risk` and the number of deaths (events) `n.event` at all times in the data set.

```
> fit = survfit(Surv(tongue,delta)~1,conf.type="plain")
> summary(fit)
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     80       2    0.975  0.0175       0.9408        1.000
    3     78       3    0.938  0.0271       0.8845        0.991
    4     75       2    0.913  0.0316       0.8506        0.974
    5     73       2    0.888  0.0353       0.8183        0.957
    8     71       1    0.875  0.0370       0.8025        0.947
   10     69       1    0.862  0.0386       0.7868        0.938
   12     68       1    0.850  0.0400       0.7712        0.928
   13     67       3    0.812  0.0438       0.7257        0.898
   16     64       2    0.786  0.0460       0.6961        0.876
   18     62       1    0.774  0.0470       0.6815        0.866
   23     61       1    0.761  0.0479       0.6670        0.855
   24     60       1    0.748  0.0487       0.6527        0.844
   26     59       2    0.723  0.0503       0.6243        0.821
```

27	57	2	0.697	0.0516	0.5963	0.799
28	55	1	0.685	0.0522	0.5825	0.787
30	54	3	0.647	0.0537	0.5414	0.752
32	51	1	0.634	0.0541	0.5279	0.740
41	50	1	0.621	0.0545	0.5145	0.728
42	49	1	0.609	0.0549	0.5011	0.716
51	48	1	0.596	0.0552	0.4879	0.704
56	47	1	0.583	0.0554	0.4747	0.692
62	45	1	0.570	0.0557	0.4612	0.680
65	44	1	0.557	0.0559	0.4478	0.667
67	43	1	0.544	0.0561	0.4345	0.654
69	41	1	0.531	0.0563	0.4208	0.641
70	40	1	0.518	0.0564	0.4073	0.628
72	39	1	0.505	0.0565	0.3938	0.615
73	38	1	0.491	0.0566	0.3805	0.602
77	35	1	0.477	0.0567	0.3662	0.588
91	27	1	0.460	0.0573	0.3474	0.572
93	26	1	0.442	0.0577	0.3288	0.555
96	24	1	0.424	0.0582	0.3095	0.538
100	22	1	0.404	0.0586	0.2893	0.519
104	20	3	0.344	0.0594	0.2273	0.460
112	13	1	0.317	0.0604	0.1988	0.436
129	11	1	0.288	0.0614	0.1680	0.409
157	8	1	0.252	0.0634	0.1280	0.377
167	7	1	0.216	0.0638	0.0912	0.341
181	5	1	0.173	0.0640	0.0475	0.299

Estimated survival probabilities are calculated in `survival` along with estimated standard errors (via Greenwood's formula) in `std.err`. Large-sample 95% confidence intervals for $S_T(t)$ are shown for each death (event) time t.

Q: Calculate a point and interval estimate for the one-year survival probability $S_T(52)$.
A: Recall the Kaplan-Meier estimate $\widehat{S}_T(t)$ is constant except at those times t where at least one death (event) occurs. Therefore, a point estimate of the one-year survival probability $S_T(52)$ is

$$\widehat{S}_T(51) = 0.596.$$

The (estimated) standard error of this point estimate is

$$\widehat{\sigma}_{\widehat{S}_T(51)} = \sqrt{\{\widehat{S}_T(51)\}^2 \sum_{A(u)} \frac{d(u)}{n(u)[n(u)-d(u)]}},$$

where $A(u) = \{\text{all death times } u \leq 51\}$. An approximate 95% confidence interval for $S_T(52)$ is $(0.488, 0.704)$. That is, we are 95% confident the population-level one-year survival probability $S_T(52)$ is between 0.488 and 0.704. $\square$

13.5 Two-sample tests

Remark: In survival data applications, especially in clinical trials, the goal is often to compare two or more groups of individuals. If the primary endpoint is time to an event (e.g., time to death, time to relapse, etc.), then an important issue is determining if one treatment increases or decreases the distribution of this time. Let Z denote the treatment group assignment. If there are two treatments of interest, then $Z \in \{1, 2\}$.

Inference: The problem of comparing two treatments statistically can be framed as a hypothesis test. If we denote by $S_1(t)$ and $S_2(t)$ the survivor functions for treatments 1 and 2, respectively, the null hypothesis of **no treatment difference** is

$$H_0 : S_1(t) = S_2(t),$$

for all $t > 0$, or, equivalently, in terms of the hazard functions,

$$H_0 : \lambda_1(t) = \lambda_2(t),$$

for all $t > 0$, where $\lambda_j(t) = -\frac{d}{dt}\ln\{S_j(t)\}$, for $j = 1, 2$. One possible alternative hypothesis specifies that the survival time for one treatment is stochastically larger (or smaller) than the other treatment. For example, we might test H_0 against

$$H_a : S_1(t) \leq S_2(t),$$

for all $t > 0$, with strict inequality for some t, or $H_a : S_1(t) \geq S_2(t)$. A two-sided alternative specifies

$$H_a : S_1(t) \neq S_2(t),$$

for some $t > 0$.

Remark: To address the two-sample survival problem, we will make use of a **nonparametric** test; that is, we will use a test statistic whose distribution (under H_0) does not depend on the shape of the underlying survival functions (at least, not asymptotically). The most widely used test in censored survival analysis is the **logrank test** which we now describe.

Notation: Data from a two-sample censored survival analysis problem can be expressed as a sample of triplets; namely,

$$\{(X_i, \Delta_i, Z_i);\ i = 1, 2, ..., n\},$$

where $X_i = \min\{T_i, C_i\}$. Recall that for the ith individual,

$$\begin{aligned} T_i &= \text{latent } \textbf{failure} \text{ time} \\ C_i &= \text{latent } \textbf{censoring} \text{ time.} \end{aligned}$$

The failure indicator for the ith individual is given by

$$\Delta_i = \begin{cases} 1, & \text{if } T_i \leq C_i \\ 0, & \text{if } T_i > C_i \end{cases}$$

and the treatment indicator is

$$Z_i = \begin{cases} 1, & i\text{th individual in treatment group 1} \\ 2, & i\text{th individual in treatment group 2.} \end{cases}$$

Notation: Let n_1 be the number of individuals assigned to treatment 1; i.e.,

$$n_1 = \sum_{i=1}^{n} I(Z_i = 1),$$

and n_2 be the number of individuals assigned to treatment 2; i.e.,

$$n_2 = \sum_{i=1}^{n} I(Z_i = 2),$$

so that $n = n_1 + n_2$. The **number at risk** at time u from treatment 1 is denoted by $n_1(u)$; i.e.,

$$n_1(u) = \sum_{i=1}^{n} I(X_i \geq u, Z_i = 1).$$

That is, $n_1(u)$ is the number of individuals in treatment group 1 who have neither died nor have been censored at time u. Similarly,

$$n_2(u) = \sum_{i=1}^{n} I(X_i \geq u, Z_i = 2)$$

is the number at risk at time u from treatment group 2.

Notation: The **number of deaths** at time u in treatment group 1 is denoted by $d_1(u)$; i.e.,

$$d_1(u) = \sum_{i=1}^{n} I(X_i = u, \Delta_i = 1, Z_i = 1).$$

Similarly,

$$d_2(u) = \sum_{i=1}^{n} I(X_i = u, \Delta_i = 1, Z_i = 2)$$

is the number of deaths at time u in treatment group 2. The number of deaths at time u for both treatment groups is

$$d(u) = d_1(u) + d_2(u).$$

This notation allows for the possibility of having more than one death occurring at the same time (that is, "tied" survival times).

Remark: A formal derivation of the logrank test statistic, as well as asymptotic considerations, relies on **martingale theory**. We will avoid this more advanced material and take the following informal approach.

- At any time u where a death is observed; i.e., when $d(u) \geq 1$, the data available to us can be summarized in the following 2×2 table:

	Treatment 1	Treatment 2	Total
Number of deaths	$d_1(u)$	$d_2(u)$	$d(u)$
Number alive	$n_1(u) - d_1(u)$	$n_2(u) - d_2(u)$	$n(u) - d(u)$
Total	$n_1(u)$	$n_2(u)$	$n(u)$

  If $H_0 : S_1(t) = S_2(t)$ is true, then we would expect

$$d_1(u) - \frac{n_1(u)}{n(u)} d(u)$$

  to be "close" to zero (actually, its expectation is zero under H_0).

- Therefore, consider constructing this same 2×2 table at each point in time u where an event (death) occurs. That is, consider constructing a sequence of 2×2 tables, where each table in the sequence corresponds to a unique time u where $d(u) \geq 1$. Using similar logic, the sum

$$\sum_{A(u)} \left[d_1(u) - \frac{n_1(u)}{n(u)} d(u) \right]$$

  where $A(u) = \{u : d(u) \geq 1\}$ denotes the set of all distinct death times u, should be close to zero when H_0 is true (again, its expectation is equal to zero under H_0).

- We now examine what would happen if $H_0 : S_1(t) = S_2(t)$ is not true:
  - If the hazard rate for treatment 1 was **greater** than the hazard rate for treatment 2 over all u, then we would expect

$$d_1(u) - \frac{n_1(u)}{n(u)} d(u) > 0.$$

  - If the hazard rate for treatment 1 was **less** than the hazard rate for treatment 2 over all u, then we would expect

$$d_1(u) - \frac{n_1(u)}{n(u)} d(u) < 0.$$

- The last observation suggests that $H_0 : S_1(t) = S_2(t)$ should be rejected if the statistic

$$T^* = \sum_{A(u)} \left[d_1(u) - \frac{n_1(u)}{n(u)} d(u) \right],$$

  is too large or too small, depending on the alternative we are interested in.

- In order to gauge the strength of evidence against H_0, we must be able to evaluate the distribution of T^* (at least, approximately) when H_0 is true. To do this, T^* needs to be standardized appropriately. Specifically, this standardized version is the **logrank test statistic**, given by

$$T_{LR} = \frac{T^*}{\text{se}(T^*)} = \frac{\displaystyle\sum_{A(u)} \left[d_1(u) - \frac{n_1(u)}{n(u)} d(u) \right]}{\sqrt{\displaystyle\sum_{A(u)} \frac{n_1(u) n_2(u) d(u) \{n(u) - d(u)\}}{n^2(u)\{n(u) - 1\}}}}.$$

We now examine the sampling distribution of T_{LR} when H_0 is true.

Sampling distribution: We now informally argue that when $H_0 : S_1(t) = S_2(t)$ is true, the logrank test statistic $T_{LR} \sim \mathcal{AN}(0, 1)$, for large n. To see why this is true, consider again the 2×2 table:

	Treatment 1	Treatment 2	Total
Number of deaths	$d_1(u)$	$\cdot$	$d(u)$
Number alive	$\cdot$	$\cdot$	$n(u) - d(u)$
Total	$n_1(u)$	$n_2(u)$	$n(u)$

Conditional on the marginal counts, the random variable $d_1(u)$ follows a hypergeometric distribution with probability mass function

$$P\{d_1(u) = d\} = \frac{\dbinom{n_1(u)}{d} \dbinom{n_2(u)}{d(u) - d}}{\dbinom{n(u)}{d(u)}}.$$

Thus, the conditional mean and variance of $d_1(u)$ are

$$\frac{n_1(u)}{n(u)} d(u)$$

and

$$\frac{n_1(u) n_2(u) d(u) \{n(u) - d(u)\}}{n^2(u)\{n(u) - 1\}},$$

respectively. It can be shown that

$$T^* = \sum_{A(u)} \left[d_1(u) - \frac{n_1(u)}{n(u)} d(u) \right]$$

is the sum of uncorrelated pieces $d_1(u) - \frac{n_1(u)}{n(u)}d(u)$, each with mean zero under H_0 (not intuitive) and that the sum

$$\sum_{A(u)} \frac{n_1(u)n_2(u)d(u)\{n(u) - d(u)\}}{n^2(u)\{n(u) - 1\}}$$

is the variance of T^* when H_0 is true (also not intuitive). With both of these results in place, it follows that, under $H_0 : S_1(t) = S_2(t)$, the logrank test statistic $T_{LR} \sim \mathcal{AN}(0, 1)$ by a version of the Central Limit Theorem for martingale type data.

Implementation: To test

$$\begin{gathered} H_0 : S_1(t) = S_2(t) \\ \text{versus} \\ H_a : S_1(t) \neq S_2(t), \end{gathered}$$

an approximate level α rejection region is

$$\text{RR} = \{T_{LR} : |T_{LR}| > z_{\alpha/2}\},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of a $\mathcal{N}(0, 1)$ distribution. One-sided tests use a suitably adjusted rejection region.

- If we were interested in showing that treatment 1 is better (i.e., longer survival times) than treatment 2, we would use

  $$\text{RR} = \{T_{LR} : T_{LR} < -z_\alpha\}.$$

  The form of this rejection region makes sense because under $H_a : S_1(t) \geq S_2(t)$, we would expect the observed number of deaths from treatment 1, $d_1(u)$, to be **less than** that expected under H_0. This will encourage T_{LR} to be negative.

- If we wanted to show treatment 2 is better (i.e., longer survival times), we would we would use

  $$\text{RR} = \{T_{LR} : T_{LR} > z_\alpha\}.$$

  If $H_a : S_1(t) \leq S_2(t)$ is true, we would expect $d_1(u)$ to be **greater than** that expected under H_0. This will encourage T_{LR} to be positive.

Note: To "derive" the form of the logrank test, we have summarized the data using only 2×2 tables at the distinct death times. In constructing the logrank test statistic, we never made any assumptions regarding the shape of the underlying survival distributions. This explains why this test is **nonparametric** in nature.

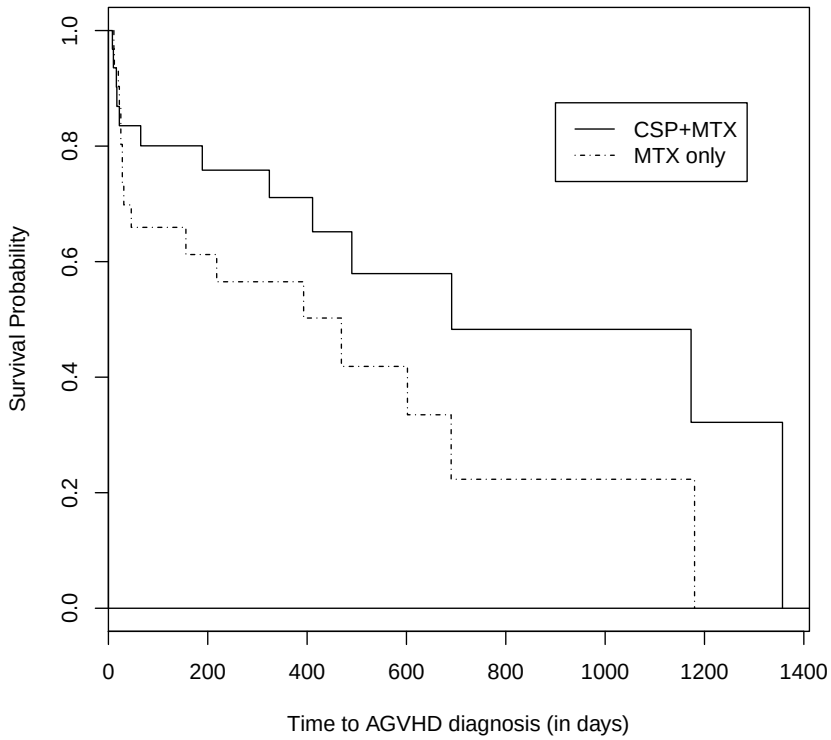

Figure 13.11: RCT data. Kaplan-Meier estimates for the time to diagnosis of AGVHD for two treatment groups.

Example 13.1 (continued). We now revisit the randomized clinical trial data for 64 patients assigned to

- Group 1: Cyclosporine and methotrexate (CSP+MTX)
- Group 2: Methotrexate only (MTX).

Recall the primary endpoint was the time from treatment assignment until diagnosis of AGVHD. Figure 13.11 (above) shows the Kaplan-Meier estimates of the survivor functions. We now use the logrank test to infer what these estimates say about the true survivor functions $S_1(t)$ and $S_2(t)$.

Analysis: In particular, suppose that we wanted to test whether the adminstration of cyclosporine and methotrexate prolonged the time to AGVHD diagnosis (when compared to methotrexate only); i.e., we want to test

$$\begin{gathered} H_0 : S_1(t) = S_2(t) \\ \text{versus} \\ H_a : S_1(t) \geq S_2(t), \end{gathered}$$

for all t (with strict inequality in H_a for some t). To do group summaries, we can use the `survfit` function in R:

```
fit.1 = survfit(Surv(agvhd.times,cens.agvhd.times)~treatment,conf.int=0.95)
> fit.1
             n events median 0.95LCL 0.95UCL
treatment=1 32     13    691     411      NA
treatment=2 32     17    469     156      NA
```

This table gives point estimates of the median time to AGVHD diagnosis along with large-sample 95% confidence intervals for the population median (using Greenwood's formula). Note that for these data, the sample sizes and numbers of events are too small to calculate the upper endpoints of both intervals.

To perform the logrank test, we use the `survdiff` function:

```
fit.2 = survdiff(Surv(agvhd.times,cens.agvhd.times)~treatment)
> fit.2
             N Observed Expected (O-E)^2/E (O-E)^2/V
treatment=1 32       13     17.4      1.09      2.74
treatment=2 32       17     12.6      1.50      2.74

Chisq = 2.7 on 1 degrees of freedom, p = 0.1
```

Note: The `fit.2` output gives the square of the logrank statistic, that is,

$$T_{LR}^2 = \left(\frac{\displaystyle\sum_{A(u)} \left[d_1(u) - \frac{n_1(u)}{n(u)} d(u) \right]}{\sqrt{\displaystyle\sum_{A(u)} \frac{n_1(u) n_2(u) d(u) \{n(u) - d(u)\}}{n^2(u)\{n(u)-1\}}}} \right)^2.$$

Because our alternative hypothesis is $H_a : S_1(t) \geq S_2(t)$, we know that

$$T_{LR}^2 = 2.74 \implies T_{LR} = -\sqrt{2.74} = -1.66.$$

Therefore, a (large-sample) level $\alpha = 0.05$ test would reject $H_0 : S_1(t) = S_2(t)$ in favor of $H_a : S_1(t) \geq S_2(t)$ because

$$\text{RR} = \{T_{LR} < -z_{0.05}\} = \{T_{LR} < -1.65\}.$$

We have just enough evidence to conclude the time to AGVHD diagnosis for treatment group 1 (cyclosporine and methotrexate) is stochastically larger than the time to AGVHD diagnosis for treatment group 2 (methotrexate only). $\square$