STAT 513

THEORY OF STATISTICAL INFERENCE

Fall, 2011

Lecture Notes

Joshua M. Tebbs

Department of Statistics

University of South Carolina

Contents

10 Hypothesis Testing

Complementary reading: Chapter 10 (WMS).

10.1 Introduction and review

PREVIEW: Classical statistical inference deals with making statements about population (model) parameters. The two main areas of statistical inference are **estimation** (point estimation and confidence intervals) and **hypothesis testing**. Point and interval estimation were discussed CH8-9 (WMS). This chapter deals with hypothesis testing.

Example 10.1. Actuarial data reveal that the claim amount for a "standard class" of policy holders, denoted by Y (measured in \$1000s), follows an exponential distribution with mean $\theta > 0$. Suppose that we adopt this model for Y and that we observe an iid sample of claims, denoted by $Y_1, Y_2, ..., Y_n$. Recall the following facts from STAT 512:

1. A sufficient statistic for θ is
$$T = \sum_{i=1}^{n} Y_i.$$
2. The maximum likelihood estimator (MLE) for θ is
$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$
3. The minimum variance unbiased estimator (MVUE) for θ is $\overline{Y}$.
4. The quantity
$$Q = \frac{2T}{\theta} \sim \chi^2(2n),$$
and therefore is a pivot. This is an exact (finite sample) result; i.e., $Q \sim \chi^2(2n)$ exactly for all n.
5. The quantity
$$Z = \frac{\overline{Y} - \theta}{S/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1),$$

as $n \to \infty$, and therefore Z is a large sample pivot. This means that $Z \sim \mathcal{AN}(0,1)$, when n is large. The larger the n, the better the approximation.

INTERVAL ESTIMATION: We have at our disposal two pivots, namely,

$$Q = \frac{2T}{\theta} \sim \chi^2(2n)$$

and

$$Z = \frac{\overline{Y} - \theta}{S/\sqrt{n}} \sim \mathcal{AN}(0,1).$$

The (exact) confidence interval for θ arising from Q is

$$\left(\frac{2T}{\chi^2_{2n,\alpha/2}}, \frac{2T}{\chi^2_{2n,1-\alpha/2}} \right),$$

where $\chi^2_{2n,1-\alpha/2}$ and $\chi^2_{2n,\alpha/2}$ denote the lower and upper $\alpha/2$ quantiles of a $\chi^2(2n)$ distribution, respectively. The (approximate) confidence interval for θ arising from Z is

$$\overline{Y} \pm z_{\alpha/2} \left(\frac{S}{\sqrt{n}} \right),$$

where S is the sample standard deviation and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution.

SIMULATION EXERCISE: To compare the coverage probabilities of the exact and approximate intervals, we will use **Monte Carlo simulation**. In particular, we use R to generate $B = 10,000$ iid samples

$$Y_1, Y_2, ..., Y_n \sim \text{exponential}(\theta),$$

with $n = 10$.

- For each of the $B = 10,000$ samples, we will keep track of the values of T, $\overline{Y}$, and S. We will then compute the exact and approximate 95 percent confidence intervals for θ with each sample (that is, $1 - \alpha = 0.95$).
- Therefore, at the end of the simulation, we will have generated $B = 10,000$ exact intervals and $B = 10,000$ approximate intervals.

- We can then compute the proportion of the intervals (both exact and approximate) which contain θ. For purposes of illustration, we take $\theta = 10$. Because we are computing 95 percent confidence intervals, we would expect this proportion to be close to $1 - \alpha = 0.95$.
- We then repeat this simulation exercise for $n = 30$, $n = 100$, and $n = 1000$. Here are the results:

Interval	$n = 10$	$n = 30$	$n = 100$	$n = 1000$
Exact	0.953	0.949	0.952	0.951
Approximate	0.868	0.915	0.940	0.951

Table 10.1: Monte Carlo simulation. Coverage probabilities for exact and approximate 95 percent confidence intervals for an exponential mean θ, when $\theta = 10$.

DISCUSSION: As we can see, regardless of the sample size n, the exact interval produces a coverage probability that hovers around the nominal $1-\alpha = 0.95$ level, as expected. On the other hand, the coverage probability of the approximate interval is much lower than the nominal $1 - \alpha = 0.95$ level when n is small, although, as n increases, the coverage probability does get closer to the nominal level.

MORAL: We will discuss two types of statistical inference procedures: those that are **exact** and those that are **approximate**. Exact procedures are based on exact distributional results. Approximate procedures are typically based on large sample distributional results (e.g., Central Limit Theorem, Delta Method, Slutsky's Theorem, etc.).

- In some problems, exact inference may not be available or the exact distributional results needed may be so intractable that they are not helpful. In these instances, approximate procedures can be valuable.
- Approximate procedures are based on the (rather nonsensical) notion that the sample size $n \to \infty$. However, these procedures often do confer acceptable results for reasonably sized samples.

PREVIEW: Suppose your colleague claims that the mean claim amount θ for a new class of customers is larger than the mean amount for the standard class of customers, known to be θ_0. How can we determine (statistically) if there is evidence to support this claim? Here, it makes sense to think of two competing hypotheses:

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta > \theta_0.$$

- H_0 says that the mean claim amount for the new class of customers, θ, is the same as the mean claim amount for the standard class, θ_0.
- H_a says that the mean claim amount for the new class of customers, θ, is larger than the mean claim amount for the standard class, θ_0, that is, your colleague's claim is correct.
- Based on a sample of claim amounts $Y_1, Y_2, ..., Y_n$ from the new class of customers, how should we formally decide between H_0 and H_a? This question can be answered by performing a hypothesis test.

10.2 The elements of a hypothesis test

TERMINOLOGY: A **hypothesis test** is an inferential technique which pits two competing hypotheses versus each other. The goal is to decide which hypothesis is more supported by the observed data. The four parts of a hypothesis test are

1. the null hypothesis, H_0
2. the alternative hypothesis, H_a
3. the test statistic
4. the rejection region.

TERMINOLOGY: The **null hypothesis** H_0 states the value of the parameter to be tested. For example, if our colleague in Example 10.1 wants to compare the mean claim amount of the new class θ to the mean claim amount for the standard class (known to be $\theta_0 = 10$, say), then the null hypothesis would be

$$H_0 : \theta = 10.$$

In this course, we will usually take the null hypothesis to be **sharp**; that is, there is only one value of the parameter θ possible under H_0.

TERMINOLOGY: The **alternative hypothesis** H_a describes what values of θ we are interested in testing H_0 against. For example, if our colleague in Example 10.1 believed that the mean claim amount for the new class of customers was

- greater than $\theta_0 = 10$, s/he would use

$$H_a : \theta > 10.$$

- less than $\theta_0 = 10$, s/he would use

$$H_a : \theta < 10.$$

- different than $\theta_0 = 10$, s/he would use

$$H_a : \theta \neq 10.$$

The alternative hypothesis H_a is sometimes called the **researcher's hypothesis**, since it is often the hypothesis the researcher wants to conclude is supported by the data.

NOTE: The first two examples of H_a above are called **one-sided** alternatives. The last example is called a **two-sided** alternative. One-sided alternatives state pointedly which direction we are testing H_0 against. A two-sided alternative does not specify this.

TERMINOLOGY: A **test statistic** is a statistic that is used to test H_0 versus H_a. We make our decision by comparing the observed value of the test statistic to its sampling distribution under H_0.

- If the observed value of the test statistic is consistent with its sampling distribution under H_0, then this is not evidence for H_a.
- If the observed value of the test statistic is not consistent with its sampling distribution under H_0, and it is more consistent with the sampling distribution under H_a, then this is evidence for H_a.

TERMINOLOGY: The **rejection region**, denoted by RR, specifies the values of the test statistic for which H_0 is rejected. The rejection region is usually located in tails of the test statistic's sampling distribution computed under H_0. This is why we take H_0 to be sharp, namely, so that we can construct a single sampling distribution.

PREVAILING RULE: In any hypothesis test, if the test statistic falls in rejection region, then we reject H_0.

STATES OF NATURE: Table 10.2 summarizes the four possible outcomes from performing a hypothesis test.

	Decision: Reject H_0	Decision: Do not reject H_0
Truth: H_0	Type I Error	correct decision
Truth: H_a	correct decision	Type II Error

Table 10.2: States of nature in testing $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ (or any other H_a).

TERMINOLOGY: **Type I Error**: *Rejecting* H_0 *when* H_0 *is true.* The probability of Type I Error is denoted by α. Notationally,

$$\begin{aligned}\alpha = P(\text{Type I Error}) &= P(\text{Reject } H_0 | H_0 \text{ is true})\\ &= P(\text{Reject } H_0 | \theta = \theta_0).\end{aligned}$$

The Type I Error probability α is also called the **significance level** for the test. We would like α to be small. It is common to choose this value up front.

TERMINOLOGY: **Type II Error**: *Not rejecting H_0 when H_a is true.* The probability of Type II Error is denoted by β. Notationally,

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

REMARK: Obviously, $H_a : \theta \neq \theta_0$ (or any other H_a) can be true in many ways, so we can compute β for different values of θ under H_a. Specifically, the probability of Type II error, when $\theta = \theta_a \in H_a$, is

$$\beta = \beta(\theta_a) = P(\text{Do not reject } H_0 | \theta = \theta_a).$$

That is, this probability will be different for different values of $\theta_a \in H_a$. Ideally, we would like β to be small for all $\theta_a \in H_a$.

Example 10.2. Suppose that industrial accidents occur according to a Poisson process with mean $\theta = 20$ per site per year. New safety measures have been put in place to decrease the number of accidents at industrial sites all over the US. Suppose that after implementation of the new measures, we will observe the number of accidents for a sample of $n = 10$ sites. Denote these data by $Y_1, Y_2, ..., Y_{10}$. We are interested in testing

$$H_0 : \theta = 20$$

versus

$$H_a : \theta < 20.$$

To perform the test, suppose we use the test statistic

$$T = \sum_{i=1}^{10} Y_i$$

and the rejection region RR $= \{t : t \leq 175\}$.

QUESTIONS:

(a) What is the distribution of T when H_0 is true?

(b) What is $\alpha = P(\text{Type I Error})$ for this RR?

(c) Suppose that $\theta = 18$, that is, H_a is true. What is the probability of Type II Error when using this RR?

Solutions:

(a) Recall that the sum of n iid Poisson(θ) random variables is distributed as Poisson($n\theta$). Therefore, when $H_0 : \theta = 20$ is true,

$$T \sim \text{Poisson}(200).$$

Note that this is an exact distributional result; i.e., all of the calculations that follow are exact and not approximate.

(b) The probability of Type I Error is

$$\begin{aligned}\alpha &= P(\text{Reject } H_0 | H_0 \text{ is true}) \\ &= P(T \leq 175 | \theta = 20) \\ &= \sum_{t=0}^{175} \frac{200^t e^{-200}}{t!} \approx 0.0394.\end{aligned}$$

I found this probability using the `ppois(175,200)` command in R.

(c) First, note that when $\theta = 18$,

$$T \sim \text{Poisson}(180).$$

Therefore, the probability of Type II Error when $\theta = 18$ is

$$\begin{aligned}\beta = \beta(18) &= P(\text{Do not reject } H_0 | \theta = 18) \\ &= P(T > 175 | \theta = 18) \\ &= \sum_{t=176}^{\infty} \frac{180^t e^{-180}}{t!} \approx 0.6272.\end{aligned}$$

I found this probability using the `1-ppois(175,180)` command in R.

DISCUSSION:

- For the rejection region RR $= \{t : t \leq 175\}$, the probability of Type I Error is small ($\alpha \approx 0.0394$). This assures us that if H_0 is really true (and that the new safety measures, in fact, did not work), then we are not likely to reject H_0.
- However, if implementing the safety measures did work and the mean number of accidents per site/per year actually decreased to $\theta = 18$ (i.e., H_a is true), then we are still likely not to reject H_0 since $\beta = \beta(18) \approx 0.6272$.

- Note that

$$P(\text{Do not reject } H_0|\theta = 18) \approx 0.6272 \Longleftrightarrow P(\text{Reject } H_0|\theta = 18) \approx 0.3728.$$

In other words, we would have only about a 37 percent chance of correctly concluding that the safety measures actually worked. □

Example 10.3. Suppose that $Y_1, Y_2, ..., Y_{25}$ is an iid sample of $n = 25$ observations from a $\mathcal{N}(\theta, \sigma_0^2)$ distribution, where $\sigma_0^2 = 100$ is known. We would like to test

$$H_0 : \theta = 75$$

$$\text{versus}$$

$$H_a : \theta > 75.$$

To perform the test, suppose we use the test statistic

$$\overline{Y} = \frac{1}{25}\sum_{i=1}^{25} Y_i$$

and the rejection region RR $= \{\overline{y} : \overline{y} > k\}$, where k is a constant.

Questions:

(a) What is the distribution of $\overline{Y}$ when H_0 is true?

(b) Find the value of k that provides a level $\alpha = 0.10$ test.

(c) Suppose that $\theta = 80$, that is, H_a is true. What is the probability of Type II Error when using this RR?

Solutions:

(a) In general, recall that $\overline{Y}$ is normally distributed with mean θ and variance σ^2/n, that is,

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

Therefore, when $H_0 : \theta = 75$ is true,

$$\overline{Y} \sim \mathcal{N}(75, 4).$$

Note that this is an exact distributional result; i.e., all of the calculations that follow are exact and not approximate.

(b) To find the value of k, we set

$$\begin{aligned}\alpha = 0.10 &= P(\text{Reject } H_0 | H_0 \text{ is true}) \\ &= P(\overline{Y} > k | \theta = 75) \\ &= P\left(Z > \frac{k - 75}{2}\right),\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. Therefore, because $z_{0.10} = 1.28$, this means

$$\frac{k-75}{2} = 1.28 \quad \Longrightarrow \quad k = 77.56.$$

The rejection region $\text{RR} = \{\overline{y} : \overline{y} > 77.56\}$ confers a Type I Error probability (significance level) of $\alpha = 0.10$.

(c) First, note that when $\theta = 80$,

$$\overline{Y} \sim \mathcal{N}(80, 4).$$

Therefore, the probability of Type II Error, when $\theta = 80$, is

$$\begin{aligned}\beta = \beta(80) &= P(\text{Do not reject } H_0 | \theta = 80) \\ &= P(\overline{Y} < 77.56 | \theta = 80) \\ &= P\left(Z < \frac{77.56 - 80}{2}\right) = P(Z < -1.22) \approx 0.1112.\end{aligned}$$

I found this probability using the `pnorm(-1.22,0,1)` command in R. □

Example 10.4. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $n = 100$. We would like to test

$$H_0 : p = 0.10$$

versus

$$H_a : p < 0.10.$$

To perform the test, suppose we use the test statistic

$$T = \sum_{i=1}^{100} Y_i$$

and the rejection region $\text{RR} = \{t : t \leq k\}$, where k is a constant.

QUESTIONS:

(a) What is the distribution of T when H_0 is true?

(b) Is it possible to find an exact level $\alpha = 0.05$ rejection region?

(c) With $k = 5$, find the probability of Type II Error when $p = 0.05$.

SOLUTIONS:

(a) In general, we recall that if $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, then the (sufficient) statistic

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

Therefore, when H_0 is true and $n = 100$, we have $T \sim b(100, 0.10)$.

(b) The value of k is chosen so that

$$\begin{aligned} \alpha &= P(\text{Reject } H_0 | H_0 \text{ is true}) \\ &= P(T \leq k | p = 0.10) \\ &= \sum_{t=0}^{k} \underbrace{\binom{100}{t}(0.10)^t(1-0.10)^{100-t}}_{b(100,0.10)\text{ pmf}}. \end{aligned}$$

The R command `pbinom(k,100,0.10)` gives the following:

$$\begin{aligned} k = 3 &\implies \alpha = 0.0078 \\ k = 4 &\implies \alpha = 0.0237 \\ k = 5 &\implies \alpha = 0.0576 \\ k = 6 &\implies \alpha = 0.1172. \end{aligned}$$

We can not get an exact level $\alpha = 0.05$ rejection region of the form $\text{RR} = \{t : t \leq k\}$.

(c) If $k = 5$ and $p = 0.05$, our level $\alpha = 0.0576$ rejection region is $\text{RR} = \{t : t \leq 5\}$ and $T \sim b(100, 0.05)$. Therefore,

$$\begin{aligned} \beta = \beta(0.05) &= P(\text{Do not reject } H_0 | p = 0.05) \\ &= P(T > 5 | p = 0.05) \\ &= \sum_{t=6}^{100} \underbrace{\binom{100}{t}(0.05)^t(1-0.05)^{100-t}}_{b(100,0.05)\text{ pmf}} \approx 0.3840. \end{aligned}$$

I found this probability using the `1-pbinom(5,100,0.05)` command in R. □

10.3 Common large sample tests

REMARK: The term "large sample" is used to describe hypothesis tests that are constructed using asymptotic (large sample) theory, so the following tests are approximate for "large" sample sizes. We present large sample hypothesis tests for

1. one population mean μ
2. one population proportion p
3. the difference of two population means $\mu_1 - \mu_2$
4. the difference of two population proportions $p_1 - p_2$.

TEST STATISTIC: In each of these situations, we will use a point estimator $\widehat{\theta}$ which satisfies

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. Recall that

$$\sigma_{\widehat{\theta}} = \sqrt{V(\widehat{\theta})}$$

denotes the **standard error** of $\widehat{\theta}$. In most cases, the estimated standard error $\widehat{\sigma}_{\widehat{\theta}}$ must be used in place of $\sigma_{\widehat{\theta}}$. The estimated standard error $\widehat{\sigma}_{\widehat{\theta}}$ is simply a point estimator for the true standard error $\sigma_{\widehat{\theta}}$. In fact, if

$$\frac{\sigma_{\widehat{\theta}}}{\widehat{\sigma}_{\widehat{\theta}}} \xrightarrow{p} 1,$$

as $n \to \infty$, then

$$Z^* = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}_{\widehat{\theta}}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, by Slutsky's Theorem.

TWO-SIDED TEST: Suppose that we would like to test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta \neq \theta_0.$$

This is called a **two-sided test** because H_a does not specify a direction indicating departure from H_0. Therefore, large values of

$$Z = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}},$$

in either direction, are evidence against H_0. Note that, for n large, $Z \sim \mathcal{AN}(0,1)$ when $H_0 : \theta = \theta_0$ is true. Therefore,

$$\text{RR} = \{z : |z| > z_{\alpha/2}\}$$

is an approximate level α rejection region. That is, we will reject H_0 whenever $Z > z_{\alpha/2}$ or $Z < -z_{\alpha/2}$. For example, if $\alpha = 0.05$, then $z_{\alpha/2} = z_{0.025} = 1.96$.

ONE-SIDED TESTS: Suppose that we would like to test

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

This is called a **one-sided test** because H_a indicates a specific direction indicating a departure from H_0. In this case, only large values of

$$Z = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}},$$

are evidence against H_0. Therefore,

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

is an approximate level α rejection region. That is, we will reject H_0 whenever $Z > z_\alpha$. For example, if $\alpha = 0.05$, then $z_\alpha = z_{0.05} = 1.65$. By an analogous argument, the one-sided test of

$$H_0 : \theta = \theta_0$$
$$\text{versus}$$
$$H_a : \theta < \theta_0$$

can be performed using

$$\text{RR} = \{z : z < -z_\alpha\}$$

as an approximate level α rejection region.

10.3.1 One population mean

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a distribution with mean μ and variance σ^2 and that interest lies in testing

$$H_0 : \mu = \mu_0$$
$$\text{versus}$$
$$H_a : \mu \neq \mu_0$$

(or any other H_a). In this situation, we identify

$$\begin{aligned} \theta &= \mu \\ \widehat{\theta} &= \overline{Y} \\ \sigma_{\widehat{\theta}} &= \frac{\sigma}{\sqrt{n}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \frac{S}{\sqrt{n}}, \end{aligned}$$

where S denotes the sample standard deviation. Therefore, if σ^2 is known, we use

$$Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}.$$

Otherwise, we use

$$Z = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}.$$

Both statistics have large sample $\mathcal{N}(0, 1)$ distributions when $H_0 : \mu = \mu_0$ is true.

10.3.2 One population proportion

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample and that interest lies in testing

$$H_0 : p = p_0$$
$$\text{versus}$$
$$H_a : p \neq p_0$$

(or any other H_a). In this situation, we identify

$$\begin{aligned}\theta &= p\\ \widehat{\theta} &= \widehat{p}\\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{p(1-p)}{n}}\\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.\end{aligned}$$

To perform this test, there are two candidate test statistics. The first is

$$Z_W = \frac{\widehat{p} - p_0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}},$$

which arises from the theory we have just developed. A second test statistic is

$$Z_S = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

The test statistic Z_S uses the standard error

$$\widehat{\sigma}_{\widehat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$$

which is the correct standard error when $H_0: p = p_0$ is true. For theoretical reasons, Z_W is called a **Wald statistic** and Z_S is called a **score statistic**. Both have large sample $\mathcal{N}(0,1)$ distributions when $H_0: p = p_0$ is true. The score statistic Z_S is known to have better properties in small (i.e., finite) samples; i.e., it possesses a true significance level which is often closer to the nominal level α. The Wald statistic is often liberal, possessing a true significance level larger than the nominal level.

10.3.3 Difference of two population means

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$ are iid with mean μ_1 and variance σ_1^2

Sample 2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$ are iid with mean μ_2 and variance σ_2^2,

and that interest lies in testing

$$H_0 : \mu_1 - \mu_2 = d_0$$

versus

$$H_a : \mu_1 - \mu_2 \neq d_0$$

(or any other H_a), where d_0 is a known constant. Note that taking $d_0 = 0$ allows one to test the equality of μ_1 and μ_2. In this situation, we identify

$$\begin{aligned} \theta &= \mu_1 - \mu_2 \\ \widehat{\theta} &= \overline{Y}_{1+} - \overline{Y}_{2+} \\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}. \end{aligned}$$

If σ_1^2 and σ_2^2 are both known (which would be unlikely), then we would use

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Otherwise, we use

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

Both statistics have large sample $\mathcal{N}(0,1)$ distributions when $H_0 : \mu_1 - \mu_2 = d_0$ is true.

10.3.4 Difference of two population proportions

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1:	$Y_{11}, Y_{12}, ..., Y_{1n_1}$	are iid Bernoulli(p_1)
Sample 2:	$Y_{21}, Y_{22}, ..., Y_{2n_2}$	are iid Bernoulli(p_2),

and that interest lies in testing

$$H_0 : p_1 - p_2 = d_0$$

versus

$$H_a : p_1 - p_2 \neq d_0$$

(or any other H_a), where d_0 is a known constant. Note that taking $d_0 = 0$ allows one to test the equality of p_1 and p_2. In this situation, we identify

$$\begin{aligned}
\theta &= p_1 - p_2 \\
\widehat{\theta} &= \widehat{p}_1 - \widehat{p}_2 \\
\sigma_{\widehat{\theta}} &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\
\widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.
\end{aligned}$$

The Wald statistic is

$$Z_W = \frac{(\widehat{p}_1 - \widehat{p}_2) - d_0}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}}.$$

A score statistic is available when $d_0 = 0$. It is given by

$$Z_S = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where

$$\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}$$

is the **pooled sample proportion**, as it estimates the common $p_1 = p_2 = p$ under H_0. Both statistics have large sample $\mathcal{N}(0,1)$ distributions when $H_0 : p_1 - p_2 = d_0$ is true. As in the one-sample problem, the score statistic performs better in small samples.

10.4 Sample size calculations

IMPORTANCE: We now address the problem of sample size determination, restricting attention to one-sample settings. We assume that the estimator $\widehat{\theta}$ satisfies

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{AN}(0,1),$$

for large n, where $\sigma_{\widehat{\theta}}$ is the standard error of $\widehat{\theta}$. Recall that $\widehat{\theta}$, and consequently its standard error $\sigma_{\widehat{\theta}}$, depends on n, the sample size. We focus on the one-sided test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta > \theta_0,$$

that employs the level α rejection region

$$\text{RR} = \{z : z > z_\alpha\} \Longleftrightarrow \{\boldsymbol{y} : \widehat{\theta} > k\},$$

where k is chosen so that $P_{H_0}(\widehat{\theta} > k) \equiv P(\widehat{\theta} > k|H_0 \text{ is true}) = \alpha$.

SETTING: Our goal is to determine the sample size n that confers a specified Type II Error probability β. However, β is a function θ, so we must specify a particular value of θ to consider. Because the alternative hypothesis is of the form $H_a : \theta > \theta_0$, we are interested in a value $\theta_a > \theta_0$; i.e.,

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

where $\Delta > 0$ is the **practically important difference** that we wish to detect.

IMPORTANT: To derive a general formula for the sample size in a particular problem, we exploit the following two facts:

- when H_0 is true, our level α rejection region $\text{RR} = \{z : z > z_\alpha\}$ implies that
$$\frac{k - \theta_0}{\sigma_{\widehat{\theta}}} = z_\alpha.$$
- when H_a is true and $\theta_a = \theta_0 + \Delta$, then for a specified value of β, it follows that
$$\frac{k - \theta_a}{\sigma_{\widehat{\theta}}} = -z_\beta;$$
see Figure 10.5, pp 508 (WMS). These two formulae provide the basis for calculating the necessary sample size n. When a two-sided alternative $H_a : \theta \neq \theta_0$ is specified, the only change is that we replace z_α with $z_{\alpha/2}$.

POPULATION MEAN: For the one-sample test regarding a population mean, that is, of $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$, we have

$$\frac{k - \mu_0}{\sigma/\sqrt{n}} = z_\alpha.$$

When $\mu_a = \mu_0 + \Delta$, then for a specified value of β, we have

$$\frac{k - \mu_a}{\sigma/\sqrt{n}} = -z_\beta.$$

Solving these two equations simultaneously for n gives

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

where $\Delta = \mu_a - \mu_0$. Note that the population variance σ^2 must be specified in advance. In practice, we must provide a "guess" or an estimate of its value. This guess may be available from preliminary studies or from other historical information.

Example 10.5. A marine biologist, interested in the distribution of the size of a particular type of anchovy, would like to test

$$H_0 : \mu = 20$$

$$\text{versus}$$

$$H_a : \mu > 20,$$

where μ denotes the mean anchovy length (measured in cm). She would like to perform this test using $\alpha = 0.05$. Furthermore, when $\mu = \mu_a = 22$, she would like the probability of Type II Error to be only $\beta = 0.1$. What sample size should she use? Based on previous studies, a guess of $\sigma \approx 2.5$ is provided.

SOLUTION. We have $\mu_0 = 20$ and $\mu_a = 22$ so that $\Delta = 2$. We have $z_\alpha = z_{0.05} = 1.65$ and $z_\beta = z_{0.10} = 1.28$. Thus, the desired sample size is

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\Delta^2} = \frac{(1.65 + 1.28)^2 (2.5)^2}{2^2} \approx 13.41.$$

Therefore, she should collect 14 anchovies. □

POPULATION PROPORTION: For the one-sample test regarding a population proportion, that is, $H_0 : p = p_0$ versus $H_a : p > p_0$, it follows that

$$\frac{k - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = z_\alpha.$$

When $p_a = p_0 + \Delta$, then for a specified value of β, we have

$$\frac{k - p_a}{\sqrt{\frac{p_a(1-p_a)}{n}}} = -z_\beta.$$

Eliminating the common k in these two equations and solving for n produces

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

where $\Delta = p_a - p_0$.

Example 10.6. Researchers are planning a Phase III clinical trial to determine the probability of response, p, to a new drug treatment. It is believed that the standard treatment produces a positive response in 35 percent of the population. To determine if the new treatment increases the probability of response, the researchers would like to test, at the $\alpha = 0.05$ level,

$$H_0 : p = 0.35$$

versus

$$H_a : p > 0.35.$$

In addition, they would like to detect a "clinically important" increase in the response probability to $p = p_a = 0.40$ with probability 0.80 (so that the Type II Error probability $\beta = 0.20$). The clinically important difference $\Delta = p_a - p_0 = 0.05$ is a value that represents "a practically important increase" for the manufacturers of the new drug. What is the minimum sample size that should be used in the Phase III trial?

Solution. The desired sample size is

$$\begin{aligned} n &= \left[\frac{z_{0.05}\sqrt{p_0(1-p_0)} + z_{0.20}\sqrt{p_a(1-p_a)}}{\Delta}\right]^2 \\ &= \left[\frac{1.65\sqrt{0.35(1-0.35)} + 0.84\sqrt{0.40(1-0.40)}}{0.05}\right]^2 \approx 574.57. \end{aligned}$$

Thus, the researchers will need to recruit $n = 575$ patients for the Phase III trial. □

10.5 Confidence intervals and hypothesis tests

REVELATION: There is an elegant duality between confidence intervals and hypothesis tests. In a profound sense, they are essentially the same thing, as we now illustrate. Suppose that we have a point estimator, say, $\widehat{\theta}$, which satisfies

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{N}(0,1).$$

Using Z as a pivot, it follows that

$$\widehat{\theta} \pm z_{\alpha/2}\sigma_{\widehat{\theta}}$$

is a $100(1-\alpha)$ percent confidence interval for θ.

REMARK: In what follows, we assume that $\sigma_{\widehat{\theta}}$ does not depend on θ (although the following conclusions hold even if it does). If $\sigma_{\widehat{\theta}}$ depends on other nuisance parameters, without loss, we assume that these parameters are known.

HYPOTHESIS TEST: The two-sided level α hypothesis test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta \neq \theta_0$$

employs the rejection region

$$\text{RR} = \{z : |z| > z_{\alpha/2}\}$$

which means that H_0 is not rejected when

$$-z_{\alpha/2} < \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} < z_{\alpha/2}.$$

However, algebraically, the last inequality can be rewritten as

$$\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}} < \theta_0 < \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}},$$

which we recognize as the set of all θ_0 that fall between the $100(1-\alpha)$ percent confidence interval limits.

PUNCHLINE: The hypothesis $H_0 : \theta = \theta_0$ is not rejected in favor of $H_a : \theta \neq \theta_0$, at significance level α, whenever θ_0 is contained in the $100(1-\alpha)$ percent confidence interval for θ. If θ_0 is not contained in the $100(1-\alpha)$ percent confidence interval for θ, then this is the same as rejecting H_0 at level α.

10.6 Probability values (p-values)

REMARK: When performing a hypothesis test, simply saying that we "reject H_0" or that we "do not reject H_0" is somewhat uninformative. A probability value (p-value) provides a numerical measure of how much evidence we have against H_0.

TERMINOLOGY: The **probability value** for a hypothesis test specifies the smallest value of α for which H_0 is rejected. Thus, if the probability value is less than (or equal to) α, we reject H_0. If the probability value is greater than α, we do not reject H_0.

REMARK: Probability values are computed in a manner consistent with the alternative hypothesis H_a. Since the probability value is viewed as a measure of how much evidence we have against H_0, *it is always computed under the assumption that H_0 is true.*

Example 10.7. Suppose that $Y_1, Y_2, ..., Y_{100}$ is an iid $\mathcal{N}(\mu, \sigma_0^2)$ sample, where $\sigma_0^2 = 100$ is known, and that we want to test

$$H_0 : \mu = 75$$

versus

$$H_a : \mu > 75.$$

Suppose that the sample mean is $\overline{y} = 76.42$, and, thus, the one sample z statistic is

$$z = \frac{\overline{y} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{76.42 - 75}{10/\sqrt{100}} = 1.42.$$

Since our alternative is one-sided, we would use the rejection region $\text{RR} = \{z : z > z_\alpha\}$, where z_α denotes the upper α quantile of the standard normal distribution.

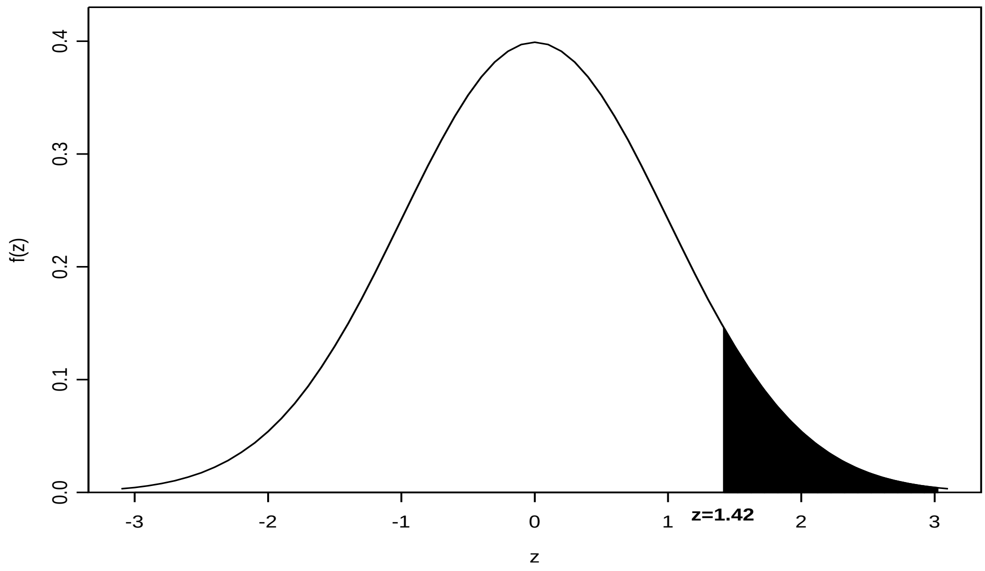

Figure 10.1: $\mathcal{N}(0,1)$ density with one-sided probability value $P(Z > 1.42) = 0.0778$.

α	Test statistic	Rejection region	Reject H_0?
$\alpha = 0.05$	$z = 1.42$	$\{z : z > 1.65\}$	no
$\alpha = 0.10$	$z = 1.42$	$\{z : z > 1.28\}$	yes

From the table, we note that

$$1.28 = z_{0.10} < z = 1.42 < z_{0.05} = 1.65.$$

Therefore, the probability value is somewhere between 0.05 and 0.10. In fact, observing that our alternative is one-sided, we see that

$$\text{p-value} = P(Z > 1.42) = 0.0778$$

(see Figure 10.1). Therefore, if $\alpha < 0.0778$, we would not reject H_0. On the other hand, if $\alpha \geq 0.0778$, we would reject H_0. Remember, the probability value is the "borderline" value of α for which H_0 is rejected. $\square$

Example 10.8. It has been suggested that less than 20 percent of all individuals who sign up for an extended gym membership continue to use the gym regularly six months after joining. Suppose that Y denotes the number of members who use a certain gym regularly (i.e., at least 3 times per week on average) six months after joining, to be observed from a sample of $n = 50$ members. Assume that $Y \sim b(50, p)$ and that we are to test

$$\begin{gathered} H_0 : p = 0.20 \\ \text{versus} \\ H_a : p < 0.20. \end{gathered}$$

If $Y = y = 6$, the exact probability value is

$$\begin{aligned} \text{p-value} &= P(Y \leq 6) \\ &= \sum_{y=0}^{6} \underbrace{\binom{50}{y}(0.20)^y(1-0.20)^{50-y}}_{b(50,0.20)\text{ pmf}} \approx 0.1034, \end{aligned}$$

computed using the `pbinom(6,50,0.20)` command in R. This is somewhat strong evidence against H_0, although it is "not enough" at the standard $\alpha = 0.05$ level of significance. Instead of using the exact probability value, we could also compute the approximate probability value as

$$\begin{aligned} \text{p-value} &= P(\widehat{p} < 0.12) \\ &\approx P\left(Z < \frac{0.12 - 0.20}{\sqrt{\frac{0.20(1-0.20)}{50}}}\right) \\ &= P(Z < -1.41) = 0.0793. \end{aligned}$$

As you can see, there is a mild discrepancy here in the exact and approximate probability values. Approximate results should always be interpreted with caution. □

REMARK: In a profound sense, a probability value, say, P, is really a random variable. This should be obvious since it depends on a test statistic, which, in turn, is computed from a sample of random variables $Y_1, Y_2, ..., Y_n$. In the light of this, it seems logical to think about the distribution of P. If the test statistic has a continuous distribution, then **when H_0 is true**, the probability value $P \sim \mathcal{U}(0, 1)$. This is a theoretical result which would be proven in a more advanced course.

10.7 Small sample hypothesis tests using the t distribution

GOAL: We now focus on small sample hypothesis tests for

- a single population mean μ
- the difference of two population means $\mu_1 - \mu_2$.

In the one-sample problem, we know that when $H_0 : \mu = \mu_0$ is true,

$$Z = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1),$$

as $n \to \infty$, by the Central Limit Theorem and Slutsky's Theorem. Therefore, Z can be used as a large sample test statistic to test $H_0 : \mu = \mu_0$. However, the large sample $\mathcal{N}(0, 1)$ distribution may be inaccurate when n is small. This occurs when the underlying distribution is highly skewed and/or when S is not a good estimator of σ.

10.7.1 One-sample test

SETTING: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, where both parameters μ and σ^2 are unknown, and that we want to test

$$H_0 : \mu = \mu_0$$

$$\text{versus}$$

$$H_a : \mu \neq \mu_0$$

(or any other H_a). When $H_0 : \mu = \mu_0$ is true, the **one-sample t-statistic**

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

Therefore, to perform a level α (two-sided) test, we use the rejection region

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

Probability values are also computed from the $t(n - 1)$ distribution. One-sided tests use a suitably-adjusted rejection region.

Table 10.3: Crab temperature data. These observations are modeled as $n = 8$ iid realizations from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

25.8	24.6	26.1	24.9	25.1	25.3	24.0	24.5

Example 10.9. A researcher observes a sample of $n = 8$ crabs and records the body temperature of each (in degrees C); see Table 10.3. She models these observations as an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. She would like to test, at level $\alpha = 0.05$,

$$H_0 : \mu = 25.4$$

versus

$$H_a : \mu < 25.4.$$

The level $\alpha = 0.05$ rejection region is

$$\text{RR} = \{t : t < -t_{7,0.05} = -1.895\}.$$

From the data in Table 10.3, we compute $\overline{y} = 25.0$ and $s = 0.69$; thus, the value of the one-sample t-statistic is

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{25.0 - 25.4}{0.69/\sqrt{8}} = -1.64.$$

Therefore, we do not have sufficient evidence to reject H_0 at the $\alpha = 0.05$ level since our test statistic t does not fall in RR. Equivalently, the probability value is

$$\text{p-value} = P[t(7) \leq -1.64] \approx 0.073,$$

which is not smaller than $\alpha = 0.05$. I used the R command `pt(-1.64,7)` to compute this probability. □

10.7.2 Two-sample test

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$ are iid with mean μ_1 and variance σ_1^2

Sample 2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$ are iid with mean μ_2 and variance σ_2^2, and that interest lies in testing

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

(or any other H_a), where d_0 is a known constant. When the population variances are equal; that is, when $\sigma_1^2 = \sigma_2^2$, we know that

$$\frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

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. Therefore, to perform a level α (two-sided) test, we use the test statistic

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

and the rejection region

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

Probability values are also computed from the $t(n_1 + n_2 - 2)$ distribution. One-sided tests use a suitably-adjusted rejection region.

REMARK: When $\sigma_1^2 \neq \sigma_2^2$; that is, when the population variances are **not** equal, we can use the modified t-statistic given by

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

Under H_0, the distribution of this modified t-statistic is approximated by a $t(\nu)$ distribution, where the degrees of freedom

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

10.8 Hypothesis tests for variances

10.8.1 One-sample test

SETTING: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, where both parameters are unknown, and that interest lies in testing

$$H_0 : \sigma^2 = \sigma_0^2$$
$$\text{versus}$$
$$H_a : \sigma^2 \neq \sigma_0^2,$$

(or any other H_a), where σ_0^2 is a specified value. When H_0 is true; i.e., when $\sigma^2 = \sigma_0^2$, the statistic

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

Therefore, a level α (two-sided) rejection region is

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

Probability values are also computed from the $\chi^2(n-1)$ distribution. One-sided tests use a suitably-adjusted rejection region.

10.8.2 Two-sample test

SETTING: Suppose that we have two **independent** samples:

$$\begin{aligned} &\text{Sample 1}: \quad Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \text{iid } \mathcal{N}(\mu_1, \sigma_1^2) \\ &\text{Sample 2}: \quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \text{iid } \mathcal{N}(\mu_2, \sigma_2^2), \end{aligned}$$

and that interest lies in testing

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$\text{versus}$$
$$H_a : \sigma_1^2 \neq \sigma_2^2,$$

(or any other H_a). Recall from Chapter 7 (WMS) that, in general,

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

However, note that when $H_0: \sigma_1^2 = \sigma_2^2$ is true, F reduces algebraically to

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

Therefore, a level α (two-sided) rejection region is

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

Probability values are also computed from the $F(n_1-1, n_2-1)$ distribution. One-sided tests use a suitably-adjusted rejection region.

10.9 Power, the Neyman-Pearson Lemma, and UMP tests

10.9.1 Power

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$ and that we use a level α rejection region to test $H_0: \theta = \theta_0$ versus a suitable alternative. The **power function** of the test, denoted by $K(\theta)$, is given by

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

That is, the power function gives the probability of rejecting H_0 as a function of θ.

- If $\theta = \theta_0$, that is H_0 is true, then $K(\theta_0) = \alpha$, the probability of Type I Error.

- For values of θ that are "close" to θ_0, one would expect the power to be smaller, than, say, when θ is far away from θ_0. This makes sense intuitively; namely, it is more difficult to detect a small departure from H_0 (i.e., to reject H_0) than it is to detect a large departure from H_0.

- The shape of the power function always depends on the alternative hypothesis.

NOTE: If θ_a is a value of θ in the alternative space; that is, if $\theta_a \in H_a$, then

$$K(\theta_a) = 1 - \beta(\theta_a).$$

Proof. This follows directly from the complement rule; that is,

$$\begin{aligned} K(\theta_a) &= P(\text{Reject } H_0|\theta = \theta_a) \\ &= 1 - P(\text{Do not reject } H_0|\theta = \theta_a) = 1 - \beta(\theta_a). \quad \square \end{aligned}$$

Example 10.10. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\theta, \sigma_0^2)$ sample, where σ_0^2 is known, and that we would like to test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta > \theta_0.$$

Suppose that we use the level α rejection region $\text{RR} = \{z : z > z_\alpha\}$, where

$$Z = \frac{\overline{Y} - \theta_0}{\sigma_0/\sqrt{n}}$$

and z_α denotes the upper α quantile of the standard normal distribution. The power function for the test, for $\theta > \theta_0$, is given by

$$\begin{aligned} K(\theta) = P(\text{Reject } H_0|\theta) &= P(Z > z_\alpha|\theta) \\ &= P\left(\left.\frac{\overline{Y} - \theta_0}{\sigma_0/\sqrt{n}} > z_\alpha\right|\theta\right) \\ &= P\left[\left.\overline{Y} > \theta_0 + z_\alpha\left(\frac{\sigma_0}{\sqrt{n}}\right)\right|\theta\right] \\ &= P\left[Z > \frac{\theta_0 + z_\alpha\left(\frac{\sigma_0}{\sqrt{n}}\right) - \theta}{\sigma_0/\sqrt{n}}\right] \\ &= 1 - F_Z\left[\frac{\theta_0 + z_\alpha\left(\frac{\sigma_0}{\sqrt{n}}\right) - \theta}{\sigma_0/\sqrt{n}}\right], \end{aligned}$$

where $F_Z(\cdot)$ denotes the $\mathcal{N}(0, 1)$ cumulative distribution function (cdf). Note that the power when $H_0 : \theta = \theta_0$ is true is

$$K(\theta_0) = 1 - F_Z(z_\alpha) = 1 - (1 - \alpha) = \alpha.$$

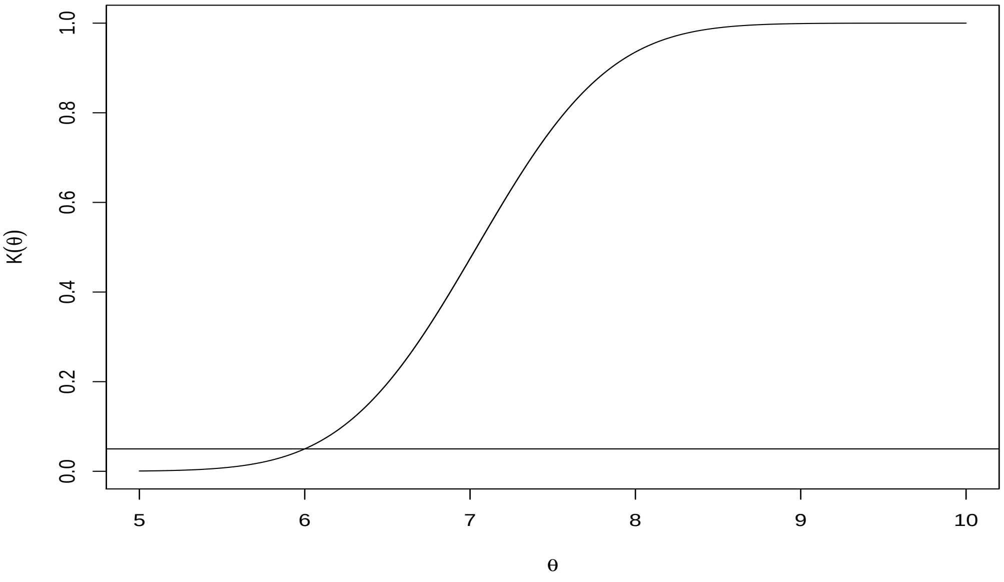

Figure 10.2: Power function $K(\theta)$ in Example 10.10 with $\alpha = 0.05$, $\theta_0 = 6$, $\sigma_0^2 = 4$, and $n = 10$. A horizontal line at $\alpha = 0.05$ is drawn.

ILLUSTRATION: Figure 10.2 displays the graph of $K(\theta)$ when $\alpha = 0.05$, $\theta_0 = 6$, $\sigma_0^2 = 4$, and $n = 10$. That is, we are testing

$$H_0 : \theta = 6$$

versus

$$H_a : \theta > 6.$$

We make the following observations.

- Note that $K(6) = 0.05$, that is, the power of the test when $H_0 : \theta = 6$ is true is equal to $\alpha = 0.05$.
- Note that $K(\theta)$ is an increasing function of θ. Therefore, the probability of rejecting H_0 increases as θ increases. For example, $K(6.5) \approx 0.1965$, $K(7) \approx 0.4746$, $K(8) \approx 0.9354$, $K(9) \approx 0.9990$, etc. □

10.9.2 The Neyman-Pearson Lemma

TERMINOLOGY: In this course, we will usually take the null hypothesis to be sharp, or **simple**; that is, there is just one value of θ possible under H_0. The alternative may be simple or **composite**. Here is an example of a simple-versus-simple test:

$$H_0 : \theta = 5$$

$$\text{versus}$$

$$H_a : \theta = 6.$$

Here is an example of a simple-versus-composite test:

$$H_0 : \theta = 5$$

$$\text{versus}$$

$$H_a : \theta > 5.$$

Note that there are an infinite number of values of θ specified in a composite alternative hypothesis. In this example, H_a consists of any value of θ larger than 5.

GOAL: For a level α simple-versus-simple test, we seek the **most powerful rejection region**; i.e., the rejection region that maximizes the probability of rejecting H_0 when H_a is true. The Neyman-Pearson Lemma tells us how to find this "most powerful test."

RECALL: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y; \theta)$. The **likelihood function** for θ is given by

$$L(\theta) = L(\theta|\boldsymbol{y}) = L(\theta|y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f_Y(y_i; \theta).$$

NEYMAN-PEARSON LEMMA: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y; \theta)$, and let $L(\theta)$ denote the likelihood function. Consider the following simple-versus-simple hypothesis test:

$$H_0 : \theta = \theta_0$$

$$\text{versus}$$

$$H_a : \theta = \theta_a.$$

The level α test that maximizes the power when $H_a : \theta = \theta_a$ is true uses the rejection region

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

where k is chosen so that

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

This is called the **most-powerful level α test** for H_0 versus H_a.

Example 10.11. Suppose that Y is a single observation (i.e., an iid sample of size $n = 1$) from an exponential distribution with mean θ. Using this single observation, we would like to test

$$H_0 : \theta = 2$$

$$\text{versus}$$

$$H_a : \theta = 3.$$

Use the Neyman-Pearson Lemma to find the most powerful level $\alpha = 0.10$ test.

Solution. Because the sample size is $n = 1$, the likelihood function $L(\theta)$ is simply

$$L(\theta) = f_Y(y; \theta) = \frac{1}{\theta} e^{-y/\theta},$$

the pdf of Y. To use the Neyman-Pearson Lemma, we first form the ratio

$$\begin{aligned} \frac{L(\theta_0)}{L(\theta_a)} = \frac{L(2)}{L(3)} &= \frac{\frac{1}{2}e^{-y/2}}{\frac{1}{3}e^{-y/3}} \\ &= \frac{3}{2} e^{-y/6}. \end{aligned}$$

Therefore, the Neyman-Pearson Lemma says that the most-powerful level $\alpha = 0.10$ test is created by choosing k such that

$$P\left(\frac{3}{2} e^{-Y/6} < k \middle| \theta = 2 \right) = 0.10.$$

This is not a friendly probability calculation to make; e.g., do you know the distribution of $\frac{3}{2}e^{-Y/6}$? There is no need to worry, because we can re-write the event $\{\frac{3}{2}e^{-Y/6} < k\}$ in a more friendly way. Note that

$$\begin{aligned}\frac{3}{2}e^{-Y/6} < k &\iff e^{-Y/6} < \frac{2k}{3}\\ &\iff -Y/6 < \ln\left(\frac{2k}{3}\right)\\ &\iff Y > -6\ln\left(\frac{2k}{3}\right) \equiv k',\ \text{say.}\end{aligned}$$

Thus, we have changed the problem to now choosing k' so that

$$P(Y > k'|\theta = 2) = 0.10.$$

This is an easy probability calculation to make. In fact, when $\theta = 2$ (i.e., H_0 is true), then $Y \sim$ exponential(2) and therefore

$$0.10 \overset{\text{set}}{=} P(Y > k'|\theta = 2) = \int_{k'}^{\infty} \frac{1}{2}e^{-y/2}dy \implies k' = 4.6052.$$

I used the `qexp(0.90,1/2)` command in R to find $k' = 4.6052$, the 90th percentile of $Y \sim$ exponential(2). Therefore, the most-powerful level $\alpha = 0.10$ test uses the rejection region

$$\text{RR} = \{y : y > 4.6052\}.$$

That is, we reject $H_0 : \theta = 2$ in favor of $H_a : \theta = 3$ whenever $Y > 4.6052$.

QUESTION. What is the power this test when H_a is true?

SOLUTION. If $H_a : \theta = 3$ is true, then $Y \sim$ exponential(3). Therefore,

$$\begin{aligned}K(3) &= P(\text{Reject } H_0|\theta = 3)\\ &= P(Y > 4.6052|\theta = 3)\\ &= \int_{4.6052}^{\infty} \frac{1}{3}e^{-y/3}dy \approx 0.2154.\end{aligned}$$

I used the `1-pexp(4.6052,1/3)` command in R to find this probability.

REMARK: Note that even though we have found the most powerful level $\alpha = 0.10$ test of H_0 versus H_a, the test is not all that powerful; we have only about a 21.5 percent chance of correcting rejecting H_0 when H_a is true. Of course, this should not be surprising, given that we have just a single observation Y. We are trying to make a decision with very little information about θ. □

Example 10.12. Suppose that $Y_1, Y_2, ..., Y_{10}$ is an iid sample of Poisson(θ) observations and that we want to test

$$H_0 : \theta = 1$$
$$\text{versus}$$
$$H_a : \theta = 2.$$

Find the most-powerful level α test.

Solution. The likelihood function for θ is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{10} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= \frac{\theta^{\sum_{i=1}^{10} y_i} e^{-10\theta}}{\prod_{i=1}^{10} y_i!} \\ &= \frac{\theta^{u} e^{-10\theta}}{\prod_{i=1}^{10} y_i!}, \end{aligned}$$

where the sufficient statistic $u = \sum_{i=1}^{10} y_i$. Now, form the ratio

$$\begin{aligned} \frac{L(\theta_0)}{L(\theta_a)} = \frac{L(1)}{L(2)} &= \frac{1^u e^{-10(1)} / \prod_{i=1}^{10} y_i!}{2^u e^{-10(2)} / \prod_{i=1}^{10} y_i!} \\ &= \frac{1}{2^u e^{-10}}. \end{aligned}$$

Therefore, the Neyman-Pearson Lemma says that the most-powerful level α test is created by choosing k such that

$$P\left(\frac{1}{2^U e^{-10}} < k \middle| \theta = 1 \right) = \alpha.$$

This is not a friendly probability calculation, so let's rewrite the event $\{\frac{1}{2^U e^{-10}} < k\}$. Note that

$$\begin{aligned} \frac{1}{2^U e^{-10}} < k &\iff 2^U e^{-10} > \frac{1}{k} \\ &\iff 2^U > \frac{e^{10}}{k} \\ &\iff U \ln 2 > 10 - \ln k \\ &\iff U > \frac{10 - \ln k}{\ln 2} \equiv k', \text{ say.} \end{aligned}$$

Thus, we have changed the problem to now choosing k' so that

$$P(U > k' | \theta = 1) = \alpha.$$

This is an easier probability to handle, because we know that when $H_0 : \theta = 1$ is true, the sufficient statistic

$$U = \sum_{i=1}^{10} Y_i \sim \text{Poisson}(10).$$

Because k' is not an integer, we need to solve the equation

$$\alpha = P(U > k'|\theta = 1) = P(U \geq m|\theta = 1)$$

for m, where $m = [k'] + 1$ and $[\cdot]$ is the greatest integer function. Consider the following table:

m	α
14	0.1355
15	0.0835
16	0.0487
17	0.0270
18	0.0143

I used the R command `1-ppois(m-1,10)` to find the α entries in this table. The Neyman-Pearson Lemma says that, for example, the most-powerful level $\alpha = 0.0487$ test uses the rejection region

$$\text{RR} = \{u : u \geq 16\}.$$

As another example, the most-powerful level $\alpha = 0.0143$ test uses the rejection region

$$\text{RR} = \{u : u \geq 18\}.$$

QUESTION. What is the power of the level $\alpha = 0.0487$ test when H_a is true?

SOLUTION. When $H_a : \theta = 2$ is true, we know that $U \sim \text{Poisson}(20)$. Therefore,

$$\begin{aligned} K(2) &= P(\text{Reject } H_0|\theta = 2) \\ &= P(U \geq 16|\theta = 2) \\ &= \sum_{j=16}^{\infty} \frac{20^j e^{-20}}{j!} \approx 0.8435. \end{aligned}$$

I used the `1-ppois(15,20)` command in R to find this probability. □

RESULT: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$ and let U be a sufficient statistic. The rejection region for the most powerful level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_a$ always depends on U.

Proof. From the Factorization Theorem, we can write

$$\frac{L(\theta_0)}{L(\theta_a)} = \frac{g(u;\theta_0)h(\boldsymbol{y})}{g(u;\theta_a)h(\boldsymbol{y})} = \frac{g(u;\theta_0)}{g(u;\theta_a)},$$

where g and h are nonnegative functions. By the Neyman-Pearson Lemma, the most-powerful level α rejection region is

$$\text{RR} = \left\{\boldsymbol{y} : \frac{L(\theta_0)}{L(\theta_a)} < k\right\} = \left\{\boldsymbol{y} : \frac{g(u;\theta_0)}{g(u;\theta_a)} < k\right\},$$

where k is chosen so that $P(\text{Reject } H_0|H_0 \text{ is true}) = \alpha$. Clearly, this rejection region depends on the sufficient statistic U. $\square$

10.9.3 Uniformly most powerful (UMP) tests

REMARK: For a simple-versus-simple test, the Neyman-Pearson Lemma shows us explicitly how to derive the most-powerful level α rejection region. We now discuss simple-versus-composite tests; e.g., $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ and $H_0 : \theta = \theta_0$ versus $H_a : \theta < \theta_0$.

TERMINOLOGY: When a test maximizes the power for all θ in the alternative space; i.e., for all $\theta \in H_a$, it is called the **uniformly most powerful (UMP) level α test**. In other words, if $K_U(\theta)$ denotes the power function for the UMP level α test of H_0 versus H_a, and if $K_{U^*}(\theta)$ denotes the power function for some other level α test, then $K_U(\theta) \geq K_{U^*}(\theta)$ for all $\theta \in H_a$.

FINDING UMP TESTS: Suppose that our goal is to find the UMP level α test of

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta > \theta_0.$$

Instead of considering this simple-versus-composite test, we first "pretend" like we have the level α simple-versus-simple test

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

where $\theta_a > \theta_0$ is arbitrary. If we can then show that neither the test statistic nor the rejection region for the most powerful level α simple-versus-simple test depends on θ_a, then the test with the same rejection region will be UMP level α for the simple-versus-composite test $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$.

CURIOSITY: Why does this work? Essentially we are showing that for a given θ_a, the level α simple-versus-simple test is most powerful, by appealing to the Neyman-Pearson Lemma. However, since the value θ_a is arbitrary and since the most powerful RR is free of θ_a, this same test must be most powerful level α for every value of $\theta_a > \theta_0$; i.e., it must be the **uniformly** most powerful (UMP) level α test for all $\theta > \theta_0$.

Example 10.13. Suppose that $Y_1, Y_2, ..., Y_{15}$ is an iid sample from a Rayleigh distribution with pdf

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

Find the UMP level $\alpha = 0.05$ test of

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

Solution. We begin by using the Neyman-Pearson Lemma to find the most-powerful level $\alpha = 0.05$ test for

$$H_0 : \theta = 1$$
$$\text{versus}$$
$$H_a : \theta = \theta_a,$$

where $\theta_a > 1$. The likelihood function is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{15} \frac{2y_i}{\theta} e^{-y_i^2/\theta} \\ &= \left(\frac{2}{\theta}\right)^{15} \prod_{i=1}^{15} y_i \; e^{-\sum_{i=1}^{15} y_i^2/\theta} \\ &= \left(\frac{2}{\theta}\right)^{15} \prod_{i=1}^{15} y_i \; e^{-u/\theta}, \end{aligned}$$

where the sufficient statistic $u = \sum_{i=1}^{15} y_i^2$. Now, form the ratio

$$\begin{aligned} \frac{L(\theta_0)}{L(\theta_a)} = \frac{L(1)}{L(\theta_a)} &= \frac{2^{15} \prod_{i=1}^{15} y_i \; e^{-u}}{\left(\frac{2}{\theta_a}\right)^{15} \prod_{i=1}^{15} y_i \; e^{-u/\theta_a}} \\ &= \theta_a^{15} e^{-u\left(1-\frac{1}{\theta_a}\right)}. \end{aligned}$$

Therefore, the Neyman-Pearson Lemma says that the most-powerful level $\alpha = 0.05$ test is created by choosing k such that

$$P\left[\theta_a^{15} e^{-U\left(1-\frac{1}{\theta_a}\right)} < k \middle| \theta = 1\right] = 0.05,$$

where $U = \sum_{i=1}^{15} Y_i^2$. This is not a friendly calculation. However, note that

$$\begin{aligned} \theta_a^{15} e^{-U\left(1-\frac{1}{\theta_a}\right)} < k &\iff e^{-U\left(1-\frac{1}{\theta_a}\right)} < \frac{k}{\theta_a^{15}} \\ &\iff -U\left(1-\frac{1}{\theta_a}\right) < \ln\left(\frac{k}{\theta_a^{15}}\right) \\ &\iff U > \frac{-\ln\left(\frac{k}{\theta_a^{15}}\right)}{1-\frac{1}{\theta_a}} \equiv k', \text{ say.} \end{aligned}$$

Thus, the problem has now changed to choosing k' so that

$$P(U > k' | \theta = 1) = 0.05.$$

This is an easier probability to handle, because we can find the distribution of the sufficient statistic U. In fact, a simple transformation argument shows that

$$Y \sim \text{Rayleigh}(\theta) \Longrightarrow W = Y^2 \sim \text{exponential}(\theta),$$

see, e.g., Exercise 9.34 (WMS, pp 458). Therefore,

$$U = \sum_{i=1}^{15} Y_i^2 = \sum_{i=1}^{15} W_i \sim \text{gamma}(15, \theta).$$

When $H_0 : \theta = 1$ is true, $U \sim \text{gamma}(15, 1)$. Therefore, we choose k' so that

$$0.05 \overset{\text{set}}{=} P(U > k'|\theta = 1) = \int_{k'}^{\infty} \frac{1}{\Gamma(15)} u^{14} e^{-u} du \implies k' = 21.8865.$$

I used the R command `qgamma(0.95,15,1)` to find $k' = 21.8865$, the 95th percentile of $U \sim \text{gamma}(15, 1)$. The Neyman-Pearson Lemma allows us to conclude that the most powerful level $\alpha = 0.05$ test of

$$H_0 : \theta = 1$$

versus

$$H_a : \theta = \theta_a$$

uses the rejection region

$$\text{RR} = \{u : u > 21.8865\}.$$

KEY POINT: Note that neither the test statistic $U = \sum_{i=1}^{15} Y_i^2$ nor the rejection region $\text{RR} = \{u : u > 21.8865\}$ depends on the specific value of θ_a in this simple alternative. Therefore, this RR is the most powerful rejection region for any $\theta_a > 1$, that is,

$$\text{RR} = \{u : u > 21.8865\}$$

is the UMP level $\alpha = 0.05$ rejection region for testing

$$H_0 : \theta = 1$$

versus

$$H_a : \theta > 1.$$

QUESTION. What is the power function $K(\theta)$ for this test?

SOLUTION. Note that, in general, $U \sim \text{gamma}(15, \theta)$. Therefore,

$$\begin{aligned} K(\theta) = P(\text{Reject } H_0|\theta) &= P(U > 21.8865|\theta) \\ &= \int_{21.8865}^{\infty} \frac{1}{\Gamma(15)\theta^{15}} u^{14} e^{-u/\theta} du \\ &= 1 - F_U(21.8865), \end{aligned}$$

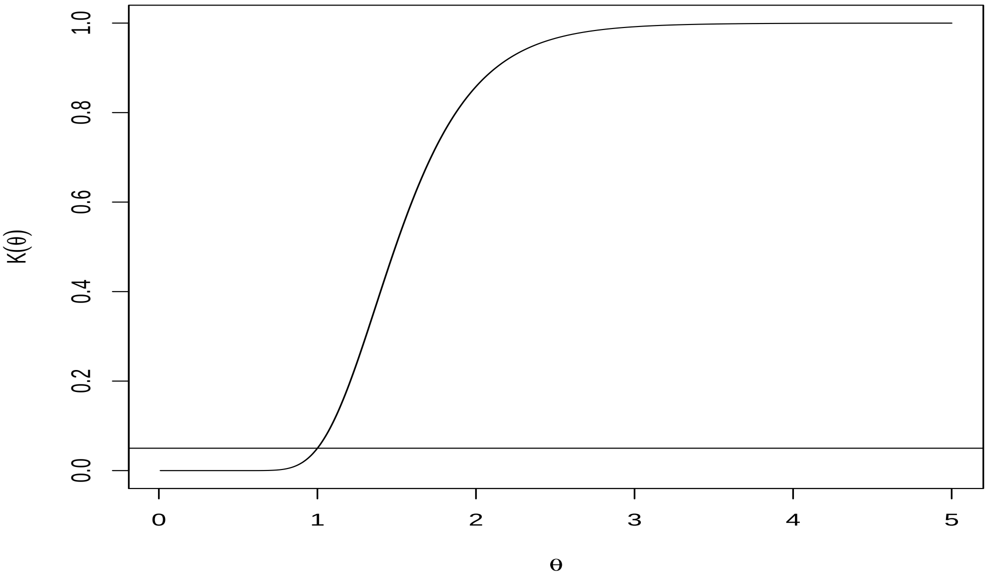

Figure 10.3: Power function $K(\theta)$ in Example 10.13 with $\alpha = 0.05$, $\theta_0 = 1$, and $n = 15$. A horizontal line at $\alpha = 0.05$ is drawn.

where $F_U(\cdot)$ is the cumulative distribution function (cdf) of $U \sim \text{gamma}(15, \theta)$. This cdf does not exist in closed form, but it be calculated in R; see Figure 10.3.

REMARK: UMP level α tests do not always exist. For example, a two-sided test $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ never has a UMP rejection region. This is because

- the power function of the UMP level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta < \theta_0$ will be larger than the power function of the UMP level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ when $\theta < \theta_0$.
- the power function of the UMP level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ will be larger than the power function of the UMP level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta < \theta_0$ when $\theta > \theta_0$.

For two-sided alternatives, the class of level α tests, say, $\mathcal{C}$, is too large, and finding one rejection region that uniformly beats all other level α rejection regions is impossible.

10.10 Likelihood ratio tests

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y; \theta)$, where the parameter $\theta \in \Omega$. We call Ω the **parameter space**; that is, Ω represents the set of all values that θ (scalar or vector) can assume. For example, if

- $Y \sim b(1, \theta) \Longrightarrow \Omega = \{\theta : 0 < \theta < 1\}$
- $Y \sim \text{exponential}(\theta) \Longrightarrow \Omega = \{\theta : \theta > 0\}$
- $Y \sim \text{gamma}(\alpha, \beta) \Longrightarrow \Omega = \{\boldsymbol{\theta} = (\alpha, \beta)' : \alpha > 0,\ \beta > 0\}$
- $Y \sim \mathcal{N}(\mu, \sigma^2) \Longrightarrow \Omega = \{\boldsymbol{\theta} = (\mu, \sigma^2)' : -\infty < \mu < \infty,\ \sigma^2 > 0\}$.

TERMINOLOGY: Suppose that we partition Ω into two sets Ω_0 and Ω_a, that is, we write

$$\Omega = \Omega_0 \cup \Omega_a,$$

where Ω_0 and Ω_a are mutually exclusive. A hypothesis test can be stated very generally as $H_0 : \theta \in \Omega_0$ versus $H_a : \theta \in \Omega_a$. We call Ω_0 the **null space** and Ω_a the **alternative space**.

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y; \theta)$, where $\theta \in \Omega$. A level α **likelihood ratio test** (LRT) for

$$H_0 : \theta \in \Omega_0$$

versus

$$H_a : \theta \in \Omega_a$$

employs the test statistic

$$\lambda = \frac{L(\widehat{\Omega}_0)}{L(\widehat{\Omega})} \equiv \frac{\sup_{\theta \in \Omega_0} L(\theta)}{\sup_{\theta \in \Omega} L(\theta)}$$

and uses the rejection region

$$\text{RR} = \{\lambda : \lambda \leq k\},$$

where k is chosen such that

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

From the definition, we see that $0 \leq \lambda \leq 1$, because $L(\cdot)$ is positive and $\Omega_0 \subset \Omega$. Also,

- $L(\widehat{\Omega}_0)$ is the likelihood function evaluated at the maximum likelihood estimator (MLE) over Ω_0, the "restricted" parameter space.
- $L(\widehat{\Omega})$ is the likelihood function evaluated at the MLE over Ω, the "unrestricted" parameter space.

TECHNICAL NOTE: If H_0 is a **composite** hypothesis, we define

$$\alpha = \sup_{\theta \in \Omega_0} P(\text{Reject } H_0|\theta) = \sup_{\theta \in \Omega_0} K(\theta),$$

where $K(\theta)$ denotes the power function.

Example 10.14. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(θ) distribution. Find the level α likelihood ratio test (LRT) for

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

Solution. Here, the restricted parameter space is $\Omega_0 = \{\theta_0\}$, that is, Ω_0 contains only one value of θ. The alternative parameter space is $\Omega_a = \{\theta : \theta > 0,\ \theta \neq \theta_0\}$, and the unrestricted parameter space is $\Omega = \{\theta : \theta > 0\}$. Note that $\Omega = \Omega_0 \cup \Omega_a$. The likelihood function for θ is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} e^{-y_i/\theta} \\ &= \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} y_i/\theta} \\ &= \theta^{-n} e^{-u/\theta}, \end{aligned}$$

where the sufficient statistic $u = \sum_{i=1}^{n} y_i$. Over the restricted (null) space, we have

$$L(\widehat{\Omega}_0) = \sup_{\theta \in \Omega_0} L(\theta) = L(\theta_0),$$

because Ω_0 contains only the singleton θ_0. Over the unrestricted space,

$$L(\widehat{\Omega}) = \sup_{\theta \in \Omega} L(\theta) = L(\widehat{\theta}),$$

where $\widehat{\theta}$ is the maximum likelihood estimator (MLE) of θ. Recall that for the exponential(θ) model, the MLE is

$$\widehat{\theta} = \overline{Y}.$$

Therefore, the likelihood ratio test statistic is

$$\lambda = \frac{L(\widehat{\Omega}_0)}{L(\widehat{\Omega})} = \frac{L(\theta_0)}{L(\overline{y})} = \frac{\theta_0^{-n} e^{-u/\theta_0}}{\overline{y}^{-n} e^{-u/\overline{y}}}.$$

Because $u = \sum_{i=1}^n y_i = n\overline{y}$, we can rewrite λ as

$$\lambda = \left(\frac{\overline{y}}{\theta_0}\right)^n \frac{e^{-u/\theta_0}}{e^{-n\overline{y}/\overline{y}}} = \left(\frac{e}{\theta_0}\right)^n \overline{y}^n e^{-n\overline{y}/\theta_0}.$$

Therefore, to find the level α LRT, we would choose k so that

$$P\left[\left(\frac{e}{\theta_0}\right)^n \overline{Y}^n e^{-n\overline{Y}/\theta_0} \leq k \,\middle|\, \theta = \theta_0\right] = \alpha.$$

This is an unfriendly request, so let's approach the problem of choosing k in another way.

EXCURSION: For $a > 0$, define the function

$$g(a) = \left(\frac{e}{\theta_0}\right)^n a^n e^{-na/\theta_0}$$

so that

$$\ln g(a) = \ln c_0 + n \ln a - \frac{na}{\theta_0},$$

where the constant $c_0 = (e/\theta_0)^n$. Note that

$$\frac{\partial \ln g(a)}{\partial a} = \frac{n}{a} - \frac{n}{\theta_0}.$$

If we set this derivative equal to 0 and solve for a, we get the first order critical point

$$a = \theta_0.$$

This value of a maximizes $\ln g(a)$ because

$$\frac{\partial^2 \ln g(a)}{\partial a^2} = -\frac{n}{a^2} < 0,$$

by the Second Derivative Test. Also, note that

$$\frac{\partial \ln g(a)}{\partial a} = \begin{cases} \frac{n}{a} - \frac{n}{\theta_0} > 0, & \text{if } a < \theta_0 \\ \frac{n}{a} - \frac{n}{\theta_0} < 0, & \text{if } a > \theta_0, \end{cases}$$

so $\ln g(a)$ is strictly increasing for $a < \theta_0$ and strictly decreasing for $a > \theta_0$. However, because the log function is 1:1, all of these findings apply to the function $g(a)$ as well:

- $g(a)$ is strictly increasing when $a < \theta_0$.
- $g(a)$ is strictly decreasing when $a > \theta_0$.
- $g(a)$ is maximized when $a = \theta_0$.

Therefore, there exist constants $c_1 < c_2$ such that

$$g(a) \leq k \Longleftrightarrow a \leq c_1 \text{ or } a \geq c_2.$$

This is easy to see from sketching a graph of $g(a)$, for $a > 0$.

LRT: Now, returning to the problem at hand, we need to choose k so that

$$P\left[\left(\frac{e}{\theta_0}\right)^n \overline{Y}^n e^{-n\overline{Y}/\theta_0} \leq k \middle| \theta = \theta_0\right] = \alpha.$$

The recent excursive argument should convince you that this is equivalent to choosing c_1 and c_2 so that

$$P(\{\overline{Y} \leq c_1\} \cup \{\overline{Y} \geq c_2\} | \theta = \theta_0) = \alpha.$$

However, because $c_1 < c_2$, the sets $\{\overline{Y} \leq c_1\}$ and $\{\overline{Y} \geq c_2\}$ must be mutually exclusive. By Kolmolgorov's third axiom of probability, we have

$$\begin{aligned} \alpha &= P(\{\overline{Y} \leq c_1\} \cup \{\overline{Y} \geq c_2\} | \theta = \theta_0) \\ &= P(\overline{Y} \leq c_1 | \theta = \theta_0) + P(\overline{Y} \geq c_2 | \theta = \theta_0). \end{aligned}$$

We have changed the problem to now specifying the constants c_1 and c_2 that satisfy this most recent expression. This is a much friendlier request because the distribution of $\overline{Y}$ is tractable; in fact, a simple mgf argument shows that, in general,

$$\overline{Y} \sim \text{gamma}\left(n, \frac{\theta}{n}\right),$$

Therefore, when $H_0 : \theta = \theta_0$ is true, we can take c_1 and c_2 to satisfy

$$\begin{aligned}\int_0^{c_1} \frac{1}{\Gamma(n)\left(\frac{\theta_0}{n}\right)^n} a^{n-1} e^{-a/\left(\frac{\theta_0}{n}\right)} da &= \alpha/2 \\ \int_{c_2}^{\infty} \frac{1}{\Gamma(n)\left(\frac{\theta_0}{n}\right)^n} a^{n-1} e^{-a/\left(\frac{\theta_0}{n}\right)} da &= \alpha/2,\end{aligned}$$

that is, c_1 is the lower $\alpha/2$ quantile of the gamma$(n, \theta_0/n)$ distribution and c_2 is the corresponding upper $\alpha/2$ quantile. R makes getting these quantiles simple. It is possible to get closed-form expressions for c_1 and c_2. In fact, it can be shown that

$$\begin{aligned}c_1 &= \left(\frac{\theta_0}{2n}\right)\chi^2_{2n,1-\alpha/2} \\ c_2 &= \left(\frac{\theta_0}{2n}\right)\chi^2_{2n,\alpha/2},\end{aligned}$$

where $\chi^2_{2n,1-\alpha/2}$ and $\chi^2_{2n,\alpha/2}$ are the lower and upper $\alpha/2$ quantiles of the $\chi^2(2n)$ distribution. Therefore, the level α likelihood ratio test (LRT) uses the rejection region

$$\text{RR} = \{\overline{y} : \overline{y} \leq c_1 \text{ or } \overline{y} \geq c_2\}.$$

ILLUSTRATION: Suppose that $\alpha = 0.05$, $\theta_0 = 10$, and $n = 20$, so that

$$\begin{aligned}c_1 &= \left(\frac{10}{40}\right)\chi^2_{40,0.975} = 6.1083 \\ c_2 &= \left(\frac{10}{40}\right)\chi^2_{40,0.025} = 14.8354.\end{aligned}$$

Therefore, the level $\alpha = 0.05$ LRT employs the rejection region

$$\text{RR} = \{\overline{y} : \overline{y} \leq 6.1083 \text{ or } \overline{y} \geq 14.8354\}.$$

For this rejection region, the power function is given by

$$\begin{aligned}K(\theta) &= P(\text{Reject } H_0|\theta) \\ &= P(\overline{Y} \leq c_1|\theta) + P(\overline{Y} \geq c_2|\theta) \\ &= \int_0^{c_1} \frac{1}{\Gamma(20)\left(\frac{\theta}{20}\right)^{20}} a^{20-1} e^{-a/\left(\frac{\theta}{20}\right)} da + \int_{c_2}^{\infty} \frac{1}{\Gamma(20)\left(\frac{\theta}{20}\right)^{20}} a^{20-1} e^{-a/\left(\frac{\theta}{20}\right)} da\end{aligned}$$

This power function is shown in Figure 10.4. □

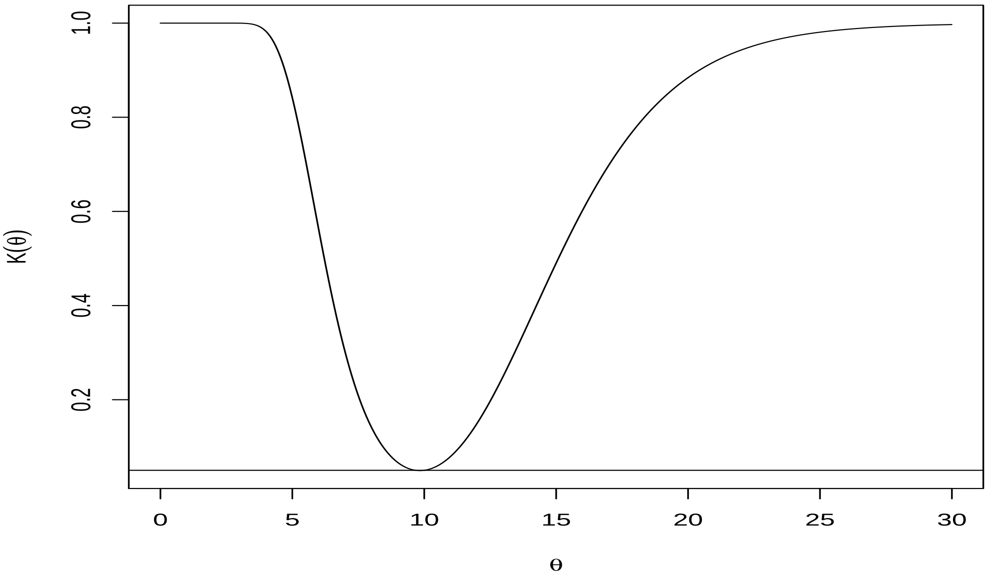

Figure 10.4: Power function $K(\theta)$ in Example 10.14 with $\alpha = 0.05$, $\theta_0 = 10$, and $n = 20$. A horizontal line at $\alpha = 0.05$ is drawn.

REMARK: In Example 10.14, we were fortunate to know the sampling distribution of $\overline{Y}$ when H_0 was true. In other situations, the distribution of the test statistic may be intractable. When this occurs, the following large-sample result can prove to be useful.

ASYMPTOTIC RESULT: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$, where $\theta \in \Omega$, and that we are to test

$$H_0 : \theta \in \Omega_0$$

versus

$$H_a : \theta \in \Omega_a.$$

Under certain "regularity conditions" (which we will omit), it follows that, under H_0,

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

as $n \to \infty$, where ν is the difference between the number of free parameters specified by $\theta \in \Omega_0$ and the number of free parameters specified in by $\theta \in \Omega$. The term "free parameters" will become clear in the next example.

Example 10.15. 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?"*

Based on the data collected, here is the table that cross-classifies all 2,864 subjects by care denial (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

Are HIV-infected individuals in certain insurance groups more likely to be denied dental care? To answer this, we would like to test

$$H_0 : p_1 = p_2 = p_3 = p_4$$

$$\text{versus}$$

$$H_a : H_0 \text{ not true},$$

where p_i denotes the population proportion of subjects in insurance group i who are denied dental care. Perform a level $\alpha = 0.05$ LRT to test this claim.

Solution. We will assume that the four groups of individuals (stratified by insurance type) are independent and denote by

$$Y_i = \text{ number of individuals in the } i\text{th insurance group denied dental care,}$$

for $i = 1, 2, 3, 4$. Treating the column totals as fixed, we assume that

$$Y_i \sim b(n_i, p_i).$$

That is, within the ith insurance group, we are envisaging each patient as a "trial;" if a patient is denied dental care, s/he is treated as a "success." Here, note that

$$\begin{aligned}\Omega_0 &= \{\boldsymbol{\theta} : \boldsymbol{\theta} \in [0,1]^4 : p_1 = p_2 = p_3 = p_4\} \\ \Omega &= \{\boldsymbol{\theta} : \boldsymbol{\theta} \in [0,1]^4\},\end{aligned}$$

where $\boldsymbol{\theta} = (p_1, p_2, p_3, p_4)'$, and $\Omega_a = \Omega - \Omega_0$.

- Under $H_0 : p_1 = p_2 = p_3 = p_4$, each of the parameters is the same. Therefore, only 1 of the parameters is allowed to vary (i.e., once we know 1, the other 3 are uniquely determined).
- Under H_a, all 4 parameters allowed to vary freely.
- Therefore, the difference in the number of free parameters is $\nu = 4 - 1 = 3$.

The likelihood function of $\boldsymbol{\theta} = (p_1, p_2, p_3, p_4)'$ is given by the product of the four binomial probability mass functions; i.e.,

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

MLE under H_0:

When H_0 is true, that is, when $\boldsymbol{\theta} \in H_0$, then

$$p_1 = p_2 = p_3 = p_4 = p,$$

say, and the likelihood function L reduces to

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

The loglikelihood function is given by

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

where the constant $c = \prod_{i=1}^4 \binom{n_i}{y_i}$. 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 under H_0, we set this partial derivative equal to zero and solve for p. That is,

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

It is straightforward to show that $\partial^2/\partial p^2 \ln L(\widehat{p}) < 0$ so that $\widehat{p}$ maximizes $\ln L(p)$ by the Second Derivative Test.

Unrestricted MLE:

Maximizing $L(p_1, p_2, p_3, p_4)$ over the unrestricted space Ω is just as easy. To do this, we write

$$\begin{aligned}
L(p_1, p_2, p_3, p_4) &= \prod_{i=1}^4 \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}\\
&= \prod_{i=1}^4 \binom{n_i}{y_i} \prod_{i=1}^4 p_i^{y_i} \prod_{i=1}^4 (1-p_i)^{n_i - y_i},
\end{aligned}$$

so that the loglikelihood function is

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

The unrestricted maximum likelihood estimator of $\boldsymbol{\theta} = (p_1, p_2, p_3, p_4)'$ is obtained by solving the system

$$\begin{aligned}
\frac{\partial \ln L(p_1, p_2, p_3, p_4)}{\partial p_1} &\overset{\text{set}}{=} 0\\
\frac{\partial \ln L(p_1, p_2, p_3, p_4)}{\partial p_2} &\overset{\text{set}}{=} 0\\
\frac{\partial \ln L(p_1, p_2, p_3, p_4)}{\partial p_3} &\overset{\text{set}}{=} 0\\
\frac{\partial \ln L(p_1, p_2, p_3, p_4)}{\partial p_4} &\overset{\text{set}}{=} 0,
\end{aligned}$$

for p_1, p_2, p_3, and p_4, producing maximum likelihood estimators $\widehat{p}_1$, $\widehat{p}_2$, $\widehat{p}_3$, and $\widehat{p}_4$, respectively. This system of partial derivatives becomes

$$\begin{aligned}
\frac{y_1}{p_1} - \frac{n_1 - y_1}{1 - p_1} &\overset{\text{set}}{=} 0 \\
\frac{y_2}{p_2} - \frac{n_2 - y_2}{1 - p_2} &\overset{\text{set}}{=} 0 \\
\frac{y_3}{p_3} - \frac{n_3 - y_3}{1 - p_3} &\overset{\text{set}}{=} 0 \\
\frac{y_4}{p_4} - \frac{n_4 - y_4}{1 - p_4} &\overset{\text{set}}{=} 0.
\end{aligned}$$

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**.

LRT statistic:

The likelihood ratio statistic is given by

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

To find the (exact) level $\alpha = 0.05$ LRT, one would have to specify the value of k that satisfies

$$P\left\{ \frac{\left(\frac{\sum_{i=1}^4 Y_i}{\sum_{i=1}^4 n_i}\right)^{\sum_{i=1}^4 Y_i} \left[1 - \left(\frac{\sum_{i=1}^4 Y_i}{\sum_{i=1}^4 n_i}\right)\right]^{\sum_{i=1}^4 (n_i - Y_i)}}{\prod_{i=1}^4 \left(\frac{Y_i}{n_i}\right)^{Y_i} \prod_{i=1}^4 \left[1 - \left(\frac{Y_i}{n_i}\right)\right]^{n_i - Y_i}} \leq k \;\middle|\; p_1 = p_2 = p_3 = p_4 \right\} = 0.05.$$

Because this is an intractable request, it is more sensible to use the large sample χ^2 approximation to $-2 \ln \lambda$. Under H_0,

$$-2 \ln \lambda \xrightarrow{d} \chi^2(\nu = 3),$$

as $\min_i n_i \to \infty$. Therefore, the statistic $-2\ln\lambda$ follows an approximate $\chi^2(3)$ distribution when H_0 is true. Now,

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

Furthermore,

$$0.05 = P(\lambda \leq k|H_0) = P(-2\ln\lambda \geq k'|H_0)$$

so taking $k' = \chi^2_{3,\alpha}$ makes

$$\text{RR} = \{\lambda : -2\ln\lambda \geq \chi^2_{3,\alpha}\}$$

an approximate level α rejection region for testing H_0 versus H_a.

DENTAL CARE DATA: For the dental care data, we have $\nu = 3$, so the approximate level $\alpha = 0.05$ rejection region is

$$\text{RR} = \{\lambda : -2\ln\lambda \geq \chi^2_{3,0.05} = 7.8147\}.$$

The binomial stratum counts are $y_1 = 49$, $y_2 = 142$, $y_3 = 181$, and $y_4 = 175$. The stratum sample sizes are $n_1 = 658$, $n_2 = 839$, $n_3 = 811$, and $n_4 = 556$. It is straightforward (but tedious) to calculate

$$\begin{aligned}-2\ln\lambda &= -2\ln\left\{\frac{\left(\frac{547}{2864}\right)^{547}\left(\frac{2317}{2864}\right)^{2317}}{\left(\frac{49}{658}\right)^{49}\left(\frac{142}{839}\right)^{142}\left(\frac{181}{811}\right)^{181}\left(\frac{175}{556}\right)^{175}\left(\frac{609}{658}\right)^{609}\left(\frac{697}{839}\right)^{697}\left(\frac{630}{811}\right)^{630}\left(\frac{381}{556}\right)^{381}}\right\}\\ &\approx 127.7924.\end{aligned}$$

We therefore reject H_0 because the test statistic $-2\ln\lambda = 127.7924$ falls in the rejection region. In fact, the probability value is

$$\text{p-value} = P[\chi^2(3) > 127.7924] < 0.0000000000000001,$$

indicating that the evidence against H_0 is indeed overwhelming. Based on these data, there is clear evidence that HIV-infected individuals in certain insurance groups are more likely to be denied dental care.

11 Linear Regression Models

Complementary reading: Chapter 11 and Appendix A (WMS).

11.1 Introduction

IMPORTANCE: A problem that often arises in economics, engineering, medicine, and other areas is that of investigating the mathematical relationship between two (or more) variables. In such settings, the goal is often to model a continuous random variable Y as a function of one or more **independent variables**, say, $x_1, x_2, ..., x_k$. Mathematically, we can express this model as

$$Y = g(x_1, x_2, ..., x_k) + \epsilon,$$

where $g : \mathcal{R}^k \to \mathcal{R}$, and where the random variable ϵ satisfies certain conditions. This is called a **regression model**.

- The presence of the random error term ϵ conveys the fact that the relationship between the dependent variable Y and the independent variables through $g(x_1, x_2, ..., x_k)$ is not perfect.

- The independent variables $x_1, x_2, ..., x_k$ are assumed to be **fixed** (not random), and they are measured without error. If $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$, then

$$\begin{aligned} E(Y|x_1, x_2, ..., x_k) &= g(x_1, x_2, ..., x_k) \\ V(Y|x_1, x_2, ..., x_k) &= \sigma^2. \end{aligned}$$

LINEAR MODELS: In this course, we will consider models of the form

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}_{g(x_1, x_2, ..., x_k)} + \epsilon,$$

that is, g is a linear function of the **regression parameters** β_0, β_1, ..., β_k. We call this a **linear regression model**.

REMARK: In some problems, a **nonlinear regression model** may be appropriate. For example, suppose that Y measures plant growth (in cm, say) and x denotes time. We would expect the relationship to eventually "level off" as x gets large, as plants can not continue to grow forever. A popular model for this situation is the nonlinear model

$$Y = \underbrace{\frac{\beta_0}{1+\beta_1 e^{\beta_2 x}}}_{g(x)} + \epsilon.$$

Note that, if $\beta_2 < 0$, then

$$\lim_{x\rightarrow\infty} g(x) = \lim_{x\rightarrow\infty}\left(\frac{\beta_0}{1+\beta_1 e^{\beta_2 x}}\right) = \beta_0.$$

Therefore, if $\beta_2 < 0$, this g function has a horizontal asymptote at $y = \beta_0$, a characteristic that is consistent with the data we would likely observe.

DESCRIPTION: We call a regression model a **linear regression model** if the regression parameters enter the g function in a linear fashion. For example, each of the models is a linear regression model:

$$\begin{aligned}
Y &= \underbrace{\beta_0 + \beta_1 x}_{g(x)} + \epsilon \\
Y &= \underbrace{\beta_0 + \beta_1 x + \beta_2 x^2}_{g(x)} + \epsilon \\
Y &= \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}_{g(x_1,x_2)} + \epsilon.
\end{aligned}$$

These should be contrasted with the nonlinear model above, where the regression parameters β_0, β_1, and β_2 enter the g function nonlinearly. The term "linear" does not refer to the shape of the regression function g. It refers to the manner in which the regression parameters β_0, β_1, ..., β_k enter the g function.

GOALS: We will restrict attention to linear (regression) models. Our goals are to

- obtain estimates of the regression parameters and study the sampling distributions of these estimators
- perform statistical inference for the regression parameters and functions of them
- make predictions about future values of Y based on an estimated model.

11.2 Simple linear regression

TERMINOLOGY: A **simple linear regression model** includes only one independent variable x. The model is of the form

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

The regression function $g(x) = \beta_0 + \beta_1 x$ is a straight line with intercept β_0 and slope β_1. If $E(\epsilon) = 0$, then β_1 quantifies the change in $E(Y)$ brought about by a one-unit change in x.

TERMINOLOGY: When we say, "fit a regression model," we mean that we would like to estimate the regression parameters in the model with the observed data. Suppose that we collect (x_i, Y_i), $i = 1, 2, ..., n$, and postulate the simple linear regression model

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

for each $i = 1, 2, ..., n$. Our first goal is to estimate β_0 and β_1. Formal assumptions for the error terms ϵ_i will be given later.

11.2.1 Least squares estimation

LEAST SQUARES: A widely-accepted method of estimating the model parameters β_0 and β_1 is that of least squares. The **method of least squares** says to choose the values of β_0 and β_1 that minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2.$$

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 $Q(\beta_0, \beta_1)$. A two-variable minimization exercise can be used to find expressions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Taking partial derivatives of $Q(\beta_0, \beta_1)$, we obtain

$$\begin{aligned}
\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) \overset{\text{set}}{=} 0 \\
\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)x_i \overset{\text{set}}{=} 0.
\end{aligned}$$

Solving for β_0 and β_1 gives the **least squares estimators**

$$\begin{aligned}\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1\overline{x}\\ \widehat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i-\overline{x})(Y_i-\overline{Y})}{\sum_{i=1}^n (x_i-\overline{x})^2}.\end{aligned}$$

11.2.2 Properties of the least squares estimators

INTEREST: We wish to investigate the properties of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as estimators of the true regression parameters 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, we need to formally state our assumptions on the error terms ϵ_i. Specifically, we will assume that $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. This means that

- $E(\epsilon_i) = 0$, for $i = 1, 2, ..., n$
- $V(\epsilon_i) = \sigma^2$, for $i = 1, 2, ..., n$, that is, the variance is constant
- the random variables ϵ_i are independent
- the random variables ϵ_i are normally distributed.

OBSERVATION: Under the assumption that $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, it follows that

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

In addition, the random variables Y_i are **independent**. They are not identically distributed because the mean $\beta_0 + \beta_1 x_i$ is different for each x_i.

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,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

since

$$\begin{aligned}\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y}) &= \sum_{i=1}^n (x_i - \overline{x})Y_i - \sum_{i=1}^n (x_i - \overline{x})\overline{Y} \\ &= \sum_{i=1}^n (x_i - \overline{x})Y_i - \overline{Y}\sum_{i=1}^n (x_i - \overline{x})\end{aligned}$$

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 that $\widehat{\beta}_1$ 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

$$\begin{aligned}E(\widehat{\beta}_1) = E\left(\sum_{i=1}^n c_i Y_i\right) &= \sum_{i=1}^n c_i E(Y_i) \\ &= \sum_{i=1}^n c_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i.\end{aligned}$$

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 that $\widehat{\beta}_0$ is unbiased, we first note that

$$E(\widehat{\beta}_0) = E(\overline{Y} - \widehat{\beta}_1 \overline{x}) = E(\overline{Y}) - \overline{x}E(\widehat{\beta}_1).$$

However, $E(\widehat{\beta}_1) = \beta_1$ and

$$\begin{aligned} E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\beta_0 + \frac{1}{n}\sum_{i=1}^{n}\beta_1 x_i \\ &= \beta_0 + \beta_1\overline{x}. \end{aligned}$$

Therefore,

$$\begin{aligned} E(\widehat{\beta}_0) &= E(\overline{Y}) - \overline{x}E(\widehat{\beta}_1) \\ &= \beta_0 + \beta_1\overline{x} - \beta_1\overline{x} = \beta_0, \end{aligned}$$

as claimed. We have shown that the least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased.

NOTE: It is important to note that the only assumption we used in the preceding argument was that $E(\epsilon_i) = 0$. Therefore, a sufficient condition for the least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to be unbiased is that $E(\epsilon_i) = 0$. □

Fact 2. The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have the following characteristics:

$$\begin{aligned} V(\widehat{\beta}_0) &= \sigma^2\left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \\ V(\widehat{\beta}_1) &= \sigma^2\left[\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \\ \text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) &= \sigma^2\left[\frac{-\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]. \end{aligned}$$

REMARK: For these formulae to hold, we need to use the assumptions that $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$, and ϵ_i independent (i.e., 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

$$\begin{aligned} V(\widehat{\beta}_0) &= V(\overline{Y} - \widehat{\beta}_1 \overline{x}) \\ &= V(\overline{Y}) + \overline{x}^2 V(\widehat{\beta}_1) - 2\overline{x}\text{Cov}(\overline{Y}, \widehat{\beta}_1). \end{aligned}$$

Note that

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

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,

$$\begin{aligned} V(\widehat{\beta}_0) = \frac{\sigma^2}{n} + \sigma^2\left[\frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] &= \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \\ &= \sigma^2\left[\frac{\sum_{i=1}^n (x_i - \overline{x})^2 + n\overline{x}^2}{n\sum_{i=1}^n (x_i - \overline{x})^2}\right] \\ &= \sigma^2\left[\frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2}\right], \end{aligned}$$

as claimed. Finally, the covariance between $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is given by

$$\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \text{Cov}(\overline{Y} - \widehat{\beta}_1 \overline{x}, \widehat{\beta}_1) = \text{Cov}(\overline{Y}, \widehat{\beta}_1) - \overline{x}V(\widehat{\beta}_1).$$

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 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 $i = 1, 2, ..., n$. However, under our model assumptions,

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

Therefore, $\widehat{\beta}_1$ is normally distributed since 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. Therefore, $\widehat{\beta}_0$ is normally distributed as well. Note that we have used the normality assumption on the errors ϵ_i to prove this fact. □

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 that

$$\widehat{\beta}_0 \sim \mathcal{N}(\beta_0, c_{00}\sigma^2) \quad \text{and} \quad \widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2),$$

where

$$c_{00} = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2} \quad \text{and} \quad c_{11} = \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

11.2.3 Estimating the error variance

REVIEW: In the simple linear regression model

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

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, 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**.

NOTE: In the simple linear regression model, we 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. We define the ith **residual** by

$$e_i = Y_i - \widehat{Y}_i.$$

We define the **error (residual) sum of squares** by

$$\text{SSE} \equiv \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

Fact 4. In the simple linear regression model,

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

is an unbiased estimator of σ^2, that is, $E(\widehat{\sigma}^2) = \sigma^2$.

Proof. See WMS, pp 580-581. We will prove this later under a more general setting. □

NOTATION: Your authors denote the unbiased estimator of σ^2 by S^2. I don't like this notation because we have always used S^2 to denote the sample variance of $Y_1, Y_2, ..., Y_n$.

Fact 5. If $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, then

$$\frac{\text{SSE}}{\sigma^2} = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

The proof of this fact is beyond the scope of this course.

Fact 6. If $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, then $\widehat{\sigma}^2$ is **independent** of both $\widehat{\beta}_0$ and $\widehat{\beta}_1$. The proof of this fact is also beyond the scope of this course.

11.2.4 Inference for β_0 and β_1

INTEREST: In the simple linear regression model

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

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid}\, \mathcal{N}(0, \sigma^2)$, the regression parameters β_0 and β_1 are unknown. It is therefore of interest to (a) construct confidence intervals and (b) perform hypothesis tests for these parameters. In practice, 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 usually less meaningful, unless one is explicitly interested in the mean of Y when $x = 0$.

INFERENCE FOR β_1: Under our model assumptions, recall that the least squares estimator

$$\widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2),$$

where $c_{11} = 1/\sum_{i=1}^n (x_i - \overline{x})^2$. Standardizing, we have

$$Z = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\sigma^2}} \sim \mathcal{N}(0, 1).$$

Recall also that

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_1$, it follows that Z and W are also independent. Therefore,

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\widehat{\sigma}^2}} = \frac{(\widehat{\beta}_1 - \beta_1)/\sqrt{c_{11}\sigma^2}}{\sqrt{\frac{(n-2)\widehat{\sigma}^2}{\sigma^2}\Big/(n-2)}} \sim t(n-2).$$

Because $t \sim t(n-2)$, t is a pivot and we can write

$$P\left(-t_{n-2,\alpha/2} < \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{c_{11}\widehat{\sigma}^2}} < t_{n-2,\alpha/2}\right) = 1 - \alpha,$$

where $t_{n-2,\alpha/2}$ denotes the upper $\alpha/2$ quantile of the $t(n-2)$ distribution. Rearranging the event inside the probability symbol, we have

$$P\left(\widehat{\beta}_1 - t_{n-2,\alpha/2}\sqrt{c_{11}\widehat{\sigma}^2} < \beta_1 < \widehat{\beta}_1 + t_{n-2,\alpha/2}\sqrt{c_{11}\widehat{\sigma}^2}\right) = 1 - \alpha,$$

which shows that

$$\widehat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{c_{11}\widehat{\sigma}^2}.$$

is a $100(1-\alpha)$ percent confidence interval for β_1. If our interest was to test

$$H_0 : \beta_1 = \beta_{1,0}$$

versus

$$H_a : \beta_1 \neq \beta_{1,0},$$

where $\beta_{1,0}$ is a fixed 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: A completely analogous argument shows that

$$t = \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{c_{00}\widehat{\sigma}^2}} \sim t(n-2),$$

where $c_{00} = \sum_{i=1}^n x_i^2/n\sum_{i=1}^n(x_i - \overline{x})^2$. Therefore, a $100(1-\alpha)$ percent confidence interval for β_0 is

$$\widehat{\beta}_0 \pm t_{n-2,\alpha/2}\sqrt{c_{00}\widehat{\sigma}^2}.$$

In addition, a level α test of

$$H_0 : \beta_0 = \beta_{0,0}$$

versus

$$H_a : \beta_0 \neq \beta_{0,0}$$

can be performed using

$$t = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\sqrt{c_{00}\widehat{\sigma}^2}}$$

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.

11.2.5 Confidence intervals for $E(Y|x^*)$

INTEREST: 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 first consider constructing confidence intervals for linear parametric functions of the form

$$\theta = a_0\beta_0 + a_1\beta_1,$$

where a_0 and a_1 are fixed constants.

ESTIMATION: Using the least squares estimators of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as point estimators for β_0 and β_1, respectively, a point estimator for θ is

$$\widehat{\theta} = a_0\widehat{\beta}_0 + a_1\widehat{\beta}_1.$$

It is easy to see that $\widehat{\theta}$ is an unbiased estimator for θ since

$$E(\widehat{\theta}) = a_0E(\widehat{\beta}_0) + a_1E(\widehat{\beta}_1) = a_0\beta_0 + a_1\beta_1 = \theta.$$

It is also possible to show that

$$V(\widehat{\theta}) \equiv \sigma^2_{\widehat{\theta}} = \sigma^2 \left[\frac{\frac{a_0^2}{n}\sum_{i=1}^n x_i^2 + a_1^2 - 2a_0a_1\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right].$$

Since $\widehat{\theta}$ is a linear combination of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, both of which are normally distributed, it follows that

$$\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2_{\widehat{\theta}}).$$

INFERENCE: The variance $\sigma^2_{\widehat{\theta}}$ depends on the unknown parameter σ^2. An estimate of $\sigma^2_{\widehat{\theta}}$ is given by

$$\widehat{\sigma}^2_{\widehat{\theta}} = \widehat{\sigma}^2 \left[\frac{\frac{a_0^2}{n}\sum_{i=1}^n x_i^2 + a_1^2 - 2a_0a_1\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right],$$

where

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}.$$

Because $\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2_{\widehat{\theta}})$, we have by standardization,

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{N}(0,1).$$

Recall also that

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, it is independent of $\widehat{\theta}$ and hence Z and W are independent. Therefore,

$$t = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}_{\widehat{\theta}}} = \frac{(\widehat{\theta} - \theta)/\sigma_{\widehat{\theta}}}{\sqrt{\frac{(n-2)\widehat{\sigma}^2}{\sigma^2}/(n-2)}} \sim t(n-2).$$

Since t is a pivotal quantity, a $100(1-\alpha)$ percent confidence interval for θ is

$$\widehat{\theta} \pm t_{n-2,\alpha/2}\widehat{\sigma}_{\widehat{\theta}}.$$

In addition, tests of hypotheses concerning θ use the $t(n-2)$ distribution.

SPECIAL CASE: A special case of the preceding result is estimating the mean value of Y for a fixed value of x, say, x^*. In our simple linear regression model, we know that

$$E(Y|x^*) = \beta_0 + \beta_1 x^*,$$

which is just a linear combination of the form $\theta = a_0\beta_0 + a_1\beta_1$, where $a_0 = 1$ and $a_1 = x^*$. Therefore,

$$\widehat{\theta} \equiv \widehat{E(Y|x^*)} = \widehat{\beta}_0 + \widehat{\beta}_1 x^*$$

is an unbiased estimator of $\theta \equiv E(Y|x^*) = \beta_0 + \beta_1 x^*$ and its variance is

$$V(\widehat{\theta}) = \sigma^2_{\widehat{\theta}} = \sigma^2 \left[\frac{\frac{a_0^2}{n}\sum_{i=1}^n x_i^2 + a_1^2 - 2a_0a_1\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right].$$

Applying the preceding general results to this special case, a $100(1-\alpha)$ **percent confidence interval** for $E(Y|x^*) = \beta_0 + \beta_1 x^*$ is given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}.$$

NOTE: The confidence interval for $E(Y|x^*) = \beta_0 + \beta_1 x^*$ will be different for different values of x^*; see pp 597 (WMS). It is easy to see that the width of the confidence interval will be smallest when $x^* = \overline{x}$ and will increase as the distance between x^* and $\overline{x}$ increases. That is, more precise inference for $\theta = E(Y|x^*) = \beta_0 + \beta_1 x^*$ will result when x^* is close to $\overline{x}$. When x^* is far away from $\overline{x}$, our precision may not be adequate. It is sometimes desired to estimate $E(Y|x^*) = \beta_0 + \beta_1 x^*$ for a value of x^* outside the range of x values in the observed data. This is called **extrapolation**. In order for these inferences to be valid, we must believe that the model is accurate even for x values outside the range where we have observed data. In some situations, this may be reasonable; in others, we may have no basis for making such a claim without data to support it.

11.2.6 Prediction intervals for Y^*

PREDICTION: For some research questions, we may not be interested in the mean $E(Y|x^*) = \beta_0 + \beta_1 x^*$, but rather in the actual value of Y we may observe when $x = x^*$. On the surface, this may sound like the same problem, but they are very different.

EXAMPLE: Suppose that we have adopted the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where 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 those 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 the random variable Y, say Y^*, that comes from a distribution with mean $E(Y|x^* = 700)$. Estimating $E(Y|x^* = 700)$ is much easier than predicting Y^*.

GOAL: Our goal is to construct a **prediction interval** for a new value of Y, which we denote by Y^*. Our point predictor for Y^*, when $x = x^*$, is

$$\widehat{Y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

This point predictor is the same as 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 Y^*, not estimating $E(Y|x^*)$. We call $\widehat{Y}^*$ a **prediction** to make the distinction clear.

TERMINOLOGY: Define the random variable

$$U = Y^* - \widehat{Y}^*.$$

We call U the **prediction error**. Note that

$$\begin{aligned} E(U) = E(Y^* - \widehat{Y}^*) &= E(Y^*) - E(\widehat{Y}^*) \\ &= (\beta_0 + \beta_1 x^*) - E(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \\ &= (\beta_0 + \beta_1 x^*) - (\beta_0 + \beta_1 x^*) = 0. \end{aligned}$$

That is, the prediction error U is an unbiased estimator of 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 model assumptions, we know that $V(Y^*) = \sigma^2$. 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 value Y^* is a new value of Y, and, hence, is independent of $Y_1, Y_2, ..., Y_n$. Therefore,

$$\begin{aligned} V(U) &= V(Y^* - \widehat{Y}^*) = V(Y^*) + V(\widehat{Y}^*) - 2\text{Cov}(Y^*, \widehat{Y}^*) \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]. \end{aligned}$$

We finally note that 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. We have shown that

$$U = Y^* - \widehat{Y}^* \sim \mathcal{N}\left\{0,\ \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]\right\}.$$

Standardizing, we have

$$Z = \frac{Y^* - \widehat{Y}^*}{\sqrt{\sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} \sim \mathcal{N}(0,1).$$

Recall also that

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, it is independent of $\widehat{Y}^*$ and hence Z and W are independent. Therefore,

$$t = \frac{Z}{\sqrt{W/(n-2)}} = \frac{Y^* - \widehat{Y}^*}{\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}} \sim t(n-2).$$

Using t as a pivot, 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 $t_{n-2,\alpha/2}$ denotes the upper $\alpha/2$ quantile of the $t(n-2)$ distribution. Rearranging the event inside the probability symbol, we have

$$P\left(\widehat{Y}^* - t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]} < Y^* \right.$$
$$\left. < \widehat{Y}^* + t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}\right) = 1 - \alpha.$$

We call

$$\widehat{Y}^* \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}$$

is a $100(1-\alpha)$ **percent prediction interval** for Y^*.

NOTE: It is of interest to compare the confidence interval for $E(Y|x^*)$, given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]},$$

to the prediction interval for Y^*, given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}.$$

As we can see, 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.

11.2.7 Example

Example 11.1. A botanist is studying the absorption of salts by living plant cells. She prepares $n = 30$ dishes containing potato slices and adds a bromide solution to each dish. She waits a duration of time x (measured in hours) and then analyzes the potato slices for absorption of bromide ions (y, measured in mg/1000g). Here are the data.

Dish	x	y	Dish	x	y	Dish	x	y
1	16.4	5.2	11	65.5	15.3	21	121.6	23.0
2	18.2	1.0	12	68.6	11.2	22	121.8	22.3
3	21.6	4.8	13	75.4	16.9	23	122.4	24.6
4	22.3	2.7	14	76.3	12.3	24	124.4	22.4
5	24.1	1.1	15	88.0	15.3	25	128.0	28.1
6	29.7	3.5	16	92.0	19.9	26	128.0	20.5
7	34.6	8.7	17	96.6	21.1	27	131.2	26.5
8	35.2	10.1	18	98.1	19.5	28	140.7	31.3
9	56.5	11.4	19	103.9	20.7	29	145.8	29.1
10	58.7	10.8	20	115.9	22.4	30	149.5	32.6

Table 11.1: Botany data. Absorption of bromide ions (y, measured in mg/1000g) and time (x, measured in hours).

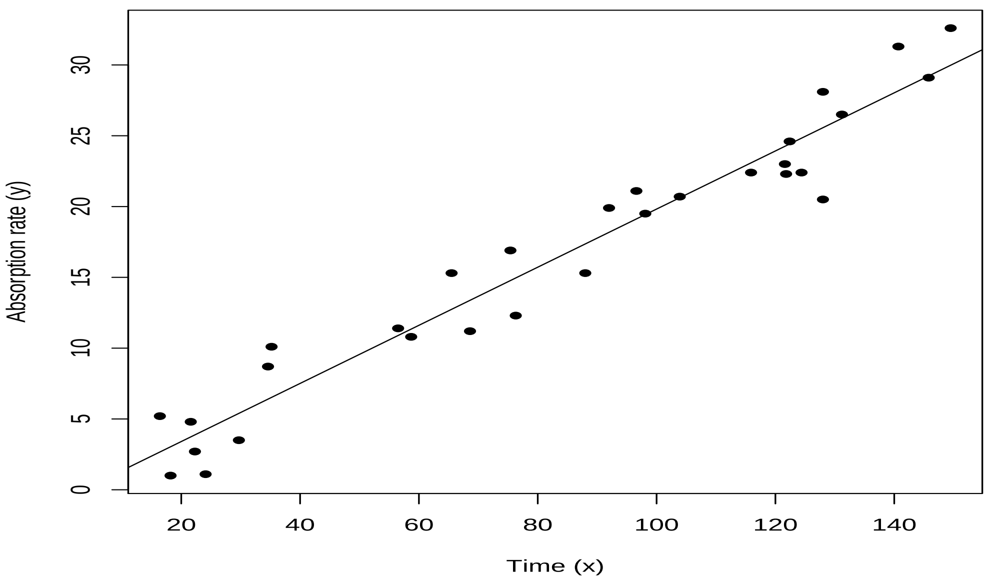

Figure 11.1: Botany data. Absorption of bromide ions (y, measured in mg/1000g) versus time (x, measured in hours). The least squares regression line has been superimposed.

REGRESSION MODEL: From the scatterplot in Figure 11.1, the linear regression model

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

for $i = 1, 2, ..., 30$, appears to be appropriate. Fitting this model in R, we get the output:

```
> summary(fit)
Call: lm(formula = absorp ~ time)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.700374   0.894462  -0.783     0.44
time         0.205222   0.009509  21.582   <2e-16 ***

Residual standard error: 2.236 on 28 degrees of freedom
Multiple R-squared: 0.9433,  Adjusted R-squared: 0.9413
F-statistic: 465.8 on 1 and 28 DF,  p-value: < 2.2e-16
```

OUTPUT: The `Estimate` output gives the least squares estimates $\widehat{\beta}_0 \approx -0.700$ and $\widehat{\beta}_1 \approx 0.205$. The equation of the least squares regression line is therefore

$$\widehat{Y} = -0.700 + 0.205x,$$

or, in other words,

$$\widehat{\texttt{ABSORPTION}} = -0.700 + 0.205\texttt{TIME}.$$

The `Std.Error` output gives

$$\begin{aligned} 0.894462 &= \widehat{\text{se}}(\widehat{\beta}_0) = \sqrt{c_{00}\widehat{\sigma}^2} \\ 0.009509 &= \widehat{\text{se}}(\widehat{\beta}_1) = \sqrt{c_{11}\widehat{\sigma}^2}, \end{aligned}$$

which are the estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively, where

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{30-2} = (2.236)^2 \approx 5.00$$

is the square of the `Residual standard error`. The `t value` output gives the t statistics

$$\begin{aligned} t = -0.783 &= \frac{\widehat{\beta}_0 - 0}{\sqrt{c_{00}\widehat{\sigma}^2}} \\ t = 21.582 &= \frac{\widehat{\beta}_1 - 0}{\sqrt{c_{11}\widehat{\sigma}^2}}, \end{aligned}$$

which test $H_0 : \beta_0 = 0$ versus $H_a : \beta_0 \neq 0$ and $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, respectively. Two-sided probability values are in `Pr(>|t|)`. We see that

- there is insufficient evidence against $H_0 : \beta_0 = 0$ (p-value = 0.44).
- there is strong evidence against $H_0 : \beta_1 = 0$ (p-value < 0.0001). This means that the absorption rate Y is (significantly) linearly related to duration time x.

CONFIDENCE INTERVALS: Ninety-five percent confidence intervals for β_0 and β_1 are

$$\begin{aligned} \widehat{\beta}_0 \pm t_{28,0.025}\widehat{\text{se}}(\widehat{\beta}_0) &\Longrightarrow -0.700 \pm 2.048(0.894) \Longrightarrow (-2.53, 1.13) \\ \widehat{\beta}_1 \pm t_{28,0.025}\widehat{\text{se}}(\widehat{\beta}_1) &\Longrightarrow 0.205 \pm 2.048(0.010) \Longrightarrow (0.18, 0.23). \end{aligned}$$

We are 95 percent confident that β_0 is between -2.53 and 1.13. We are 95 percent confident that β_1 is between 0.18 and 0.23.

PREDICTION: Suppose that we are interested estimating $E(Y|x)$ and predicting a new Y when $x^* = 80$ hours. We use R to compute the following:

```
> predict(fit,data.frame(time=80),level=0.95,interval="confidence")
     fit      lwr      upr
15.71735 14.87807 16.55663
> predict(fit,data.frame(time=80),level=0.95,interval="prediction")
     fit      lwr      upr
15.71735 11.06114 20.37355
```

- Note that

$$\widehat{E(Y|x^*)} = \widehat{Y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^* = -0.700 + 0.205(80) \approx 15.71735.$$

- A 95 percent **confidence interval** for $E(Y|x^* = 80)$ is $(14.88, 16.56)$. When the duration time is $x = 80$ hours, we are 95 percent confident that the mean absorption is between 14.88 and 16.56 mg/1000g.

- A 95 percent **prediction interval** for Y^*, when $x = 80$, is $(11.06, 20.37)$. When the duration time is $x = 80$ hours, we are 95 percent confident that the absorption for a new dish will be between 11.06 and 20.37 mg/1000g. □

11.3 Correlation

RECALL: In the simple linear regression model

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

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, it is assumed that the independent variable x is fixed. This assumption is plausible in designed experiments, say, where the investigator has control over which values of x will be included in the experiment. For example,

- x = dose of a drug, Y = change in blood pressure for a human subject

- x = concentration of toxic substance, Y = number of mutant offspring observed for a pregnant rat
- x = time, Y = absorption of bromide ions.

In other settings, it is unreasonable to think that the researcher can "decide" beforehand which values of x will be observed. Consider the following examples:

- X = weight, Y = height of a human subject
- X = average heights of plants in a plot, Y = yield
- X = STAT 513 HW score, Y = STAT 513 final exam score.

In each of these instances, the independent variable X is best regarded as **random**. It is unlikely that the researcher can control (fix) its value.

IMPORTANT: When both X and Y are best regarded as random, it is conventional to model the observed data as realizations of (X, Y), a bivariate random vector. A popular model for (X, Y) is the bivariate normal distribution.

RECALL: The random vector (X, Y) is said to have a **bivariate normal distribution** if its (joint) pdf is given by

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}e^{-Q/2}$$

for all $(x, y)' \in \mathcal{R}^2$, where

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right].$$

Under the bivariate normal model, recall from STAT 511 that

$$E(Y|X) = \beta_0 + \beta_1 X,$$

where

$$\begin{aligned}\beta_0 &= \mu_Y - \beta_1\mu_X \\ \beta_1 &= \rho\left(\frac{\sigma_Y}{\sigma_X}\right).\end{aligned}$$

IMPORTANT: Note that because

$$\beta_1 = \rho\left(\frac{\sigma_Y}{\sigma_X}\right),$$

the correlation ρ and the (population) slope parameter β_1 have the same sign.

ESTIMATION: Suppose that $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is an iid sample of size n from a bivariate normal distribution with marginal means μ_X and μ_Y, marginal variances σ_X^2 and σ_Y^2, and correlation ρ. The likelihood function is given by

$$\begin{aligned} L(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) &= \prod_{i=1}^{n} f_{X,Y}(x_i, y_i) \\ &= \left(\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\right)^n e^{-\sum_{i=1}^n Q_i/2}, \end{aligned}$$

where

$$Q_i = \frac{1}{1-\rho^2}\left[\left(\frac{x_i-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x_i-\mu_X}{\sigma_X}\right)\left(\frac{y_i-\mu_Y}{\sigma_Y}\right) + \left(\frac{y_i-\mu_Y}{\sigma_Y}\right)^2\right].$$

The maximum likelihood estimators are

$$\widehat{\mu}_X = \overline{X}, \qquad \widehat{\mu}_Y = \overline{Y}, \qquad \widehat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2, \qquad \widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2,$$

and

$$\widehat{\rho} = r = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

HYPOTHESIS TEST: In the bivariate normal model, suppose that it is desired to test

$$H_0: \rho = 0$$

versus

$$H_a: \rho \neq 0.$$

Since ρ and β_1 always have the same sign, mathematically, this is equivalent to testing

$$H_0: \beta_1 = 0$$

versus

$$H_a: \beta_1 \neq 0.$$

That is, we can use the statistic

$$t = \frac{\widehat{\beta}_1}{\sqrt{c_{11}\widehat{\sigma}^2}}$$

to test $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$. A level α rejection region is

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

One sided tests can be performed similarly.

RESULT: Simple calculations show that

$$t = \frac{\widehat{\beta}_1}{\sqrt{c_{11}\widehat{\sigma}^2}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Therefore, the test of $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$ (or any other suitable H_a) can be performed using only the calculated value of r.

REMARK: Even though the tests

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0$$

and

$$H_0 : \rho = 0$$

versus

$$H_a : \rho \neq 0$$

are carried out in the exact same manner, it is important to remember that the interpretation of the results is very different, depending on which test we are performing.

- In the first test, we are determining whether or not there is a **linear relationship** between Y and x. The independent variable x is best regarded as fixed.
- In the second test, we are actually determining whether or not the random variables X and Y are **independent**. Recall that in the bivariate normal model,

$$X \text{ and } Y \text{ independent} \iff \rho = 0.$$

REMARK: In some problems, it may be of interest to test

$$H_0 : \rho = \rho_0$$
$$\text{versus}$$
$$H_a : \rho \neq \rho_0$$

(or any other suitable H_a), where $\rho_0 \neq 0$. In this case, there is no equivalence between the two tests (as when $\rho_0 = 0$) that we saw before. We are forced to use a different test (i.e., one that is based on large sample theory).

ASYMPTOTIC RESULT: Suppose that $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is an iid sample of size n from a bivariate normal distribution with marginal means μ_X and μ_Y, marginal variances σ_X^2 and σ_Y^2, and correlation ρ. Let

$$r = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

denote the maximum likelihood estimator of ρ. For large n, the statistic

$$W = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \sim \mathcal{AN}\left[\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right].$$

IMPLEMENTATION: This asymptotic result above can be used to test

$$H_0 : \rho = \rho_0$$
$$\text{versus}$$
$$H_a : \rho \neq \rho_0$$

(or any other suitable H_a), where $\rho_0 \neq 0$. The test statistic is the standardized value of W, computed under H_0, that is,

$$Z = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{1/\sqrt{n-3}}.$$

An approximate level α rejection region is

$$\text{RR} = \{z : |z| > z_{\alpha/2}\},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. One sided tests can be performed similarly.

11.4 Multiple linear regression models

11.4.1 Introduction

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 basic 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 $p = k + 1$ regression parameters β_0, β_1, ..., β_k. These are unknown 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 contributes a response Y and a value of each of the independent variables $x_1, x_2, ..., x_k$.
- We continue to assume that $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$.
- We also assume that the independent variables $x_1, x_2, ..., x_k$ are fixed and measured without error. Therefore, Y is normally distributed with

$$\begin{aligned} E(Y|x_1, x_2, ..., x_k) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \\ V(Y|x_1, x_2, ..., x_k) &= \sigma^2. \end{aligned}$$

PREVIEW: To fit the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

we will still use the **method of least squares**. However, simple computing formulae for the least squares estimators of β_0, β_1, ..., β_k are no longer available (as they were in the simple linear regression model). It is advantageous to express multiple linear regression models in terms of matrices and vectors. This greatly streamlines notation and makes calculations tractable.

11.4.2 Matrix representation

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

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

With these definitions, the model above can be expressed equivalently as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

In this equivalent 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 regression parameters
- $\boldsymbol{\epsilon}$ is an $n \times 1$ (random) vector of unobserved errors.

LEAST SQUARES: The notion of least squares is the same as it was in the simple linear regression model. To fit a multiple linear regression model, we want to find the values of $\beta_0, \beta_1, ..., \beta_k$ that minimize

$$Q(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})]^2,$$

or, in matrix notation, the value of $\boldsymbol{\beta}$ that minimizes

$$Q = Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

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 $\beta_0, \beta_1, ..., \beta_k$ and set these equal to zero; i.e.,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial Q}{\partial \beta_0} \\ \frac{\partial Q}{\partial \beta_1} \\ \vdots \\ \frac{\partial Q}{\partial \beta_k} \end{pmatrix} \overset{\text{set}}{=} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

These are called the **normal equations**. Solving the normal equations for $\beta_0, \beta_1, ..., \beta_k$ gives the least squares estimators, which we denote by $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$.

NORMAL EQUATIONS: Using the calculus of matrices makes this much easier; in particular, the normal equations above can be expressed as

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}.$$

Provided that $\mathbf{X}'\mathbf{X}$ is full rank, the (unique) solution is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

This is the **least squares estimator** of $\boldsymbol{\beta}$. The fitted regression model is

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}},$$

or, equivalently,

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}.$$

NOTE: For the least squares estimator $\widehat{\boldsymbol{\beta}}$ to be unique, we need $\mathbf{X}$ to be of **full column rank**; i.e., $r(\mathbf{X}) = p = k + 1$. That is, there must be no linear dependencies among the columns of $\mathbf{X}$. If $r(\mathbf{X}) < p$, then $\mathbf{X}'\mathbf{X}$ does not have a unique inverse. In this case, the normal equations can not be solved uniquely. We will henceforth assume that $\mathbf{X}$ is of full column rank.

11.4.3 Random vectors: Important results

IMPORTANCE: Because multiple linear regression models are best presented in terms of (random) vectors and matrices, it is important to extend the notions of mean, variance, and covariance to random vectors. Doing so allows us to examine sampling distributions and the resulting inference that arises in multiple linear regression models.

TERMINOLOGY: Suppose that $Z_1, Z_2, ..., Z_n$ are random variables. We call

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}$$

a **random vector**. The multivariate probability density function (pdf) of $\mathbf{Z}$ is denoted by $f_{\mathbf{Z}}(\mathbf{z})$. The function $f_{\mathbf{Z}}(\mathbf{z})$ describes probabilistically how the random variables $Z_1, Z_2, ..., Z_n$ are jointly distributed.

- If $Z_1, Z_2, ..., Z_n$ are independent variables, then

  $$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} f_{Z_i}(z_i),$$

  where $f_{Z_i}(z_i)$ is the marginal pdf of Z_i.

- If $Z_1, Z_2, ..., Z_n$ are iid from a common marginal pdf, say, $f_Z(z)$, then

  $$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} f_Z(z_i).$$

TERMINOLOGY: Suppose $Z_1, Z_2, ..., Z_n$ are random variables with means $E(Z_i) = \mu_i$ and variances $V(Z_i) = \sigma_i^2$, for $i = 1, 2, ..., n$, and covariances $\text{Cov}(Z_i, Z_j) = \sigma_{ij}$ for $i \neq j$. The **mean** of a random vector $\mathbf{Z}$ is given by

$$E(\mathbf{Z}) = E\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} E(Z_1) \\ E(Z_2) \\ \vdots \\ E(Z_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}.$$

The **variance** of $\mathbf{Z}$ is

$$V(\mathbf{Z}) = V\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix} = \mathbf{V}.$$

- $\mathbf{V}$ is an $n \times n$ matrix. It is also called the **variance-covariance matrix** of $\mathbf{Z}$.
- $\mathbf{V}$ consists of the n variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ on the diagonal and the $2\binom{n}{2}$ covariance terms $\text{Cov}(Z_i, Z_j)$, for $i \neq j$, on the off-diagonal.
- Since $\text{Cov}(Z_i, Z_j) = \text{Cov}(Z_j, Z_i)$, $\mathbf{V}$ is **symmetric**; i.e., $\mathbf{V}' = \mathbf{V}$.

TERMINOLOGY: Suppose that

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}$$

are random vectors. The **covariance** between $\mathbf{Y}$ and $\mathbf{Z}$ is

$$\text{Cov}(\mathbf{Y}, \mathbf{Z}) = \begin{pmatrix} \text{Cov}(Y_1, Z_1) & \text{Cov}(Y_1, Z_2) & \cdots & \text{Cov}(Y_1, Z_m) \\ \text{Cov}(Y_2, Z_1) & \text{Cov}(Y_2, Z_2) & \cdots & \text{Cov}(Y_2, Z_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_n, Z_1) & \text{Cov}(Y_n, Z_2) & \cdots & \text{Cov}(Y_n, Z_m) \end{pmatrix}_{n\times m}.$$

RESULTS: Suppose $\mathbf{Z}$ is a random vector with mean $E(\mathbf{Z}) = \boldsymbol{\mu}$ and variance-covariance matrix $V(\mathbf{Z}) = \mathbf{V}$. Suppose $\mathbf{a}$ is a nonrandom (constant) vector and that $\mathbf{A}$ and $\mathbf{B}$ are nonrandom (constant) matrices.

1. $E(\mathbf{a}+\mathbf{BZ}) = \mathbf{a}+\mathbf{B}E(\mathbf{Z}) = \mathbf{a}+\mathbf{B}\boldsymbol{\mu}$
2. $V(\mathbf{a}+\mathbf{BZ}) = \mathbf{B}V(\mathbf{Z})\mathbf{B}' = \mathbf{BVB}'$
3. $\text{Cov}(\mathbf{AY},\mathbf{BZ}) = \mathbf{A}\text{Cov}(\mathbf{Y},\mathbf{Z})\mathbf{B}'$.

TERMINOLOGY: Let $\mathbf{Y}$ be an $n\times 1$ random vector with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{V}$. Let $\mathbf{A}$ be an $n \times n$ nonrandom matrix. We call $\mathbf{Y}'\mathbf{AY}$ a **quadratic form**. The mean of a quadratic form is

$$E(\mathbf{Y}'\mathbf{AY}) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{AV}),$$

where $\text{tr}(\cdot)$ means "trace," that is, $\text{tr}(\mathbf{AV})$ is the sum of the diagonal elements of $\mathbf{AV}$.

REMARK: It is important to see that a quadratic form $\mathbf{Y}'\mathbf{AY}$ is a scalar random variable. Therefore, its mean $E(\mathbf{Y}'\mathbf{AY})$ is a scalar constant. Quadratic forms are important in the theory of linear (regression) models. It turns out that **sums of squares** (which appear in analysis of variance tables) can always be written as quadratic forms.

11.4.4 Multivariate normal distribution

TERMINOLOGY: Suppose that $Z_1, Z_2, ..., Z_n$ are iid $\mathcal{N}(0,1)$ random variables. The joint pdf of $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)'$, for all $\mathbf{z} \in \mathcal{R}^n$, is given by

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} f_Z(z_i) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n z_i^2/2} = (2\pi)^{-n/2}\exp(-\mathbf{z}'\mathbf{z}/2). \end{aligned}$$

If $\mathbf{Z}$ has a pdf given by $f_{\mathbf{Z}}(\mathbf{z})$, we say that $\mathbf{Z}$ has a **standard multivariate normal distribution**; i.e., a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{I}$. Here,

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

That is, $\mathbf{0}$ is an $n \times 1$ zero vector and $\mathbf{I}$ is the $n \times n$ identity matrix. We write $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$. Note that

$$Z_1, Z_2, ..., Z_n \sim \text{iid } \mathcal{N}(0,1) \Longleftrightarrow \mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}).$$

TERMINOLOGY: The random vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$ is said to have a **multivariate normal distribution** with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{V}$ if its joint 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 $\mathbf{y} \in \mathcal{R}^n$. We write $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$.

FACTS:

- If $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)' \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$, then $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, for each $i = 1, 2, ..., n$.
- If $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{a}_{m \times 1}$ and $\mathbf{B}_{m \times n}$ are nonrandom, then

$$\mathbf{U} = \mathbf{a} + \mathbf{B}\mathbf{Y} \sim \mathcal{N}_m(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{V}\mathbf{B}').$$

APPLICATION: 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 $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$. Equivalently, we can write this model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I})$. 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

$$V(\mathbf{Y}) = V(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = V(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}.$$

Because $\mathbf{Y}$ is a linear combination of $\boldsymbol{\epsilon}$, which is normally distributed by assumption, it follows that

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).\ \Box$$

11.4.5 Estimating the error variance

REVIEW: Consider the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Recall that the least squares estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Our next task is to estimate the error variance σ^2.

TERMINOLOGY: We define the **error (residual) sum of squares** as

$$\begin{aligned} \text{SSE} &= (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= (\mathbf{Y} - \widehat{\mathbf{Y}})'(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{e}'\mathbf{e}. \end{aligned}$$

- The $n \times 1$ vector $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ contains the least squares **fitted values**.
- The $n \times 1$ vector $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ contains the least squares **residuals**.

TERMINOLOGY: Consider the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and define

$$\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

$\mathbf{M}$ is called the **hat matrix**. Many important quantities in linear regression can be written as functions of the hat matrix. For example, the vector of fitted values can be written as

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{M}\mathbf{Y}.$$

The vector of residuals can be written as

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{M}\mathbf{Y} = (\mathbf{I} - \mathbf{M})\mathbf{Y}.$$

The error (residual) sum of squares can be written as

$$\text{SSE} = (\mathbf{Y} - \widehat{\mathbf{Y}})'(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}.$$

Note that SSE $= \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}$ is a quadratic form.

FACTS: The matrix $\mathbf{M}$ possesses the following properties:

- $\mathbf{M}$ is symmetric, i.e., $\mathbf{M}' = \mathbf{M}$.
- $\mathbf{M}$ is idempotent, i.e., $\mathbf{M}^2 = \mathbf{M}$.
- $\mathbf{M}\mathbf{X} = \mathbf{X}$, i.e., $\mathbf{M}$ projects each column of $\mathbf{X}$ onto itself.

RESULT: Consider the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Let $p = k + 1$ denote the number of regression parameters in the model. The quantity

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n - p}$$

is an **unbiased estimator** of σ^2, that is, $E(\widehat{\sigma}^2) = \sigma^2$.

Proof. Recall that SSE $= \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}$. Because $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $V(\mathbf{Y}) = \sigma^2\mathbf{I}$, we have

$$\begin{aligned} E(\text{SSE}) &= E[\mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}] \\ &= (\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{M})\mathbf{X}\boldsymbol{\beta} + \text{tr}[(\mathbf{I} - \mathbf{M})\sigma^2\mathbf{I}]. \end{aligned}$$

The first term $(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{M})\mathbf{X}\boldsymbol{\beta} = 0$ because

$$(\mathbf{I} - \mathbf{M})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{M}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}.$$

Because the $\text{tr}(\cdot)$ function is linear,

$$\begin{aligned} \text{tr}[(\mathbf{I} - \mathbf{M})\sigma^2\mathbf{I}] &= \sigma^2[\text{tr}(\mathbf{I}) - \text{tr}(\mathbf{M})] \\ &= \sigma^2\{n - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\}. \end{aligned}$$

Since $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ for any 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

$$\begin{aligned} \sigma^2\{n - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\} &= \sigma^2\{n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]\} \\ &= \sigma^2[n - \text{tr}(\mathbf{I}_p)] = \sigma^2(n-p), \end{aligned}$$

since $\mathbf{I}_p = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$ is $p \times p$. We have shown that $E(\text{SSE}) = \sigma^2(n-p)$. Thus,

$$E(\widehat{\sigma}^2) = E\left(\frac{\text{SSE}}{n-p}\right) = \frac{\sigma^2(n-p)}{n-p} = \sigma^2,$$

showing that $\widehat{\sigma}^2$ is an unbiased estimator of σ^2. □

RESULT: Consider the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Let $p = k+1$ denote the number of regression parameters in the model. Under these model assumptions,

$$\frac{\text{SSE}}{\sigma^2} = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

The proof of this result is beyond the scope of this course.

11.4.6 Sampling distribution of $\widehat{\boldsymbol{\beta}}$

GOAL: 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 investigate the **sampling distribution** of the least squares estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

MEAN AND VARIANCE: The mean of $\widehat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}. \end{aligned}$$

This shows that $\widehat{\boldsymbol{\beta}}$ is an **unbiased estimator** of $\boldsymbol{\beta}$. The variance of $\widehat{\boldsymbol{\beta}}$ is

$$\begin{aligned} V(\widehat{\boldsymbol{\beta}}) &= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

NORMALITY: Since $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is a linear combination of $\mathbf{Y}$, which is (multivariate) normal under our model assumptions, it follows that $\widehat{\boldsymbol{\beta}}$ is normally distributed as well. Therefore, we have shown that

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]. \ \Box$$

IMPLICATIONS: The following results are direct consequences of our recent discussion:

1. $E(\widehat{\beta}_j) = \beta_j$, for $j = 0, 1, ..., k$; that is, the least squares estimators are unbiased.

2. $V(\widehat{\beta}_j) = c_{jj}\sigma^2$, for $j = 0, 1, ..., k$, where

$$c_{jj} = (\mathbf{X}'\mathbf{X})^{-1}_{jj}$$

is the corresponding jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. An estimate of $V(\widehat{\beta}_j)$ is

$$\widehat{V}(\widehat{\beta}_j) = c_{jj}\widehat{\sigma}^2 = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{jj},$$

where

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-p}.$$

3. $\text{Cov}(\widehat{\beta}_i, \widehat{\beta}_j) = c_{ij}\sigma^2$, where

$$c_{ij} = (\mathbf{X}'\mathbf{X})^{-1}_{ij}$$

is the corresponding ith row, jth column entry of $(\mathbf{X}'\mathbf{X})^{-1}$, for $i, j = 0, 1, ..., k$. An estimate of $\text{Cov}(\widehat{\beta}_i, \widehat{\beta}_j)$ is

$$\widehat{\text{Cov}}(\widehat{\beta}_i, \widehat{\beta}_j) = c_{ij}\widehat{\sigma}^2 = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{ij}.$$

4. Marginally, $\widehat{\beta}_j \sim \mathcal{N}(\beta_j, c_{jj}\sigma^2)$, for $j = 0, 1, ..., k$.

11.4.7 Inference for regression parameters

IMPORTANCE: Consider our 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 $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Confidence intervals and hypothesis tests for β_j can help us assess the importance of using the independent variable x_j in a model with the other independent variables. That is, inference regarding β_j is always **conditional** on the other variables being included in the model.

CONFIDENCE INTERVALS: Since $\widehat{\beta}_j \sim \mathcal{N}(\beta_j, c_{jj}\sigma^2)$, for $j = 0, 1, 2, ..., k$, it follows, from standardization, that

$$Z_j = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\sigma^2}} \sim \mathcal{N}(0, 1).$$

Recall also that

$$W = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

Because $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_j$, it follows that Z and W are also independent. Therefore,

$$t = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\widehat{\sigma}^2}} = \frac{(\widehat{\beta}_j - \beta_j)/\sqrt{c_{jj}\sigma^2}}{\sqrt{\frac{(n-p)\widehat{\sigma}^2}{\sigma^2}/(n-p)}} \sim t(n-p).$$

Because $t \sim t(n-p)$, t is a pivot and we can write

$$P\left(-t_{n-p,\alpha/2} < \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}\widehat{\sigma}^2}} < t_{n-p,\alpha/2}\right) = 1 - \alpha,$$

where $t_{n-p,\alpha/2}$ denotes the upper $\alpha/2$ quantile of the $t(n-p)$ distribution. Rearranging the event inside the probability symbol, we have

$$P\left(\widehat{\beta}_j - t_{n-p,\alpha/2}\sqrt{c_{jj}\widehat{\sigma}^2} < \beta_j < \widehat{\beta}_j + t_{n-p,\alpha/2}\sqrt{c_{jj}\widehat{\sigma}^2}\right) = 1 - \alpha.$$

This shows that

$$\widehat{\beta}_j \pm t_{n-p,\alpha/2}\sqrt{c_{jj}\widehat{\sigma}^2}.$$

is a $100(1-\alpha)$ **percent confidence interval** for β_j.

HYPOTHESIS TESTS: Suppose that we want to test

$$H_0 : \beta_j = \beta_{j,0}$$

versus

$$H_a : \beta_j \neq \beta_{j,0},$$

where $\beta_{j,0}$ is a fixed value (often, $\beta_{j,0} = 0$). We 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.

11.4.8 Confidence intervals for $E(Y|\mathbf{x}^*)$

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 learned how to obtain confidence intervals for the mean response $E(Y|x^*) = \beta_0 + \beta_1 x^*$. Extending this to multiple linear regression models is straightforward.

GOAL: 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 $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, or, equivalently,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Our goal is to construct confidence intervals for linear parametric functions of the form

$$\theta = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k = \mathbf{a}'\boldsymbol{\beta},$$

where

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

INFERENCE: A point estimator for $\theta = \mathbf{a}'\boldsymbol{\beta}$ is

$$\widehat{\theta} = \mathbf{a}'\widehat{\boldsymbol{\beta}},$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. It is easy to see that $\widehat{\theta}$ is an **unbiased estimator** for θ since

$$E(\widehat{\theta}) = E(\mathbf{a}'\widehat{\boldsymbol{\beta}}) = \mathbf{a}'E(\widehat{\boldsymbol{\beta}}) = \mathbf{a}'\boldsymbol{\beta} = \theta.$$

The variance of $\widehat{\theta}$ is given by

$$V(\widehat{\theta}) = V(\mathbf{a}'\widehat{\boldsymbol{\beta}}) = \mathbf{a}'V(\widehat{\boldsymbol{\beta}})\mathbf{a} = \mathbf{a}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a} = \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}.$$

Since $\widehat{\theta} = \mathbf{a}'\widehat{\boldsymbol{\beta}}$ is a linear combination of $\widehat{\boldsymbol{\beta}}$, which is normally distributed, $\widehat{\theta}$ is also normally distributed. Therefore, we have shown that

$$\widehat{\theta} \sim \mathcal{N}[\theta, \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}].$$

Standardizing, we have

$$Z = \frac{\widehat{\theta} - \theta}{\sqrt{\sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}} \sim \mathcal{N}(0, 1).$$

It also follows that

$$t = \frac{\widehat{\theta} - \theta}{\sqrt{\widehat{\sigma}^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}} \sim t(n-p),$$

where $p = k + 1$ and

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-p}.$$

Since t is a pivotal quantity, a $100(1-\alpha)$ percent confidence interval for $\theta = \mathbf{a}'\boldsymbol{\beta}$ is

$$\widehat{\theta} \pm t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}.$$

In addition, tests of hypotheses concerning θ use the $t(n-p)$ distribution.

SPECIAL CASE: A special case of the preceding result is estimating the mean value of Y for a fixed value of $\mathbf{x} = (x_1, x_2, ..., x_k)'$, say,

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_k^* \end{pmatrix}.$$

In our multiple linear regression model, we know that

$$E(Y|\mathbf{x}^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_k x_k^*,$$

which is just a linear combination of the form $\theta = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k = \mathbf{a}'\boldsymbol{\beta}$, where

$$\mathbf{a} = \begin{pmatrix} 1 \\ x_1^* \\ \vdots \\ x_k^* \end{pmatrix}.$$

Therefore,

$$\widehat{\theta} \equiv \widehat{E(Y|\mathbf{x}^*)} = \widehat{\beta}_0 + \widehat{\beta}_1 x^* + \widehat{\beta}_2 x_2^* + \cdots + \widehat{\beta}_k x_k^* = \mathbf{a}'\widehat{\boldsymbol{\beta}}$$

is an unbiased estimator of $\theta = E(Y|\mathbf{x}^*)$, and its variance is

$$V(\widehat{\theta}) = \sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a},$$

where $\mathbf{a}$ is as given above. Applying the preceding general results to this special case, a $100(1-\alpha)$ **percent confidence interval** for $E(Y|\mathbf{x}^*)$, the mean of Y when $\mathbf{x} = \mathbf{x}^*$, is

$$\widehat{\theta} \pm t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}.$$

11.4.9 Prediction intervals for Y^*

RECALL: In the simple linear regression model, we learned how to obtain prediction intervals for a new response Y^*. Extending this to multiple linear regression models is straightforward.

GOAL: 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 $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, or, equivalently,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose that we would like to predict the value of a new response Y^*, for a fixed value of $\mathbf{x} = (x_1, x_2, ..., x_k)'$, say,

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_k^* \end{pmatrix}.$$

Our point predictor for Y^*, based on the least squares fit, is

$$\widehat{Y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \widehat{\beta}_2 x_2^* + \cdots + \widehat{\beta}_k x_k^* = \mathbf{a}'\widehat{\boldsymbol{\beta}},$$

where $\mathbf{a} = (1, x_1^*, x_2^*, ..., x_k^*)'$ and $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Define the error in prediction by $U = Y^* - \widehat{Y}^*$. Analogously to the simple linear regression case,

$$U = Y^* - \widehat{Y}^* \sim \mathcal{N}\{0, \sigma^2[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]\}.$$

Using the fact that $(n-p)\widehat{\sigma}^2/\sigma^2 \sim \chi^2(n-p)$, it follows that

$$t = \frac{Y^* - \widehat{Y}^*}{\sqrt{\widehat{\sigma}^2\,[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]}} \sim t(n-p).$$

Therefore,

$$\widehat{Y}^* \pm t_{n-p,\alpha/2}\sqrt{\widehat{\sigma}^2\,[1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}]},$$

is a $100(1-\alpha)$ **percent prediction interval** for Y^*.

REMARK: Comparing the prediction interval for Y^* to the analogous $100(1-\alpha)$ percent confidence interval for $E(Y|\mathbf{x}^*)$, we see that the intervals are again identical except the prediction interval has an extra "1" in the estimated standard error. This results from the extra variability that arises when predicting Y^* as opposed to estimating $E(Y|\mathbf{x}^*)$.

11.4.10 Example

Example 11.2. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study from the LaTrobe Valley of Victoria, Australia, samples of cheddar cheese were analyzed for their chemical composition and were subjected to taste tests. For each specimen, the taste Y was obtained by combining the scores from several tasters. Data were collected on the following variables:

$$\begin{aligned} Y &= \text{ taste score (TASTE)} \\ x_1 &= \text{ concentration of acetic acid (ACETIC)} \\ x_2 &= \text{ concentration of hydrogen sulfide (H2S)} \\ x_3 &= \text{ concentration of lactic acid (LACTIC).} \end{aligned}$$

Variables `ACETIC` and `H2S` were both measured on the log scale. The variable `LACTIC` has not been transformed. Table 11.2 contains concentrations of the various chemicals in $n = 30$ specimens of cheddar cheese and the observed taste score.

Specimen	TASTE	ACETIC	H2S	LACTIC	Specimen	TASTE	ACETIC	H2S	LACTIC
1	12.3	4.543	3.135	0.86	16	40.9	6.365	9.588	1.74
2	20.9	5.159	5.043	1.53	17	15.9	4.787	3.912	1.16
3	39.0	5.366	5.438	1.57	18	6.4	5.412	4.700	1.49
4	47.9	5.759	7.496	1.81	19	18.0	5.247	6.174	1.63
5	5.6	4.663	3.807	0.99	20	38.9	5.438	9.064	1.99
6	25.9	5.697	7.601	1.09	21	14.0	4.564	4.949	1.15
7	37.3	5.892	8.726	1.29	22	15.2	5.298	5.220	1.33
8	21.9	6.078	7.966	1.78	23	32.0	5.455	9.242	1.44
9	18.1	4.898	3.850	1.29	24	56.7	5.855	10.20	2.01
10	21.0	5.242	4.174	1.58	25	16.8	5.366	3.664	1.31
11	34.9	5.740	6.142	1.68	26	11.6	6.043	3.219	1.46
12	57.2	6.446	7.908	1.90	27	26.5	6.458	6.962	1.72
13	0.7	4.477	2.996	1.06	28	0.7	5.328	3.912	1.25
14	25.9	5.236	4.942	1.30	29	13.4	5.802	6.685	1.08
15	54.9	6.151	6.752	1.52	30	5.5	6.176	4.787	1.25

Table 11.2: Cheese data. `ACETIC`, `H2S`, and `LACTIC` are independent variables. The response variable is `TASTE`.

REGRESSION MODEL: Suppose the researchers postulate that each of the three chemical composition variables x_1, x_2, and x_3 is important in describing the taste. In this case, they might initially consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. We now use R to fit this model using the method of least squares. Here is the output:

```
> summary(fit)
Call: lm(formula = taste ~ acetic + h2s + lactic)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -28.877     19.735  -1.463  0.15540
acetic         0.328      4.460   0.074  0.94193
h2s            3.912      1.248   3.133  0.00425 **
lactic        19.670      8.629   2.279  0.03109 *

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518,     Adjusted R-squared: 0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.810e-06
```

OUTPUT: The `Estimate` output gives the values of the least squares estimates:

$$\widehat{\beta}_0 \approx -28.877 \qquad \widehat{\beta}_1 \approx 0.328 \qquad \widehat{\beta}_2 \approx 3.912 \qquad \widehat{\beta}_3 \approx 19.670.$$

Therefore, the fitted least squares regression model is

$$\widehat{Y} = -28.877 + 0.328x_1 + 3.912x_2 + 19.670x_3,$$

or, in other words,

$$\widehat{\texttt{TASTE}} = -28.877 + 0.328\texttt{ACETIC} + 3.912\texttt{H2S} + 19.670\texttt{LACTIC}.$$

The `Std.Error` output gives

$$\begin{aligned}
19.735 &= \widehat{\text{se}}(\widehat{\beta}_0) = \sqrt{c_{00}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{00}} \\
4.460 &= \widehat{\text{se}}(\widehat{\beta}_1) = \sqrt{c_{11}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{11}} \\
1.248 &= \widehat{\text{se}}(\widehat{\beta}_2) = \sqrt{c_{22}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{22}} \\
8.629 &= \widehat{\text{se}}(\widehat{\beta}_3) = \sqrt{c_{33}\widehat{\sigma}^2} = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{33}},
\end{aligned}$$

where

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{30-4} = (10.13)^2 \approx 102.63$$

is the square of the `Residual standard error`. The `t value` output gives the t statistics

$$\begin{aligned}
t = -1.463 &= \frac{\widehat{\beta}_0 - 0}{\sqrt{c_{00}\widehat{\sigma}^2}} \\
t = 0.074 &= \frac{\widehat{\beta}_1 - 0}{\sqrt{c_{11}\widehat{\sigma}^2}} \\
t = 3.133 &= \frac{\widehat{\beta}_2 - 0}{\sqrt{c_{22}\widehat{\sigma}^2}} \\
t = 2.279 &= \frac{\widehat{\beta}_3 - 0}{\sqrt{c_{33}\widehat{\sigma}^2}}.
\end{aligned}$$

These t statistics can be used to test $H_0 : \beta_i = 0$ versus $H_0 : \beta_i \neq 0$, for $i = 0, 1, 2, 3$. Two-sided probability values are in `Pr(>|t|)`. At the $\alpha = 0.05$ level,

- we do not reject $H_0 : \beta_0 = 0$ (p-value = 0.155). **Interpretation:** In the model which includes all three independent variables, the intercept term β_0 is not statistically different from zero.

- we do not reject $H_0 : \beta_1 = 0$ (p-value = 0.942). **Interpretation:** `ACETIC` does not significantly add to a model that includes `H2S` and `LACTIC`.

- we reject $H_0 : \beta_2 = 0$ (p-value = 0.004). **Interpretation:** `H2S` does significantly add to a model that includes `ACETIC` and `LACTIC`.

- we reject $H_0 : \beta_3 = 0$ (p-value = 0.031). **Interpretation:** `LACTIC` does significantly add to a model that includes `ACETIC` and `H2S`.

CONFIDENCE INTERVALS: Ninety-five percent confidence intervals for the regression parameters β_0, β_1, β_2, and β_3, respectively, are

$$
\begin{aligned}
\widehat{\beta}_0 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_0) &\implies -28.877 \pm 2.056(19.735) \implies (-69.45, 11.70) \\
\widehat{\beta}_1 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_1) &\implies 0.328 \pm 2.056(4.460) \implies (-8.84, 9.50) \\
\widehat{\beta}_2 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_2) &\implies 3.912 \pm 2.056(1.248) \implies (1.35, 6.48) \\
\widehat{\beta}_3 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_3) &\implies 19.670 \pm 2.056(8.629) \implies (1.93, 37.41).
\end{aligned}
$$

PREDICTION: Suppose that we are interested estimating $E(Y|\mathbf{x}^*)$ and predicting a new Y when `ACETIC` $= 5.5$, `H2S` $= 6.0$, and `LACTIC` $= 1.4$, so that

$$\mathbf{x}^* = \begin{pmatrix} 5.5 \\ 6.0 \\ 1.4 \end{pmatrix}.$$

We use R to compute the following:

```
> predict(fit,data.frame(acetic=5.5,h2s=6.0,lactic=1.4),level=0.95,interval="confidence")
     fit      lwr     upr
23.93552 20.04506 27.82597
> predict(fit,data.frame(acetic=5.5,h2s=6.0,lactic=1.4),level=0.95,interval="prediction")
     fit      lwr      upr
23.93552 2.751379 45.11966
```

- Note that

$$
\begin{aligned}
\widehat{E(Y|\mathbf{x}^*)} = \widehat{Y}^* &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \widehat{\beta}_2 x_2^* + \widehat{\beta}_3 x_3^* \\
&= -28.877 + 0.328(5.5) + 3.912(6.0) + 19.670(1.4) \approx 23.936.
\end{aligned}
$$

- A 95 percent **confidence interval** for $E(Y|\mathbf{x}^*)$ is $(20.05, 27.83)$. When `ACETIC` $=$ 5.5, `H2S` $= 6.0$, and `LACTIC` $= 1.4$, we are 95 percent confident that the mean taste rating is between 20.05 and 27.83.

- A 95 percent **prediction interval** for Y^*, when $\mathbf{x} = \mathbf{x}^*$, is $(2.75, 45.12)$. When `ACETIC` $= 5.5$, `H2S` $= 6.0$, and `LACTIC` $= 1.4$, we are 95 percent confident that the taste rating for a new cheese specimen will be between 2.75 and 45.12.

11.5 The analysis of variance for linear regression

IMPORTANCE: The fit of a linear regression model (simple or linear) can be summarized in an **analysis of variance (ANOVA)** table. An ANOVA table provides a partition of the variability in the observed data. This partition, in turn, allows us to assess the overall fit of the model.

MODEL: Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, and let $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ denote the hat matrix. Recall that $\widehat{\mathbf{Y}} = \mathbf{MY}$ and $\mathbf{e} = (\mathbf{I} - \mathbf{M})\mathbf{Y}$ denote the vectors of least squares fitted values and residuals, respectively.

SUMS OF SQUARES: Start with the simple quadratic form $\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{IY}$. Note that

$$\begin{aligned}
\mathbf{Y}'\mathbf{Y} &= \mathbf{Y}'(\mathbf{M} + \mathbf{I} - \mathbf{M})\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{MY} + \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{MMY} + \mathbf{Y}'(\mathbf{I} - \mathbf{M})(\mathbf{I} - \mathbf{M})\mathbf{Y} \\
&= \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} + \mathbf{e}'\mathbf{e}.
\end{aligned}$$

This equation can be expressed equivalently as

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

TERMINOLOGY: We call

- $\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} Y_i^2$ the **uncorrected total** sum of squares
- $\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} = \sum_{i=1}^{n} \widehat{Y}_i^2$ the **uncorrected regression (model)** sum of squares
- $\mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$ the **error (residual)** sum of squares.

CORRECTED VERSIONS: When we fit a linear regression model, we are often interested in the regression coefficients that are attached to independent variables; i.e.,

$\beta_1, \beta_2, ..., \beta_k$. We generally are not interested in the intercept term β_0, the overall mean of Y (ignoring the independent variables). Therefore, it is common to "remove" the effects of fitting the intercept term β_0. This removal is accomplished by subtracting $n\overline{Y}^2$ from both sides of the last equation. This gives

$$\sum_{i=1}^n Y_i^2 - n\overline{Y}^2 = \sum_{i=1}^n \widehat{Y}_i^2 - n\overline{Y}^2 + \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2,$$

or, equivalently,

$$\underbrace{\sum_{i=1}^n (Y_i - \overline{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}_{\text{SSE}}.$$

We call

- SST the **corrected total** sum of squares
- SSR the **corrected regression (model)** sum of squares
- SSE the **error (residual)** sum of squares.

QUADRATIC FORMS: To enhance our understanding of the partitioning of sums of squares, we express the SST = SSR + SSE partition in terms of quadratic forms. The basic **uncorrected** partition is given by

$$\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{M}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}.$$

To write the corrected partition, we subtract $n\overline{Y}^2 = \mathbf{Y}'n^{-1}\mathbf{J}\mathbf{Y}$ from both sides of the last equation, 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. This gives

$$\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'n^{-1}\mathbf{J}\mathbf{Y} = \mathbf{Y}'\mathbf{M}\mathbf{Y} - \mathbf{Y}'n^{-1}\mathbf{J}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}$$

or, equivalently,

$$\underbrace{\mathbf{Y}'(\mathbf{I} - n^{-1}\mathbf{J})\mathbf{Y}}_{\text{SST}} = \underbrace{\mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y}}_{\text{SSR}} + \underbrace{\mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}}_{\text{SSE}}.$$

ANOVA TABLE: The general form of an ANOVA table for linear regression (simple or multiple) is given below:

Source	df	SS	MS	F
Regression	$p-1$	SSR	$\text{MSR} = \frac{\text{SSR}}{p-1}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	$n-p$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-p}$	
Total	$n-1$	SST		

NOTES:

- The corrected partition SSR + SSE = SST appears in the column labeled "SS" (**sum of squares**).
- The column labeled "df" gives the **degrees of freedom** for each quadratic form. Mathematically,
$$\begin{aligned} p-1 &= r(\mathbf{M}-n^{-1}\mathbf{J}) \\ n-p &= r(\mathbf{I}-\mathbf{M}) \\ n-1 &= r(\mathbf{I}-n^{-1}\mathbf{J}). \end{aligned}$$
That is, the degrees of freedom are the ranks of the quadratic form matrices in
$$\mathbf{Y}'(\mathbf{I}-n^{-1}\mathbf{J})\mathbf{Y} = \mathbf{Y}'(\mathbf{M}-n^{-1}\mathbf{J})\mathbf{Y} + \mathbf{Y}'(\mathbf{I}-\mathbf{M})\mathbf{Y}.$$
Note also that the degrees of freedom add down (as the SS do).
- The column labeled "MS" contains the **mean squares**
$$\begin{aligned} \text{MSR} &= \frac{\text{SSR}}{p-1} \\ \text{MSE} &= \frac{\text{SSE}}{n-p}. \end{aligned}$$
That is, the mean squares are the SS divided by the corresponding degrees of freedom. Note that
$$\widehat{\sigma}^2 = \text{MSE} = \frac{\text{SSE}}{n-p}$$
is our unbiased estimator of the error variance σ^2 in the underlying model.
- The ANOVA table F statistic will be discussed next.

F STATISTIC: The F statistic in the ANOVA table is used to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

versus

$$H_a : \text{at least one of the } \beta_j \text{ is nonzero.}$$

In other words, F tests whether or not at least one of the independent variables $x_1, x_2, ..., x_k$ is important in describing the response Y. If H_0 is rejected, we do not know which one or how many of the β_j's are nonzero; only that at least one is. In this light, one could argue that this test is not all that meaningful.

JUSTIFICATION: When H_0 is true,

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2(p-1), \quad \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-p),$$

and SSR and SSE are independent. These facts would be proven in a more advanced course. Therefore, when H_0 is true,

$$F = \frac{\frac{\text{SSR}/\sigma^2}{p-1}}{\frac{\text{SSE}/\sigma^2}{n-p}} = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} = \frac{\text{MSR}}{\text{MSE}} \sim F(p-1, n-p).$$

The test above uses a one-sided, upper tail rejection region. Specifically, a level α rejection region is

$$\text{RR} = \{F : F > F_{p-1,n-p,\alpha}\},$$

where $F_{p-1,n-p,\alpha}$ denotes 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.

TERMINOLOGY: Since

$$\text{SST} = \text{SSR} + \text{SSE},$$

the proportion of the total variation in the data explained by the model is

$$R^2 = \frac{\text{SSR}}{\text{SST}}.$$

The statistic R^2 is called the **coefficient of determination**. The larger the R^2, the more variation that is being explained by the regression model.

Example 11.2 (continued). In Example 11.2, we fit the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. The ANOVA table, obtained using SAS, is shown below.

```
Analysis of Variance
                                 Sum of           Mean
 Source             DF          Squares         Square    F Value    Pr > F
 Model               3       4994.50861     1664.83620      16.22    <.0001
 Error              26       2668.37806      102.62993
 Corrected Total    29       7662.88667
```

The F statistic is used to test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

versus

$$H_a : \text{at least one of the } \beta_j \text{ is nonzero.}$$

ANALYSIS: Based on the F statistic ($F = 16.22$), and the corresponding probability value (p-value < 0.0001), we conclude that at least one of `ACETIC`, `H2S`, and `LACTIC` is important in describing taste (that is, we reject H_0). The coefficient of determination is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{4994.51}{7662.89} \approx 0.652.$$

That is, about 65.2 percent of the variability in the taste data is explained by the independent variables. If we analyze these data using R, we get the following:

```
anova.fit<-anova(lm(taste˜acetic+h2s+lactic))
anova.fit
Response: taste
           Df  Sum Sq  Mean Sq  F value     Pr(>F)
acetic      1 2314.14  2314.14  22.5484  6.528e-05 ***
h2s         1 2147.11  2147.11  20.9209  0.0001035 ***
lactic      1  533.26   533.26   5.1959  0.0310870 *
Residuals  26 2668.38   102.63
```

NOTE: The convention used by R is to "split up" the (corrected) regression sum of squares

$$\text{SSR} = 4994.50861$$

into sums of squares for each of the three independent variables `ACETIC`, `H2S`, and `LACTIC`, as they are added sequentially to the model (these are called **sequential sums of squares**). The sequential sums of squares for the independent variables add to the SSR (up to rounding error) for the model, that is,

$$\begin{aligned}\text{SSR} = 4994.51 &= 2314.14 + 2147.11 + 533.26 \\ &= \text{SS}(\texttt{ACETIC}) + \text{SS}(\texttt{H2S}) + \text{SS}(\texttt{LACTIC}).\end{aligned}$$

In words,

- SS(`ACETIC`) is the sum of squares added when compared to a model that includes only an intercept term.
- SS(`H2S`) is the sum of squares added when compared to a model that includes an intercept term and `ACETIC`.
- SS(`LACTIC`) is the sum of squares added when compared to a model that includes an intercept term, `ACETIC`, and `H2S`.

11.6 Reduced versus full model testing

SETTING: Consider the (full) multiple 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 $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, or, equivalently,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. We now consider the question of whether or not a smaller model is adequate for the data. That is, can we remove some of the independent variables and write a smaller model that does just as well at describing the data as the full model?

REMARK: Besides their ease of interpretation, smaller models confer statistical benefits. Remember that for each additional independent variable we add to the model, there is an associated regression parameter that has to be estimated. For each additional regression parameter that we have to estimate, we lose a degree of freedom for error. Remember that MSE, our estimator for the error variance σ^2 uses the degrees of freedom for error in its computation. Thus, the fewer error degrees of freedom we have, the less precise estimate we have of σ^2. With an imprecise estimate of σ^2, hypothesis tests, confidence intervals, and prediction intervals are less informative.

TERMINOLOGY: We call

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_g x_{ig} + \beta_{g+1} x_{i(g+1)} + \cdots + \beta_k x_{ik} + \epsilon_i$$

the **full model** because it includes all of the independent variables $x_1, x_2, ..., x_k$. We call

$$Y_i = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_g x_{ig} + \epsilon_i$$

a **reduced model** because it includes only the independent variables $x_1, x_2, ..., x_g$, where $g < k$, that is, independent variables $x_{g+1}, x_{g+2}, ..., x_k$ are not included in the reduced model.

MATRIX NOTATION: In matrix notation, the full model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1g} & x_{1(g+1)} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2g} & x_{2(g+1)} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{ng} & x_{n(g+1)} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_g \\ \beta_{g+1} \\ \vdots \\ \beta_k \end{pmatrix}.$$

In matrix notation, the reduced model is

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where

$$\mathbf{X}_0 = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1g} \\ 1 & x_{21} & x_{22} & \cdots & x_{2g} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{ng} \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_g \end{pmatrix}.$$

That is, the matrix $\mathbf{X}_0$ is simply $\mathbf{X}$ with the last $(k-g)$ columns removed.

TESTING PROBLEM: In order to determine whether or not the extra independent variables $x_{g+1}, x_{g+2}, ..., x_k$ should be included in the regression, we are interested in testing the reduced model versus the full model, that is,

$$H_0 : \mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

$$\text{versus}$$

$$H_a : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

In terms of the regression parameters in the full model, we are essentially testing

$$H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$

$$\text{versus}$$

$$H_a : \text{not } H_0.$$

INTUITION: Define the hat matrices for the reduced and full models by $\mathbf{M}_0 = \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'$ and $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, respectively. We know that

$$\begin{aligned} \text{SSR}_F &= \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y} \\ \text{SSR}_R &= \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y} \end{aligned}$$

are the (corrected) regression sum of squares for the full and reduced models, respectively. Since the regression sum of squares SSR can never decrease by adding independent variables, it follows that

$$\text{SSR}_F = \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y} \geq \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y} = \text{SSR}_R.$$

In the light of this, our intuition should suggest the following:

- If $\text{SSR}_F = \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y}$ and $\text{SSR}_R = \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y}$ are "close," then the additional independent variables $x_{g+1}, x_{g+2}, ..., x_k$ do not add too much to the regression, and the reduced model is adequate at describing the data.

- if $\text{SSR}_F = \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y}$ and $\text{SSR}_R = \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y}$ are not "close," then the additional independent variables $x_{g+1}, x_{g+2}, ..., x_k$ add a significant amount to the regression. This suggests that the reduced model does an insufficient job of describing the data when compared to the full model.

- We therefore make our decision by examining the size of

  $$\text{SSR}_F - \text{SSR}_R = \mathbf{Y}'(\mathbf{M} - n^{-1}\mathbf{J})\mathbf{Y} - \mathbf{Y}'(\mathbf{M}_0 - n^{-1}\mathbf{J})\mathbf{Y} = \mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}.$$

  If this difference is "large," then the reduced model does not do a good job of describing the data (when compared to the full model).

- We are assuming that the full model already does a good job of describing the data; we are trying to find a smaller model that does just as well.

TEST STATISTIC: Theoretical arguments in linear models show that when the reduced model is correct,

$$F = \frac{\mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}/(k-g)}{\text{MSE}_F} \sim F(k-g, n-p),$$

where $p = k + 1$ and MSE_F is the mean squared error computed from the full model. Therefore, a level α rejection region for testing

$$H_0 : \mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

versus

$$H_a : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

is given by

$$\text{RR} = \{F : F > F_{k-g,n-p,\alpha}\},$$

where $F_{k-g,n-p,\alpha}$ is the upper α quantile of the $F(k-g, n-p)$ distribution.

Example 11.2 (continued). In Example 11.2, consider the full model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

Suppose we believe that a simple linear regression model with `ACETIC` (x_1) only does just as well as the full model at describing `TASTE`. In this case, the reduced model is

$$Y_i = \gamma_0 + \gamma_1 x_{i1} + \epsilon_i.$$

IMPLEMENTATION: To test the reduced model versus the full model, we first compute the ANOVA tables from both model fits. The ANOVA table from the full model fit (using SAS) is

```
Analysis of Variance: Full Model
                                Sum of           Mean
 Source             DF         Squares         Square    F Value    Pr > F
 Model               3      4994.50861     1664.83620      16.22    <.0001
 Error              26      2668.37806      102.62993
 Corrected Total    29      7662.88667
```

The ANOVA table from the reduced model fit (using SAS) is

```
Analysis of Variance: Reduced Model
                                Sum of           Mean
 Source             DF         Squares         Square    F Value    Pr > F
 Model               1      2314.14151     2314.14151      12.11    0.0017
 Error              28      5348.74515      191.02661
 Corrected Total    29      7662.88667
```

Therefore, the difference in the (corrected) regression sum of squares is

$$\begin{aligned}\mathbf{Y}'(\mathbf{M}-\mathbf{M}_0)\mathbf{Y} &= \text{SSR}_F - \text{SSR}_R \\ &= 4994.50861 - 2314.14151 = 2680.367\end{aligned}$$

and the test statistic is

$$F = \frac{\mathbf{Y}'(\mathbf{M}-\mathbf{M}_0)\mathbf{Y}/(k-g)}{\text{MSE}_F} = \frac{2680.367/(3-1)}{102.62993} \approx 13.058.$$

A level $\alpha = 0.05$ rejection region is

$$\text{RR} = \{F : F > F_{2,26,0.05} = 3.369\}.$$

I used the R command `qf(0.95,2,26)` to compute $F_{2,26,0.05}$. Because the test statistic F falls in the rejection region, we reject H_0 at the $\alpha = 0.05$ level. We conclude that the reduced model does not do as well as the full model in describing `TASTE`. The probability value for the test is

$$\text{p-value} = P(F_{2,26} > 13.058) \approx 0.0001,$$

computed using the `1-pf(13.058,2,26)` in R.

IMPORTANT: It is interesting to note that the sum of squares

$$\begin{aligned} 2680.367 &= \mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y} \\ &= \text{SS}(\texttt{H2S}) + \text{SS}(\texttt{LACTIC}) = 2147.11 + 533.26. \end{aligned}$$

That is, we can obtain $\mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}$ by adding the sequential sum of squares corresponding to the independent variables not in the reduced model.

REMARK: It is possible to implement this test completely in R. Here is the output:

```
> fit.full<-lm(taste~acetic+h2s+lactic)
> fit.reduced<-lm(taste~acetic)
> anova(fit.reduced,fit.full,test="F")
Model 1: taste ~ acetic
Model 2: taste ~ acetic + h2s + lactic
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     28 5348.7
2     26 2668.4  2    2680.4 13.058 0.0001186 ***
```

ANALYSIS: R's convention is to produce the F statistic

$$F = \frac{\mathbf{Y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}/(k-g)}{\text{MSE}_F} = \frac{2680.367/(3-1)}{102.62993} \approx 13.058$$

automatically with the corresponding p-value in `Pr(>F)`.

12 An Introduction to Bayesian Inference

Complementary reading: Chapter 16 (WMS).

12.1 Introduction

THE BIG PICTURE: Statistical inference deals with drawing conclusions, after observing numerical data, about quantities that are not observed (e.g., model parameters, etc.). For example, if $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, we may be interested in using the data $\boldsymbol{Y} = \boldsymbol{y}$ to make statements about the values of μ and σ^2, the parameters which describe the distribution (i.e., the population) from which the data $\boldsymbol{y}$ are taken. We can make these statements using point estimates, confidence interval estimates, or by performing hypothesis tests that are pertinent to the problem at hand.

CLASSICAL APPROACH: Up until now, in your exposure to statistics, you have most likely been taught exclusively the **classical** or **frequentist** approach to inference; that is, you have been taught to regard the model parameter θ (scalar or vector-valued) as a fixed, but unknown value and to use the data $\boldsymbol{Y} = \boldsymbol{y}$ to make some statement about θ. This classical approach can be summarized as follows:

1. Treat the parameter θ as a fixed (but unknown) quantity.
2. Assume that $Y_1, Y_2, ..., Y_n$ is a sample (perhaps an iid sample) from the probability distribution $f_Y(y; \theta)$, where $\theta \in \Omega$.
3. Observe the data $\boldsymbol{Y} = \boldsymbol{y}$.
4. Draw inference about θ based on the observed data $\boldsymbol{y}$.

Example 12.1. In a public-health study, researchers would like to learn about the prevalence of HIV in Houston, TX, among heterosexual male intravenous drug users (IVDUs) not receiving treatment for their addiction. In this study, the goal is to estimate θ, the (unknown) proportion of HIV positives in this population. A sample of n individuals will be obtained from the population and the positive/negative statuses of the individuals $Y_1, Y_2, ..., Y_n$ will be modeled as iid Bernoulli(θ) observations, where $0 < \theta < 1$. Under the classical approach, the prevalence θ is regarded as fixed (but unknown), and the data $\boldsymbol{Y} = \boldsymbol{y}$ are used to draw inference about θ.

BAYESIAN APPROACH: Instead of treating model parameters as fixed quantities and modeling only the data $\boldsymbol{Y}$, the Bayesian sets up a full probability model; that is, a **joint probability distribution** for the data $\boldsymbol{Y}$ and the model parameters in θ.

- The model for the θ should be consistent with our prior knowledge of the underlying scientific problem and of the data collection process. For example, in Example 12.1, what prior knowledge might we have about θ, the probability of HIV infection?
- In the Bayesian approach, unobserved model parameters are not treated as fixed quantities; rather, they themselves are modeled as random quantities which vary according to a probability distribution; this is called the **prior distribution**.
- The prior distribution reflects (or models) our prior beliefs about θ. The Bayesian approach allows us to then incorporate this knowledge into the inferential procedure. We now describe the mathematics of how this is done.

12.2 Bayesian posteriors

IMPORTANT: One primary goal of any Bayesian analysis is to obtain the **posterior distribution** for θ. The posterior distribution combines our a priori knowledge about θ and information in the observed data $\boldsymbol{Y}$. We now present a general algorithm on how to find the posterior distribution in any problem.

1. Start by choosing a **prior distribution** for θ, say, $\theta \sim g(\theta)$. This distribution reflects our a priori knowledge regarding θ. We will discuss methods for choosing $g(\theta)$ in due course.

2. Construct the **conditional distribution** $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)$. For example, if $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y_i;\theta)$, the distribution of $\boldsymbol{Y}$, conditional on θ, is given by

$$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} f_Y(y_i;\theta).$$

   Note that this is simply $L(\theta|\boldsymbol{y})$, the likelihood function of θ.

3. Find the **joint distribution** of $\boldsymbol{Y}$ and θ; this is

$$f_{\boldsymbol{Y},\theta}(\boldsymbol{y},\theta) = f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)g(\theta).$$

   This follows from the definition of a conditional distribution in STAT 511 (remembering that θ is regarded as a random variable).

4. Compute $m_{\boldsymbol{Y}}(\boldsymbol{y})$, the **marginal distribution** of the data $\boldsymbol{Y}$; this is given by

$$m_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_{\theta} f_{\boldsymbol{Y},\theta}(\boldsymbol{y},\theta)d\theta.$$

5. The **posterior distribution** is the conditional distribution of θ, given $\boldsymbol{Y} = \boldsymbol{y}$. Again, from the definition of conditional distributions, the posterior is

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

   Under the Bayesian framework, all inference regarding θ (e.g., estimation, testing, etc.) is conducted using the posterior distribution $g(\theta|\boldsymbol{y})$.

REMARK: The Bayesian approach allows the researcher to incorporate prior information about θ. Clearly, in many problems, this would be desirable. For example, if we are talking about HIV infection in Houston, we know that, at least, the prevalence θ should not be large. In fact, a wide variety of estimates of HIV prevalence have appeared in the literature, ranging up to about 3 million infected in the United States (this is about one percent nationwide). If Houston and this male IVDU cohort "follows the pattern" of this nationwide estimate, taking a Bayesian approach affords the researcher the flexibility to incorporate this prior information into the analysis. On the other hand, the classical approach does not allow one to exploit this prior information.

Example 12.2. In our Houston HIV example (see Example 12.1), suppose that we model the positive/negative statuses $Y_1, Y_2, ..., Y_n$, conditional on θ, as iid Bernoulli(θ) observations, where $0 < \theta < 1$. Since we are considering HIV prevalence, we know θ is likely small, so we decide to model θ as a realization from a beta(α, β) distribution, where $\alpha < \beta$; this is the prior distribution. This prior distribution is reasonable since

- the support of a beta random variable is $R = (0, 1)$ which coincides with the Bernoulli(θ) parameter space; i.e., $\Omega = \{\theta : 0 < \theta < 1\}$.
- the beta(α, β) family is very flexible; that is, the pdf can assume many different shapes by changing α and β; furthermore, taking $\alpha < \beta$ provides a pdf that is concentrated closer to $\theta = 0$ than to $\theta = 1$.
- the beta(α, β) distribution turns out to be a **conjugate prior** for the Bernoulli likelihood (we'll explain this later).

Following the previously-mentioned steps, we now derive $g(\theta|\boldsymbol{y})$, the posterior distribution of θ.

1. The **prior** distribution for θ is $\theta \sim$ beta(α, β); thus, for $0 < \theta < 1$,
$$g(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$
2. The **conditional** distribution of the data $\boldsymbol{Y}$, given θ, is, for $y_i = 0, 1$,
$$\begin{aligned} f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} f_Y(y_i;\theta) &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} = \theta^u(1-\theta)^{n-u}, \end{aligned}$$
where the sufficient statistic $u = \sum_{i=1}^n y_i$.
3. The **joint** distribution of $\boldsymbol{Y}$ and θ, for values of $y_i = 0, 1$ and $0 < \theta < 1$, is
$$\begin{aligned} f_{\boldsymbol{Y},\theta}(\boldsymbol{y},\theta) &= f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)g(\theta) \\ &= \theta^u(1-\theta)^{n-u} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{u+\alpha-1}(1-\theta)^{n+\beta-u-1}. \end{aligned}$$

4. The **marginal** distribution of the data $\boldsymbol{Y}$, for $y_i = 0, 1$, is

$$\begin{aligned} m_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_\theta f_{\boldsymbol{Y},\theta}(\boldsymbol{y},\theta)d\theta &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{u+\alpha-1}(1-\theta)^{n+\beta-u-1}d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \underbrace{\theta^{u+\alpha-1}(1-\theta)^{n+\beta-u-1}}_{\text{beta}(u+\alpha,n+\beta-u)\text{ kernel}}d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(u+\alpha)\Gamma(n+\beta-u)}{\Gamma(n+\alpha+\beta)}, \end{aligned}$$

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

5. The **posterior** distribution $g(\theta|\boldsymbol{y})$ is, for $0 < \theta < 1$, given by

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

Note that $g(\theta|\boldsymbol{y})$ is the beta$(u+\alpha, n+\beta-u)$ pdf; that is, the posterior distribution of θ, given the data $\boldsymbol{y}$, is beta$(u+\alpha, n+\beta-u)$, where $u = \sum_{i=1}^n y_i$.

UPDATE: We started by modeling the unknown prevalence θ as a beta(α, β) random variable, we then observed the data $\boldsymbol{y}$ from the study, and we finally updated our prior beliefs, based on the data $\boldsymbol{y}$, to arrive at the posterior distribution of θ. Schematically,

$$\underbrace{\theta \sim \text{beta}(\alpha,\beta)}_{\text{prior distribution}} \implies \text{Observe data } \boldsymbol{y} \implies \underbrace{\theta \sim \text{beta}(u+\alpha, n+\beta-u)}_{\text{posterior distribution}}.$$

We see that the posterior distribution depends on (a) the prior distribution through the values of α and β, and on (b) the data $\boldsymbol{y}$ through the sufficient statistic $u = \sum_{i=1}^n y_i$. It is also interesting to note that both the prior and posterior distributions are members of the beta family (this is due to conjugacy).

ILLUSTRATION: In our Houston HIV example, suppose that $n = 100$; that is, we observe 100 IVDU subjects, and that our prior distribution is $\theta \sim \text{beta}(1, 19)$; that is, a beta distribution with $\alpha = 1$ and $\beta = 19$. This may be a reasonable prior distribution since the prior mean $E(\theta) = 0.05$, which is "small," consistent (at least) with our prior

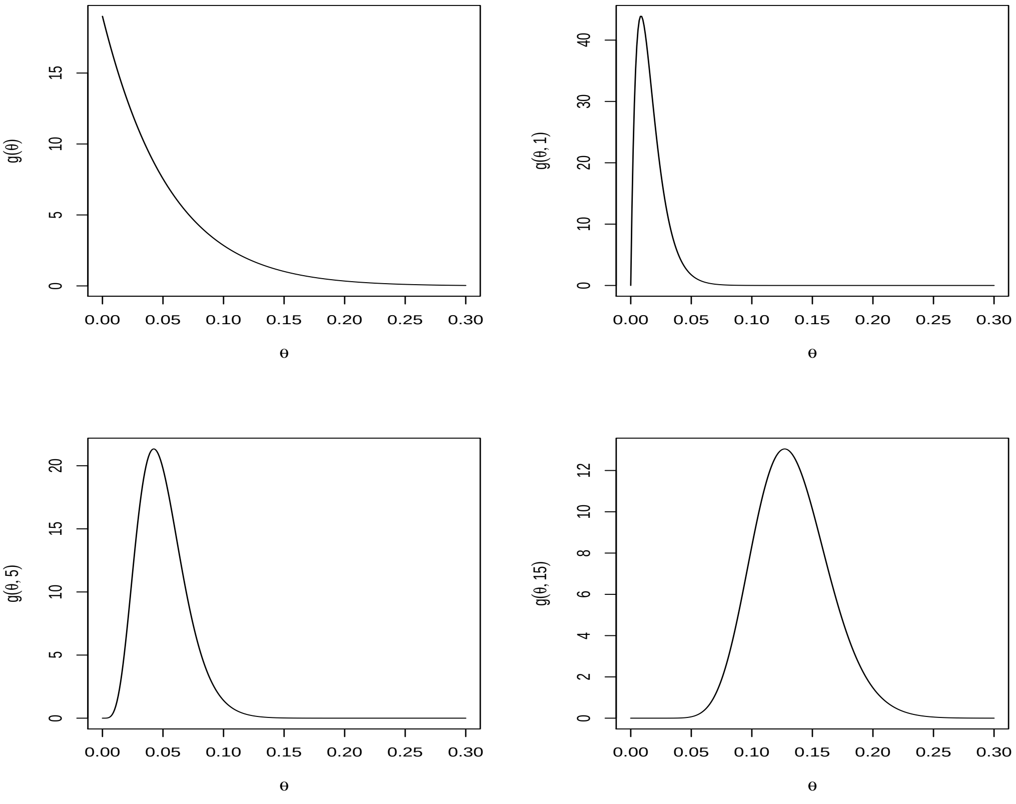

Figure 12.1: Binomial-beta Bayesian prior and posteriors in Example 12.2. Upper left: $\theta \sim \text{beta}(1, 19)$, prior; Upper right: Posterior distribution of θ when $u = 1$, $\text{beta}(2, 118)$; Lower left: Posterior distribution of θ when $u = 5$, $\text{beta}(6, 114)$; Lower right: Posterior distribution of θ when $u = 15$, $\text{beta}(16, 104)$. The sufficient statistic is $u = \sum_{i=1}^{100} y_i$.

belief that θ is likely not large. In Figure 12.1, we depict this prior distribution of θ (upper left) and posterior distributions based on three different values of $u = \sum_{i=1}^{100} y_i$. Consider the following table, which describes three possible realizations of this study (that is, three different values of u).

Prior, $g(\theta)$	Observed data	Posterior, $g(\theta\|\boldsymbol{y})$
$\text{beta}(1, 19)$	$u = 1$	$\text{beta}(2, 118)$
$\text{beta}(1, 19)$	$u = 5$	$\text{beta}(6, 114)$
$\text{beta}(1, 19)$	$u = 15$	$\text{beta}(16, 104)$

NOTE: Figure 12.1 illustrates the effect that the observed data $\boldsymbol{y}$ (through the sufficient statistic) can have on the posterior distribution $g(\theta|\boldsymbol{y})$. For $u = 1$, the posterior is left-shifted from the prior. For $u = 5$, the posterior remains located in a similar position as the prior, although the variability has been reduced. For $u = 15$, the posterior is notably right-shifted from the prior. □

Example 12.3. An animal biologist is interested in modeling the number of rat pups per mother, Y, for *Rattus rattus*, commonly known as the "black rat." Suppose that $Y_1, Y_2, ..., Y_n$ denote litter sizes for a sample of n rat mothers and assume that $Y_1, Y_2, ..., Y_n$, conditional on θ, is an iid sample from a Poisson distribution with mean θ. In turn, θ is modeled as a gamma(α, β) random variable. A Bayesian approach is taken to exploit the information from previous rat studies; to be specific, it is known that the mean number of pups per litter is around 5-7, but can be as high as 20. A gamma prior distribution is reasonable since

- the support of a gamma random variable is $R = (0, \infty)$ which coincides with the Poisson(θ) parameter space; i.e., $\Omega = \{\theta : \theta > 0\}$.
- the gamma(α, β) family is very flexible; that is, the pdf can assume many different shapes by changing α and β; right skewed is consistent with prior knowledge.
- the gamma(α, β) distribution is a **conjugate prior** for the Poisson likelihood.

We derive $g(\theta|\boldsymbol{y})$, the posterior distribution of θ, following the steps mentioned previously.

1. The **prior** distribution for θ is $\theta \sim$ gamma(α, β); thus, for $\theta > 0$,
$$g(\theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\theta^{\alpha-1}e^{-\theta/\beta}.$$

2. The **conditional** distribution of the data $\boldsymbol{Y}$, given θ, is, for $y_i = 0, 1, ...,$
$$\begin{aligned} f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta) = \prod_{i=1}^n f_Y(y_i;\theta) &= \prod_{i=1}^n \frac{\theta^{y_i}e^{-\theta}}{y_i!} \\ &= \frac{\theta^{\sum_{i=1}^n y_i}e^{-n\theta}}{\prod_{i=1}^n y_i!} = \frac{\theta^u e^{-n\theta}}{\prod_{i=1}^n y_i!}, \end{aligned}$$
where the sufficient statistic $u = \sum_{i=1}^n y_i$.

3. The **joint** distribution of $\boldsymbol{Y}$ and θ, for values of $y_i = 0, 1, ...,$ and $\theta > 0$, is

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

4. The **marginal** distribution of the data $\boldsymbol{Y}$, for $y_i = 0, 1, ...,$ is

$$\begin{aligned} m_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_\theta f_{\boldsymbol{Y},\theta}(\boldsymbol{y},\theta)d\theta &= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha \prod_{i=1}^n y_i!}\theta^{u+\alpha-1}e^{-\theta/(n+1/\beta)^{-1}}d\theta \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha \prod_{i=1}^n y_i!}\int_0^\infty \underbrace{\theta^{u+\alpha-1}e^{-\theta/(n+1/\beta)^{-1}}}_{\text{gamma}[u+\alpha,(n+1/\beta)^{-1}]\text{ kernel}} d\theta \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha \prod_{i=1}^n y_i!}\Gamma(u+\alpha)\left(\frac{1}{n+1/\beta}\right)^{u+\alpha}, \end{aligned}$$

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

5. The **posterior** distribution $g(\theta|\boldsymbol{y})$ is, for $\theta > 0$, given by

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

Note that $g(\theta|\boldsymbol{y})$ is the gamma$[u+\alpha,(n+1/\beta)^{-1}]$ pdf; that is, the posterior distribution of θ, given the data $\boldsymbol{y}$, is gamma$[u+\alpha,(n+1/\beta)^{-1}]$, where $u = \sum_{i=1}^n y_i$.

UPDATE: We started by modeling the unknown mean θ as a gamma(α,β) random variable, we then observed the data $\boldsymbol{y}$ from the study, and we finally updated our prior beliefs, based on the data $\boldsymbol{y}$, to arrive at the posterior distribution of θ. Schematically,

$$\underbrace{\theta \sim \text{gamma}(\alpha,\beta)}_{\text{prior distribution}} \Longrightarrow \text{Observe data } \boldsymbol{y} \Longrightarrow \underbrace{\theta \sim \text{gamma}[u+\alpha,(n+1/\beta)^{-1}]}_{\text{posterior distribution}}.$$

We see that the posterior distribution depends on (a) the prior distribution through the values of α and β, and on (b) the data $\boldsymbol{y}$ through the sufficient statistic $u = \sum_{i=1}^n y_i$. It is also interesting to note that both the prior and posterior distributions are members of the gamma family (due to conjugacy).

ILLUSTRATION: In our rat pup example, suppose that $n = 10$; that is, we observe the litter sizes of 10 rat mothers, and that our prior distribution is $\theta \sim \text{gamma}(2,3)$; that is, a gamma distribution with $\alpha = 2$ and $\beta = 3$. These choices of α and β provide a prior mean $E(\theta) = 6$, which is consistent with our prior knowledge. In Figure 12.2, we depict this prior distribution (upper left) and posterior distributions based on three different values of $u = \sum_{i=1}^{10} y_i$. Consider the following table, which describes three possible realizations of this study (that is, three different values of u).

Prior, $g(\theta)$	Observed data	Posterior, $g(\theta\vert\boldsymbol{y})$
$\text{gamma}(2,3)$	$u = 32$	$\text{gamma}(34, 0.0968)$
$\text{gamma}(2,3)$	$u = 57$	$\text{gamma}(59, 0.0968)$
$\text{gamma}(2,3)$	$u = 90$	$\text{gamma}(92, 0.0968)$

NOTE: Figure 12.2 illustrates the effect that the observed data $\boldsymbol{y}$ (through the sufficient statistic u) can have on the posterior distribution $g(\theta|\boldsymbol{y})$. Similarly to Example 12.2, we see the posterior distributions are much less variable than the prior, with central locations depending heavily on the observed data $\boldsymbol{y}$. □

REMARK: We have presented a 5-step algorithm to construct the posterior distribution of θ, given the data $\boldsymbol{Y} = \boldsymbol{y}$. It turns out that Step 4, the step that deals with deriving the marginal distribution $m_{\boldsymbol{Y}}(\boldsymbol{y})$, is not really needed. In addition, when a sufficient statistic $U = U(\boldsymbol{Y})$ is available, the posterior calculation becomes even easier. Starting with Step 3, we have the joint distribution of $\boldsymbol{Y}$ and θ, given by

$$f_{\boldsymbol{Y},\theta}(\boldsymbol{y},\theta) = f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)g(\theta).$$

Suppose that $U = U(\boldsymbol{Y})$ is a sufficient statistic for θ. By the Factorization Theorem, we know that the likelihood function $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)$ can be written as

$$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta) = k_1(u,\theta)k_2(\boldsymbol{y}),$$

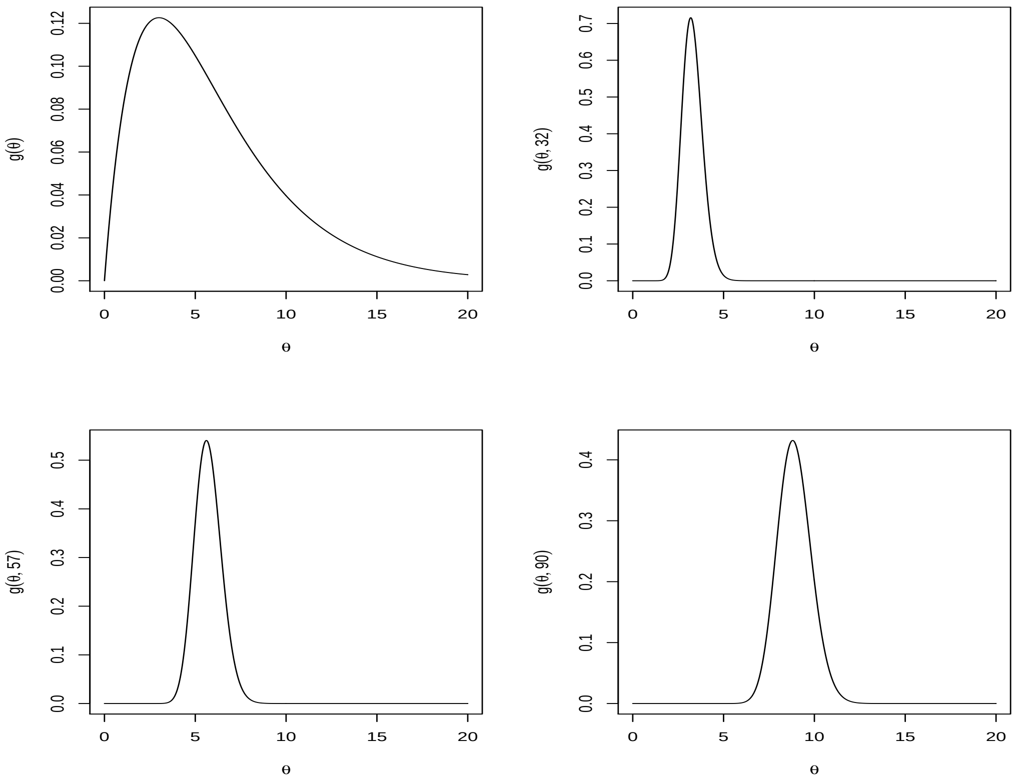

Figure 12.2: Poisson-gamma Bayesian prior and posteriors in Example 12.3. Upper left: $\theta \sim \text{gamma}(2, 3)$, prior; Upper right: Posterior distribution of θ when $u = 32$, $\text{gamma}(34, 0.0968)$; Lower left: Posterior distribution of θ when $u = 57$, $\text{gamma}(59, 0.0968)$; Lower right: Posterior distribution of θ when $u = 90$, $\text{gamma}(92, 0.0968)$. The sufficient statistic is $u = \sum_{i=1}^{10} y_i$.

where k_1 and k_2 are both nonnegative functions; k_1 depends on θ and the sufficient statistic u, and k_2 is free of θ. Therefore, the joint distribution of $\boldsymbol{Y}$ and θ can be written as

$$f_{\boldsymbol{Y},\theta}(\boldsymbol{y}, \theta) = k_1(u, \theta)g(\theta)k_2(\boldsymbol{y}).$$

Therefore, the posterior distribution satisfies

$$g(\theta|\boldsymbol{y}) = \frac{f_{\boldsymbol{Y},\theta}(\boldsymbol{y}, \theta)}{m_{\boldsymbol{Y}}(\boldsymbol{y})} = \frac{k_1(u, \theta)g(\theta)k_2(\boldsymbol{y})}{m_{\boldsymbol{Y}}(\boldsymbol{y})} \propto k_1(u, \theta)g(\theta).$$

This result should convince us of two important facts.

- The posterior distribution $g(\theta|\boldsymbol{y})$ is always a function of the sufficient statistic U.
- Because the posterior distribution $g(\theta|\boldsymbol{y})$ is a bona fide density function (remember, it is regarded as a function of θ), the factor $k_2(\boldsymbol{y})/m_{\boldsymbol{Y}}(\boldsymbol{y})$, which is free of θ, is simply the "right constant" that makes $g(\theta|\boldsymbol{y})$ integrate to 1.

In the light of these two findings, we can present a "shortcut" algorithm to construct the posterior distribution $g(\theta|\boldsymbol{y})$ when a sufficient statistic U exists. Note that there is no harm in denoting the posterior distribution by $g(\theta|u)$ since it must depend on u.

1. Start by choosing a **prior** distribution for θ, say, $\theta \sim g(\theta)$. This step is unchanged from before.
2. Find the **conditional** distribution of the sufficient statistic U, given θ; denote this distribution by $f_{U|\theta}(u|\theta)$. This step should be simple if you remember the distribution of sufficient statistics (you can quickly derive this distribution otherwise).
3. Write the **joint** distribution of U and θ; this is
$$f_{U,\theta}(u,\theta) = f_{U|\theta}(u|\theta)g(\theta).$$
4. The **posterior** distribution $g(\theta|u)$ is proportional to the joint distribution $f_{U,\theta}(u,\theta)$, that is,
$$g(\theta|u) \propto f_{U,\theta}(u,\theta) = f_{U|\theta}(u|\theta)g(\theta).$$
Therefore, all you have to do is examine $f_{U|\theta}(u|\theta)g(\theta)$ and classify the part of this function that depends on θ as the kernel of a well-known distribution (e.g., beta, gamma, normal, etc.). Because the posterior $g(\theta|u)$ is proportional to this kernel, the posterior distribution must match the distribution identified by the kernel.

Example 12.4. We now illustrate this "shortcut" posterior construction method using (a) the binomial-beta model in Example 12.2 and (b) the Poisson-gamma model in Example 12.3.

- In Example 12.2, conditional of θ, $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli distribution with mean θ. The sufficient statistic is

  $$U = \sum_{i=1}^{n} Y_i \sim b(n, \theta)$$

  so that

  $$f_{U|\theta}(u|\theta) = \binom{n}{u} \theta^u (1-\theta)^{n-u},$$

  for $u = 0, 1, ..., n$. In turn, θ follows a beta(α, β) prior distribution. Therefore,

  $$\begin{aligned} g(\theta|u) \propto f_{U|\theta}(u|\theta)g(\theta) &= \binom{n}{u} \theta^u (1-\theta)^{n-u} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \binom{n}{u} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\theta^{u+\alpha-1}(1-\theta)^{n+\beta-u-1}}_{\text{beta}(u+\alpha, n+\beta-u) \text{ kernel}}. \end{aligned}$$

  We can immediately deduce that the posterior distribution of θ, given the data (through the sufficient statistic u), is beta with parameters $u + \alpha$ and $n + \beta - u$. This was our same finding in Example 12.2.

- In Example 12.3, conditional of θ, $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson distribution with mean θ. The sufficient statistic is

  $$U = \sum_{i=1}^{n} Y_i \sim \text{Poisson}(n\theta)$$

  so that

  $$f_{U|\theta}(u|\theta) = \frac{(n\theta)^u e^{-n\theta}}{u!},$$

  for $u = 0, 1, ...,$. In turn, θ follows a gamma(α, β) prior distribution. Therefore,

  $$\begin{aligned} g(\theta|u) \propto f_{U|\theta}(u|\theta)g(\theta) &= \frac{(n\theta)^u e^{-n\theta}}{u!} \times \frac{1}{\Gamma(\alpha)\beta^\alpha} \theta^{\alpha-1} e^{-\theta/\beta} \\ &= \frac{n^u}{u!\Gamma(\alpha)\beta^\alpha} \underbrace{\theta^{u+\alpha-1} e^{-\theta/(n+1/\beta)^{-1}}}_{\text{gamma}[u+\alpha, (n+1/\beta)^{-1}] \text{ kernel}}. \end{aligned}$$

  We can immediately deduce that the posterior distribution of θ, given the data (through the sufficient statistic u), is gamma with parameters $u+\alpha$ and $(n+1/\beta)^{-1}$. This was our same finding in Example 12.3.

12.3 Prior model selection

DISCUSSION: In Example 12.3, recall that we made the following assumptions:

- $Y_1, Y_2, ..., Y_n$ are iid Poisson(θ).
- The prior distribution for θ is $\theta \sim \text{gamma}(\alpha, \beta)$.
- Recall also that the posterior distribution of θ, given the data $\boldsymbol{y}$, is also gamma (but with "updated" shape and scale parameters).

Suppose that, in Example 12.3, we instead took θ to have a $\mathcal{N}(\mu_0, \sigma_0^2)$ prior distribution, where both μ_0 and σ_0^2 are known. For this choice, it is easy to show that the joint distribution of $\boldsymbol{Y}$ and θ is, for values of $y_i = 0, 1, ...,$ and for $\theta \in \mathcal{R}$, given by

$$f_{\boldsymbol{Y},\theta}(\boldsymbol{y}, \theta) = \frac{\theta^u e^{-[n\theta + (\theta-\mu_0)^2/2\sigma_0^2]}}{\sqrt{2\pi}\sigma_0 \prod_{i=1}^n y_i!},$$

where the sufficient statistic $u = \sum_{i=1}^n y_i$. The marginal distribution of the data $\boldsymbol{Y}$ is

$$m_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_\theta f_{\boldsymbol{Y},\theta}(\boldsymbol{y}, \theta) d\theta = \int_{-\infty}^{\infty} \frac{\theta^u e^{-[n\theta + (\theta-\mu_0)^2/2\sigma_0^2]}}{\sqrt{2\pi}\sigma_0 \prod_{i=1}^n y_i!} d\theta.$$

Unfortunately, this marginal distribution does not exist in closed form (we can't get a closed form antiderivative of the integrand above). Therefore, the posterior distribution will not exist in closed form either.

QUESTION: In Example 12.3, why is it that when $\theta \sim \text{gamma}(\alpha, \beta)$, the posterior $g(\theta|\boldsymbol{y})$ exists in closed form (and is a gamma pdf), but when $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, it does not?

12.3.1 Conjugate priors

TERMINOLOGY: Let $\mathcal{F} = \{f_Y(y; \theta); \theta \in \Omega\}$ 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 $g(\theta|\boldsymbol{y}) \in \mathcal{G}$, for all $f_Y(y; \theta) \in \mathcal{F}$ and for all priors $g(\theta) \in \mathcal{G}$.

Table 12.1: Some common conjugate families.

Family	Parameter	Conjugate family	Prior hyperparameters
binomial(n, p)	p	beta(α, β)	α, β
Poisson(λ)	λ	gamma(α, β)	α, β
$\mathcal{N}(\theta, \sigma_0^2)$	θ	$\mathcal{N}(\mu, \tau^2)$	μ, τ^2
$\mathcal{N}(\mu_0, \sigma^2)$	σ^2	Inverse gamma(α, β)	α, β
exponential$(1/\theta)$	θ	gamma(α, β)	α, β
multinomial$(n, \boldsymbol{p})$	$\boldsymbol{p}$	Dirichlet$(\alpha_1, \alpha_2, ..., \alpha_{k+1})$	$\alpha_1, \alpha_2, ..., \alpha_{k+1}$

TERMINOLOGY: The parameters that index the prior distribution are called **hyperparameters**. For example, the beta prior has two hyperparameters, α and β. The Bayesian approach we have outlined so far (often called the "classical Bayesian approach") requires that the researcher specifies the values of all hyperparameters. There are more advanced Bayesian approaches that do not require prior model hyperparameter selection.

CONJUGACY: The basic justification for the use of conjugate prior distributions is that they simplify the computations and that one can write out a closed-form expression for the posterior distribution $g(\theta|\boldsymbol{y})$. Thus, in single-parameter problems, conjugate priors are chosen often for convenience. However, in multiple-parameter problems, conjugate priors may not exist. Conceptually, there is nothing to prevent one from using a nonconjugate prior in the general Bayesian approach. In this case, although it may not be possible to write out a closed-form expression for $g(\theta|\boldsymbol{y})$, it is generally possible to approximate it numerically using Bayesian simulation techniques.

12.3.2 Noninformative priors

TERMINOLOGY: When there is a general lack of a priori knowledge about the parameters of interest, prior models can be difficult to choose. It might also be desired for the prior distribution $g(\theta)$ to play a minimal role in determining the posterior distribution

$g(\theta|\boldsymbol{y})$, and, hence, the resulting inference. Such distributions are called **noninformative priors**; they are also referred to as "vague," "diffuse," or "flat" priors. The rationale for using a noninformative prior is to "let the data speak for themselves" and to have the prior distribution contribute only minimally.

Example 12.5. Consider the following two research situations:

- A researcher is interested in reporting an estimate for p, the proportion of individuals who experience an allergic reaction to a new drug. Since the drug is new, we might have a genuine lack of knowledge as to where (the distribution of) p is likely located. In this case, one could noninformatively take $p \sim \text{beta}(1,1)$; that is, a beta prior with parameters $\alpha = \beta = 1$. Recall that the beta$(1,1)$ distribution is the same as a $\mathcal{U}(0,1)$ distribution. This distribution is flat over $\Omega = \{p : 0 < p < 1\}$, so its contribution to the posterior $g(p|\boldsymbol{y})$ will be minimal.

- A medical investigation is undertaken to learn about the relationship between brain lesion frequency for patients with advanced multiple sclerosis. A Poisson(λ) model is assumed for Y, the number of brain lesions per subject. A largely noninformative prior for λ is a gamma$(\alpha = 1/2, \beta = 100)$ distribution; this distribution is relatively flat over $\Omega = \{\lambda : \lambda > 0\}$, so it will not have a large effect in determining the posterior $g(\lambda|\boldsymbol{y})$. □

JEFFREYS' PRIORS: One approach used to elicit a noninformative prior distribution is due to Jeffreys, whose "principle" leads to specifying $g(\theta) \propto [J(\theta)]^{1/2}$, where

$$J(\theta) = -E\left[\left.\frac{\partial^2 \log f_Y(Y;\theta)}{\partial\theta^2}\right|\theta\right]$$

and $f_Y(y;\theta)$ denotes the conditional pdf (pmf) for Y, given θ.

Example 12.6. Suppose that $Y_1, Y_2, ..., Y_n$, conditional on θ, are iid Bernoulli(θ). Derive Jeffreys' prior.

Solution. The probability mass function for $Y \sim \text{Bernoulli}(\theta)$, for $y = 0, 1$, is given by $f_Y(y;\theta) = \theta^y(1-\theta)^{1-y}$, and thus,

$$\log f_Y(y;\theta) = y \log\theta + (1-y)\log(1-\theta).$$

The first and second derivatives of $\log f_Y(y;\theta)$ are

$$\frac{\partial \log f_Y(y;\theta)}{\partial\theta} = \frac{y}{\theta} - \frac{1-y}{1-\theta}$$

and

$$\frac{\partial^2 \log f_Y(y;\theta)}{\partial\theta^2} = -\frac{y}{\theta^2} - \frac{1-y}{(1-\theta)^2}.$$

Thus,

$$\begin{aligned} J(\theta) = -E\left[\left.\frac{\partial^2 \log f_Y(Y;\theta)}{\partial\theta^2}\right|\theta\right] &= E\left[\frac{Y}{\theta^2} + \frac{1-Y}{(1-\theta)^2}\right] \\ &= \frac{1}{\theta} + \frac{1}{(1-\theta)} = \frac{1}{\theta(1-\theta)}. \end{aligned}$$

Thus, Jeffreys' prior is taken as

$$g(\theta) \propto [J(\theta)]^{1/2} = \left[\frac{1}{\theta(1-\theta)}\right]^{1/2} = \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where $\alpha = \beta = 1/2$. Of course, we recognize this as the beta(1/2,1/2) kernel; thus, the beta(1/2,1/2) distribution is the Jeffreys' noninformative prior for θ. □

BAYESIAN CRITICISMS: Some classical (i.e., non-Bayesian) statisticians seem to be bothered by the fact that the prior distribution for the parameter θ needs to specified beforehand by the researcher.

- Of course, classical statisticians are fine with choosing a model for $\boldsymbol{Y}$, but, for some reason, they cringe at the thought of adding an additional model for θ. I would argue that in almost every real problem, the researcher will have some information regarding θ that can be conveniently included in the statistical model.

- Furthermore, if the amount of data observed is large; that is, large enough for the likelihood to dominate the prior model, the prior's contribution will be small. If this occurs, then the posterior distribution is likely not to be affected too greatly by the chosen prior model, unless the prior is just terribly misspecified.

- More advanced modeling problems, tackled from a non-Bayesian point of view, can prove to be very difficult and very messy. These problems are often times much easier addressed by attacking them from a Bayesian point of view. Usually, much more computation is needed (to simulate posteriors), but this is hardly a big deal.

12.4 Point estimation

REMARK: The Bayesian is primarily interested in the posterior distribution $g(\theta|\boldsymbol{y})$ because, by combining the prior model and information from the likelihood function, the posterior distribution contains all the information regarding θ. However, in practice, numerical summaries of the posterior distribution $g(\theta|\boldsymbol{y})$ are often desired.

POSTERIOR POINT ESTIMATION: As in the classical framework, we now consider constructing a point estimator for θ, focusing on measures of central location. To describe the location of the posterior distribution $g(\theta|\boldsymbol{y})$, the commonly-used numerical summaries of location are the mean, mode, and median.

- The **posterior mean** of θ is given by

$$\widehat{\theta}_B = E(\theta|\boldsymbol{Y} = \boldsymbol{y}) = \int_\theta \theta g(\theta|\boldsymbol{y})d\theta.$$

  That is, $\widehat{\theta}_B$ is the mean of θ, computed using the posterior distribution.

- The **posterior mode** of θ is given by

$$\widehat{\theta}_B^* = \arg\max_\theta g(\theta|\boldsymbol{y}).$$

  That is, $\widehat{\theta}_B^*$ is the value of θ that maximizes the function $g(\theta|\boldsymbol{y})$.

- The **posterior median** $\widetilde{\theta}_B$ solves the equation

$$0.5 = P(\theta \leq \widetilde{\theta}_B|\boldsymbol{y}) = \int_{-\infty}^{\widetilde{\theta}_B} g(\theta|\boldsymbol{y})d\theta.$$

  That is, $\widetilde{\theta}_B$ the 50th percentile of the posterior distribution.

Example 12.7. Recall Example 12.2, where we considered the prevalence of HIV infection among male IVDU subjects. In that example, the prior distribution was $\theta \sim \text{beta}(1, 19)$. Based on 100 male positive/negative statuses, modeled as an iid Bernoulli sample, we found the posterior distribution to be $\theta \sim \text{beta}(u + 1, 119 - u)$,

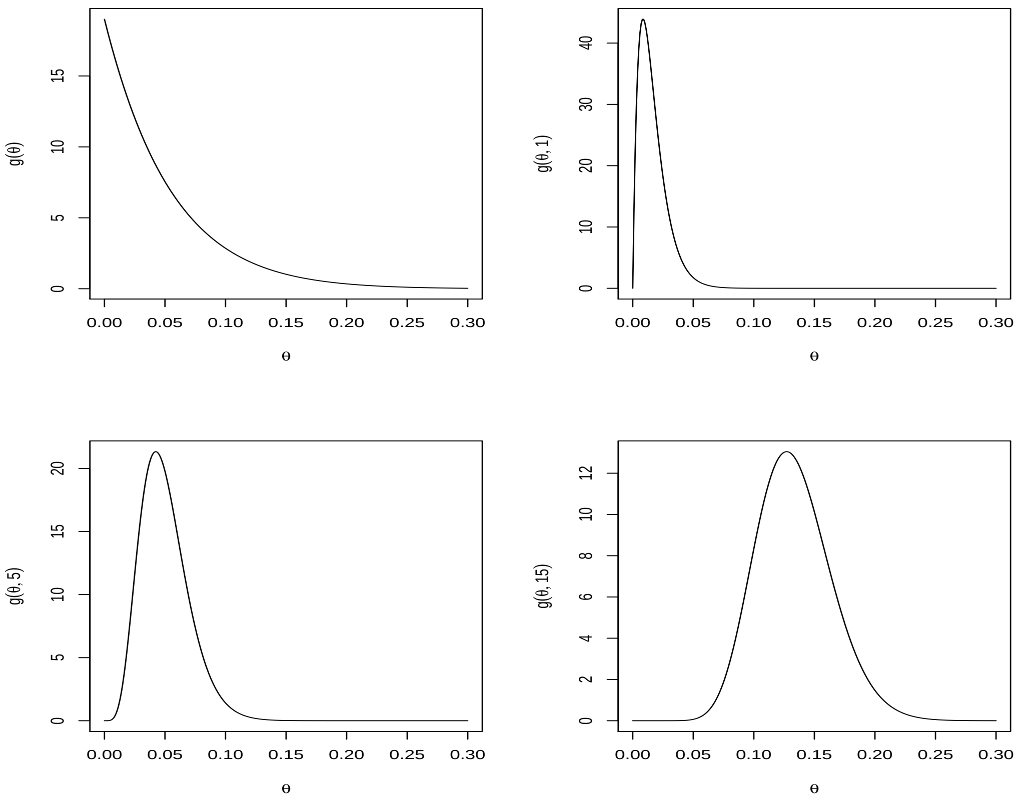

Figure 12.3: Binomial-beta Bayesian prior and posteriors in Example 12.2. Upper left: $\theta \sim \text{beta}(1,19)$, prior; Upper right: Posterior distribution of θ when $u = 1$, $\text{beta}(2,118)$; Lower left: Posterior distribution of θ when $u = 5$, $\text{beta}(6,114)$; Lower right: Posterior distribution of θ when $u = 15$, $\text{beta}(16,104)$. The sufficient statistic is $u = \sum_{i=1}^{100} y_i$.

where $u = \sum_{i=1}^{100} y_i$ denotes the total number of positives among the 100 male subjects. The table below gives the values of the posterior mean ($\widehat{\theta}_B$), mode ($\widehat{\theta}_B^*$), and median ($\widetilde{\theta}_B$) for three different values of u. The figure above depicts the three posterior distributions.

Prior, $g(\theta)$	Data	Posterior, $g(\theta\|u)$	$\widehat{\theta}_B$	$\widehat{\theta}_B^*$	$\widetilde{\theta}_B$	MLE
$\text{beta}(1,19)$	$u = 1$	$\text{beta}(2,118)$	0.0167	0.0085	0.0141	0.0100
$\text{beta}(1,19)$	$u = 5$	$\text{beta}(6,114)$	0.0500	0.0424	0.0475	0.0500
$\text{beta}(1,19)$	$u = 15$	$\text{beta}(16,104)$	0.1333	0.1271	0.1313	0.1500

COMPUTATIONS: To compute the posterior **mean** $\widehat{\theta}_B$, note that the posterior distribution is $\theta \sim \text{beta}(u+1, 119-u)$ so that

$$\widehat{\theta}_B = E(\theta|\boldsymbol{Y} = \boldsymbol{y}) = \frac{u+1}{120}.$$

To compute the posterior **mode** $\widehat{\theta}_B^*$, we need to find the value of θ that maximizes the $\text{beta}(u+1, 119-u)$ posterior density; i.e., the value of θ that maximizes

$$g(\theta|\boldsymbol{y}) = \frac{\Gamma(120)}{\Gamma(u+1)\Gamma(120-u)}\theta^u(1-\theta)^{118-u},$$

for fixed u. Because the log function is increasing, the value of θ that maximizes $g(\theta|\boldsymbol{y})$ is the same value of θ that maximizes

$$\log g(\theta|\boldsymbol{y}) = \log\Gamma(120) - \log\Gamma(u+1) - \log\Gamma(120-u) + u\log\theta + (118-u)\log(1-\theta).$$

To find the maximizer, we set the derivative of the log posterior equal to zero and solve for θ; that is, we solve the following equation for θ:

$$0 \overset{\text{set}}{=} \frac{\partial}{\partial\theta}\log g(\theta|\boldsymbol{y}) = \frac{u}{\theta} - \frac{118-u}{1-\theta} \Longrightarrow \theta = \frac{u}{118} \equiv \widehat{\theta}_B^*.$$

To find the posterior median $\widetilde{\theta}_B$, we solve

$$0.5 = P(\theta \leq \widetilde{\theta}_B|\boldsymbol{y}) = \int_0^{\widetilde{\theta}_B} \frac{\Gamma(120)}{\Gamma(u+1)\Gamma(120-u)}\theta^u(1-\theta)^{118-u}d\theta,$$

for $\widetilde{\theta}_B$. This is just the 50th percentile of the $\text{beta}(u+1, 119-u)$ distribution, which can be easily found using software (e.g., using the `qbeta` function in R).

REMARK: In the general binomial-beta Bayesian setting (outlined in Example 12.2), since the posterior distribution $\theta \sim \text{beta}(u+\alpha, n-u+\beta)$, the posterior mean is easily computed; that is,

$$\widehat{\theta}_B \equiv E(\theta|\boldsymbol{Y} = \boldsymbol{y}) = \int_0^1 \theta g(\theta|\boldsymbol{y})d\theta = \frac{u+\alpha}{n+\alpha+\beta}.$$

It is insightful to note that we can express $\widehat{\theta}_B$ as

$$\widehat{\theta}_B = \frac{u+\alpha}{n+\alpha+\beta} = \left(\frac{n}{n+\alpha+\beta}\right)\left(\frac{u}{n}\right) + \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)\left(\frac{\alpha}{\alpha+\beta}\right),$$

a **weighted average** of the maximum likelihood estimate $\widehat{\theta} = u/n$ and the prior mean $\alpha/(\alpha+\beta)$, where the weights are $n/(n+\alpha+\beta)$ and $(\alpha+\beta)/(n+\alpha+\beta)$, respectively. Note that the prior mean receives less weight as n increases. This makes sense intuitively; namely, if we have a larger sample size, we should weight the maximum likelihood estimate more and the prior mean less. On the other hand, if n is small, then the maximum likelihood estimate may not possess enough information about θ; in this case, we would want to weigh our prior beliefs more heavily.

12.5 Interval estimation

DIATRIBE: From a classical point of view, we have, at great length (mostly in STAT 512/513) discussed the construction and interpretation of **confidence intervals**.

- Recall that a $100(1-\alpha)$ percent confidence interval for θ includes this fixed parameter $100(1-\alpha)$ percent of the time in repeated sampling. The word "confidence" is carefully (perplexingly?) chosen so that we don't say "the probability that our computed interval contains θ is $1-\alpha$."

- The classical statistician regards θ as a fixed constant. Therefore, if the observed confidence interval is $(4.12, 12.39)$, say, it does not make sense to even think about $P(4.12 < \theta < 12.39)$ since the "event" inside the probability symbol does not contain anything random.

- Instead, we remember that if we were to repeat the experiment or study over and over again, each time under identical conditions, our interval estimation procedure would produce intervals that contain θ approximately $100(1-\alpha)$ percent of the time, and the computed interval we obtained (from the actual study) is just an example of one of these potentially observed intervals.

- Personally, I think this notion is rather confusing, since, in practice (that is, with real data), we only get to see the one observed confidence interval, not a large number of them. In addition, novice statistics students (and even experienced scientists) find the notion of "repeated sampling," inherent to the notion of a sampling distribution, to be both frustrating and unintuitive. Unfortunately, the correct interpretation of confidence intervals relies on this notion.

- For what it is worth, I find probability values (p-values), which are also based on the notion of repeated sampling, to be equally as confusing.

RECALL: Bayesian point estimators are often a measure of central tendency of the posterior distribution $g(\theta|\boldsymbol{y})$, such as a mean, median, mode, or perhaps even some other functional. It is also important to report posterior uncertainty. This can be done by using **credible intervals** (also known as **posterior probability intervals**). Such intervals are the Bayesian analogues of classical (frequentist) confidence intervals. However, their interpretation, as we will see now, is quite different.

CREDIBLE INTERVALS: If $g(\theta|\boldsymbol{y})$ is the posterior distribution of θ, given the data $\boldsymbol{Y} = \boldsymbol{y}$, then for any interval $A = (\theta_L, \theta_U)$, the **credible probability** of A is

$$P(\theta_L < \theta < \theta_U|\boldsymbol{y}) = \int_{\theta_L}^{\theta_U} g(\theta|\boldsymbol{y})d\theta.$$

If $g(\theta|\boldsymbol{y})$ is a discrete posterior distribution, we simply replace the integral with a sum. If $P(\theta_L < \theta < \theta_U|\boldsymbol{y}) = 1 - \alpha$, then we call (θ_L, θ_U) a $100(1-\alpha)$ **percent credible interval**. We interpret a $100(1-\alpha)$ percent credible interval $A = (\theta_L, \theta_U)$ as follows:

"The probability that θ is between θ_L and θ_U is $1 - \alpha$."

This is true since $g(\theta|\boldsymbol{y})$ is the (updated) probability distribution of θ given the data $\boldsymbol{y}$. Note that the interpretation of a Bayesian credible interval is far more straightforward than the interpretation of a classical confidence interval. However, the ease of interpretation comes with additional assumptions. The Bayesian model requires the specification of a prior distribution $g(\theta)$.

CONSTRUCTION: A $100(1-\alpha)$ percent credible interval results when the credible probability of $A = (\theta_L, \theta_U)$ is $1 - \alpha$. Here are two popular ways to construct $100(1-\alpha)$ percent credible intervals:

- Simply take the endpoints of A to be the lower and upper $\alpha/2$ quantiles of $g(\theta|\boldsymbol{y})$; this is called an **equal tail (ET) credible interval**.

- Take A to be the region of values that contain $100(1-\alpha)$ percent of the posterior probability in such a way that the posterior density $g(\theta|\boldsymbol{y})$ within the region A is never lower than outside A; this is called a **highest posterior density (HPD) credible interval**.

REMARK: Intuitively, if $g(\theta|\boldsymbol{y})$ is unimodal and symmetric, then the ET and HPD intervals are the same interval. The ET interval is often preferred in practice because it is easier to compute.

Example 12.8. Recall Example 12.2, where we considered the prevalence of HIV infection among male IVDU subjects. In that example, the prior distribution was $\theta \sim \text{beta}(1,19)$. Based on 100 male positive/negative statuses, modeled as an iid Bernoulli sample, we found the posterior distribution to be $\theta \sim \text{beta}(u+1, 119-u)$, where $u = \sum_{i=1}^{100} y_i$. The table below gives the 95 percent equal tail (ET) credible interval and the large-sample 95 percent Wald interval for three different values of u. The figure on the next page depicts the three posterior distributions.

Prior, $g(\theta)$	Data	Posterior, $g(\theta|u)$	95% ET interval	95% Wald interval
beta$(1,19)$	$u=1$	beta$(2,118)$	(0.0020, 0.0459)	$(-0.0010, 0.0295)$
beta$(1,19)$	$u=5$	beta$(6,114)$	(0.0187, 0.0953)	(0.0073, 0.0927)
beta$(1,19)$	$u=15$	beta$(16,104)$	(0.0788, 0.1994)	(0.0800, 0.2200)

COMPUTATIONS: Credible intervals come directly from the posterior distribution $g(\theta|u)$. Because the posterior distribution of $\theta \sim \text{beta}(u+1, 119-u)$, to compute the 95 percent equal tail (ET) interval, we need to find the values of θ_L and θ_U that satisfy the following integral equations:

$$0.025 = P(\theta < \theta_L|u) = \int_0^{\theta_L} g(\theta|u)d\theta = \int_0^{\theta_L} \underbrace{\frac{\Gamma(120)}{\Gamma(u+1)\Gamma(120-u)}\theta^u(1-\theta)^{118-u}}_{\text{beta}(u+1,119-u)\text{ density}} d\theta$$

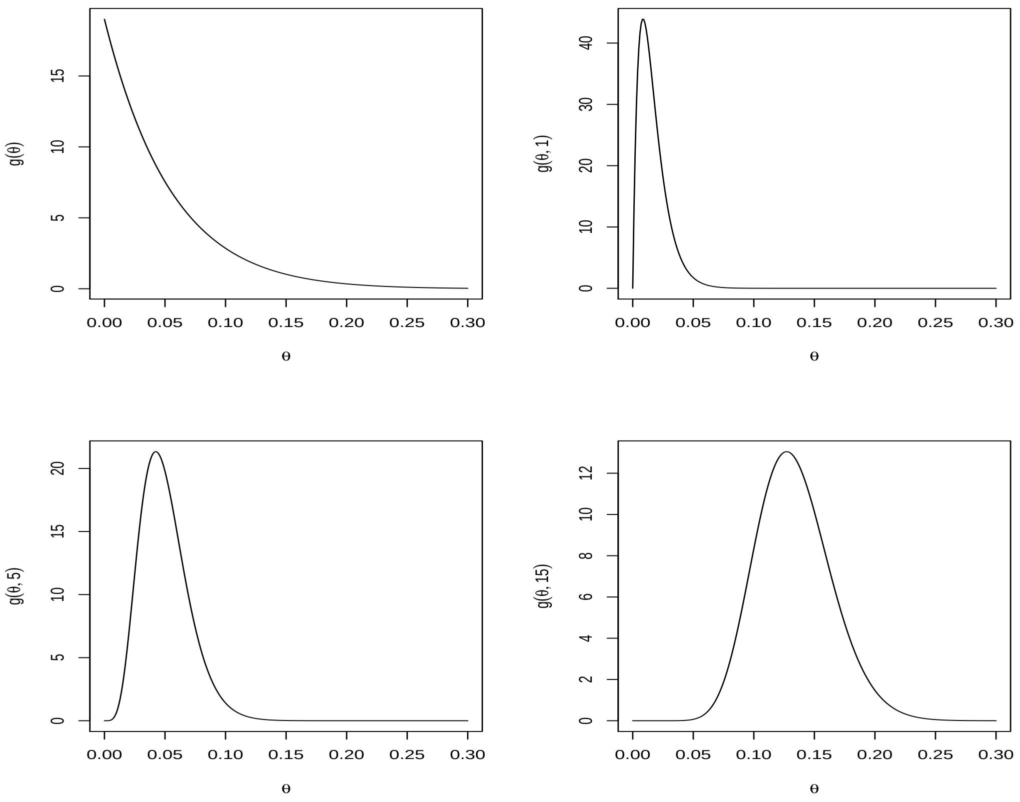

Figure 12.4: Binomial-beta Bayesian prior and posteriors in Example 12.2. Upper left: $\theta \sim \text{beta}(1, 19)$, prior; Upper right: Posterior distribution of θ when $u = 1$, $\text{beta}(2, 118)$; Lower left: Posterior distribution of θ when $u = 5$, $\text{beta}(6, 114)$; Lower right: Posterior distribution of θ when $u = 15$, $\text{beta}(16, 104)$. The sufficient statistic is $u = \sum_{i=1}^{100} y_i$.

and

$$0.025 = P(\theta > \theta_U | u) = \int_{\theta_U}^{1} g(\theta|u) d\theta = \int_{\theta_U}^{1} \frac{\Gamma(120)}{\Gamma(u+1)\Gamma(120-u)} \theta^u (1-\theta)^{118-u} d\theta,$$

respectively. In other words, we need to find the lower and upper 0.025 quantiles of the $\text{beta}(u+1, 119-u)$ distribution. Computing these quantiles is easily done using software (e.g., using the `qbeta` function in R). The classical 95 percent Wald interval for θ is computed as usual; i.e.,

$$\widehat{\theta} \pm 1.96\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{100}}.$$

- Perhaps the most glaring deficiency with the classical Wald interval is that its lower endpoint can be negative; that is, the interval itself can extend outside the parameter space $(0,1)$. On the other hand, Bayesian credible intervals can never extend beyond the parameter space because the posterior distribution is guaranteed to never fall outside of it.

- The prior information "pulls" the ET interval in the direction of the prior mean, here, $E(p) = 0.05$. This is evident by looking at the ET intervals based on $u = 1$, which is pulled to the right, and $u = 15$, which is pulled to the left.

12.6 Hypothesis tests

REMARK: On an introductory level, hypothesis testing is far less formal within a Bayesian framework than it is within the classical framework. In fact, many Bayesians do not even advocate their use, and, instead, simply summarize the posterior distributions (e.g., with point/interval estimates) without applying a formal test. Thus, our discussion of hypothesis tests will not be as rigorous as it was in a classical context (Chapter 10).

SETTING: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$, where the prior distribution $\theta \sim g(\theta)$. Within a Bayesian framework, suppose that we would like to test the hypothesis

$$H_0 : \theta \in \Omega_0$$

versus

$$H_a : \theta \in \Omega_a,$$

where $\Omega = \Omega_0 \cup \Omega_a$. As we have already learned, for the Bayesian, all inference is carried out using the posterior distribution $g(\theta|\boldsymbol{y})$. This is a valid probability distribution, so the probabilities

$$P(H_0 \text{ is true}|\boldsymbol{y}) = P(\theta \in \Omega_0|\boldsymbol{y}) = \int_{\Omega_0} g(\theta|\boldsymbol{y})d\theta$$

and

$$P(H_a \text{ is true}|\boldsymbol{y}) = P(\theta \in \Omega_a|\boldsymbol{y}) = \int_{\Omega_a} g(\theta|\boldsymbol{y})d\theta$$

make perfect sense and can be computed exactly. As for a decision rule, the Bayesian can choose to reject H_0 when

$$P(\theta \in \Omega_0|\boldsymbol{y}) < P(\theta \in \Omega_a|\boldsymbol{y});$$

i.e., reject H_0 when $P(\theta \in \Omega_0|\boldsymbol{y}) < 0.5$. If one wants to heavily guard against erroneously rejecting H_0, one can choose this threshold probability to be very small, say, 0.1 or 0.01.

REMARK: It is worth noting that statements like $P(H_0 \text{ is true}|\boldsymbol{y})$ and $P(H_a \text{ is true}|\boldsymbol{y})$ make no sense to the classical statistician because θ is a nonrandom quantity. Taking this perspective, the classical statistician is interested in using tests with good size and power properties. On the other hand, concepts like "Type I Error probability" (i.e., size) and power do not make sense to the Bayesian.

Example 12.9. Recall Example 12.2, where we considered the prevalence of HIV infection among male IVDU subjects. In that example, the prior distribution was $\theta \sim \text{beta}(1, 19)$. Based on 100 positive/negative statuses, modeled as an iid Bernoulli sample, we found the posterior distribution to be $\theta \sim \text{beta}(u + 1, 119 - u)$, where $u = \sum_{i=1}^{100} y_i$. Suppose that we would like to test $H_0 : \theta \leq 0.05$ versus $H_a : \theta > 0.05$. The table below gives the posterior probabilities $P(\theta \leq 0.05|u)$ for three different values of u.

Prior, $g(\theta)$	Data	Posterior, $g(\theta\|u)$	$P(\theta \leq 0.05\|u)$
$\text{beta}(1, 19)$	$u = 1$	$\text{beta}(2, 118)$	0.9837
$\text{beta}(1, 19)$	$u = 5$	$\text{beta}(6, 114)$	0.5502
$\text{beta}(1, 19)$	$u = 15$	$\text{beta}(16, 104)$	0.0003

COMPUTATIONS: The posterior probability $P(\theta \leq 0.05|u)$ is computed using the posterior distribution $g(\theta|u)$. Because the posterior distribution of $\theta \sim \text{beta}(u+1, 119-u)$,

$$P(\theta \leq 0.05|u) = \int_0^{0.05} \underbrace{\frac{\Gamma(120)}{\Gamma(u+1)\Gamma(120-u)}\theta^u(1-\theta)^{118-u}}_{\text{beta}(u+1,119-u)\text{ density}} d\theta,$$

which, again, is easily computed using software (e.g., using the `pbeta` function in R).

13 Survival Analysis

13.1 Introduction

INTRODUCTION: 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, machine/part failure, insurance claim, natural disaster, eradication of infection, etc.

- In chronic disease clinical trials; e.g., trials involving cancer, AIDS, diabetes, cardiovascular disease, etc., the primary endpoint (variable) of interest may be time to death, time to relapse of disease, etc. For such trials, we are usually interested in comparing the distribution of the time to event among competing 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, since 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 standard statistical techniques to be used. This area of (bio)statistics is called **survival analysis**.

Example 13.1. A randomized clinical trial involves 64 cancer patients with severe aplastic anemia. This condition occurs when an individual's bone marrow stops making enough new blood cells (this is a very 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 assigned to one of two treatment groups:

Table 13.1: Time to diagnosis of severe AGVHD for cancer patients. Starred subjects represent censored observations.

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

- 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. An important endpoint (variable) is the time from treatment assignment until the diagnosis of a life-threatening stage of acute graft versus host disease (AGVHD), a frequent complication where the donor's bone marrow cells attack the patient's organs and tissue. Table 13.1 presents these times (in days) for the 64 patients. In this trial, only 30 of the 64 patents actually reached the endpoint (i.e., were diagnosed with AGVHD). The remaining 34 patients were censored (i.e., they were never diagnosed with AGVHD).

- How should we model the diagnosis times? How should we compare the two groups?
- What effects do censoring/staggered entry have on the resulting analysis?
- Figure 13.1 displays estimates of the survivor distributions. It appears that those in the CSP+MTX group have a longer time to diagnosis of AGVHD. How are these estimates constructed? Is the difference between the two groups significant?

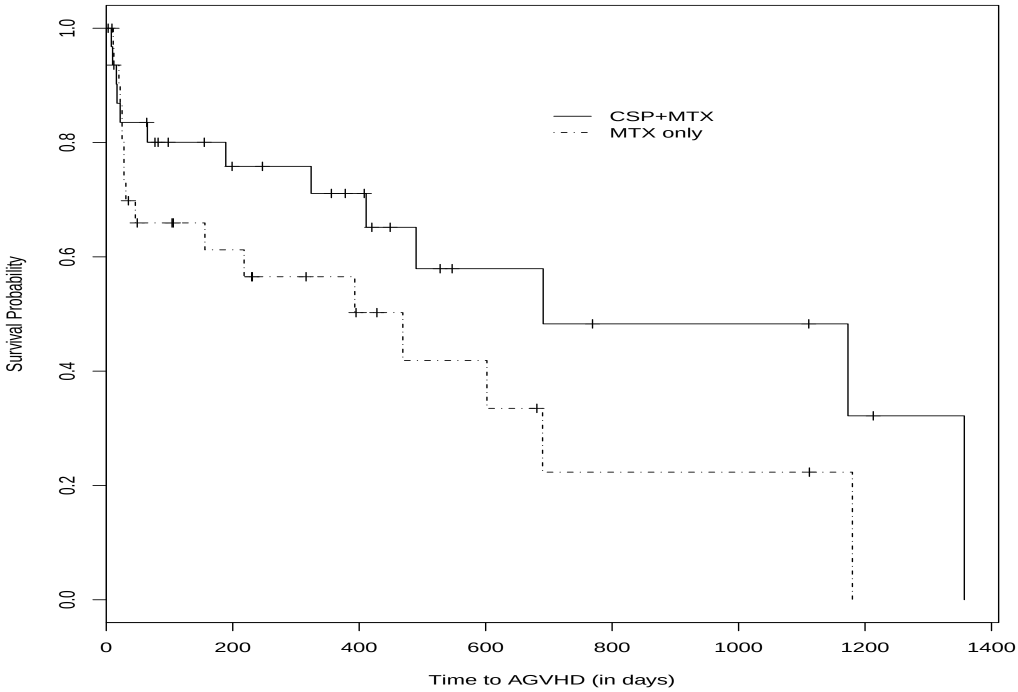

Figure 13.1: Kaplan-Meier survival function estimates of the time to diagnosis of AGVHD for two treatment groups.

13.2 Describing the distribution of time to an event

TERMINOLOGY: Let the random variable T denote the time to an event. It is understood to mean that T is a positive random variable, for which there is an unambiguous start (point of infection, start of treatment, etc.) and end (death, diagnosis, etc.) with the period in between corresponding to T. Random variables T with positive support are called **lifetime random variables**.

- survival time (from birth to death)
- the time from treatment of infection/disease to death (this may be tricky if individuals die from "other causes;" more about this later)
- the time to diagnosis of a more severe condition (e.g., AIDS, etc.)

NOTE: 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 with the infection. 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.

DESCRIPTION: We now describe some different, but equivalent, ways of defining the distribution of T. In our discussion, we assume that T is continuous.

- The cumulative distribution function (cdf)

$$F_T(t) = P(T \leq t).$$

- The survivor function

$$S_T(t) = P(T > t) = 1 - F_T(t).$$

- The probability density function

$$f_T(t) = \frac{d}{dt}F_T(t) = -\frac{d}{dt}S_T(t).$$

  Also, recall that

$$F_T(t) = \int_0^t f_T(u)du$$

  and

$$S_T(t) = \int_t^\infty f_T(u)du.$$

Example 13.2. A simple parametric model for T is the exponential distribution. Recall that if $T \sim$ exponential(β), 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

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

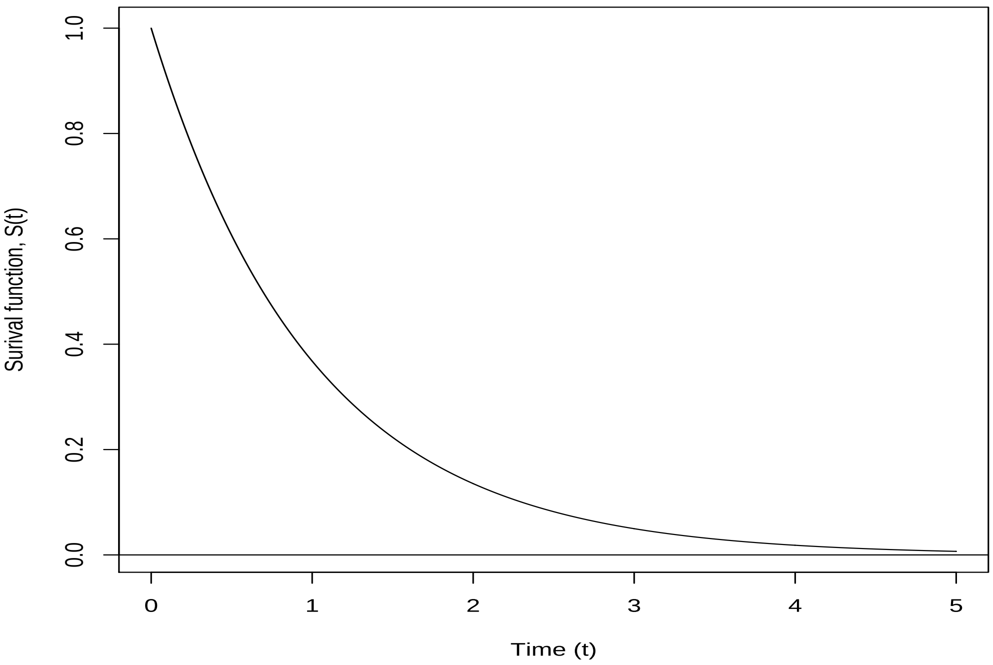

Figure 13.2: *The survivor function of* $T \sim$ exponential(1).

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 when $\beta = 1$. Note that $S_T(1) = e^{-1} \approx 0.367$; i.e., only 36.7 percent of the population "survives" one year, say (if T is measured in years). Also,

$$S_T(\phi_{0.5}) = e^{-\phi_{0.5}} = 0.5 \implies \phi_{0.5} = S_T^{-1}(0.5) = \ln 2 \approx 0.693;$$

i.e., the median survival $\phi_{0.5} = 0.693$ years.

TERMINOLOGY: We say that the distribution of a survival time T_1 is **stochastically larger** than another survival time T_2, and write $T_1 \geq_{\rm st} T_2$, if the survival function of T_1 is greater than or equal to the survival function of T_2 for all t; that is,

$$S_1(t) = P(T_1 > t) \geq P(T_2 > t) = S_2(t), \text{ for all } t \geq 0.$$

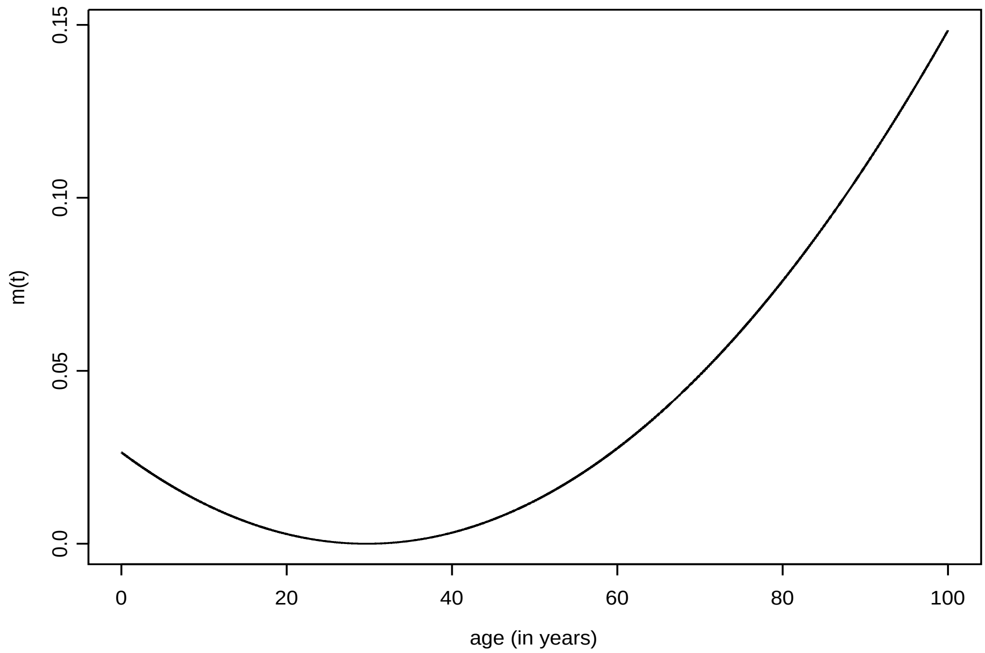

Figure 13.3: Mortality rate for a human population.

TERMINOLOGY: The **mortality rate**, at time t, is the proportion of the population who fail between times t and $t+1$ among individuals alive at time t. Usually, t is taken to be an integer in terms of some unit of time (e.g., day, month, year, etc.); i.e.,

$$m_T(t) = P(t \leq T < t+1|T \geq t).$$

The mortality rate for a human population might look like Figure 13.3.

TERMINOLOGY: The **hazard rate** is just a "continuous version" of a mortality rate. Informally, the hazard rate $\lambda_T(t)$ is the limit of the mortality rate if the interval of time is taken to be arbitrarily small; i.e., the mortality rate is the **instantaneous** rate of failure at time t, given that the individual is alive at time t. That is,

$$\lambda_T(t) = \lim_{h\to 0} \frac{P(t \leq T < t+h|T \geq t)}{h}.$$

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 very important characteristic of a lifetime distribution. It indicates the way the risk of failure varies with 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 (like people). Certain types of electronic devices may actually display a decreasing hazard function.

NOTE: It is insightful to note that

$$\begin{aligned}\lambda_T(t) &= \lim_{h\to 0}\frac{P(t\leq T<t+h|T\geq t)}{h}\\ &= \lim_{h\to 0}\frac{P(t\leq T<t+h)}{hP(T\geq t)}\\ &= \frac{1}{P(T\geq t)}\lim_{h\to 0}\frac{F_T(t+h)-F(t)}{h}=\frac{f_T(t)}{S_T(t)}=\frac{-\frac{d}{dt}S_T(t)}{S_T(t)}=-\frac{d}{dt}\log\{S_T(t)\}.\end{aligned}$$

Integrating both sides of the last equation, we get

$$-\log\{S_T(t)\}=\int_0^t \lambda_T(u)du\equiv\Lambda_T(t).$$

The function $\Lambda_T(t)$ is called the **cumulative hazard function**. Consequently,

$$S_T(t)=\exp\left\{-\int_0^t \lambda_T(u)du\right\}=\exp\left\{-\Lambda_T(t)\right\}.$$

REMARK: Because of these one-to-one relationships, we can describe the distribution of the continuous survival time T by using $f_T(t)$, $F_T(t)$, $S_T(t)$, $\lambda_T(t)$, or $\Lambda_T(t)$.

Example 13.3. Suppose that $T\sim\text{Weibull}(\alpha,\beta)$, where α and β are parameters larger than zero. This model is very common in engineering, and it has also been used in medical and actuarial science applications. The pdf of T is

$$f_T(t)=\begin{cases}\left(\dfrac{\alpha}{\beta}\right)t^{\alpha-1}\exp(-t^{\alpha}/\beta), & t>0\\ 0, & \text{otherwise.}\end{cases}$$

The cdf of T is

$$F_T(t)=\begin{cases}0, & t\leq 0\\ 1-\exp(-t^{\alpha}/\beta), & t>0.\end{cases}$$

The survivor function of T is

$$S_T(t)=1-F_T(t)=\begin{cases}1, & t\leq 0\\ \exp(-t^{\alpha}/\beta), & t>0.\end{cases}$$

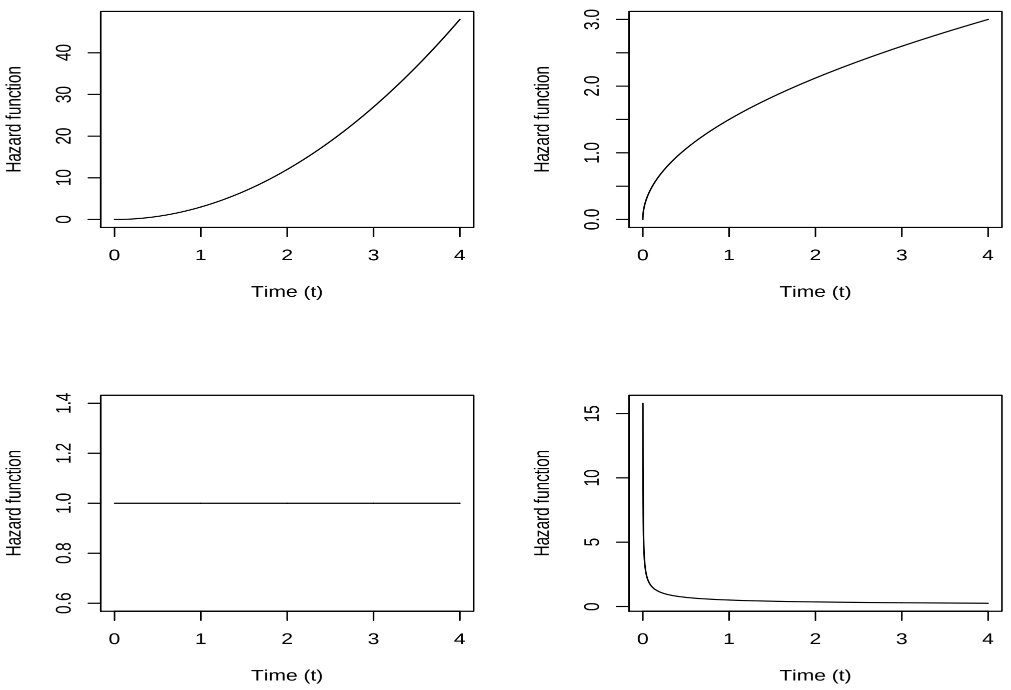

Figure 13.4: Weibull hazard functions with $\beta = 1$. Upper left: $\alpha = 3$. Upper right: $\alpha = 1.5$. Lower left: $\alpha = 1$. Lower right: $\alpha = 0.5$.

Therefore, the hazard function, for $t > 0$, is

$$\lambda_T(t) = \frac{f_T(t)}{S_T(t)} = \frac{\left(\frac{\alpha}{\beta}\right) t^{\alpha-1}\exp(-t^\alpha/\beta)}{\exp(-t^\alpha/\beta)} = \left(\frac{\alpha}{\beta}\right) t^{\alpha-1}.$$

Plots of Weibull hazard functions are given in Figure 13.4. It is easy to show

- $\lambda_T(t)$ is increasing if $\alpha > 1$, (population gets weaker with aging)
- $\lambda_T(t)$ is constant if $\alpha = 1$ (constant hazard; exponential distribution), and
- $\lambda_T(t)$ is decreasing if $\alpha < 1$ (population gets stronger with aging).

REMARK: In most clinical trials applications and research in survival analysis, it has become common to use **nonparametric** (and semiparametric) models where the shape of the distribution function is left unspecified. This is the approach we take henceforth.

13.3 Censoring and life table estimates

REMARK: Two important issues arise in survival analysis (in particular, clinical trials) when time to event data are being considered.

- Some individuals are still alive (the event of interest has not occurred) at the time of analysis. This results in **right censored data**.
- The length of follow-up varies due to **staggered entry** over calendar time. Patient time is measured from entry into the study.

In addition to censoring occurring because of insufficient follow-up (e.g., due to the study ending), it may also occur for other reasons. For example,

- loss to follow-up; e.g., the patient stops coming to the clinic or moves away
- death from other causes (competing risks).

These different forms of censoring are referred to as **random right censoring**. Random censoring creates difficulties in the analysis as is illustrated by the following example.

Example 13.4. Data from 146 individuals, who previously had myocardial infarction (MI); i.e., a heart attack, and participated in a clinical trial for an antihypertensive treatment (that is, a treatment to lower high blood pressure), are given in Table 13.2. The data have been grouped into one-year intervals, and all time is measured in terms of patient time. Here, the endpoint T is time to death.

QUESTION: How should we estimate the five-year survival rate $S_T(5)$?

SOLUTION. Two naive answers are given by

$$\frac{\text{76 deaths in 5 years}}{\text{146 individuals}} = 0.521 \Longrightarrow \widehat{S}_T(5) = 0.479.$$

$$\frac{\text{76 deaths in 5 years}}{\text{146-29 individuals}} = 0.650 \Longrightarrow \widehat{S}_T(5) = 0.350.$$

- 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 $t = 5$.

Table 13.2: Myocardial infarction data. Measured in patient time.

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

This corresponds to censoring on the right. This is probably 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. This is probably not the case either, so this estimate is overly **pessimistic**; i.e., this underestimates $S_T(5)$.

LIFE-TABLE ESTIMATES: Note that $S_T(5)$ can be expressed as

$$S_T(5) = \prod_{i=1}^{5} q_i,$$

where

$$q_i = 1 - \underbrace{P(i-1 \leq T < i | T \geq i-1)}_{\text{mortality rate at year } t=i-1},$$

for $i = 1,2,...,5$. So, we just need to estimate q_i. Note that $1-q_i$ is the **mortality rate** $m_T(t)$ at year $t = i-1$.

RIGHT CENSORING: Suppose that anyone withdrawn (censored) in an interval of time is censored at the end of that interval (right censoring). Our table then looks like

Time	$n(t)$	$d(t)$	$w(t)$	$\widehat{m}_T(t) = \frac{d(t)}{n(t)}$	$1 - \widehat{m}_T(t)$	$\widehat{S}_T^R(t) = \prod\{1 - \widehat{m}_T(t)\}$
$[0, 1)$	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 5-year survival probability, based on the life-table, would be $\widehat{S}_T^R(5) = 0.432$.

LEFT CENSORING: Suppose that anyone withdrawn (censored) in an interval of time is censored at the beginning of that interval (left censoring). Our table then looks like

Time	$n(t)$	$d(t)$	$w(t)$	$\widehat{m}_T(t) = \frac{d(t)}{n(t)-w(t)}$	$1 - \widehat{m}_T(t)$	$\widehat{S}_T^L(t) = \prod\{1 - \widehat{m}_T(t)\}$
$[0, 1)$	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 5-year survival probability, based on the life-table, would be $\widehat{S}_T^L(5) = 0.400$.

SUMMARY:

- Our (extremely) naive estimates ranged from 0.350 to 0.479.
- Our life-table estimates range from 0.400 to 0.432, depending on whether we assumed censoring occurred on the left or right of each interval.
- It is likely censoring occurs at a time inside the interval (not always on the endpoints). Thus, $\widehat{S}_T^R(5)$ and $\widehat{S}_T^L(5)$ are still too optimistic and pessimistic, respectively.

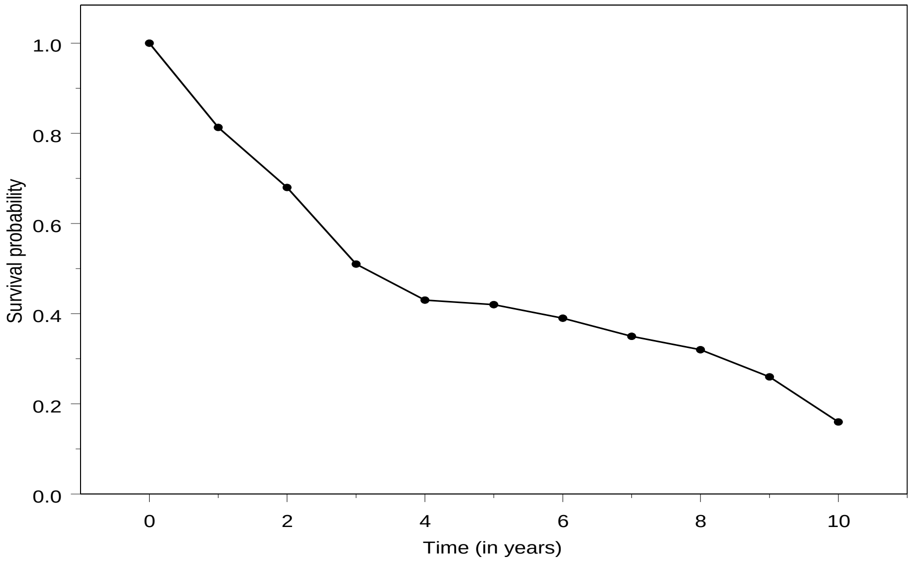

Figure 13.5: Myocardial infarction data in Example 13.4. Life-table estimate of the survival distribution $S_T(t)$.

COMPROMISE: A compromise is to use the following table:

Time	$n(t)$	$d(t)$	$w(t)$	$\widehat{m}_T(t) = \frac{d(t)}{n(t)-w(t)/2}$	$1 - \widehat{m}_T(t)$	$\widehat{S}_T(t) = \prod\{1 - \widehat{m}_T(t)\}$
$[0,1)$	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

From this table, our estimate is $\widehat{S}_T(5) = 0.417$, which is, of course, between $\widehat{S}_T^R(5)$ and $\widehat{S}_T^L(5)$. This is called the **life-table estimate**. The value $n(t) - w(t)/2$ is called the **effective sample size**. A plot of the estimated survival probabilities is in Figure 13.5.

INFERENCE: Of course, life-table estimates are computed from a sample of data, and, hence, are subject to natural sampling variability (as any other estimator is). Theoretical arguments show that, for a fixed t, $\widehat{S}_T(t)$ is approximately normal with mean $S_T(t)$ and variance which is consistently estimated by

$$\widehat{\text{var}}[\widehat{S}_T(t)] = \{\widehat{S}_T(t)\}^2 \sum_{j=1}^{t} \frac{d_j}{(n_j - w_j/2)(n_j - d_j - w_j/2)},$$

where $n_j = n(j)$, $d_j = d(j)$, and $w_j = w(j)$. The formula for $\widehat{\text{var}}[\widehat{S}_T(t)]$ is called **Greenwood's formula**. An approximate (large-sample) confidence interval for $S_T(t)$ is therefore given by

$$\widehat{S}_T(t) \pm z_{\alpha/2}\widehat{\text{se}}[\widehat{S}_T(t)],$$

where $\widehat{\text{se}}[\widehat{S}_T(t)] = \widehat{\text{var}}[\widehat{S}_T(t)]^{1/2}$.

MI DATA: Consider the following table used to find estimated standard errors for the MI data in Example 13.4:

Time	$n(t)$	$d(t)$	$w(t)$	$\widehat{S}_T(t)$	$\sum_j \frac{d_j}{(n_j-w_j/2)(n_j-d_j-w_j/2)}$	$\widehat{\text{se}}[\widehat{S}_T(t)]$
$[0,1)$	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

For the MI data, an approximate 95 percent confidence interval for $S_T(5)$ is given by

$$0.417 \pm 1.96(0.044) \implies (0.331, 0.503).$$

We are 95 percent confident that the proportion of patients surviving 5 years after an MI episode is between 0.331 and 0.503.

13.4 The Kaplan-Meier estimator

NOTE: In Example 13.4, we saw that the bias in estimating the survival function (incorrectly assuming that censoring occurs at the left or right of each interval) decreases when the interval is taken to be smaller (e.g., 1 year as opposed to 5 year intervals). Thus, if the data are not grouped (i.e., we know the exact times), we could apply the life-table estimator using intervals with very small lengths.

KAPLAN-MEIER ESTIMATOR: The "limit" of the life-table estimator; i.e., when the interval lengths are taken so small that at most one observation occurs within any interval, is called the **product-limit estimator** or the **Kaplan-Meier estimator**. Kaplan and Meier (1958) derived this limiting estimator using likelihood theory (not as the limit of the life-table estimator). However, it is instructive and intuitive to consider the KM estimator as a limit of the life-table estimator.

NON-INFORMATIVE CENSORING: In order for life table estimators 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 assumption is called the **non-informative censoring** assumption. Those at risk, at any time t, should be representative of the entire population alive at the same time so that the estimated mortality rates reflect the true population mortality rates.

REMARK: If censoring only occurs because of staggered entry, then the assumption of independent censoring seems 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 processes depend on the survival time. If at all possible, censoring from these latter situations should be kept to a minimum.

Example 13.5. *Computing the Kaplan-Meier estimate.* In this example, we work with a small (fictitious) data set to illustrate how the KM estimate is computed. Suppose we have the following death and censoring times for $n = 10$ patients.

Time	4.5	7.5	8.5	11.5	13.5	15.5	16.5	17.5	19.5	21.5
Censored	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.

That is, we have 6 deaths and 4 censored observations (out of the 10 patients). Denote

$$\widehat{m}_T(t) = \frac{d(t)}{n(t)} = \frac{\text{number of deaths in an interval}}{\text{number at risk at beginning of the interval}},$$

an estimate of the mortality rate at time t. 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	1	1	1	$\frac{9}{10}$	.	.	$\frac{8}{10}$	.	.	.	$\frac{48}{70}$	.	.	.	$\frac{192}{350}$	$\frac{144}{350}$	.	.	$\frac{144}{700}$

The Kaplan-Meier (or product limit) estimator will be a **step function** taking jumps at times where an event (death) occurs. Thus, since there is at most one occurrence in any interval of time, the KM estimator of the survival function $S_T(t)$ is computed by

$$\widehat{S}_T(t) = \prod_{j:t_j \leq t} \left(1 - \frac{1}{n_j}\right),$$

where n_j is the number of individuals still at risk at time t_j. By convention, the KM estimator is taken to be **right continuous**. The KM estimate for the data in Example 13.5, along with 95 percent confidence bands, is given in Figure 13.6. The confidence bands are computed as the endpoints of

$$\widehat{S}_T(t) \pm 1.96 \times \widehat{\text{se}}[\widehat{S}_T(t)],$$

where the (estimated) standard error is computed using Greenwood's formula. R automates the entire analysis; here is the output.

```
> summary(fit)
Call: survfit(formula=Surv(survtime,status)~1,conf.type="plain")
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  4.5     10       1    0.900  0.0949       0.7141        1.000
  7.5      9       1    0.800  0.1265       0.5521        1.000
 11.5      7       1    0.686  0.1515       0.3888        0.983
 15.5      5       1    0.549  0.1724       0.2106        0.887
 16.7      4       1    0.411  0.1756       0.0673        0.756
 19.5      2       1    0.206  0.1699       0.0000        0.539
```

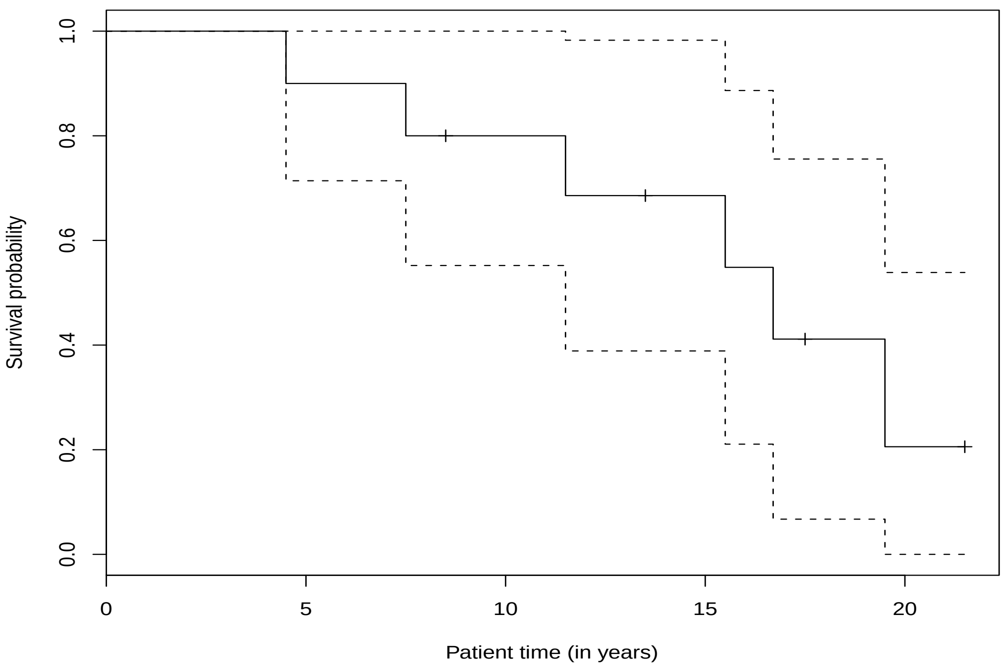

Figure 13.6: Kaplan-Meier estimate of $S_T(t)$ for the data in Example 13.5. Confidence bands have been included.

DESCRIPTION: In describing censored survival data, it is useful to conceptualize the existence of two **latent variables** for each individual corresponding to the failure time and censoring time. The term "latent" means "missing" or "not observed."

- 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 were observed until death, whereas C_i corresponds to the time that the ith individual would have been censored assuming death did not intervene.
- For example, C_i may be the time from entry into the study until the time of analysis. If censoring were to occur for other reasons, (e.g., loss to follow up, competing risks, etc.) this would have to be accounted for in the analysis.

OBSERVABLES: In actuality, 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 see whether the individual failed (died) or was censored; i.e., we get to see

$$\Delta_i = I(T_i \leq C_i) = \begin{cases} 1, & \text{if } T_i \leq C_i \\ 0, & \text{if } T_i > C_i. \end{cases}$$

Therefore, the variables (X_i, Δ_i), $i = 1, 2, ..., n$, are the observables in a survival experiment, whereas T_i and C_i are latent variables which are useful in conceptualizing the problem.

GOAL: Although not always observed, the main goal in survival analysis is to make inference about the probability distribution of the latent variable T. For example, in the one-sample problem, we are usually interested in estimating the survival 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 our 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 KM estimator for the survival distribution $S_T(t)$ is given by

$$\begin{aligned} \text{KM}(t) &= \prod_{\{i: X_i \leq t\}} \left\{ \frac{n(X_i) - 1}{n(X_i)} \right\}^{\Delta_i} \\ &= \prod_{\{i: X_i \leq t\}} \left\{ 1 - \frac{1}{n(X_i)} \right\}^{\Delta_i}. \end{aligned}$$

This is the definition of the KM estimator when there are no tied survival times in our sample. This formula emerges as the “limit” of the life-table estimator, that is, when we are allowed to partition patient time into very small intervals (as in Example 13.5) so that at most one event can occur in each interval.

DEALING WITH TIES: Let $d(t)$ denote the number of observed deaths in the sample at time t; that is,

$$d(t) = \sum_{i=1}^{n} I(X_i = t, \Delta_i = 1).$$

Generally, $d(t)$ is equal to 0 or equal to 1 with continuous survival data (where there are no ties). More generally, however, $d(t)$ may be greater than 1 when ties are allowed. In this situation, we can write the KM estimator as

$$\text{KM}(t) = \prod_{A(u)} \left\{1 - \frac{d(u)}{n(u)}\right\},$$

where $A(u)$ is the set of all death times u less than or equal to t. A consistent estimator of the variance of the KM estimator is also taken as the limit of **Greenwood's formula**; in particular,

$$\widehat{\text{var}}[\text{KM}(t)] = \{\text{KM}(t)\}^2 \sum_{A(u)} \left[\frac{d(u)}{n(u)\{n(u) - d(u)\}}\right].$$

Thus, for a fixed t, because $\text{KM}(t)$ is approximately normal in large samples, a $100(1-\alpha)$ percent confidence interval for the survival function $S_T(t)$ is given by

$$\text{KM}(t) \pm z_{\alpha/2}\widehat{\text{se}}\{\text{KM}(t)\},$$

where $\widehat{\text{se}}\{\text{KM}(t)\} = \widehat{\text{var}}[\text{KM}(t)]^{1/2}$.

Example 13.6. In this example, we simulate a set of censored survival data for $n = 100$ individuals and plot the resulting KM estimate of $S_T(t)$.

- We assume that the true survival times are exponential with mean $\beta = 5$ years. We generate $T_i \sim$ iid exponential(5).
- We assume that the true censoring times are exponential with mean $\beta = 10$ years. We generate $C_i \sim$ iid exponential(10).
- Note that the censoring time distribution is stochastically larger than the survival time distribution. Because the observed time is $X_i = \min\{T_i, C_i\}$, this means that fewer observations will be censored.

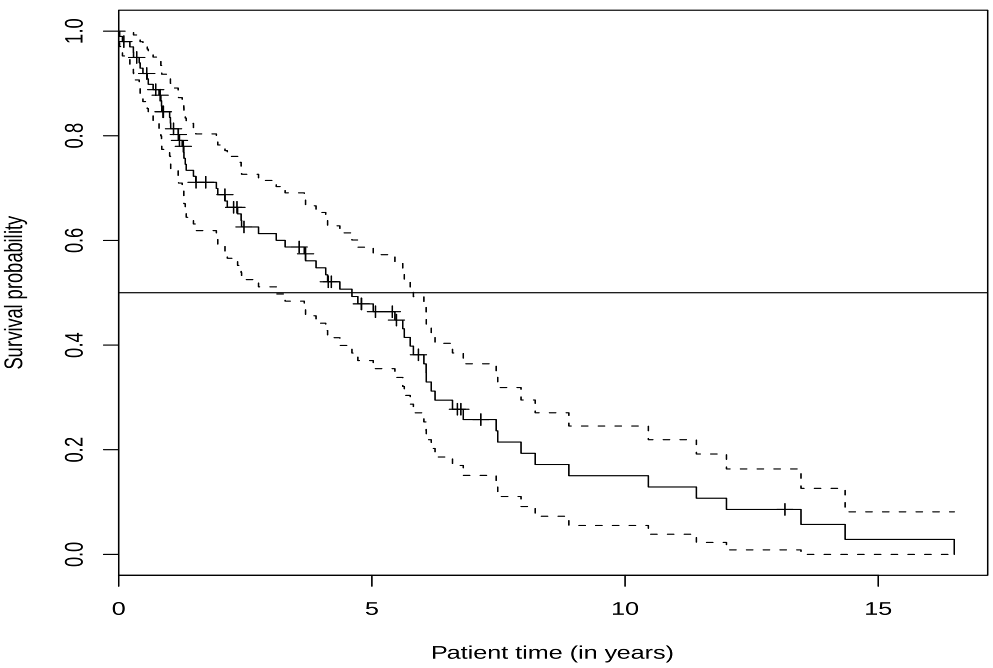

Figure 13.7: KM estimate of $S_T(t)$ using simulated data in Example 13.6. Confidence bands are included.

RESULTS: I have displayed below the results for the first 5 individuals:

```
  survtime.1.5. censtime.1.5. obstime.1.5.
1         6.805         9.380        6.805
2         6.592         9.552        6.592
3         6.678         5.919        5.919
4         5.640        14.223        5.640
5         5.829         3.690        3.690
```

- Note that `obstime` (X_i) is the minimum of `survtime` (T_i) and `censtime` (C_i).
- The KM estimate of $S_T(t)$ is given in Figure 13.7. Note that in this example, the true $S_T(t)$ is

$$S_T(t) = \begin{cases} 1, & t \leq 0 \\ e^{-t/5}, & t > 0. \end{cases}$$

- Note that the true median survival time is

$$\phi_{0.5} = S_T^{-1}(0.5) = 5\ln(2) \approx 3.468.$$

Example 13.7. Sickle-Santanello et al. (1988, *Cytometry*) present data on $n = 80$ male subjects with advanced tongue cancer. There were actually two types of cancerous tumors examined in the study, but for the purposes of this discussion, we will not distinguish between the two tumor types. The endpoint was time to death (measured in weeks from entry into the study). Among the 80 subjects, there were 52 deaths and 28 individuals censored. Below is the R output from the analysis. The KM estimate of $S_T(t)$ is displayed in Figure 13.8.

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     79       1    0.987  0.0126       0.9627        1.000
    3     78       3    0.949  0.0247       0.9010        0.998
    4     75       2    0.924  0.0298       0.8656        0.982
    5     73       2    0.899  0.0339       0.8322        0.965
    8     71       1    0.886  0.0357       0.8160        0.956
   10     69       1    0.873  0.0375       0.7998        0.947
   12     68       1    0.860  0.0391       0.7839        0.937
   13     67       3    0.822  0.0432       0.7372        0.906
   16     64       2    0.796  0.0455       0.7070        0.885
   18     62       1    0.783  0.0465       0.6921        0.875
   23     61       1    0.771  0.0475       0.6774        0.864
   24     60       1    0.758  0.0484       0.6628        0.853
   26     59       2    0.732  0.0501       0.6338        0.830
   27     57       2    0.706  0.0515       0.6054        0.807
   28     55       1    0.693  0.0521       0.5913        0.796
   30     54       3    0.655  0.0538       0.5495        0.760
   32     51       1    0.642  0.0542       0.5358        0.748
   41     50       1    0.629  0.0546       0.5221        0.736
   42     49       1    0.616  0.0550       0.5086        0.724
   51     48       1    0.604  0.0554       0.4951        0.712
   56     47       1    0.591  0.0556       0.4817        0.700
   62     45       1    0.578  0.0559       0.4680        0.687
   65     44       1    0.564  0.0562       0.4543        0.675
   67     43       1    0.551  0.0564       0.4408        0.662
   69     41       1    0.538  0.0566       0.4270        0.649
   70     40       1    0.524  0.0568       0.4132        0.636
   72     39       1    0.511  0.0569       0.3995        0.622
   73     38       1    0.498  0.0569       0.3859        0.609
   77     35       1    0.483  0.0571       0.3715        0.595
   91     27       1    0.465  0.0577       0.3524        0.578
   93     26       1    0.448  0.0582       0.3335        0.562
   96     24       1    0.429  0.0587       0.3139        0.544
  100     22       1    0.409  0.0592       0.2935        0.525
  104     20       3    0.348  0.0600       0.2304        0.466
  112     13       1    0.321  0.0610       0.2016        0.441
```

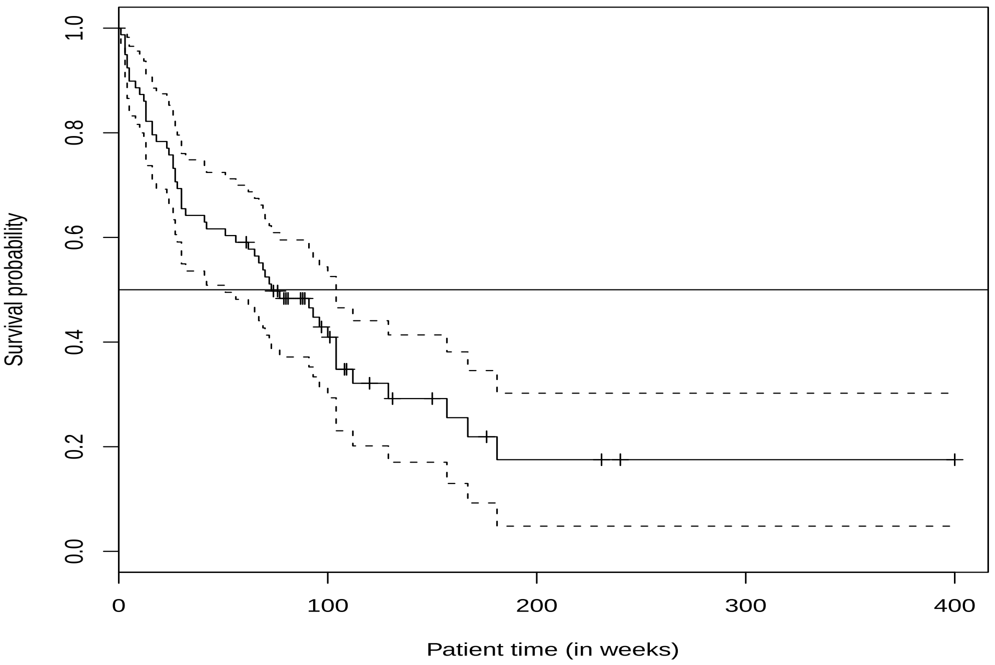

Figure 13.8: Tongue cancer data. KM estimate of $S_T(t)$ in Example 13.7. Confidence bands are included.

```
  129      11         1     0.292  0.0621          0.1703           0.414
  157       8         1     0.256  0.0642          0.1298           0.381
  167       7         1     0.219  0.0645          0.0925           0.346
  181       5         1     0.175  0.0648          0.0482           0.302
```

Question: From these data, what is an estimate of the one year survival probability? two year survival probability? That is, what are $\widehat{S}_T(52)$ and $\widehat{S}_T(104)$?

Answers: From the R output, we have

$$\widehat{S}_T(51) = 0.604.$$

Because $\widehat{S}_T(t)$ remains constant for all $t \in [51, 56)$, this is also our estimate for $S_T(52)$. A 95 percent confidence interval for the one year survival probability $S_T(52)$ is $(0.4951, 0.712)$. An estimate of the two year survival probability $S_T(104)$ is 0.348 (95 percent CI = 0.2304 to 0.466).

13.5 Two-sample tests

GOAL: 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., death, 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 $Z \in \{1, 2\}$.

TWO-SAMPLE PROBLEM: The problem of comparing two treatments can be posed as a hypothesis test. If we denote by $S_1(t)$ and $S_2(t)$ the survival 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 $\lambda_j(t) = -\frac{d}{dt}\log\{S_j(t)\}$, 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 $t > 0$.

APPROACH: To address the two sample survival problem, we will make use of a **non-parametric** 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

$$P\{d_1(u) = d\} = \frac{\binom{n_1(u)}{d}\binom{n_2(u)}{d(u)-d}}{\binom{n(u)}{d(u)}}.$$

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.

TESTING PROCEDURE: To test

$$H_0 : S_1(t) = S_2(t)$$

versus

$$H_0 : S_1(t) \neq S_2(t),$$

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, then we would reject H_0 when $T_{LR} < -z_\alpha$, since, under $H_a : S_1(t) \geq S_2(t)$, we would expect the observed number of deaths from treatment 1 to be less than that expected under H_0.

- If we wanted to show that treatment 2 is better than treatment 1 (insofar as prolonging survival), then we would reject H_0 when $T_{LR} > z_\alpha$, since, under $H_a : S_1(t) \leq S_2(t)$, we would expect the observed number of deaths from treatment 1 to be larger than that expected under H_0.

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.

Example 13.8. Highly active antiretroviral therapy (HAART) is the combination of several antiretroviral medications used to slow the rate at which HIV makes copies of itself (multiplies) in the body. Is a combination of antiretroviral medications more effective than using just one medication (monotherapy) in the treatment of HIV? In a two-group clinical trial involving patients with advanced AIDS, 24 patients receive a standard monotherapy (treatment 1) and 24 patients receive a new HAART (treatment 2). Death/censoring times, measured in days, are given Table 13.3. The Kaplan-Meier estimates for these data (by treatment) are given in Figure 13.9.

Table 13.3: Time to death in patients with advanced AIDS. Measured in days. Starred subjects represent censored observations.

Standard treatment				HAART			
14	333	706	1730	64	863	1873	2380
17	444	909	1834	178	998	1993	2680
128	558	1213	2244*	478	1205	1999	2696
129	568	1216*	2246	533	1232	2140	2896
164	677	1420	2565	742	1232	2204*	3223
228	702	1527	3004	756	1433	2361	3344*

OUTPUT: The `fit.1` output gives point and confidence interval estimates for the median survival times; estimated standard errors are computed using Greenwood's formula.

```
> fit.1
Call: survfit(formula = Surv(survtime, delta) ~ treat)
        records n.max n.start events median 0.95LCL 0.95UCL
treat=1      24    24      24     22    704     558    1730
treat=2      24    24      24     22   1653    1205    2380
```

Therefore,

- we are 95 percent confident that the median survival time for treatment group 1 (monotherapy) is between 558 and 1730 days.
- we are 95 percent confident that the median survival time for treatment group 2 (HAART) is between 1205 and 2380 days.

```
> fit.2
Call: survdiff(formula = Surv(survtime, delta) ~ treat)
         N Observed Expected (O-E)^2/E (O-E)^2/V
treat=1 24       22     15.8      2.44      3.98
treat=2 24       22     28.2      1.36      3.98
Chisq = 4  on 1 degrees of freedom, p = 0.0461
```

OUTPUT: The `fit.2` output gives the value of the square of the logrank statistic, that is, it gives

$$T_{LR}^2 = \left(\frac{\sum_{A(u)} \left[d_1(u) - \frac{n_1(u)}{n(u)} d(u) \right]}{\sqrt{\sum_{A(u)} \frac{n_1(u)n_2(u)d(u)\{n(u)-d(u)\}}{n^2(u)\{n(u)-1\}}}} \right)^2 .$$

To test

$$H_0 : S_1(t) = S_2(t)$$

versus

$$H_a : S_1(t) \neq S_2(t),$$

an approximate level α rejection region is

$$\text{RR} = \{T_{LR} : |T_{LR}| > z_{\alpha/2}\} = \{T_{LR} : T_{LR}^2 > \chi^2_{1,\alpha}\},$$

where $\chi^2_{1,\alpha}$ is the upper α quantile of a $\chi^2(1)$ distribution. Recall that $Z \sim \mathcal{N}(0,1)$ implies that $Z^2 \sim \chi^2(1)$.

ANALYSIS: With the HAART data, we find

$$T_{LR}^2 = 3.98 \text{ (p-value} = 0.0461).$$

At the $\alpha = 0.05$ level, we have sufficient evidence to reject $H_0 : S_1(t) = S_2(t)$ in favor of the two-sided alternative $H_a : S_1(t) \neq S_2(t)$. That is, there is significant evidence that the two survivor functions are different.

NOTE: Suppose that, a priori, we had specified a one-sided alternative

$$H_a : S_1(t) \leq S_2(t),$$

that is, patients on treatment 2 (HAART) had a longer survival time on average. Noting that the observed number of deaths for treatment 1 (22) is larger than the expected number of deaths under H_0 (15.8), we know that $T_{LR} = +\sqrt{3.98} \approx 1.995$ (that is, T_{LR} is positive; not negative). Thus, at the $\alpha = 0.05$ level, we would reject $H_0 : S_1(t) = S_2(t)$ in favor of $H_a : S_1(t) \leq S_2(t)$ since $T_{LR} = 1.995 \geq z_{0.05} = 1.645$. $\square$

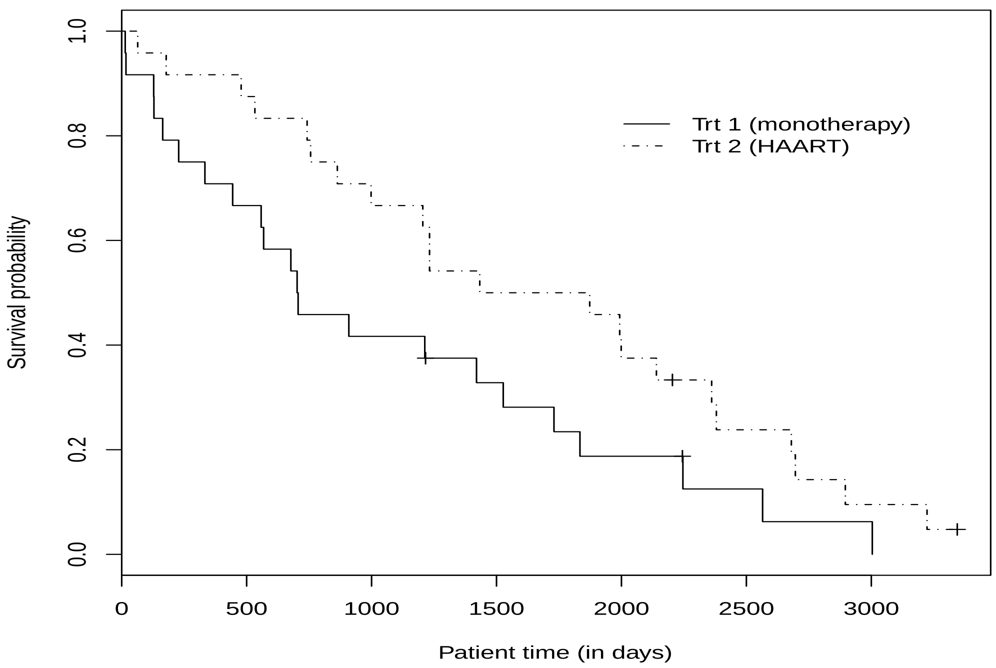

Figure 13.9: Kaplan-Meier estimates for AIDS patients in Example 13.8.

13.6 Power and sample size considerations for two-sample tests

IMPORTANT: Thus far, we have only considered the distribution of the logrank test statistic T_{LR} under the null hypothesis $H_0 : S_1(t) = S_2(t)$. However, we know that in order to assess statistical sensitivity, we must also consider the **power** of the test, or the probability of rejecting H_0 under some feasible "clinically important" alternative hypothesis.

PROPORTIONAL HAZARDS: One popular way of specifying a clinically important alternative is to make a **proportional hazards** assumption. Denote the hazard functions for treatments 1 and 2 by $\lambda_1(t)$ and $\lambda_2(t)$, respectively. The proportional hazards assumption means

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \exp(\eta), \text{ for all } t \geq 0.$$

We parameterize through the use of $\exp(\eta)$, since a hazard ratio must be **positive** and

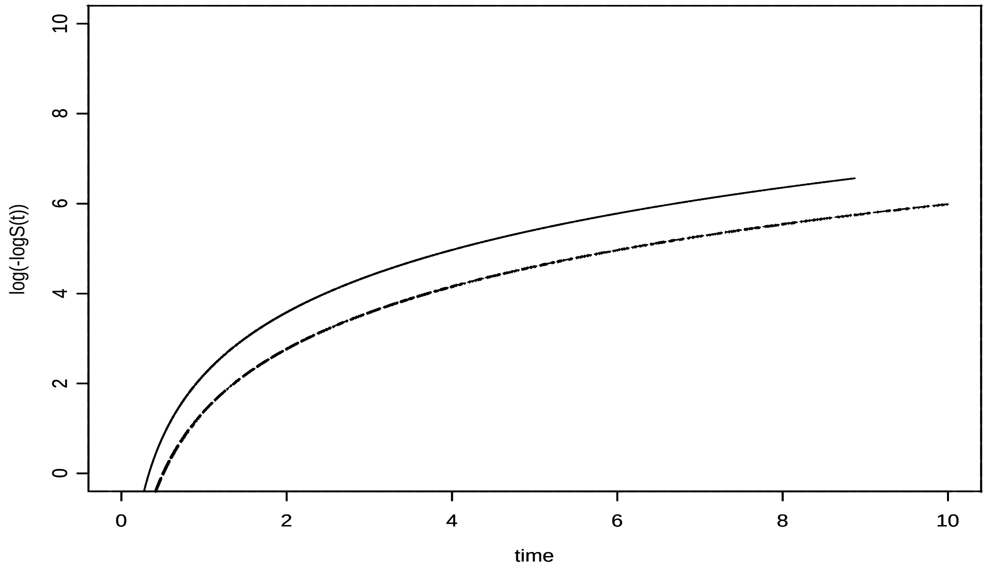

Figure 13.10: Two survivor functions plotted on the $\log\{-\log\}$ scale. These survivor functions satisfy the proportional hazards condition.

the $\eta = 0$ case would correspond to equal hazards for both treatments, which corresponds to the null hypothesis $H_0 : S_1(t) = S_2(t)$. Using the above parameterization,

- $\eta > 0 \Longrightarrow$ individuals on treatment 1 have higher rate of failure (they die faster)
- $\eta = 0 \Longrightarrow$ null hypothesis $H_0 : S_1(t) = S_2(t)$ is true
- $\eta < 0 \Longrightarrow$ individuals on treatment 1 have lower rate of failure (they live longer).

REMARK: Under a proportional hazards assumption, it can be shown that

$$\log\{-\log S_1(t)\} = \log\{-\log S_2(t)\} + \eta.$$

This relationship suggests that if we plot two survival curve estimates (e.g., the Kaplan-Meier estimates) on a $\log\{-\log\}$ scale, then we can assess the suitability of a proportional hazards assumption. If, in fact, proportional hazards was reasonable, we would expect to see something approximately like that in Figure 13.10.

SPECIAL CASE: When the survival distributions are **exponential**, so that the hazard functions are constant, we automatically have proportional hazards since

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\lambda_1}{\lambda_2}$$

is free of t. The **median survival time** of an exponential random variable with hazard λ is

$$m = \ln(2)/\lambda.$$

Therefore, the ratio of the median survival times for two treatments, under the exponential assumption, is

$$\frac{m_1}{m_2} = \frac{\ln(2)/\lambda_1}{\ln(2)/\lambda_2} = \frac{\lambda_2}{\lambda_1}.$$

That is, the ratio of the medians for two exponentially distributed random variables is inversely proportional to the ratio of the hazards. This result is very useful when one is trying to elicit clinically important differences. Investigators (e.g., physicians, etc.) are readily aware of the meaning of "median survival." On the other hand, hazard ratios are somewhat more difficult for them to understand.

LOGRANK LINK: Theoretical arguments show that the logrank test is the **most powerful** test among all nonparametric tests to detect alternatives which follow a proportional-hazards relationship. Therefore, the proportional hazards assumption not only has a simple interpretation for describing departures from the null hypothesis $H_0 : S_1(t) = S_2(t)$, but it also has nice statistical properties associated with the use of the logrank test.

POWER: In order to compute the power and the necessary sample sizes for a survival study, we need to know the distribution of the logrank test statistic T_{LR} under a specific alternative hypothesis H_a. For a proportional hazards alternative

$$H_a : \frac{\lambda_1(t)}{\lambda_2(t)} = \exp(\eta_A),$$

for $t > 0$, the logrank test statistic

$$T_{LR} \sim \mathcal{AN}\{[d\theta(1-\theta)]^{1/2}\eta_A, 1\},$$

where d is the total number of deaths from both treatments and θ is the proportion of individuals randomized to treatment 1. Unless otherwise stated, we will assume that $\theta = 0.5$. Theoretical arguments show that for a level α (two-sided) test to have power $1 - \beta$ in detecting the alternative η_A, we must have

$$\frac{\eta_A d^{1/2}}{2} \overset{\text{set}}{=} z_{\alpha/2} + z_\beta.$$

Solving for d, we get

$$d = \frac{4(z_{\alpha/2} + z_\beta)^2}{\eta_A^2}.$$

For example, if $\alpha = 0.05$ and $1 - \beta = 0.90$, then

$$d = \frac{4(1.96 + 1.28)^2}{\eta_A^2}.$$

Consider the following table of hazard ratios $\exp(\eta_A)$:

Hazard ratio, $\exp(\eta_A)$	No. of deaths, d
2.00	88
1.50	256
1.25	844
1.10	4,623

NOTE: As $\exp(\eta_A)$ becomes closer to one, we are trying to detect smaller differences between the hazard functions $\lambda_1(t)$ and $\lambda_2(t)$; thus, we will (intuitively) need a larger sample size to detect these smaller departures from H_0.

SAMPLE SIZE CALCULATIONS: During the design stage, we must ensure that a sufficient number of individuals are entered into a study and are followed long enough so that the requisite numbers of deaths are attained. One straightforward approach is to just continue the study until we obtain the required number of deaths.

Example 13.9. Suppose that patients with advanced lung cancer historically have a median survival of 6 months. We have a new treatment (treatment 2) which, if it increases median survival to 9 months, would be considered clinically important. We would like to detect such a difference with 90 percent power using a level $\alpha = 0.05$ two sided test. If both survival distributions are approximately exponential, then the clinically important hazard ratio is

$$\frac{\lambda_1}{\lambda_2} = \frac{m_2}{m_1} = \frac{9}{6}.$$

Thus, $\eta_A = \ln(9/6) = 0.4055$. With these criteria, we would need to observe

$$d = \frac{4(1.96 + 1.28)^2}{(0.4055)^2} \approx 256 \text{ deaths.}$$

Therefore, to attain the desired goals, we could, for example, enter 500 patients, randomize 250 to each treatment group, and follow these patients until we have a total of 256 deaths. Note that I have chosen "500" arbitrarily here. □

REMARK: In most survival applications involving time to death endpoints, arbitrarily picking a number of individuals and waiting for d deaths will not be adequate for the proper planning of the study. Instead, one usually needs to specify (to the investigators) the following:

- the number of patients
- the accrual period
- the length of follow-up time.

We have shown that to obtain reasonable approximations for the power, we need the expected number of events (deaths), computed under the alternative hypothesis, to be

$$d = \frac{4(z_{\alpha/2} + z_\beta)^2}{\eta_A^2};$$

i.e., we must compute the expected number of deaths separately for each treatment group, under the assumption that the alternative is true. The sum of these expected values, from both treatments, should be equal to d.

NOTE: To compute the expected number of deaths, we will assume that censoring is due to lack of follow-up resulting from staggered entry. If we, additionally, have other forms of censoring (e.g., competing risks, loss to follow-up, etc.), then the computations which follow would have to be modified.

NOTATION: In order to more thoroughly plan a study with a survival endpoint, we define the following notation:

- A is the **accrual period**; that is, the calendar period of time that patients are entering the study (e.g., January 1, 2011 through December 31, 2013)
- F is the **follow-up period**; that is, the calendar period of time after accrual has ended (before the final analysis is conducted)
- $L = A + F$ denotes the total calendar time of the study from the time the study opens until the final analysis
- $a(u)$ is the **accrual rate** at calendar time u; more precisely,

$$a(u) = \lim_{h \to 0} \left\{ \frac{\text{expected number of patients entering between } [u, u+h)}{h} \right\}$$

- The total expected number of patients in the study is then given by

$$\int_0^A a(u)du.$$

  If we have a **constant accrual rate** (this is a common assumption made in the design stage), then $a(u) = a$ and the total expected number of patients is aA.

DESIGN: Suppose we have a "clean" investigation where there is no loss to follow-up and no competing risks. If $a(u)$ is the accrual rate onto a study, randomized equally to two treatments, then the **expected number of deaths** for treatment 1 is

$$d_1 = \int_0^A \frac{a(u)}{2} F_1(L - u)du,$$

where $F_1(\cdot)$ is the cumulative distribution function for the survival time for treatment 1. To see why this makes sense, note that

- we would expect $\frac{a(u)}{2}du$ patients to enter in the interval of time from u to $u + du$.
- Of these patients, the proportion $F_1(L - u)$ are expected to die by the end of the study (i.e., at time L).

- This number summed (i.e., integrated) over u, for values $u \in [0, A]$, yields the expected number of deaths on treatment 1.

Similarly, the expected number of deaths for treatment 2 is

$$d_2 = \int_0^A \frac{a(u)}{2} F_2(L-u)du,$$

where $F_2(\cdot)$ is the cumulative distribution function for the survival time for treatment 2. The sum of these expected values, from both treatments, should equal $d_1 + d_2$; thus, we are to set

$$d_1 + d_2 = \frac{4(z_{\alpha/2} + z_\beta)^2}{\eta_A^2}.$$

Note that the number of deaths can be affected by the accrual rate, the accrual period (sample size), the follow-up period, and the failure rate (survival distribution). Some (or all) of these factors can be controlled by the investigator and have to be considered during the design stage.

Example 13.10. Suppose that the accrual rate is constant at a patients per year, and that we randomize equally to two treatments ($\theta = 0.5$), so that the accrual rate is $a/2$ patients per year for each treatment. Also, suppose that the survival distribution for treatment j is exponential with hazard ratio λ_j; $j = 1, 2$. We have

$$\begin{aligned} d_j &= \int_0^A \frac{a}{2}\left[1 - e^{-\lambda_j(L-u)}\right] du \\ &= \frac{a}{2}\left[A - \frac{e^{-\lambda_j L}}{\lambda_j}\left(e^{\lambda_j A} - 1\right)\right], \end{aligned}$$

for $j = 1, 2$. Suppose that during the design stage, we expect $a = 100$ patients per year to be recruited into the study. Suppose that the median survival for treatment 1 is 4 years; thus,

$$4 = \ln(2)/\lambda_1 \Longrightarrow \lambda_1 \approx 0.173.$$

We desire the new treatment 2 to increase median survival to 6 years (so that $\lambda_2 \approx 0.116$). If this happens, we want to have 90 percent power to detect it using a logrank test at the $\alpha = 0.05$ (two-sided) level of significance. With these medians, the hazard ratio is

$$\frac{\lambda_2}{\lambda_1} = \frac{6}{4} \quad \Longrightarrow \quad \eta_A = \ln(6/4).$$

Therefore, the total number of deaths must be

$$d = \frac{4(z_{\alpha/2} + z_\beta)^2}{\eta_A^2} = \frac{4(1.96 + 1.28)^2}{\{\ln(6/4)\}^2} \approx 256$$

so that

$$d_1(A, L) + d_2(A, L) = 256.$$

I have emphasized that d_1 and d_2 depend on our choice of the accrual period A and the length of the study L. According to our calculations, we need A and L to satisfy

$$\frac{100}{2}\left[A - \frac{e^{-0.173L}}{0.173}\left(e^{0.173A} - 1\right)\right] + \frac{100}{2}\left[A - \frac{e^{-0.116L}}{0.116}\left(e^{0.116A} - 1\right)\right] = 256.$$

NOTE: There are many (A, L) combinations that satisfy this equation. To find a particular solution, one possibility is to take the accrual period and the length of follow-up to be equal; i.e., take $A = L$. This will minimize the total length of the study. When $A = L$, the equation above has one solution; it is

$$A = L \approx 6.98 \text{ years}.$$

If the accrual rate is $a = 100$ patients/year, this would require 698 patients. □

13.7 More than two treatment groups

SETTING: We now extend our previous survival discussion to the case where our investigation involves $k > 2$ treatments. The data from such an investigation can be represented as a sample of triplets; namely,

$$\{(X_i, \Delta_i, Z_i);\ i = 1, 2, ..., n\},$$

where $X_i = \min\{T_i, C_i\}$, the failure indicator

$$\Delta_i = \begin{cases} 1, & \text{if } T_i \le C_i \\ 0, & \text{if } T_i > C_i, \end{cases}$$

and $Z_i = j$, for $j \in \{1, 2, ..., k\}$, corresponding to the treatment group to which the ith individual was assigned.

LOGRANK TEST: Let $S_j(t) = P(T \geq t|Z = j)$ denote the survival distribution for the jth treatment. We would now like to test

$$H_0 : S_1(t) = S_2(t) = \cdots = S_k(t)$$
$$\text{versus}$$
$$H_a : H_0 \text{ not true.}$$

The k-sample test we now describe is a direct generalization of the logrank test in the two sample problem (i.e., when $k = 2$). At any time u where $d(u) \geq 1$, we can envisage our data as a $2 \times k$ contingency table (like the one below), where, recall, $n_j(u)$ and $d_j(u)$ denote the number of individuals at risk and the number of deaths at time u from treatment group j, respectively:

	Treatment 1	Treatment 2	$\cdots$	Treatment k	Total
Number of deaths	$d_1(u)$	$d_2(u)$	$\cdots$	$d_k(u)$	$d(u)$
Number alive	$n_1(u) - d_1(u)$	$n_2(u) - d_2(u)$	$\cdots$	$n_k(u) - d_k(u)$	$n(u) - d(u)$
Total	$n_1(u)$	$n_2(u)$	$\cdots$	$n_k(u)$	$n(u)$

GENERALIZATION: We now consider a vector of observed number of deaths minus the expected number of deaths under H_0 for each treatment group j, i.e.,

$$\boldsymbol{d}(u) = \begin{pmatrix} d_1(u) - \frac{n_1(u)}{n(u)} d(u) \\ d_2(u) - \frac{n_2(u)}{n(u)} d(u) \\ \vdots \\ d_k(u) - \frac{n_k(u)}{n(u)} d(u) \end{pmatrix}_{k \times 1}.$$

Note that the sum of the elements in this vector is zero. If we condition on the marginal counts in the $2 \times k$ table, then the $k \times 1$ vector

$$(d_1(u), d_2(u), ..., d_k(u))'$$

follows a multivariate hypergeometric distribution. Of particular interest for us is that, conditional on the marginal counts, for $j = 1, 2, ..., k$,

$$E_C\{d_j(u)\} = \frac{n_j(u)}{n(u)} d(u),$$

where I have used the notation $E_C(\cdot)$ to denote conditional expectation, and

$$V_C\{d_j(u)\} = \frac{d(u)\{n(u)-d(u)\}n_j(u)\{n(u)-n_j(u)\}}{n^2(u)\{n(u)-1\}}.$$

Also, for $j \neq j'$,

$$\text{Cov}_C\{d_j(u), d_{j'}(u)\} = -\frac{d(u)\{n(u)-d(u)\}n_j(u)n_{j'}(u)}{n^2(u)\{n(u)-1\}}.$$

Now, consider the $(k-1)\times 1$ vector $\boldsymbol{S}$, defined by

$$\boldsymbol{S} = \begin{pmatrix} \sum_{A(u)}\left\{d_1(u) - \frac{n_1(u)}{n(u)}d(u)\right\} \\ \sum_{A(u)}\left\{d_2(u) - \frac{n_2(u)}{n(u)}d(u)\right\} \\ \vdots \\ \sum_{A(u)}\left\{d_{k-1}(u) - \frac{n_{k-1}(u)}{n(u)}d(u)\right\} \end{pmatrix}_{(k-1)\times 1},$$

where $A(u)$ is the set of death times u for all treatments. Note that we need only consider this $(k-1)$-dimensional vector, since the sum of all k elements of $\boldsymbol{d}(u)$ is zero, and, hence, one of the elements is extraneous. The corresponding $(k-1)\times(k-1)$ covariance matrix of $\boldsymbol{S}$ is given by $\boldsymbol{V} = (v_{jj'})$, for $j, j' = 1, 2, ..., k-1$, where

$$v_{jj} = \sum_{A(u)} \frac{d(u)\{n(u)-d(u)\}n_j(u)\{n(u)-n_j(u)\}}{n^2(u)\{n(u)-1\}},$$

for $j = 1, 2, ..., k-1$ (these are the diagonal elements of $\boldsymbol{V}$), and

$$v_{jj'} = -\sum_{A(u)} \frac{d(u)\{n(u)-d(u)\}n_j(u)n_{j'}(u)}{n^2(u)\{n(u)-1\}},$$

for $j \neq j'$ (these are covariance terms).

LOGRANK TEST: The k-sample logrank test statistic is the quadratic form

$$T_{LR} = \boldsymbol{S}'\boldsymbol{V}^{-1}\boldsymbol{S}.$$

Under $H_0 : S_1(t) = S_2(t) = \cdots = S_k(t)$, the logrank statistic T_{LR} has an approximate χ^2 distribution with $k-1$ degrees of freedom. If H_0 was true, then we would expect the elements in the vector $\boldsymbol{S}$ to be near zero; in this case, the quadratic form T_{LR} would also be near zero (so that, under H_0, T_{LR} would be small). If, however, H_0 was not true, then we would expect some of the elements of $\boldsymbol{S}$ to (perhaps greatly) deviate from zero; in this case, T_{LR} would be large. Therefore, to test

$$H_0 : S_1(t) = S_2(t) = \cdots = S_k(t)$$

versus

$$H_a : H_0 \text{ not true,}$$

we use the approximate level α rejection region RR $= \{T_{LR} : T_{LR} > \chi^2_{k-1,\alpha}\}$, where $\chi^2_{k-1,\alpha}$ is the upper α quantile of the $\chi^2(k-1)$ distribution.

Table 13.4: Time to death in patients with stage II breast cancer. Measured in days. Starred subjects represent censored observations.

Intensive CAF		Low dose CAF		Standard CAF	
501	4610	1959	357	4067	974
1721	665	354	1666	3494	4205
4280	3660	1157	1464	1323	2734
3350	2067	95	3146	1992	3634
3142	3260	2729	76	1482	3302
4167	653	2385	1199	1305	3436
3266	3684*	625	3750	3230	4372
894	4197*	1716	1206	71	1853*
4454	2320*	2574	391*	4117	989*
2360	3905*	1169	1847*	4002	1712*

Example 13.11. A clinical trial (CALGB 8541) includes female patients with positive stage II breast cancer. The endpoint of interest is time to death and there are $k = 3$ chemotherapy treatments:

- Intensive CAF
- Low dose CAF
- Standard dose CAF,

where CAF stands for cylclophosphamide, adriamycin, and 5-fluorouracil. Data from the trial are given in Table 13.4.

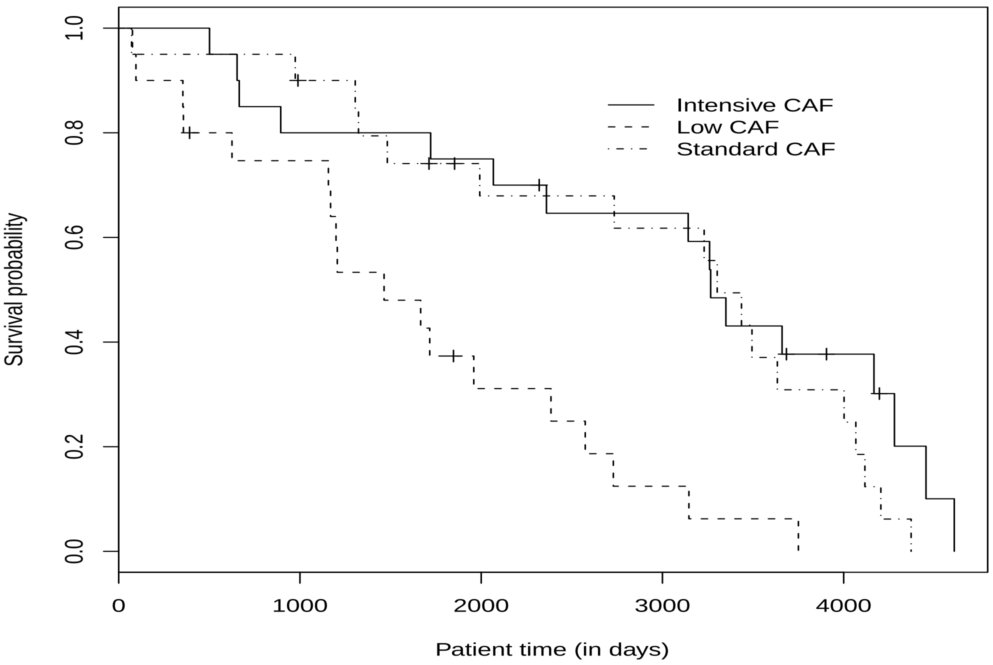

Figure 13.11: Kaplan-Meier estimates for breast cancer patients in Example 13.11.

ANALYSIS: Here is the output from the analysis of these data in R:

```
Call: survdiff(formula = Surv(survtime, delta) ~ treat)

         N Observed Expected (O-E)^2/E (O-E)^2/V

treat=1 20       16    24.21     2.784     5.784

treat=2 20       18     7.93    12.805    16.812

treat=3 20       17    18.86     0.184     0.303

Chisq = 17.7  on 2 degrees of freedom, p = 0.000141
```

The logrank test statistic is $T_{LR} = 17.7$ (p-value $= 0.000141$). Therefore, we have strong evidence against $H_0 : S_1(t) = S_2(t) = S_3(t)$, that is, there is strong evidence that the three chemotherapy treatments affect the survival rates differently.

REMARK: Note that rejection of H_0 does not tell us which survivor function estimates are statistically different. To see where the differences are explicitly, we could perform pairwise tests with $H_0 : S_1(t) = S_2(t)$, $H_0 : S_1(t) = S_3(t)$, and $H_0 : S_2(t) = S_3(t)$. □