To complete the test, we need to determine the distribution of F under H_0. First consider the sum of squares in the denominator, Q_3, which we write as:

$$Q_3/\sigma^2 = \sum_{j=1}^{b} \left\{ \frac{1}{\sigma^2} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_{\cdot j})^2 \right\}.$$

Notice, since we are discussing distributions, we are now using random variable notation. By Part (c) of Theorem 3.6.1, for $j = 1, \ldots, b$, the term within the braces has a χ^2-distribution with $n_j - 1$ degrees of freedom. Further, the samples are independent so these χ^2 random variables are independent. Hence, by Corollary 3.3.1, Q_3/σ^2 has a χ^2-distribution with $\sum_{j=1}^{b}(n_j - 1) = n - b$ degrees of freedom. By Part (b) of Theorem 3.6.1, the random variable $\overline{X}_{\cdot j}$ is independent of the sum of squares within the braces and further, by the independence of the samples, it is independent of Q_3. Thus, all b sample means are independent of Q_3. Because $\overline{X}_{\cdot\cdot} = \sum_{j=1}^{b} n_j \overline{X}_{\cdot j}$, the grand mean $\overline{X}_{\cdot\cdot}$ is a function of the b sample means, it must be independent of Q_3, also. Therefore, Q_4 is independent of Q_3. For the distribution of the numerator sum of squares, write the identity (9.2.10) as

$$Q/\sigma^2 = Q_3/\sigma^2 + Q_4/\sigma^2.$$

For the left side, under H_0, Q/σ^2 has a χ^2-distribution with $n-1$ degrees of freedom. On the right side Q_3/σ^2 has a χ^2-distribution with $n - b$ degrees of freedom and it is also independent of Q_4/σ^2. By equating the mgfs of both sides, it follows that Q_4/σ^2 has a χ^2-distribution with $(n - 1) - (n - b) = b - 1$ degrees of freedom. Therefore, under H_0, the F test statistic, (9.2.11), has a F-distribution with $b - 1$ and $n - b$ degrees of freedom.

Suppose now that we wish to compute the power of the test of H_0 against H_1 when H_0 is false, that is, when we do not have $\mu_1 = \mu_2 = \cdots = \mu_b$. In Section 9.3 we show that under H_1, Q_4/σ^2 no longer has a $\chi^2(b-1)$ distribution. Thus we cannot use an F-statistic to compute the power of the test when H_1 is true. The problem is discussed in Section 9.3.

Next, based on a simple example, we illustrate the computation of the F-test using R.

Example 9.2.1. Devore (2012), page 412, presents a data set where the response is the elastic modulus for an alloy that is cast by one of three different casting processes. The null hypothesis is that the mean of the elastic modulus is not affected by the casting process. The data are:

Cast Method	Elastic Modulus							
Permanent mold	45.5	45.3	45.4	44.4	44.6	43.9	44.6	44.0
Die cast	44.2	43.9	44.7	44.2	44.0	43.8	44.6	43.1
Plaster mold	46.0	45.9	44.8	46.2	45.1	45.5		

The data are in the file `elasticmod.rda`. The variable `elasticmod` contains the response while the variable `ind` contains the casting method (1, 2, or 3). The R code and results (test statistic F and the p-value) are:

```
oneway.test(elasticmod~ind,var.equal=T)
F = 12.565, num df = 2, denom df = 19, p-value = 0.0003336
```

With such a low p-value, the null hypothesis would be rejected and we would conclude that the casting method does have an effect on the elastic modulus. ■

In this example, the experimenter would also be interested in the pairwise comparisons of the casting methods. We consider this in Section 9.4.

EXERCISES

9.2.1. Consider the T-statistic that was derived through a likelihood ratio for testing the equality of the means of two normal distributions having common variance in Example 8.3.1. Show that T^2 is exactly the F-statistic of expression (9.2.11).

9.2.2. Under Model (9.2.1), show that the linear functions $X_{ij} - \overline{X}_{\cdot j}$ and $\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot}$ are uncorrelated.
Hint: Recall the definition of $\overline{X}_{\cdot j}$ and $\overline{X}_{\cdot\cdot}$ and, without loss of generality, we can let $E(X_{ij}) = 0$ for all i, j.

9.2.3. The following are observations associated with independent random samples from three normal distributions having equal variances and respective means μ_1, μ_2, μ_3.

I	II	III
0.5	2.1	3.0
1.3	3.3	5.1
−1.0	0.0	1.9
1.8	2.3	2.4
	2.5	4.2
		4.1

Using R or another statistical package, compute the F-statistic that is used to test $H_0 : \mu_1 = \mu_2 = \mu_3$.

9.2.4. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$. Show that

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=2}^{n}(X_i - \overline{X}')^2 + \frac{n-1}{n}(X_1 - \overline{X}')^2,$$

where $\overline{X} = \sum_{i=1}^{n} X_i/n$ and $\overline{X}' = \sum_{i=2}^{n} X_i/(n-1)$.
Hint: Replace $X_i - \overline{X}$ by $(X_i - \overline{X}') - (X_1 - \overline{X}')/n$. Show that $\sum_{i=2}^{n}(X_i - \overline{X}')^2/\sigma^2$ has a chi-square distribution with $n-2$ degrees of freedom. Prove that the two terms in the right-hand member are independent. What then is the distribution of

$$\frac{[(n-1)/n](X_1 - \overline{X}')^2}{\sigma^2}?$$

9.2.5. Using the notation of this section, assume that the means satisfy the condition that $\mu = \mu_1 + (b-1)d = \mu_2 - d = \mu_3 - d = \cdots = \mu_b - d$. That is, the last $b-1$ means are equal but differ from the first mean μ_1, provided that $d \neq 0$. Let independent random samples of size a be taken from the b normal distributions with common unknown variance σ^2.

(a) Show that the maximum likelihood estimators of μ and d are $\hat{\mu} = \overline{X}_{\cdot\cdot}$ and

$$\hat{d} = \frac{\sum_{j=2}^{b} \overline{X}_{\cdot j}/(b-1) - \overline{X}_{\cdot 1}}{b}.$$

(b) Using Exercise 9.2.4, find Q_6 and $Q_7 = c\hat{d}^2$ so that, when $d = 0$, Q_7/σ^2 is $\chi^2(1)$ and

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \overline{X}_{\cdot\cdot})^2 = Q_3 + Q_6 + Q_7.$$

(c) Argue that the three terms in the right-hand member of part (b), once divided by σ^2, are independent random variables with chi-square distributions, provided that $d = 0$.

(d) The ratio $Q_7/(Q_3 + Q_6)$ times what constant has an F-distribution, provided that $d = 0$? Note that this F is really the square of the two-sample T used to test the equality of the mean of the first distribution and the common mean of the other distributions, in which the last $b-1$ samples are combined into one.

9.2.6. On page 123 of their text, Kloke and McKean (2014) present the results of an experiment investigating 4 drugs (treatments) for their effect on lowering LDL (low density lipids) cholesterol. For the experimental design, 39 quail were randomly assigned to one of the 4 drugs. The drug was mixed in their food, but, other than this, the quail were all treated in the same way. After a specified period of time, the LDL level of each quail was determined. The first drug was a placebo, so the interest is to see if any other of the drugs resulted in lower LDL than the placebo. The data are in the file `quailldl.rda`. The first column of this matrix contains the drug indicator (1 through 4) for the quail while the second column contains the ldl level of that quail.

(a) Obtain comparison boxplots of LDL levels. Which drugs seem to result in lower LDL levels? Identify, by observation number, the outliers in the data.

(b) Compute the F-test that all mean levels of LDL are the same for all 4 drugs. Report the F-test statistic and p-value. Conclude in terms of the problem using the nominal significance level of 0.05. Use the R code in Example 9.2.1.

(c) Does your conclusion in Part (b) agree with the boxplots of Part (a)?

(d) Note that one assumption for the F-test is that the random errors e_{ij} in Model (9.2.1) are normally distributed. An estimate of e_{ij} is $x_{ij} - \overline{x}_{\cdot j}$. These are called **residuals**, i.e., what is left after the full model fit. Compute these residuals and then obtain a histogram, a boxplot, and a normal $q-q$ plot of them. Comment on the normality assumption. Use the code:

```
resd <- lm(quailmat[,2]~factor(quailmat[,1]))$resid
par(mfrow=c(2,2));hist(resd); boxplot(resd); qqnorm(resd)
```

9.2.7. Let μ_1, μ_2, μ_3 be, respectively, the means of three normal distributions with a common but unknown variance σ^2. In order to test, at the $\alpha = 5\%$ significance level, the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ against all possible alternative hypotheses, we take an independent random sample of size 4 from each of these distributions. Determine whether we accept or reject H_0 if the observed values from these three distributions are, respectively,

X_1 :	5	9	6	8
X_2 :	11	13	10	12
X_3 :	10	6	9	9

9.2.8. The driver of a diesel-powered automobile decided to test the quality of three types of diesel fuel sold in the area based on mpg. Test the null hypothesis that the three means are equal using the following data. Make the usual assumptions and take $\alpha = 0.05$.

Brand A:	38.7	39.2	40.1	38.9	
Brand B:	41.9	42.3	41.3		
Brand C:	40.8	41.2	39.5	38.9	40.3

9.3 Noncentral χ^2 and F-Distributions

Let $X_1, X_2, \ldots, X_n$ denote independent random variables that are $N(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, n$, and consider the quadratic form $Y = \sum_1^n X_i^2/\sigma^2$. If each μ_i is zero, we know that Y is $\chi^2(n)$. We shall now investigate the distribution of Y when each μ_i is not zero. The mgf of Y is given by

$$\begin{aligned} M(t) &= E\left[\exp\left(t\sum_{i=1}^{n}\frac{X_i^2}{\sigma^2}\right)\right] \\ &= \prod_{i=1}^{n} E\left[\exp\left(t\frac{X_i^2}{\sigma^2}\right)\right]. \end{aligned}$$

Consider

$$E\left[\exp\left(\frac{tX_i^2}{\sigma^2}\right)\right] = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{tx_i^2}{\sigma^2} - \frac{(x_i-\mu_i)^2}{2\sigma^2}\right] dx_i.$$

The integral exists if $t < \frac{1}{2}$. To evaluate the integral, note that

$$\frac{tx_i^2}{\sigma^2} - \frac{(x_i-\mu_i)^2}{2\sigma^2} = -\frac{x_i^2(1-2t)}{2\sigma^2} + \frac{2\mu_i x_i}{2\sigma^2} - \frac{\mu_i^2}{2\sigma^2}$$
$$= \frac{t\mu_i^2}{\sigma^2(1-2t)} - \frac{1-2t}{2\sigma^2}\left(x_i - \frac{\mu_i}{1-2t}\right)^2.$$

Accordingly, with $t < \frac{1}{2}$, we have

$$E\left[\exp\left(\frac{tX_i^2}{\sigma^2}\right)\right] = \exp\left[\frac{t\mu_i^2}{\sigma^2(1-2t)}\right]\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1-2t}{2\sigma^2}\left(x_i - \frac{\mu_i}{1-2t}\right)^2\right]dx_i.$$

If we multiply the integrand by $\sqrt{1-2t}$, $t < \frac{1}{2}$, we have the integral of a normal pdf with mean $\mu_i/(1-2t)$ and variance $\sigma^2/(1-2t)$. Thus

$$E\left[\exp\left(\frac{tX_i^2}{\sigma^2}\right)\right] = \frac{1}{\sqrt{1-2t}}\exp\left[\frac{t\mu_i^2}{\sigma^2(1-2t)}\right],$$

and the mgf of $Y = \sum_1^n X_i^2/\sigma^2$ is given by

$$M(t) = \frac{1}{(1-2t)^{n/2}}\exp\left[\frac{t\sum_1^n \mu_i^2}{\sigma^2(1-2t)}\right], \quad t < \frac{1}{2}. \tag{9.3.1}$$

A random variable that has the mgf

$$M(t) = \frac{1}{(1-2t)^{r/2}}e^{t\theta/(1-2t)}, \tag{9.3.2}$$

where $t < \frac{1}{2}$, $0 < \theta$, and r is a positive integer, is said to have a **noncentral chi-square distribution** with r degrees of freedom and noncentrality parameter θ. If one sets the noncentrality parameter $\theta = 0$, one has $M(t) = (1-2t)^{-r/2}$, which is the mgf of a random variable that is $\chi^2(r)$. Such a random variable can appropriately be called a **central chi-square variable**. We shall use the symbol $\chi^2(r,\theta)$ to denote a noncentral chi-square distribution that has the parameters r and θ; and we shall say that a random variable is $\chi^2(r,\theta)$ when that random variable has this kind of distribution. The symbol $\chi^2(r,0)$ is equivalent to $\chi^2(r)$. Thus our random variable $Y = \sum_1^n X_i^2/\sigma^2$ of this section is $\chi^2\left(n, \sum_1^n \mu_i^2/\sigma^2\right)$. The mean of Y is given by

$$E(Y) = \frac{1}{\sigma^2}\sum_{i=1}^{n}E(X_i^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\sigma^2+\mu_i^2) = n+\theta, \tag{9.3.3}$$

i.e., the mean of the central χ^2 plus the noncentrality parameter. If each μ_i is equal to zero, then Y is $\chi^2(n,0)$ or, more simply, Y is $\chi^2(n)$ with mean n.

The noncentral χ^2-variables, in which we have interest, are certain quadratic forms in normally distributed variables divided by a variance σ^2. In our example it is worth noting that the noncentrality parameter of $\sum_1^n X_i^2/\sigma^2$, which is

$\sum_1^n \mu_i^2/\sigma^2$, may be computed by replacing each X_i in the quadratic form by its mean μ_i, $i = 1, 2, \ldots, n$. This is no fortuitous circumstance; any quadratic form $Q = Q(X_1, \ldots, X_n)$ in normally distributed variables, which is such that Q/σ^2 is $\chi^2(r, \theta)$, has $\theta = Q(\mu_1, \mu_2, \ldots, \mu_n)/\sigma^2$; and if Q/σ^2 is a chi-square variable (central or noncentral) for certain real values of $\mu_1, \mu_2, \ldots, \mu_n$, it is chi-square (central or noncentral) for *all* real values of these means.

We next discuss the noncentral F-distribution. If U and V are independent and are, respectively, $\chi^2(r_1)$ and $\chi^2(r_2)$, the random variable F has been defined by $F = r_2U/r_1V$. Now suppose, in particular, that U is $\chi^2(r_1, \theta)$, V is $\chi^2(r_2)$, and U and V are independent. The distribution of the random variable r_2U/r_1V is called a **noncentral F-distribution** with r_1 and r_2 degrees of freedom with noncentrality parameter θ. Note that the noncentrality parameter of F is precisely the noncentrality parameter of the random variable U, which is $\chi^2(r_1, \theta)$. To obtain the expectation of F, use the $E(U)$ in expression (9.3.3) and the derivation of the expected value of a central F given in expression (3.6.8). These together immediately imply that

$$E(F) = \frac{r_2}{r_2 - 2}\left[\frac{r_1 + \theta}{r_1}\right], \tag{9.3.4}$$

provided, of course, that $r_2 > 2$. If $\theta > 0$ then the quantity in brackets exceeds one and, hence, the mean of the noncentral F exceeds the mean of the corresponding central F.

We next discuss the noncentral F distribution for the one-way ANOVA of the last section.

Example 9.3.1 (Noncentrality Parameter for One-way ANOVA)**.** Consider the one-way model with b levels, expression (9.2.1), with the hypotheses $H_0 : \mu_1 = \cdots = \mu_b$ versus $H_1 : \mu_j \neq \mu_{j'}$ for some $j \neq j'$. From expression (9.2.11), the F test statistic is $F = [Q_4/(b-1)]/[Q_3/(n-b)]$. In the denominator, the random variable Q_3/σ^2 is $\chi^2(n-b)$ under the full model and, hence, in particular, under H_1. It follows from Remark 9.8.3 of Section 9.8, though, that the distribution of Q_4/σ^2 is noncentral $\chi^2(b-1, \theta)$ under the full model. Recall that

$$Q_4/\sigma^2 = \frac{1}{\sigma^2}\sum_{j=1}^{b} n_j(\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot})^2.$$

Under the full model, $E(\overline{X}_{\cdot j}) = \mu_j$ and $E(\overline{X}_{\cdot\cdot}) = \sum_{j=1}^{b}(n_j/n)\mu_j$. Calling this last expectation $\overline{\mu}$, we have from the above discussion that

$$\theta = \frac{1}{\sigma^2}\sum_{j=1}^{b} n_j(\mu_j - \overline{\mu})^2. \tag{9.3.5}$$

If H_0 is true then $\mu_j \equiv \mu$, for some μ, and, hence, $\overline{\mu} = \mu$. Thus, under H_0, $\theta = 0$. Under H_1, there are distinct j and j' such that $\mu_j \neq \mu_{j'}$. In particular, then both μ_j and $\mu_{j'}$ cannot equal $\overline{\mu}$, so $\theta > 0$. Therefore, under H_1 the expectation of F exceeds the null expectation. ■

There are R commands that compute the cdf of noncentral χ^2 and F random variables. For example, suppose we want to compute $P(Y \leq y)$, where Y has a χ^2-distribution with `d` degrees of freedom and noncentrality parameter `b`. This probability is returned with the command `pchisq(y,d,b)`. The corresponding value of the pdf at y is computed by the command `dchisq(y,d,b)`. As another example, suppose we want $P(W \geq w)$, where W has an F-distribution with `n1` and `n2` degrees of freedom and noncentrality parameter `theta`. This is computed by the command `1-pf(w,n1,n2,theta)`, while the command `df(w,n1,n2,theta)` computes the value of the density of W at w. Tables of the noncentral chi-square and noncentral F-distributions are available in the literature also.

EXERCISES

9.3.1. Let $Y_i,\ i = 1, 2, \ldots, n$, denote independent random variables that are, respectively, $\chi^2(r_i, \theta_i),\ i = 1, 2, \ldots, n$. Prove that $Z = \sum_1^n Y_i$ is $\chi^2\left(\sum_1^n r_i, \sum_1^n \theta_i\right)$.

9.3.2. Compute the variance of a random variable that is $\chi^2(r, \theta)$.

9.3.3. Three different medical procedures (A, B, and C) for a certain disease are under investigation. For the study, $3m$ patients having this disease are to be selected and m are to be assigned to each procedure. This common sample size m must be determined. Let μ_1, μ_2, and μ_3, be the means of the response of interest under treatments A, B, and C, respectively. The hypotheses are: $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 : \mu_j \neq \mu_{j'}$ for some $j \neq j'$. To determine m, from a pilot study the experimenters use a guess of 30 of σ^2 and they select the significance level of 0.05. They are interested in detecting the pattern of means: $\mu_2 = \mu_1 + 5$ and $\mu_3 = \mu_1 + 10$.

(a) Determine the noncentrality parameter under the above pattern of means.

(b) Use the R function `pf` to determine the powers of the F-test to detect the above pattern of means for $m = 5$ and $m = 10$.

(c) Determine the smallest value of m so that the power of detection is at least 0.80.

(d) Answer (**a**)–(**c**) if $\sigma^2 = 40$.

9.3.4. Show that the square of a noncentral T random variable is a noncentral F random variable.

9.3.5. Let X_1 and X_2 be two independent random variables. Let X_1 and $Y = X_1 + X_2$ be $\chi^2(r_1, \theta_1)$ and $\chi^2(r, \theta)$, respectively. Here $r_1 < r$ and $\theta_1 \leq \theta$. Show that X_2 is $\chi^2(r - r_1, \theta - \theta_1)$.

9.4 Multiple Comparisons

For this section, consider the one-way ANONA model with b treatments as described in expression (9.2.1) of Section 9.2. In that section, we developed the F-test

of the hypotheses of equal means, (9.2.2). In practice, besides this test, statisticians usually want to make pairwise comparisons of the form $\mu_j - \mu_{j'}$. This is often called the **Second Stage Analysis**, while the F-test is consider the **First Stage Analysis**. The analysis for such comparisons usually consists of confidence intervals for the differences $\mu_j - \mu_{j'}$ and μ_j is **declared different** from $\mu_{j'}$ if 0 is not in the confidence interval. The random samples for treatments j and j' are: $X_{1j}, \ldots, X_{n_j j}$ from the $N(\mu_j, \sigma^2)$ distribution and $X_{1j'}, \ldots, X_{n_{j'} j'}$ from the $N(\mu_{j'}, \sigma^2)$ distribution, which are independent random samples. Based on these samples the estimator of $\mu_j - \mu_{j'}$ is $\overline{X}_{\cdot j} - \overline{X}_{\cdot j'}$. Further in the one-way analysis, an estimator of σ^2 is the full model estimator $\hat{\sigma^2}_\Omega$ defined in expression (9.2.7). As discussed in Section 9.2, $(n-b)\hat{\sigma^2}_\Omega/\sigma^2$ has a $\chi^2(n-b)$ distribution which is independent of all the sample means $\overline{X}_{\cdot j}$. Hence, for a specified α it follows as in (4.2.13) of Chapter 4 that

$$\overline{X}_{\cdot j} - \overline{X}_{\cdot j'} \pm t_{\alpha/2, n-b} \hat{\sigma}_\Omega \sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}} \tag{9.4.1}$$

is a $(1-\alpha)100\%$ confidence interval for $\mu_j - \mu_{j'}$.

We often want to make many pairwise comparisons, though. For example, the first treatment might be a placebo or represent the standard treatment. In this case, there are $b-1$ pairwise comparisons of interest. On the other hand, we may want to make all $\binom{b}{2}$ pairwise comparisons. In making so many comparisons, while each confidence interval, (9.4.1), has confidence $(1-\alpha)$, it would seem that the overall confidence diminishes. As we next show, this **slippage** of overall confidence is true. These problems are often called **Multiple Comparison Problems** (**MCP**). In this section, we present several MCP procedures.

Bonferroni Multiple Comparison Procedure

It is easy to motivate the **Bonferroni Procedure** while, at the same time, showing the slippage of confidence. This procedure is quite general and can be used in many settings not just the one-way design. So suppose we have k parameters θ_i with $(1-\alpha)100\%$ confidence intervals I_i, $i = 1, \ldots, k$, where $0 < \alpha < 1$ is given. Then the overall confidence is $P(\theta_1 \in I_1, \ldots, \theta_k \in I_k)$. Using the method of complements, DeMorgan's Laws, and Boole's inequality, expression (1.3.7) of Chapter 1, we have

$$\begin{aligned} P(\theta_1 \in I_1, \ldots, \theta_k \in I_k) &= 1 - P\left(\cup_{i=1}^k \theta_i \notin I_i\right) \\ &\geq 1 - \sum_{i=1}^k P(\theta_i \notin I_i) = 1 - k\alpha. \end{aligned} \tag{9.4.2}$$

The quantity $1 - k\alpha$ is the lower bound on the slippage of confidence. For example, if $k = 20$ and $\alpha = 0.05$ then the overall confidence may be 0. The Bonferroni procedure follows from expression (9.4.2). Simply change the confidence level of each confidence interval to $[1 - (\alpha/k)]$. Then the overall confidence is at least $1 - \alpha$.

For our one-way analysis, suppose we have k differences of interest. Then the

Bonferroni confidence interval for $\mu_j - \mu_{j'}$ is

$$\overline{X}_{\cdot j} - \overline{X}_{\cdot j'} \pm t_{\alpha/(2k),n-b}\hat{\sigma}_\Omega\sqrt{\frac{1}{n_j}+\frac{1}{n_{j'}}} \tag{9.4.3}$$

While the overall confidence of the Bonferroni procedure is at least $(1-\alpha)$, for a large number of comparisons, the lengths of its intervals are wide; i.e., a loss in precision. We offer two other procedures that, generally, lessen this effect.

The R function `mcpbon.R`[2] computes the Bonferroni procedure for all pairwise comparisons for a one-way design. The call is `mcpbon(y,ind,alpha=0.05)` where `y` is the vector of the combined samples and `ind` is the corresponding treatment vector. See Example 9.4.1 below.

Tukey's Multiple Comparison Procedure

To state **Tukey's procedure**, we first need to define the Studentized range distribution.

Definition 9.4.1. *Let $Y_1,\ldots,Y_k$ be iid $N(\mu,\sigma^2)$. Denote the range of these variables by $R = \max\{Y_i\} - \min\{Y_i\}$. Suppose mS^2/σ^2 has a $\chi^2(m)$ distribution which is independent of $Y_1,\ldots,Y_k$. Then we say that $Q = R/S$ has a* **Studentized range** *distribution with parameters k and m.* ■

The distribution of Q cannot be obtained in close form but packages such as R have functions that compute the cdf and quantiles. In R, the call `ptukey(x,k,m)` computes the cdf of Q at x, while the call `qtukey(p,k,m)` returns the pth quantile.

Consider the one-way design. First, assume that all the sample sizes are the same; i.e., for some positive integer a, $n_j = a$, for all $j = 1,\ldots,b$. Let $R = \text{Range}\{\overline{X}_{\cdot 1}-\mu_1,\ldots,\overline{X}_{\cdot b}-\mu_b\}$. Then since $\overline{X}_{\cdot 1}-\mu_1,\ldots,\overline{X}_{\cdot b}-\mu_b$ are iid $N(0,\sigma^2/a)$, the random variable $Q = R/(\hat{\sigma}_\Omega/\sqrt{a})$ has a Studentized range distribution with parameters b and $n-b$. Let $q_c = q_{1-\alpha,b,n-b}$.

$$\begin{aligned} 1-\alpha &= P(Q \le q_c) = P\left(\max\{\overline{X}_{\cdot j}-\mu_j\} - \min\{\overline{X}_{\cdot j}-\mu_j\} \le q_c\hat{\sigma}_\Omega/\sqrt{a}\right) \\ &= P\left(|(\mu_j-\mu_{j'}) - (\overline{X}_{\cdot j}-\overline{X}_{\cdot j'})| \le q_c\hat{\sigma}_\Omega/\sqrt{a}, \text{for all } j,j'\right) \end{aligned}$$

If we expand the inequality in the last statement, we obtain the $(1-\alpha)100\%$ simultaneous confidence intervals for all pairwise differences given by

$$\overline{X}_{\cdot j} - \overline{X}_{\cdot j'} \pm q_{1-\alpha,b,n-b}\frac{\hat{\sigma}_\Omega}{\sqrt{a}}, \quad \text{for all } j,j' \text{ in } 1,\ldots b. \tag{9.4.4}$$

The statistician John Tukey developed these simultaneous confidence intervals for the balanced case. For the unbalanced case, first write the error term in (9.4.4) as

$$\frac{q_{1-\alpha,b,n-b}}{\sqrt{2}}\hat{\sigma}_\Omega\sqrt{\frac{1}{a}+\frac{1}{a}}.$$

[2]Downloadable at the site listed in the Preface.

For the unbalanced case, this suggests the following intervals

$$\overline{X}_{\cdot j} - \overline{X}_{\cdot j'} \pm \frac{q_{1-\alpha,b,n-b}}{\sqrt{2}} \hat{\sigma}_\Omega \sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}, \quad \text{for all } j, j' \text{ in } 1, \ldots b. \tag{9.4.5}$$

This correction is due to Kramer and these intervals are often referred to as the Tukey-Kramer multiple comparison procedure; see Miller (1981) for discussion. These intervals do not have exact confidence $(1-\alpha)$ but studies have indicated that if the unbalance is not severe the confidence is close to $(1-\alpha)$; see Dunnett (1980). Corresponding R code is shown in Example 9.4.1.

Fisher's PLSD Multiple Comparison Procedure

The final procedure we discuss is **Fisher's Protected Least Significance Difference (PLSD)**. The setting is the general (unbalanced) one-way design (9.2.1). This procedure is a two-stage procedure. It can be used for an arbitrary umber of comparisons but we state it for all comparisons. For a specified level of significance α, Stage 1 consists of the F-test of the hypotheses of equal means, (9.2.2). If the test rejects at level α then Stage 2 consists of the usual pairwise $(1-\alpha)100\%$ confidence intervals, i.e.,

$$\overline{X}_{\cdot j} - \overline{X}_{\cdot j'} \pm t_{\alpha/2,n-b} \hat{\sigma}_\Omega \sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}, \quad \text{for all } j, j' \text{ in } 1, \ldots, b. \tag{9.4.6}$$

If the test in Stage 1 fails to reject, users sometimes perform Stage 2 using the Bonferroni procedure. Fisher's procedure does not have overall coverage $1-\alpha$, but the initial F-test offers protection. Simulation studies have shown that Fisher's procedure performs well in terms of power and level; see, for instance, Carmer and Swanson (1973) and McKean et al. (1989). The R function[3] `mcpfisher.R` computes this procedure as discussed in the next example.

Example 9.4.1 (Fast Cars)**.** Kitchens (1997) discusses an experiment concerning the speed of cars. Five cars are considered: Acura (1), Ferrari (2), Lotus (3), Porsche (4), and Viper (5). For each car, 6 runs were made, 3 in each direction. For each run, the speed recorded is the maximum speed on the run achieved without exceeding the engine's redline. The data are in the file `fastcars.rda`. Figure 9.4.1 displays the comparison boxplots of the speeds versus the cars, which shows clearly that there are differences in speed due to the car. Ferrari and Porsche seem to be the fastest but are the differences significant? We assume the one-way design (9.2.1) and use R to do the computations. Key commands and corresponding results are given next. The overall F-test of the hypotheses of equal means, (9.2.2), is quite significant: $F = 25.15$ with the p-value 0.0000. We selected the Tukey MCP at level 0.05. The command below returns all $\binom{5}{2} = 10$ pairwise comparisons, but in our summary we only list two.

```
###   Code assumes that fastcars.rda has been loaded in R
```

[3]Down loadable at the site listed in the Preface.

```
> fit <- lm(speed~factor(car))
> anova(fit)
###  F-Stat and p-value 25.145 1.903e-08
> aovfit <- aov(speed~factor(car))
> TukeyHSD(aovfit)

##   Tukey's procedures of all pairwise comparisons are computed.
##  Summary of a pertinent few
##       Cars           Mean-diff    LB CI       UB CI   Sig??
##  Porsche - Ferrari  -2.6166667  -9.0690855  3.835752  NS
##  Viper - Porsche    -7.7333333 -14.1857522 -1.280914  Sig.

##  Bonferroni
> mcpbon(speed,car)
##  Porsche - Ferrari  -2.6166667  -9.3795891  4.1462558 NS
##  Viper - Porsche    -7.7333333  -14.496255 -0.9704109 Sig.
2.197038 6.762922   0.9704109  14.49625578

##  Fisher
> mcpfisher(speed,car)
##  ftest 2.514542e+01 1.903360e-08
##  Porsche - Ferrari  -2.6166667  -7.141552   1.908219   NS
## Viper - Porsche    -7.7333333   -12.258219  -3.208448  Sig.
```

For discussion, we cite only two of Tukey's confidence intervals. As the second interval in the above printout shows, the mean speeds of both the Ferrari and Porsche are significantly faster than the mean speeds of the other cars. The difference between the Ferrari's and Porsche's mean speeds, though, is insignificant. Below the two Tukey confidence intervals, we display the results based on the Bonferroni and Fisher procedures. Note that all three procedures result in the same conclusions for these comparisons. The Bonferroni intervals are slightly larger than those of the Tukey procedure. The Fisher procedure gives the shortest intervals as expected. ■

In practice, the Tukey-Kramer procedure is often used, but there are many other multiple comparison procedures. A classical monograph on MCPs is Miller (1981) while Hus (1996) offers a more recent discussion.

EXERCISES

9.4.1. For the study discussed in Exercise 9.2.8, obtain the results of Bonferroni multiple comparison procedure using $\alpha = 0.10$. Based on this procedure, which brand of fuel if any is significantly best?

9.4.2. For the study discussed in Exercise 9.2.6, compute the Tukey-Kramer procedure. Are there any significant differences?

9.4.3. Suppose X and Y are discrete random variables that have the common range $\{1, 2, \ldots, k\}$. Let p_{1j} and p_{2j} be the respective probabilities $P(X = j)$ and

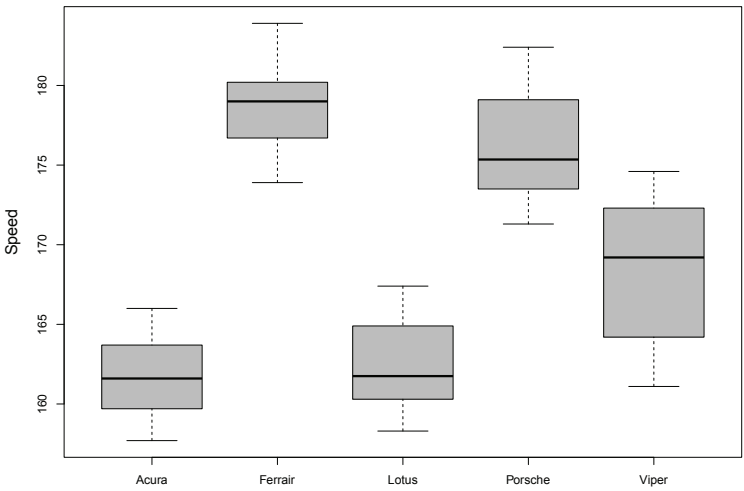

Figure 9.4.1: Boxplot of car speeds cited in Example 9.4.1.

$P(Y = j)$. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be respective independent random samples on X and Y. The samples are recorded in a $2 \times k$ contingency table of counts O_{ij}, where $O_{1j} = \#\{X_i = j\}$ and $O_{2j} = \#\{Y_i = j\}$. In Example 4.7.3, based on this table, we discussed a test that the distributions of X and Y are the same. Here we want to consider all the differences $p_{1j} - p_{2j}$ for $j = 1, \ldots, k$. Let $\hat{p}_{ij} = O_{ij}/n_i$.

(a) Determine the Bonferroni method for performing all these comparisons.

(b) Determine the Fisher method for performing all these comparisons.

9.4.4. Suppose the samples in Exercise 9.4.3 resulted in the contingency table:

	1	2	3	4	5	6	7	8	9	10
x	20	31	56	18	45	55	47	78	56	81
y	36	41	65	15	38	78	18	72	59	85

To compute (in R) the confidence intervals below, use the command `prop.test` as in Example 4.2.5.

(a) Based on the Bonferroni procedure for all 10 comparisons, compute the confidence interval for $p_{16} - p_{26}$.

(b) Based on the Fisher procedure for all 10 comparisons, compute the confidence interval for $p_{16} - p_{26}$.

9.4.5. Write an R function that computes the Fisher procedure of Exercise 9.4.3. Validate it using the data of Exercise 9.4.4.

9.4.6. Extend the Bonferroni procedure to simultaneous testing. That is, suppose we have m hypotheses of interest: H_{0i} versus H_{1i}, $i = 1, \ldots, m$. For testing H_{0i} versus H_{1i}, let $C_{i,\alpha}$ be a critical region of size α and assume H_{0i} is rejected if $\mathbf{X}_i \in C_{i,\alpha}$, for a sample $\mathbf{X}_i$. Determine a rule so that we can simultaneously test these m hypotheses with a Type I error rate less than or equal to α.

9.5 Two-Way ANOVA

Recall the one-way analysis of variance (ANOVA) problem considered in Section 9.2 which was concerned with one factor at b levels. In this section, we are concerned with the situation where we have two factors A and B with levels a and b, respectively. This is called a **two-way** analysis of variance (ANOVA). Let X_{ij}, $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, b$, denote the response for factor A at level i and factor B at level j. Denote the total sample size by $n = ab$. We shall assume that the X_{ij}s are independent normally distributed random variables with common variance σ^2. Denote the mean of X_{ij} by μ_{ij}. The mean μ_{ij} is often referred to as the mean of the (i, j)th cell. For our first model, we consider the **additive model** where

$$\mu_{ij} = \overline{\mu} + (\overline{\mu}_{i\cdot} - \overline{\mu}) + (\overline{\mu}_{\cdot j} - \overline{\mu}) \, ; \tag{9.5.1}$$

that is, the mean in the (i, j)th cell is due to additive effects of the levels, i of factor A and j of factor B, over the average (constant) $\overline{\mu}$. Let $\alpha_i = \overline{\mu}_{i\cdot} - \overline{\mu}$, $i = 1, \ldots, a$; $\beta_j = \overline{\mu}_{\cdot j} - \overline{\mu}$, $j = 1, \ldots, b$; and $\mu = \overline{\mu}$. Then the model can be written more simply as

$$\mu_{ij} = \mu + \alpha_i + \beta_j, \tag{9.5.2}$$

where $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$. We refer to this model as being a **two-way** additive ANOVA model.

For example, take $a = 2$, $b = 3$, $\mu = 5$, $\alpha_1 = 1$, $\alpha_2 = -1$, $\beta_1 = 1$, $\beta_2 = 0$, and $\beta_3 = -1$. Then the cell means are

		Factor B		
		1	2	3
Factor A	1	$\mu_{11} = 7$	$\mu_{12} = 6$	$\mu_{13} = 5$
	2	$\mu_{21} = 5$	$\mu_{22} = 4$	$\mu_{23} = 3$

Note that for each i, the plots of μ_{ij} versus j are parallel. This is true for additive models in general; see Exercise 9.5.9. We call these plots **mean profile plots**.

Had we taken $\beta_1 = \beta_2 = \beta_3 = 0$, then the cell means would be

		Factor B		
		1	2	3
Factor A	1	$\mu_{11} = 6$	$\mu_{12} = 6$	$\mu_{13} = 6$
	2	$\mu_{21} = 4$	$\mu_{22} = 4$	$\mu_{23} = 4$

The hypotheses of interest are

$$H_{0A} : \ \alpha_1 = \cdots = \alpha_a = 0 \text{ versus } H_{1A} : \ \alpha_i \neq 0, \text{ for some } i, \tag{9.5.3}$$

and

$$H_{0B}: \beta_1 = \cdots = \beta_b = 0 \text{ versus } H_{1B}: \beta_j \neq 0, \text{ for some } j. \tag{9.5.4}$$

If H_{0A} is true, then by (9.5.2) the mean of the (i,j)th cell does not depend on the level of A. The second example above is under H_{0B}. The cell means remain the same from column to column for a specified row. We call these hypotheses **main effect** hypotheses.

Remark 9.5.1. The model just described, and others similar to it, are widely used in statistical applications. Consider a situation in which it is desirable to investigate the effects of two factors that influence an outcome. Thus the variety of a grain and the type of fertilizer used influence the yield; or the teacher and the size of the class may influence the score on a standardized test. Let X_{ij} denote the yield from the use of variety i of a grain and type j of fertilizer. A test of the hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_b = 0$ would then be a test of the hypothesis that the mean yield of each variety of grain is the same regardless of the type of fertilizer used. ■

Call the model described around expression (9.5.2) the full model. We want to determine the mles. If we write out the likelihood function, the summation in the exponent of e is

$$SS = \sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij} - \overline{\mu} - \alpha_i - \beta_j)^2.$$

The mles of α_i, β_j, and $\overline{\mu}$ minimize SS. By adding in and subtracting out, we obtain:

$$SS = \sum_{i=1}^{a}\sum_{j=1}^{b}\{[\overline{x}_{\cdot\cdot} - \overline{\mu}] - [\alpha_i - (\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})] - [\beta_j - (\overline{x}_{\cdot j} - \overline{x}_{\cdot\cdot})] + [x_{ij} - \overline{x}_{i\cdot} - \overline{x}_{\cdot j} + \overline{x}_{\cdot\cdot}]\}^2. \tag{9.5.5}$$

From expression (9.5.2), we have $\sum_i \alpha_i = \sum_j \beta_j = 0$. Further,

$$\sum_{i=1}^{a}(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot}) = \sum_{j=1}^{b}(\overline{x}_{\cdot j} - \overline{x}_{\cdot\cdot}) = 0$$

and

$$\sum_{i=1}^{a}(x_{ij} - \overline{x}_{i\cdot} - \overline{x}_{\cdot j} + \overline{x}_{\cdot\cdot}) = \sum_{j=1}^{b}(x_{ij} - \overline{x}_{i\cdot} - \overline{x}_{\cdot j} + \overline{x}_{\cdot\cdot}) = 0.$$

Therefore, in the expansion of the sum of squares, (9.5.5), all cross product terms are 0. Hence, we have the identity

$$\begin{aligned} SS &= ab[\overline{x}_{\cdot\cdot} - \overline{\mu}]^2 + b\sum_{i=1}^{a}[\alpha_i - (\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})]^2 + a\sum_{j=1}^{b}[\beta_j - (\overline{x}_{\cdot j} - \overline{x}_{\cdot\cdot})]^2 \\ &\quad + \sum_{i=1}^{a}\sum_{j=1}^{b}[x_{ij} - \overline{x}_{i\cdot} - \overline{x}_{\cdot j} + \overline{x}_{\cdot\cdot}]^2. \end{aligned} \tag{9.5.6}$$

Since these are sums of squares, the minimizing values, (mles), must be

$$\hat{\overline{\mu}} = \overline{X}_{..},\ \hat{\alpha}_i = \overline{X}_{i\cdot} - \overline{X}_{..}, \text{ and } \hat{\beta}_j = \overline{X}_{\cdot j} - \overline{X}_{..}. \tag{9.5.7}$$

Note that we have used random variable notation. So these are the maximum likelihood estimators. It then follows that the maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2_\Omega = \frac{\sum_{i=1}^a \sum_{j=1}^b [X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{..}]^2}{ab} =_{\text{dfn}} \frac{Q'_3}{ab}, \tag{9.5.8}$$

where we have defined the numerator of $\hat{\sigma}^2_\Omega$ as the quadratic form Q'_3. It follows from an advanced course in linear models that $ab\hat{\sigma}^2_\Omega/\sigma^2$ has a $\chi^2((a-1)(b-1))$ distribution.

Next we construct the likelihood ratio test for H_{0B}. Under the reduced model (full model constrained by H_{0B}), $\beta_j = 0$ for all $j = 1, \ldots, b$. To obtain the mles for the reduced model, the identity (9.5.6) becomes

$$\begin{aligned} SS &= ab[\overline{x}_{..} - \overline{\mu}]^2 + b\sum_{i=1}^a [\alpha_i - (\overline{x}_{i\cdot} - \overline{x}_{..})]^2 \\ &\quad + a\sum_{j=1}^b [\overline{x}_{\cdot j} - \overline{x}_{..}]^2 + \sum_{i=1}^a \sum_{j=1}^b [x_{ij} - \overline{x}_{i\cdot} - \overline{x}_{\cdot j} + \overline{x}_{..}]^2. \end{aligned} \tag{9.5.9}$$

Thus the mles for α_i and $\overline{\mu}$ remain the same as in the full model and the reduced model maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2_\omega = \frac{\left\{a\sum_{j=1}^b [\overline{X}_{\cdot j} - \overline{X}_{..}]^2 + \sum_{i=1}^a \sum_{j=1}^b [X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{..}]^2\right\}}{ab}. \tag{9.5.10}$$

Denote the numerator of $\hat{\sigma}^2_\omega$ by Q'. Note that it is the **residual variation** left after fitting the reduced model.

Let Λ denote the likelihood ratio test statistic for H_{0B}. Our derivation is similar to the derivation for the likelihood ratio test statistic for one-way ANOVA of Section 9.2. Hence, similar to equation (9.2.9), our likelihood ratio test statistic simplifies to

$$\Lambda^{ab/2} = \frac{\hat{\sigma}^2_\Omega}{\hat{\sigma}^2_\omega} = \frac{Q'_3}{Q'}.$$

Then, similar to the one-way derivation, the likelihood ratio test rejects H_{0B} for large values of Q'_4/Q'_3, where in this case,

$$Q'_4 = a\sum_{j=1}^b [\overline{x}_{\cdot j} - \overline{x}_{..}]^2. \tag{9.5.11}$$

Note that $Q'_4 = Q' - Q'_3$; i.e., it is the incremental increase in residual variation if we use the reduced model instead of the full model.

To obtain the null distribution of Q'_4, notice that it is the numerator of the sample variance of the random variables $\sqrt{a}\overline{X}_{\cdot 1}, \ldots, \sqrt{a}\overline{X}_{\cdot b}$. These random variables are

independent with the common $N(\sqrt{a}\mu, \sigma^2)$ distribution; see Exercise 9.5.2. Hence, by Theorem 3.6.1, Q_4'/σ^2 has $\chi^2(b-1)$ distribution. In a more advanced course, it can be further shown that Q_4' and Q_3' are independent. Hence, the statistic

$$F_B = \frac{a\sum_{j=1}^{b}[\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot}]^2/(b-1)}{\sum_{i=1}^{a}\sum_{j=1}^{b}[X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{\cdot\cdot}]^2/(a-1)(b-1)} \tag{9.5.12}$$

has an $F(b-1,(a-1)(b-1))$ under H_{0B}. Thus, a level α test is to reject H_{0B} in favor of H_{1B} if

$$F_B \geq F(\alpha, b-1, (a-1)(b-1)). \tag{9.5.13}$$

If we are to compute the power function of the test, we need the distribution of F_B when H_{0B} is not true. As we have stated above, Q_3'/σ^2, (9.5.8), has a central χ^2-distribution with $(a-1)(b-1)$ degrees of freedom under the full model, and, hence, under H_{1B}. Further, it can be shown that Q_4', (9.5.11), has a noncentral χ^2-distribution with $b-1$ degrees of freedom under H_{1B}. To compute the noncentrality parameters of Q_4'/σ^2 when H_{1B} is true, we have $E(X_{ij}) = \mu + \alpha_i + \beta_j$, $E(\overline{X}_{i\cdot}) = \mu + \alpha_i$, $E(\overline{X}_{\cdot j}) = \mu + \beta_j$, and $E(\overline{X}_{\cdot\cdot}) = \mu$. Using the general rule discussed in Section 9.4, we replace the variables in Q_4'/σ^2 with their means. Accordingly, the noncentrality parameter Q_4'/σ^2 is

$$\frac{a}{\sigma^2}\sum_{j=1}^{b}(\mu + \beta_j - \mu)^2 = \frac{a}{\sigma^2}\sum_{j=1}^{b}\beta_j^2.$$

Thus, if the hypothesis H_{0B} is not true, F has a noncentral F-distribution with $b-1$ and $(a-1)(b-1)$ degrees of freedom and noncentrality parameter $a\sum_{j=1}^{b}\beta_j^2/\sigma^2$.

A similar argument can be used to construct the likelihood ratio test statistics F_A to test H_{0A} versus H_{1A}, (9.5.3). The numerator of the F test statistic is the sum of squares among rows. The test statistic is

$$F_A = \frac{b\sum_{i=1}^{a}[\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}]^2/(a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{b}[X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{\cdot\cdot}]^2/(a-1)(b-1)} \tag{9.5.14}$$

and it has an $F(a-1,(a-1)(b-1))$ distribution under H_{0A}.

9.5.1 Interaction between Factors

The analysis of variance problem that has just been discussed is usually referred to as a *two-way classification with one observation per cell*. Each combination of i and j determines a cell; thus, there is a total of ab cells in this model. Let us now investigate another two-way classification problem, but in this case we take $c > 1$ independent observations per cell.

Let X_{ijk}, $i = 1,2,\ldots,a$, $j = 1,2,\ldots,b$, and $k = 1,2,\ldots,c$, denote $n = abc$ random variables that are independent and have normal distributions with common, but unknown, variance σ^2. Denote the mean of each X_{ijk}, $k = 1,2,\ldots,c$, by μ_{ij}.

Under the additive model, (9.5.1), the mean of each cell depended on its row and column, but often the mean is cell-specific. To allow this, consider the parameters

$$\begin{aligned}\gamma_{ij} &= \mu_{ij} - \{\mu + (\overline{\mu}_{i\cdot} - \mu) + (\overline{\mu}_{\cdot j} - \mu)\} \\ &= \mu_{ij} - \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot j} + \mu,\end{aligned}$$

for $i = 1, \ldots a, j = 1, \ldots, b$. Hence γ_{ij} reflects the specific contribution to the cell mean over and above the additive model. These parameters are called **interaction parameters**. Using the second form (9.5.2), we can write the cell means as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \tag{9.5.15}$$

where $\sum_{i=1}^{a} \alpha_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$, and $\sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0$. This model is called a **two-way** model with interaction.

For example, take $a = 2$, $b = 3$, $\mu = 5$, $\alpha_1 = 1$, $\alpha_2 = -1$, $\beta_1 = 1$, $\beta_2 = 0$, $\beta_3 = -1$, $\gamma_{11} = 1$, $\gamma_{12} = 1$, $\gamma_{13} = -2$, $\gamma_{21} = -1$, $\gamma_{22} = -1$, and $\gamma_{23} = 2$. Then the cell means are

		Factor B		
		1	2	3
Factor A	1	$\mu_{11} = 8$	$\mu_{12} = 7$	$\mu_{13} = 3$
	2	$\mu_{21} = 4$	$\mu_{22} = 3$	$\mu_{23} = 5$

If each $\gamma_{ij} = 0$, then the cell means are

		Factor B		
		1	2	3
Factor A	1	$\mu_{11} = 7$	$\mu_{12} = 6$	$\mu_{13} = 5$
	2	$\mu_{21} = 5$	$\mu_{22} = 4$	$\mu_{23} = 3$

Note that the mean profile plots for this second example are parallel, but those in the first example (where interaction is present) are not.

The derivation of the mles under the full model, (9.5.15), is quite similar to the derivation for the additive model. Letting SS denote the sums of squares in the exponent of e in the likelihood function, we obtain the following identity by adding in and subtracting out (we have omitted subscripts on the sums):

$$\begin{aligned}SS &= \sum\sum\sum (x_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ijk})^2 \\ &= \sum\sum\sum \{[x_{ijk} - \overline{x}_{ij\cdot}] - [\mu - \overline{x}_{\cdots}] - [\alpha_i - (\overline{x}_{i\cdot\cdot} - \overline{x}_{\cdots})] - [\beta_j - (\overline{x}_{\cdot j\cdot} - \overline{x}_{\cdots})] \\ &\quad - [\gamma_{ij} - (\overline{x}_{ij\cdot} - \overline{x}_{i\cdot\cdot} - \overline{x}_{\cdot j\cdot} + \overline{x}_{\cdots})]\}^2 \\ &= \sum\sum\sum [x_{ijk} - \overline{x}_{ij\cdot}]^2 + abc[\mu - \overline{x}_{\cdots}]^2 + bc\sum [\alpha_i - (\overline{x}_{i\cdot\cdot} - \overline{x}_{\cdots})]^2 + \\ &\quad ac\sum [\beta_j - (\overline{x}_{\cdot j\cdot} - \overline{x}_{\cdots})]^2 + c\sum\sum [\gamma_{ij} - (\overline{x}_{ij\cdot} - \overline{x}_{i\cdot\cdot} - \overline{x}_{\cdot j\cdot} + \overline{x}_{\cdots})]^2 \end{aligned} \tag{9.5.16}$$

where, as in the additive model, the cross product terms in the expansion are 0. Thus, the mles of μ, α_i and β_j are the same as in the additive model; the mle of γ_{ij} is $\hat{\gamma}_{ij} = \overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X}_{\cdots}$; and the mle of σ^2 is

$$\hat{\sigma}^2_{\Omega} = \frac{\sum\sum\sum [X_{ijk} - \overline{X}_{ij\cdot}]^2}{abc}. \tag{9.5.17}$$

Let Q_3'' denote the numerator of $\hat{\sigma}^2$.

The major hypotheses of interest for the interaction model are

$$H_{0AB}: \gamma_{ij} = 0 \text{ for all } i, j \text{ versus } H_{1AB}: \gamma_{ij} \neq 0, \text{ for some } i, j. \tag{9.5.18}$$

Substituting $\gamma_{ij} = 0$ in SS, it is clear that the reduced model mle of σ^2 is

$$\hat{\sigma}^2_\omega = \frac{\sum\sum\sum[X_{ijk} - \overline{X}_{ij.}]^2 + c\sum\sum[\overline{X}_{ij.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}_{...}]^2}{abc}. \tag{9.5.19}$$

Let Q'' denote the numerator of $\hat{\sigma}^2_\omega$ and let $Q_4'' = Q'' - Q_3''$. Then it follows as in the additive model that the likelihood ratio test statistic rejects H_{0AB} for large values of Q_4''/Q_3''. In a more advanced class, it is shown that the standardized test statistic

$$F_{AB} = \frac{Q_4''/[(a-1)(b-1)]}{Q_3''/[ab(c-1)]} \tag{9.5.20}$$

has under H_{0AB} an F-distribution with $(a-1)(b-1)$ and $ab(c-1)$ degrees of freedom.

If $H_{0AB}: \gamma_{ij} = 0$ is accepted, then one usually continues to test $\alpha_i = 0,\ i = 1, 2, \ldots, a$, by using the test statistic

$$F = \frac{bc\sum_{i=1}^{a}(\overline{X}_{i..} - \overline{X}_{...})^2/(a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \overline{X}_{ij.})^2/[ab(c-1)]},$$

which has a null F-distribution with $a-1$ and $ab(c-1)$ degrees of freedom. Similarly, the test of $\beta_j = 0,\ j = 1, 2, \ldots, b$, proceeds by using the test statistic

$$F = \frac{ac\sum_{j=1}^{b}(\overline{X}_{.j.} - \overline{X}_{...})^2/(b-1)}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \overline{X}_{ij.})^2/[ab(c-1)]},$$

which has a null F-distribution with $b-1$ and $ab(c-1)$ degrees of freedom.

We conclude this section with an example that serves as an illustration of two-way ANOVA along with its associated R code.

Example 9.5.1. Devore (2012), page 435, presents a study concerning the effects to the thermal conductivity of an asphalt mix due to two factors: Binder Grade at three different levels (PG58, PG64, and PG70) and Coarseness of Aggregate Content at three levels (38%, 41%, and 44%). Hence, there are $3 \times 3 = 9$ different treatments. The responses are the thermal conductivities of the mixes of asphalt at these crossed levels. Two replications were performed at each treatment. The data are:

	Coarse Aggregate Content		
Binder-Grade	38%	41%	44%
PG58	0.835	0.822	0.785
	0.845	0.826	0.795
PG64	0.855	0.832	0.790
	0.865	0.836	0.800
PG70	0.815	0.800	0.770
	0.825	0.820	0.790

The data are also in the file `conductivity.rda`. Assuming this file has been loaded into the R work area, the mean profile plot is computed by

```
interaction.plot(Binder,Aggregate,Conductivity,legend=T)
```

and it is displayed in Figure 9.5.1. Note that the mean profiles are almost parallel, a graphical indication of little interaction between the factors. The ANOVA for the study is computed by the following two commands. It yields the tabled results (which we have abbreviated). The next to last column shows the F-test statistics discussed in this section.

```
fit=lm(Conductivity ~ factor(Binder) + factor(Aggregate) +
factor(Binder)*factor(Aggregate))
anova(fit)
Analysis of Variance Table
                                  Df    Sum Sq  F value    Pr(>F)
factor(Binder)                     2 0.0020893 14.1171  0.001678
factor(Aggregate)                  2 0.0082973  56.0631 8.308e-06
factor(Binder):factor(Aggregate)   4 0.0003253  1.0991  0.413558
```

As the interaction plot suggests, interaction is not significant ($p = 0.4135$). In practice, we would accept the additive (no interaction) model. The main effects are both highly significant. So both factors have an effect on conductivity. See Devore (2012) for more discussion. ■

EXERCISES

9.5.1. For the two-way interaction model, (9.5.15), show that the following decomposition of sums of squares is true:

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk}-\overline{X}_{...})^2 &= bc\sum_{i=1}^{a}(\overline{X}_{i..}-\overline{X}_{...})^2+ac\sum_{j=1}^{b}(\overline{X}_{.j.}-\overline{X}_{...})^2\\ &\quad+c\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{X}_{ij.}-\overline{X}_{i..}-\overline{X}_{.j.}+\overline{X}_{...})^2\\ &\quad+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk}-\overline{X}_{ij.})^2;\end{aligned}$$

that is, the total sum of squares is decomposed into that due to *row* differences, that due to *column* differences, that due to *interaction*, and that *within cells*.

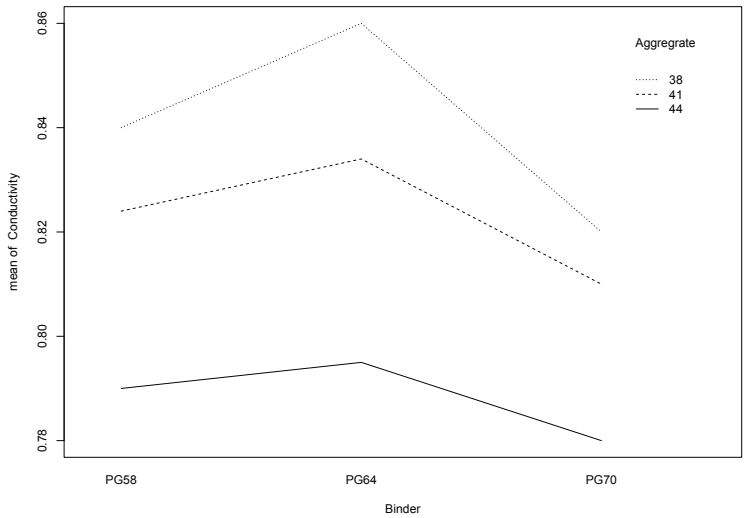

Figure 9.5.1: Mean profile plot for the study discussed in Example 9.5.1. The profiles are nearly parallel, indicating little interaction between the factors.

9.5.2. Consider the discussion above expression (9.5.14). Show that the random variables $\sqrt{a}\overline{X}_{\cdot 1}, \ldots, \sqrt{a}\overline{X}_{\cdot b}$ are independent with the common $N(\sqrt{a}\overline{\mu}, \sigma^2)$ distribution.

9.5.3. For the two-way interaction model, (9.5.15), show that the noncentrality parameter of the test statistic F_{AB} is equal to $c\sum_{j=1}^{b}\sum_{i=1}^{a}\gamma_{ij}^2/\sigma^2$.

9.5.4. Using the background of the two-way classification with one observation per cell, determine the distribution of the maximum likelihood estimators of α_i, β_j, and μ.

9.5.5. Prove that the linear functions $X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{\cdot\cdot}$ and $\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot}$ are uncorrelated, under the assumptions of this section.

9.5.6. Given the following observations associated with a two-way classification with $a = 3$ and $b = 4$, use R or another statistical package to compute the F-statistic used to test the equality of the column means ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$) and the equality of the row means ($\alpha_1 = \alpha_2 = \alpha_3 = 0$), respectively.

Row/Column	1	2	3	4
1	3.1	4.2	2.7	4.9
2	2.7	2.9	1.8	3.0
3	4.0	4.6	3.0	3.9

9.5.7. With the background of the two-way classification with $c > 1$ observations per cell, determine the distribution of the mles of α_i, β_j, and γ_{ij}.

9.5.8. Given the following observations in a two-way classification with $a = 3$, $b = 4$, and $c = 2$, compute the F-statistics used to test that all interactions are equal to zero ($\gamma_{ij} = 0$), all column means are equal ($\beta_j = 0$), and all row means are equal ($\alpha_i = 0$), respectively. Data are in the form x_{ijk}, i, j in the data set `sec951.rda`.

Row/Column	1	2	3	4
1	3.1	4.2	2.7	4.9
	2.9	4.9	3.2	4.5
2	2.7	2.9	1.8	3.0
	2.9	2.3	2.4	3.7
3	4.0	4.6	3.0	3.9
	4.4	5.0	2.5	4.2

9.5.9. For the additive model (9.5.1), show that the mean profile plots are parallel. The sample mean profile plots are given by plotting $\overline{X}_{ij\cdot}$ versus j, for each i. These offer a graphical diagnostic for interaction detection. Obtain these plots for the last exercise.

9.5.10. We wish to compare compressive strengths of concrete corresponding to $a = 3$ different drying methods (treatments). Concrete is mixed in batches that are just large enough to produce three cylinders. Although care is taken to achieve uniformity, we expect some variability among the $b = 5$ batches used to obtain the following compressive strengths. (There is little reason to suspect interaction, and hence only one observation is taken in each cell.) Data are also in the data set `sec95set2.rda`.

	Batch				
Treatment	B_1	B_2	B_3	B_4	B_5
A_1	52	47	44	51	42
A_2	60	55	49	52	43
A_3	56	48	45	44	38

(a) Use the 5% significance level and test $H_A : \alpha_1 = \alpha_2 = \alpha_3 = 0$ against all alternatives.

(b) Use the 5% significance level and test $H_B : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ against all alternatives.

9.5.11. With $a = 3$ and $b = 4$, find μ, α_i, β_j and γ_{ij} if μ_{ij}, for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$, are given by

$$\begin{matrix} 6 & 7 & 7 & 12 \\ 10 & 3 & 11 & 8 \\ 8 & 5 & 9 & 10 \end{matrix}$$

9.6 A Regression Problem

There is often interest in the relationship between two variables, for example, a student's scholastic aptitude test score in mathematics and this same student's

grade in calculus. Frequently, one of these variables, say x, is known in advance of the other and there is interest in predicting a future random variable Y. Since Y is a random variable, we cannot predict its future observed value $Y = y$ with certainty. Thus let us first concentrate on the problem of estimating the mean of Y, that is, $E(Y)$. Now $E(Y)$ is usually a function of x; for example, in our illustration with the calculus grade, say Y, we would expect $E(Y)$ to increase with increasing mathematics aptitude score x. Sometimes $E(Y) = \mu(x)$ is assumed to be of a given form, such as a linear or quadratic or exponential function; that is, $\mu(x)$ could be assumed to be equal to $\alpha + \beta x$ or $\alpha + \beta x + \gamma x^2$ or $\alpha e^{\beta x}$. To estimate $E(Y) = \mu(x)$, or equivalently the parameters α, β, and γ, we observe the random variable Y for each of n possible different values of x, say $x_1, x_2, \ldots, x_n$, which are not all equal. Once the n independent experiments have been performed, we have n pairs of known numbers $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. These pairs are then used to estimate the mean $E(Y)$. Problems like this are often classified under *regression* because $E(Y) = \mu(x)$ is frequently called a regression curve.

Remark 9.6.1. A model for the mean such as $\alpha + \beta x + \gamma x^2$ is called a **linear model** because it is linear in the parameters α, β, and γ. Thus $\alpha e^{\beta x}$ is not a linear model because it is not linear in α and β. Note that, in Sections 9.2 to 9.5, all the means were linear in the parameters and hence are linear models. ■

For the most part in this section, we consider the case in which $E(Y) = \mu(x)$ is a linear function. Denote by Y_i the response at x_i and consider the model

$$Y_i = \alpha + \beta(x_i - \overline{x}) + e_i, \quad i = 1, \ldots, n, \tag{9.6.1}$$

where $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ and $e_1, \ldots, e_n$ are iid random variables with a common $N(0, \sigma^2)$ distribution. Hence $E(Y_i) = \alpha + \beta(x_i - \overline{x})$, $\text{Var}(Y_i) = \sigma^2$, and Y_i has $N(\alpha + \beta(x_i - \overline{x}), \sigma^2)$ distribution. The major assumption is that the random errors, e_i, are iid. In particular, this means that the errors are not a function of the x_i's. This is discussed in Remark 9.6.3. First, we discuss the maximum likelihood estimates of the parameters α, β, and σ.

9.6.1 Maximum Likelihood Estimates

Assume that the n points $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ follow Model 9.6.1. So the first problem is that of fitting a straight line to the set of points; i.e., estimating α and β. As an aid to our discussion, Figure 9.6.1 shows a **scatterplot** of 60 observations $(x_1, y_1), \ldots, (x_{60}, y_{60})$ simulated from a linear model of the form (9.6.1). Our method of estimation in this section is that of maximum likelihood (mle). The joint pdf of $Y_1, \ldots, Y_n$ is the product of the individual probability density functions; that is, the likelihood function equals

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{[y_i - \alpha - \beta(x_i - \overline{x})]^2}{2\sigma^2} \right\} \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - \overline{x})]^2 \right\}. \end{aligned}$$

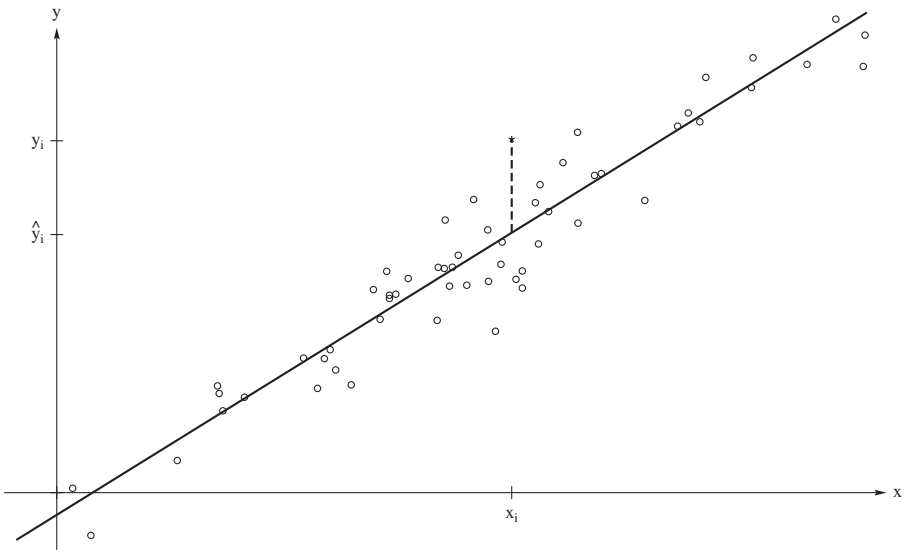

Figure 9.6.1: The plot shows the least squares fitted line (solid line) to a set of data. The dashed-line segment from $(x_i, \hat{y}_i)$ to (x_i, y_i) shows the deviation of (x_i, y_i) from its fit.

To maximize $L(\alpha, \beta, \sigma^2)$, or, equivalently, to minimize

$$-\log L(\alpha, \beta, \sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})]^2}{2\sigma^2},$$

we must select α and β to minimize

$$H(\alpha, \beta) = \sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})]^2.$$

Since $|y_i - \alpha - \beta(x_i - \overline{x})| = |y_i - \mu(x_i)|$ is the vertical distance from the point (x_i, y_i) to the line $y = \mu(x)$ (see the dashed-line segment in Figure 9.6.1), we note that $H(\alpha, \beta)$ represents the sum of the squares of those distances. Thus, selecting α and β so that the sum of the squares is minimized means that we are fitting the straight line to the data by the **method of least squares** (LS).

To minimize $H(\alpha, \beta)$, we find the two first partial derivatives,

$$\frac{\partial H(\alpha, \beta)}{\partial \alpha} = 2\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})](-1)$$

and

$$\frac{\partial H(\alpha, \beta)}{\partial \beta} = 2\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})][-(x_i - \overline{x})].$$

Setting $\partial H(\alpha, \beta)/\partial\alpha = 0$, we obtain

$$\sum_{i=1}^{n} y_i - n\alpha - \beta \sum_{i=1}^{n} (x_i - \overline{x}) = 0. \tag{9.6.2}$$

Since $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, the equation becomes $\sum_{i=1}^{n} y_i - n\alpha = 0$; hence, the mle of α is

$$\hat{\alpha} = \overline{Y}. \tag{9.6.3}$$

The equation $\partial H(\alpha, \beta)/\partial\beta = 0$ yields, with α replaced by $\overline{y}$,

$$\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) - \beta \sum_{i=1}^{n} (x_i - \overline{x})^2 = 0 \tag{9.6.4}$$

and, hence, the mle of β is

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} Y_i(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{9.6.5}$$

Equations (9.6.2) and (9.6.4) are the estimating equations for the LS solutions for this simple linear model.

The **fitted value** at the point (x_i, y_i) is given by

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}(x_i - \overline{x}), \tag{9.6.6}$$

which is shown on Figure 9.6.1. The fitted value $\hat{y}_i$ is also called the **predicted value** of y_i at x_i. The **residual** at the point (x_i, y_i) is given by

$$\hat{e}_i = y_i - \hat{y}_i, \tag{9.6.7}$$

which is also shown on Figure 9.6.1. Residual means "what is left" and the residual in regression is exactly that, i.e., what is left over after the fit. The relationship between the fitted values and the residuals are explored in Remark 9.6.3 and in Exercise 9.6.13.

To find the maximum likelihood estimator of σ^2, consider the partial derivative

$$\frac{\partial[-\log L(\alpha, \beta, \sigma^2)]}{\partial(\sigma^2)} = \frac{n}{2\sigma^2} - \frac{\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})]^2}{2(\sigma^2)^2}.$$

Setting this equal to zero and replacing α and β by their solutions $\hat{\alpha}$ and $\hat{\beta}$, we obtain

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{\alpha} - \hat{\beta}(x_i - \overline{x})]^2. \tag{9.6.8}$$

Of course, due to the invariance of mles, $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$. Note that in terms of the residuals, $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \hat{e}_i^2$. As shown in Exercise 9.6.13, the average of the residuals is 0.

Since $\hat{\alpha}$ is a linear function of independent and normally distributed random variables, $\hat{\alpha}$ has a normal distribution with mean

$$E(\hat{\alpha}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}[\alpha + \beta(x_i - \overline{x})] = \alpha$$

and variance

$$\text{var}(\hat{\alpha}) = \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 \text{var}(Y_i) = \frac{\sigma^2}{n}.$$

The estimator $\hat{\beta}$ is also a linear function of $Y_1, Y_2, \ldots, Y_n$ and hence has a normal distribution with mean

$$\begin{aligned} E(\hat{\beta}) &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})[\alpha + \beta(x_i - \overline{x})]}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \frac{\alpha\sum_{i=1}^{n}(x_i - \overline{x}) + \beta\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \beta \end{aligned}$$

and variance

$$\begin{aligned} \text{var}(\hat{\beta}) &= \sum_{i=1}^{n}\left[\frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]^2 \text{var}(Y_i) \\ &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{[\sum_{i=1}^{n}(x_i - \overline{x})^2]^2}\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \end{aligned}$$

In summary, the estimators $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of the independent normal random variables $Y_1, \ldots, Y_n$. In Exercise 9.6.4 it is further shown that the covariance between $\hat{\alpha}$ and $\hat{\beta}$ is zero. It follows that $\hat{\alpha}$ and $\hat{\beta}$ are independent random variables with a bivariate normal distribution; that is,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \text{ has a } N_2\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \sigma^2 \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{bmatrix}\right) \text{ distribution.} \tag{9.6.9}$$

Next, we consider the estimator of σ^2. It can be shown (Exercise 9.6.9) that

$$\begin{aligned} \sum_{i=1}^{n}[Y_i - \alpha - \beta(x_i - \overline{x})]^2 &= \sum_{i=1}^{n}\{(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)(x_i - \overline{x}) \\ &\quad + [Y_i - \hat{\alpha} - \hat{\beta}(x_i - \overline{x})]\}^2 \\ &= n(\hat{\alpha} - \alpha)^2 + (\hat{\beta} - \beta)^2\sum_{i=1}^{n}(x_i - \overline{x})^2 + n\hat{\sigma^2}, \end{aligned}$$

or for brevity,

$$Q = Q_1 + Q_2 + Q_3.$$

Here Q, Q_1, Q_2, and Q_3 are real quadratic forms in the variables

$$Y_i - \alpha - \beta(x_i - \overline{x}), \quad i = 1, 2, \ldots, n.$$

In this equation, Q represents the sum of the squares of n independent random variables that have normal distributions with means zero and variances σ^2. Thus Q/σ^2 has a χ^2 distribution with n degrees of freedom. Each of the random variables $\sqrt{n}(\hat{\alpha}-\alpha)/\sigma$ and $\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}(\hat{\beta}-\beta)/\sigma$ has a normal distribution with zero mean and unit variance; thus, each of Q_1/σ^2 and Q_2/σ^2 has a χ^2 distribution with 1 degree of freedom. In accordance with Theorem 9.9.2 (proved in Section 9.9), because Q_3 is nonnegative, we have that Q_1, Q_2, and Q_3 are independent and that Q_3/σ^2 has a χ^2 distribution with $n-1-1=n-2$ degrees of freedom. That is, $n\hat{\sigma}^2/\sigma^2$ has a χ^2 distribution with $n-2$ degrees of freedom.

We now extend this discussion to obtain inference for the parameters α and β. It follows from the above derivations that both the random variable T_1

$$T_1 = \frac{[\sqrt{n}(\hat{\alpha}-\alpha)]/\sigma}{\sqrt{Q_3/[\sigma^2(n-2)]}} = \frac{\hat{\alpha}-\alpha}{\sqrt{\hat{\sigma}^2/(n-2)}}$$

and the random variable T_2

$$T_2 = \frac{\left[\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}(\hat{\beta}-\beta)\right]\Big/\sigma}{\sqrt{Q_3/[\sigma^2(n-2)]}} = \frac{\hat{\beta}-\beta}{\sqrt{n\hat{\sigma}^2/[(n-2)\sum_1^n(x_i-\overline{x})^2]}} \tag{9.6.10}$$

have a t-distribution with $n-2$ degrees of freedom. These facts enable us to obtain confidence intervals for α and β; see Exercise 9.6.5. The fact that $n\hat{\sigma}^2/\sigma^2$ has a χ^2 distribution with $n-2$ degrees of freedom provides a means of determining a confidence interval for σ^2. These are some of the statistical inferences about the parameters to which reference was made in the introductory remarks of this section.

Remark 9.6.2. The more discerning reader should quite properly question our construction of T_1 and T_2 immediately above. We know that the *squares* of the linear forms are independent of $Q_3 = n\hat{\sigma}^2$, but we do not know, at this time, that the linear forms themselves enjoy this independence. A more general result is obtained in Theorem 9.9.1 of Section 9.9 and the present case is a special instance. ■

Before considering a numerical example, we discuss a diagnostic plot for the major assumption of Model 9.6.1.

Remark 9.6.3 (Diagnostic Plot Based on Fitted Values and Residuals)**.** The major assumption in the model is that the random errors $e_1, \ldots, e_n$ are iid. In particular, this means that the errors are not a function of the x_i's so that a plot of e_i versus $\alpha+\beta(x_i-\overline{x})$ should result in a random scatter. Since the errors and the parameters are unknown this plot is not possible. We have estimates, though, of these quantities, namely the residuals $\hat{e}_i$ and the fitted values $\hat{y}_i$. A diagnostic for the assumption is to plot the residuals versus the fitted values. This is called the **residual plot**. If the plot results in a random scatter, it is an indication that the model is appropriate. Patterns in the plot, though, are indicative of a poor model. Often in this later case, the patterns in the plot lead to better models. ■

As a final note, in Model 9.6.1 we have centered the x's; i.e., subtracted $\overline{x}$ from x_i. In practice, usually we do not precenter the x's. Instead, we fit the model $y_i = \alpha^* + \beta x_i + e_i$. In this case, the least squares, and hence, mles minimize the sum of squares

$$\sum_{i=1}^{n}(y_i - \alpha^* - \beta x_i)^2. \tag{9.6.11}$$

In Exercise 9.6.1, the reader is asked to show that the estimate of β remains the same as in expression (9.6.5), while $\hat{\alpha}^* = \overline{y} - \hat{\beta}\overline{x}$. We use this noncentered model in the following example.

Example 9.6.1 (Men's 1500 meters)**.** As a numerical illustration, consider data drawn from the Olympics. The response of interest is the winning time of the men's 1500 meters, while the predictor is the year of the olympics. The data were taken from Wikipedia and can be found in `olym1500mara.rda`. Assume the R vectors for the winning times and year are `time` and `year`, respectively. There are $n = 27$ data points. The top panel of Figure 9.6.2 shows a scatterplot of the data that is computed by the R command

```
par(mfrow=c(2,1));plot(time~year,xlab="Year",ylab="Winning time")
```

The winning times are steadily decreasing over time and, based on this plot, a simple linear model seems reasonable. Obviously the time for 2016 is an outlier but it is the correct time. Before proceeding to inference, though, we check the quality of the fit of the model. The following R commands obtain the least squares fit, overlaying it on the scatterplot in Figure 9.6.2, the fitted values, and the residuals. These are used to obtain the residual plot that is displayed in the bottom panel of 9.6.2.

```
fit <- lm(time~year); abline(fit)
ehat <- fit$resid; yhat <- fit$fitted.values
plot(ehat~yhat,xlab="Fitted values",ylab="Residuals")
```

Recall a "good" fit is indicated by a random scatter in the residual plot. This does not appear to be the case. There is a dependence[4] between adjacent points over time. This dependence is apparent from the scatterplot too. In a time series course, this dependence would be investigated.

Based on the dependence, the following inference is approximate. The command `summary(fit)` produces the table of coefficients:

```
             Estimate Std. Error t value Pr(>t|) |
(Intercept) 12.325411   1.039402  11.858 9.26e-12
year        -0.004376   0.000530  -8.257 1.31e-08
```

Hence, the prediction equation is $\hat{y} = 12.33 - .0044$year. Based on the slope estimate, we predict the winning time to drop by 0.004 minutes every year. For a 95% confidence interval for the slope, the t-critical value via R is `qt(.975,25)` which computes to 2.060. Using the standard error in the summary table, the following R commands compute confidence interval for the slope parameter:

```
err=0.000530*2.060;lb=-0.004376-err;ub=-0.004376+err;ci=c(lb,ub)
```

[4]This dependence is not surprising. The runners race against each other but they also try to beat the Olympic record.

```
ci; -0.0054678 -0.0032842
```

So with approximate confidence 95%, we estimate the drop in winning time to between 0.0032 to 0.0055 minutes per year.

Based on the fit, the predicted winning time for the men's 1500 meters in the 2020 Olympics is

$$\hat{y} = 12.325411 - 0.004376(2020) = 3.486. \tag{9.6.12}$$

Exercise 9.6.8 provides an estimate (predictive interval) of error for this prediction. ■

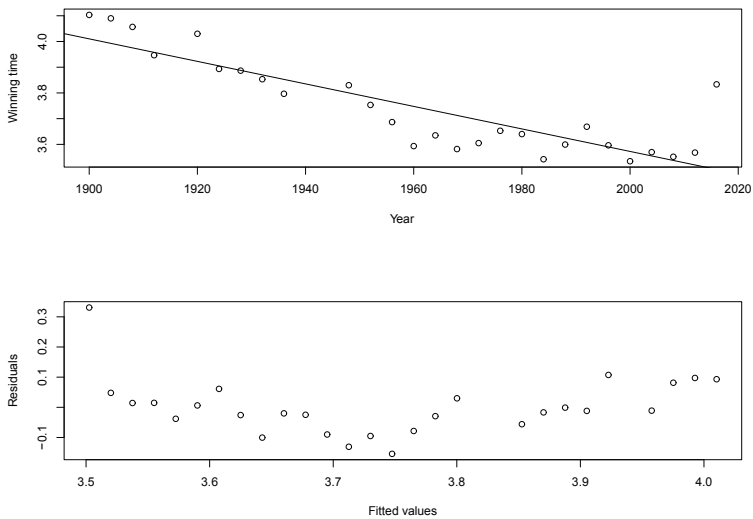

Figure 9.6.2: The top panel is the scatterplot of winning times in the men's 1500 meters versus the year of the Olympics. The least squares fit is overlaid. The bottom panel is the residual plot of the fit.

9.6.2 *Geometry of the Least Squares Fit

In the modern literature, linear models are usually expressed in terms of matrices and vectors, which we briefly introduce in this example. Furthermore, this allows us to discuss the simple geometry behind the least squares fit. Consider then Model (9.6.1). Write the vectors $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\mathbf{e} = (e_1, \ldots, e_n)'$, and $\mathbf{x}_c = (x_1 - \overline{x}, \ldots, x_n - \overline{x})'$. Let $\mathbf{1}$ denote the $n \times 1$ vector whose components are all 1. Then

Model (9.6.1) can be expressed equivalently as

$$\begin{aligned} \mathbf{Y} &= \alpha\mathbf{1} + \beta\mathbf{x}_c + \mathbf{e} \\ &= [\mathbf{1}\ \mathbf{x}_c]\begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \mathbf{e} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \end{aligned} \tag{9.6.13}$$

where $\mathbf{X}$ is the $n \times 2$ matrix with columns $\mathbf{1}$ and $\mathbf{x}_c$ and $\boldsymbol{\beta} = (\alpha, \beta)'$. Next, let $\boldsymbol{\theta} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$. Finally, let V be the two-dimensional subspace of R^n spanned by the columns of $\mathbf{X}$; i.e., V is the range of the matrix $\mathbf{X}$. Hence we can also express the model succinctly as

$$\mathbf{Y} = \boldsymbol{\theta} + \mathbf{e}, \quad \boldsymbol{\theta} \in V. \tag{9.6.14}$$

Hence, except for the random error vector $\mathbf{e}$, $\mathbf{Y}$ would lie in V. It makes sense intuitively then, as suggested by Figure 9.6.3, to estimate $\boldsymbol{\theta}$ by the vector in V that is "closest" (in Euclidean distance) to $\mathbf{Y}$, that is, by $\hat{\boldsymbol{\theta}}$, where

$$\hat{\boldsymbol{\theta}} = \text{Argmin}_{\boldsymbol{\theta} \in V} \|\mathbf{Y} - \boldsymbol{\theta}\|^2, \tag{9.6.15}$$

where the square of the **Euclidean norm** is given by $\|\mathbf{u}\|^2 = \sum_{i=1}^n u_i^2$, for $\mathbf{u} \in R^n$. As shown in Exercise 9.6.13 and depicted on the plot in Figure 9.6.3, $\hat{\boldsymbol{\theta}} = \hat{\alpha}\mathbf{1} + \hat{\beta}\mathbf{x}_c$, where $\hat{\alpha}$ and $\hat{\beta}$ are the least squares estimates given above. Also, the vector $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\boldsymbol{\theta}}$ is the vector of residuals and $n\hat{\sigma}^2 = \|\hat{\mathbf{e}}\|^2$. Also, just as depicted in Figure 9.6.3, the angle between the vectors $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{e}}$ is a right angle. In linear models, we say that $\hat{\boldsymbol{\theta}}$ is the projection of $\mathbf{Y}$ onto the subspace V.

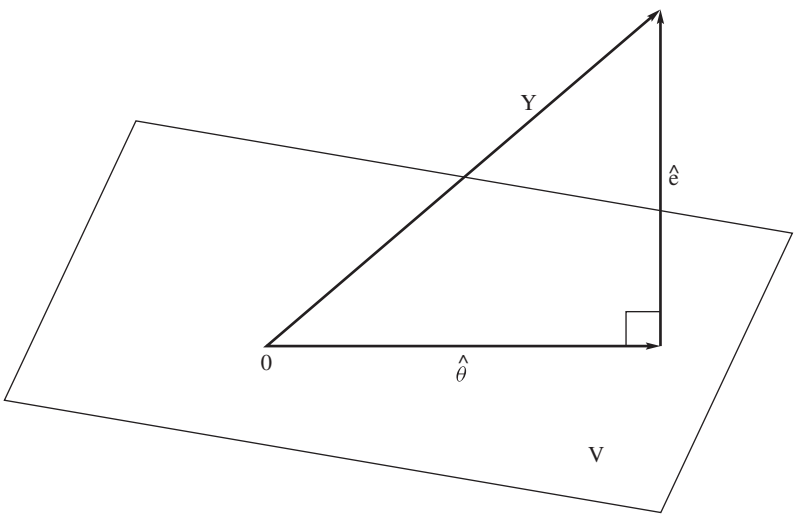

Figure 9.6.3: The sketch shows the geometry of least squares. The vector of responses is $\mathbf{Y}$, the fit is $\widehat{\boldsymbol{\theta}}$, and the vector of residuals is $\hat{\mathbf{e}}$.

EXERCISES

9.6.1. Obtain the least squares estimates for the model $y_i = \alpha^* + \beta x_i + e_i$ by minimizing the sum of squares given in expression (9.6.11). Determine the distribution of $\hat{\alpha}^*$.

9.6.2. Students' scores on the mathematics portion of the ACT examination, x, and on the final examination in the first-semester calculus (200 points possible), y, are:

x	25	20	26	26	28	28	29	32	20	25
y	138	84	104	112	88	132	90	183	100	143
x	26	28	25	31	30					
y	141	161	124	118	168					

The data are also in the rda file `regr1.rda`. Use R or another statistical package for computation and plotting.

(a) Calculate the least squares regression line for these data.

(b) Plot the points and the least squares regression line on the same graph.

(c) Obtain the residual plot and comment on the appropriateness of the model.

(d) Find 95% confidence interval for β under the usual assumptions. Comment in terms of the problem.

9.6.3 (Telephone Data)**.** Consider the data presented below. The responses (y) for this data set are the numbers of telephone calls (tens of millions) made in Belgium for the years 1950 through 1973. Time, the years, serves as the predictor variable (x). The data are discussed on page 172 of Hettmansperger and McKean (2011) and are in the file `telephone.rda`.

Year	50	51	52	53	54	55
No. Calls	0.44	0.47	0.47	0.59	0.66	0.73
Year	56	57	58	59	60	61
No. Calls	0.81	0.88	1.06	1.20	1.35	1.49
Year	62	63	64	65	66	67
No. Calls	1.61	2.12	11.90	12.40	14.20	15.90
Year	68	69	70	71	72	73
No. Calls	18.20	21.20	4.30	2.40	2.70	2.90

(a) Calculate the least squares regression line for these data.

(b) Plot the points and the least squares regression line on the same graph.

(c) What is the reason for the poor least squares fit?

9.6.4. Show that the covariance between $\hat{\alpha}$ and $\hat{\beta}$ is zero.

9.6.5. Find $(1-\alpha)100\%$ confidence intervals for the parameters α and β in Model (9.6.1).

9.6.6. Consider Model (9.6.1). Let $\eta_0 = E(Y|x = x_0 - \overline{x})$. The least squares estimator of η_0 is $\hat{\eta}_0 = \hat{\alpha} + \hat{\beta}(x_0 - \overline{x})$.

(a) Using (9.6.9), show that $\hat{\eta}_0$ is an unbiased estimator and show that its variance is given by

$$V(\hat{\eta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_1 - \overline{x})^2} \right]$$

(b) Obtain the distribution of $\hat{\eta}_0$ and use it to determine a $(1-\alpha)100\%$ confidence interval for η_0.

9.6.7. Assume that the sample $(x_1, Y_1), \ldots, (x_n, Y_n)$ follows the linear model (9.6.1). Suppose Y_0 is a future observation at $x = x_0 - \overline{x}$ and we want to determine a predictive interval for it. Assume that the model (9.6.1) holds for Y_0; i.e., Y_0 has a $N(\alpha + \beta(x_0 - \overline{x}), \sigma^2)$ distribution. We use $\hat{\eta}_0$ of Exercise 9.6.6 as our prediction of Y_0.

(a) Obtain the distribution of $Y_0 - \hat{\eta}_0$, showing that its variance is:

$$V(Y_0 - \hat{\eta}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_1 - \overline{x})^2} \right]$$

Use the fact that the future observation Y_0 is independent of the sample $(x_1, Y_1), \ldots, (x_n, Y_n)$.

(b) Determine a t-statistic with numerator $Y_0 - \hat{\eta}_0$.

(c) Now beginning with $1 - \alpha = P[-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}]$, where $0 < \alpha < 1$, determine a $(1 - \alpha)100\%$ predictive interval for Y_0.

(d) Compare this predictive interval with the confidence interval obtained in Exercise 9.6.6. Intuitively, why is the predictive interval larger?

9.6.8. In Example 9.6.1, we obtain the predicted winning time for the men's 1500 meters in the 2020 Olympics. Compute the 95% predictive interval for this prediction that is given in the last exercise. These computations are performed by the R function `cipi.R`. The call is `cipi(lm(time~year),matrix(c(1,2020),ncol=2))`. In terms of the problem, what does this predictive interval mean? Next compute the prediction for the 2024 and 2028 Olympics. Why are the intervals increasing in length?

9.6.9. Show that

$$\sum_{i=1}^n [Y_i - \alpha - \beta(x_i - \overline{x})]^2 = n(\hat{\alpha} - \alpha)^2 + (\hat{\beta} - \beta)^2 \sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{i=1}^n [Y_i - \hat{\alpha} - \hat{\beta}(x_i - \overline{x})]^2.$$

9.6.10. Let the independent random variables $Y_1, Y_2, \ldots, Y_n$ have, respectively, the probability density functions $N(\beta x_i, \gamma^2 x_i^2)$, $i = 1, 2, \ldots, n$, where the given numbers $x_1, x_2, \ldots, x_n$ are not all equal and no one is zero. Find the maximum likelihood estimators of β and γ^2.

9.6.11. Let the independent random variables $Y_1, \ldots, Y_n$ have the joint pdf

$$L(\alpha, \beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n [y_i - \alpha - \beta(x_i - \overline{x})]^2\right\},$$

where the given numbers $x_1, x_2, \ldots, x_n$ are not all equal. Let $H_0 : \beta = 0$ (α and σ^2 unspecified). It is desired to use a likelihood ratio test to test H_0 against all possible alternatives. Find Λ and see whether the test can be based on a familiar statistic.
Hint: In the notation of this section, show that

$$\sum_1^n (Y_i - \hat{\alpha})^2 = Q_3 + \widehat{\beta}^2 \sum_1^n (x_i - \overline{x})^2.$$

9.6.12. Using the notation of Section 9.2, assume that the means μ_j satisfy a linear function of j, namely, $\mu_j = c + d[j - (b+1)/2]$. Let independent random samples of size a be taken from the b normal distributions having means $\mu_1, \mu_2, \ldots, \mu_b$, respectively, and common unknown variance σ^2.

(a) Show that the maximum likelihood estimators of c and d are, respectively, $\hat{c} = \overline{X}_{\cdot\cdot}$ and

$$\hat{d} = \frac{\sum_{j=1}^b [j - (b-1)/2](\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot})}{\sum_{j=1}^b [j - (b+1)/2]^2}.$$

(b) Show that

$$\begin{aligned}\sum_{i=1}^a \sum_{j=1}^b (X_{ij} - \overline{X}_{\cdot\cdot})^2 &= \sum_{i=1}^a \sum_{j=1}^b \left[X_{ij} - \overline{X}_{\cdot\cdot} - \hat{d}\left(j - \frac{b+1}{2}\right)\right]^2 \\ &\quad + \hat{d}^2 \sum_{j=1}^b a\left(j - \frac{b+1}{2}\right)^2.\end{aligned}$$

(c) Argue that the two terms in the right-hand member of part (b), once divided by σ^2, are independent random variables with χ^2 distributions provided that $d = 0$.

(d) What F-statistic would be used to test the equality of the means, that is, $H_0 : d = 0$?

9.6.13. Consider the discussion in Section 9.6.2.

(a) Show that $\hat{\boldsymbol{\theta}} = \hat{\alpha}\mathbf{1} + \hat{\beta}\mathbf{x}_c$, where $\hat{\alpha}$ and $\hat{\beta}$ are the least squares estimators derived in this section.

(b) Show that the vector $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\boldsymbol{\theta}}$ is the vector of residuals; i.e., its ith entry is $\hat{e}_i$, (9.6.7).

(c) As depicted in Figure 9.6.3, show that the angle between the vectors $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{e}}$ is a right angle.

(d) Show that the residuals sum to zero; i.e., $\mathbf{1}'\hat{\mathbf{e}} = 0$.

9.6.14. Fit $y = a + x$ to the data

x	0	1	2
y	1	3	4

by the method of least squares.

9.6.15. Fit by the method of least squares the plane $z = a + bx + cy$ to the five points $(x, y, z) : (-1, -2, 5), (0, -2, 4), (0, 0, 4), (1, 0, 2), (2, 1, 0)$.
Let the R vectors `x,y,z` contain the values for x, y, and z. Then the LS fit is computed by `lm(z ~ x + y)`.

9.6.16. Let the 4×1 matrix $\boldsymbol{Y}$ be multivariate normal $N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, where the 4×3 matrix $\boldsymbol{X}$ equals

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & -1 & 2 \\ 1 & 0 & -3 \\ 1 & 0 & -1 \end{bmatrix}$$

and $\boldsymbol{\beta}$ is the 3×1 regression coefficient matrix.

(a) Find the mean matrix and the covariance matrix of $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

(b) If we observe $\boldsymbol{Y}'$ to be equal to $(6, 1, 11, 3)$, compute $\hat{\boldsymbol{\beta}}$.

9.6.17. Suppose $\boldsymbol{Y}$ is an $n \times 1$ random vector, $\boldsymbol{X}$ is an $n \times p$ matrix of known constants of rank p, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. Let $\boldsymbol{Y}$ have a $N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ distribution. Obtain the pdf of $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

9.6.18. Let the independent normal random variables $Y_1, Y_2, \ldots, Y_n$ have, respectively, the probability density functions $N(\mu, \gamma^2 x_i^2)$, $i = 1, 2, \ldots, n$, where the given $x_1, x_2, \ldots, x_n$ are not all equal and no one of which is zero. Discuss the test of the hypothesis $H_0 : \gamma = 1$, μ unspecified, against all alternatives $H_1 : \gamma \neq 1$, μ unspecified.

9.7 A Test of Independence

Let X and Y have a bivariate normal distribution with means μ_1 and μ_2, positive variances σ_1^2 and σ_2^2, and correlation coefficient ρ. We wish to test the hypothesis that X and Y are independent. Because two jointly normally distributed random variables are independent if and only if $\rho = 0$, we test the hypothesis $H_0 : \rho = 0$ against the hypothesis $H_1 : \rho \neq 0$. A likelihood ratio test is used. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ denote a random sample of size $n > 2$ from the

bivariate normal distribution; that is, the joint pdf of these $2n$ random variables is given by

$$f(x_1, y_1)f(x_2, y_2)\cdots f(x_n, y_n).$$

Although it is fairly difficult to show, the statistic that is defined by the likelihood ratio Λ is a function of the statistic, which is the mle of ρ, namely,

$$R = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}. \tag{9.7.1}$$

This statistic R is called the sample **correlation coefficient** of the random sample. Following the discussion after expression (5.4.5), the statistic R is a consistent estimate of ρ; see Exercise 9.7.5. The likelihood ratio principle, which calls for the rejection of H_0 if $\Lambda \leq \lambda_0$, is equivalent to the computed value of $|R| \geq c$. That is, if the absolute value of the correlation coefficient of the sample is too large, we reject the hypothesis that the correlation coefficient of the distribution is equal to zero. To determine a value of c for a satisfactory significance level, it is necessary to obtain the distribution of R, or a function of R, when H_0 is true, as we outline next.

Let $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, $n > 2$, where $x_1, x_2, \ldots, x_n$ and $\overline{x} = \sum_1^n x_i/n$ are fixed numbers such that $\sum_1^n (x_i - \overline{x})^2 > 0$. Consider the conditional pdf of $Y_1, Y_2, \ldots, Y_n$ given that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Because $Y_1, Y_2, \ldots, Y_n$ are independent and, with $\rho = 0$, are also independent of $X_1, X_2, \ldots, X_n$, this conditional pdf is given by

$$\left(\frac{1}{\sqrt{2\pi}\sigma_2}\right)^n \exp\left\{-\frac{1}{2\sigma_2^2}\sum_1^n (y_i - \mu_2)^2\right\}.$$

Let R_c be the correlation coefficient, given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, so that

$$\frac{R_c\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{9.7.2}$$

is $\hat{\beta}$, expression (9.6.5) of Section 9.6. Conditionally the mean of Y_i is μ_2; i.e., a constant. So here expression (9.7.2) has expectation 0 which implies that $E(R_c) = 0$. Next consider the t-ratio of $\hat{\beta}$ given by T_2 of expression (9.6.10) of Section 9.6. In this notation T_2 can be expressed as

$$T_2 = \frac{R_c\sqrt{\sum(Y_i - \overline{Y})^2}/\sqrt{\sum(x_i - \overline{x})^2}}{\sqrt{\frac{\sum_{i=1}^n \left\{Y_i - \overline{Y} - \left[R_c\sqrt{\sum_{j=1}^n (Y_j - \overline{Y})^2}/\sqrt{\sum_{j=1}^n (x_j - \overline{x})^2}\right](x_i - \overline{x})\right\}^2}{(n-2)\sum_{j=1}^n (x_j - \overline{x})^2}}} = \frac{R_c\sqrt{n-2}}{\sqrt{1 - R_c^2}}. \tag{9.7.3}$$

Thus T_2, given $X_1 = x_1, \ldots, X_n = x_n$, has a conditional t-distribution with $n-2$ degrees of freedom. Note that the pdf, say $g(t)$, of this t-distribution does not depend upon $x_1, x_2, \ldots, x_n$. Now the joint pdf of $X_1, X_2, \ldots, X_n$ and $R\sqrt{n-2}/\sqrt{1-R^2}$, where R is given by expression (9.7.1), is the product of $g(t)$ and the joint pdf of $X_1, \ldots, X_n$. Integration on $x_1, \ldots, x_n$ yields the marginal pdf of $R\sqrt{n-2}/\sqrt{1-R^2}$; because $g(t)$ does not depend upon $x_1, x_2, \ldots, x_n$, it is obvious that this marginal pdf is $g(t)$, the conditional pdf of $R\sqrt{n-2}/\sqrt{1-R^2}$. The change-of-variable technique can now be used to find the pdf of R.

Remark 9.7.1. Since R has, when $\rho = 0$, a conditional distribution that does not depend upon $x_1, x_2, \ldots, x_n$ (and hence that conditional distribution is, in fact, the marginal distribution of R), we have the remarkable fact that R is independent of $X_1, X_2, \ldots, X_n$. It follows that R is independent of *every function* of $X_1, X_2, \ldots, X_n$ alone, that is, a function that does not depend upon any Y_i. In like manner, R is independent of every function of $Y_1, Y_2, \ldots, Y_n$ alone. Moreover, a careful review of the argument reveals that nowhere did we use the fact that X has a normal marginal distribution. Thus, if X and Y are independent, and if Y has a normal distribution, then R has the same conditional distribution whatever the distribution of X, subject to the condition $\sum_1^n (x_i - \overline{x})^2 > 0$. Moreover, if $P[\sum_1^n (X_i - \overline{X})^2 > 0] = 1$, then R has the same marginal distribution whatever the distribution of X. ■

If we write $T = R\sqrt{n-2}/\sqrt{1-R^2}$, where T has a t-distribution with $n-2 > 0$ degrees of freedom, it is easy to show by the change-of-variable technique (Exercise 9.7.4) that the pdf of R is given by

$$h(r) = \begin{cases} \frac{\Gamma[(n-1)/2]}{\Gamma(\frac{1}{2})\Gamma[(n-2)/2]}(1-r^2)^{(n-4)/2} & -1 < r < 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{9.7.4}$$

We have now solved the problem of the distribution of R, when $\rho = 0$ and $n > 2$, or perhaps more conveniently, that of $R\sqrt{n-2}/\sqrt{1-R^2}$. The likelihood ratio test of the hypothesis $H_0 : \rho = 0$ against all alternatives $H_1 : \rho \neq 0$ may be based either on the statistic R or on the statistic $R\sqrt{n-2}/\sqrt{1-R^2} = T$, although the latter is easier to use. Therefore, a level α test is to reject $H_0 : \rho = 0$ if $|T| \geq t_{\alpha/2, n-2}$.

Remark 9.7.2. It is possible to obtain an approximate test of size α by using the fact that

$$W = \frac{1}{2}\log\left(\frac{1+R}{1-R}\right)$$

has an approximate normal distribution with mean $\frac{1}{2}\log[(1+\rho)/(1-\rho)]$ and with variance $1/(n-3)$. We accept this statement without proof. Thus a test of $H_0 : \rho = 0$ can be based on the statistic

$$Z = \frac{\frac{1}{2}\log[(1+R)/(1-R)] - \frac{1}{2}\log[(1+\rho)/(1-\rho)]}{\sqrt{1/(n-3)}}, \tag{9.7.5}$$

with $\rho = 0$ so that $\frac{1}{2}\log[(1+\rho)/(1-\rho)] = 0$. However, using W, we can also test a hypothesis like $H_0 : \rho = \rho_0$ against $H_1 : \rho \neq \rho_0$, where ρ_0 is not necessarily zero. In that case, the hypothesized mean of W is

$$\frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right).$$

Furthermore, as outlined in Exercise 9.7.6, Z can be used to obtain an asymptotic confidence interval for ρ. ■

EXERCISES

9.7.1. Show that

$$R = \frac{\sum_1^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_1^n (X_i - \overline{X})^2 \sum_1^n (Y_i - \overline{Y})^2}} = \frac{\sum_1^n X_i Y_i - n\overline{XY}}{\sqrt{\left(\sum_1^n X_i^2 - n\overline{X}^2\right)\left(\sum_1^n Y_i^2 - n\overline{Y}^2\right)}}.$$

9.7.2. A random sample of size $n = 6$ from a bivariate normal distribution yields a value of the correlation coefficient of 0.89. Would we accept or reject, at the 5% significance level, the hypothesis that $\rho = 0$?

9.7.3. Verify Equation (9.7.3) of this section.

9.7.4. Verify the pdf (9.7.4) of this section.

9.7.5. Using the results of Section 4.5, show that R, (9.7.1), is a consistent estimate of ρ.

9.7.6. By doing the following steps, determine a $(1-\alpha)100\%$ approximate confidence interval for ρ.

(a) For $0 < \alpha < 1$, in the usual way, start with $1 - \alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2})$, where Z is given by expression (9.7.5). Then isolate $h(\rho) = (1/2)\log[(1+\rho)/(1-\rho)]$ in the middle part of the inequality. Find $h'(\rho)$ and show that it is strictly positive on $-1 < \rho < 1$; hence, h is strictly increasing and its inverse function exists.

(b) Show that this inverse function is the hyperbolic tangent function given by $\tanh(y) = (e^y - e^{-y})/(e^y + e^{-y})$.

(c) Obtain a $(1-\alpha)100\%$ confidence interval for ρ.

9.7.7. The intrinsic R function `cor.test(x,y)` computes the estimate of ρ and the confidence interval in Exercise 9.7.6. Recall the baseball data which is in the file `bb.rda`.

(a) Using the baseball data, determine the estimate and the confidence interval for the correlation coefficient between height and weight for professional baseball players.

(b) Separate the pitchers and hitters and for each obtain the estimate and confidence for the correlation coefficient between height and weight. Do they differ significantly?

(c) Argue that the difference in the estimates of the correlation coefficients is the mle of $\rho_1 - \rho_2$ for two independent samples, as in Part (b).

9.7.8. Two experiments gave the following results:

n	$\overline{x}$	$\overline{y}$	s_x	s_y	r
100	10	20	5	8	0.70
200	12	22	6	10	0.80

Calculate r for the combined sample.

9.8 The Distributions of Certain Quadratic Forms

Remark 9.8.1. It is essential that the reader have the background of the multivariate normal distribution as given in Section 3.5 to understand Sections 9.8 and 9.9. ■

Remark 9.8.2. We make use of the **trace** of a square matrix. If $\mathbf{A} = [a_{ij}]$ is an $n \times n$ matrix, then we define the trace of $\mathbf{A}$, $(\operatorname{tr} \mathbf{A})$, to be the sum of its diagonal entries; i.e.,

$$\operatorname{tr} \mathbf{A} = \sum_{i=1}^{n} a_{ii}. \tag{9.8.1}$$

The trace of a matrix has several interesting properties. One is that it is a linear operator; that is,

$$\operatorname{tr} (a\mathbf{A} + b\mathbf{B}) = a \operatorname{tr} \mathbf{A} + b \operatorname{tr} \mathbf{B}. \tag{9.8.2}$$

A second useful property is: If $\mathbf{A}$ is an $n \times m$ matrix, $\mathbf{B}$ is an $m \times k$ matrix, and $\mathbf{C}$ is a $k \times n$ matrix, then

$$\operatorname{tr} (\mathbf{ABC}) = \operatorname{tr} (\mathbf{BCA}) = \operatorname{tr} (\mathbf{CAB}). \tag{9.8.3}$$

The reader is asked to prove these facts in Exercise 9.8.7. Finally, a simple but useful property is that $\operatorname{tr} a = a$, for any scalar a. ■

We begin this section with a more formal but equivalent definition of a quadratic form. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an n-dimensional random vector and let $\mathbf{A}$ be a real $n \times n$ symmetric matrix. Then the random variable $Q = \mathbf{X}'\mathbf{AX}$ is called a

quadratic form in $\mathbf{X}$. Due to the symmetry of $\mathbf{A}$, there are several ways we can write Q:

$$Q = \mathbf{X}'\mathbf{A}\mathbf{X} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}X_iX_j = \sum_{i=1}^{n} a_{ii}X_i^2 + \sum\sum_{i\neq j} a_{ij}X_iX_j \quad (9.8.4)$$

$$= \sum_{i=1}^{n} a_{ii}X_i^2 + 2\sum\sum_{i<j} a_{ij}X_iX_j. \quad (9.8.5)$$

These are very useful random variables in analysis of variance models. As the following theorem shows, the mean of a quadratic form is easily obtained.

Theorem 9.8.1. *Suppose the n-dimensional random vector $\mathbf{X}$ has mean $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}$. Let $Q = \mathbf{X}'\mathbf{A}\mathbf{X}$, where $\mathbf{A}$ is a real $n \times n$ symmetric matrix. Then*

$$E(Q) = tr\,\mathbf{A}\boldsymbol{\Sigma} + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}. \quad (9.8.6)$$

Proof: Using the trace operator and property (9.8.3), we have

$$\begin{aligned} E(Q) = E(\operatorname{tr}\mathbf{X}'\mathbf{A}\mathbf{X}) &= E(\operatorname{tr}\mathbf{A}\mathbf{X}\mathbf{X}') \\ &= \operatorname{tr}\mathbf{A}E(\mathbf{X}\mathbf{X}') \\ &= \operatorname{tr}\mathbf{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}') \\ &= \operatorname{tr}\mathbf{A}\boldsymbol{\Sigma} + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}, \end{aligned}$$

where the third line follows from Theorem 2.6.3. ■

Example 9.8.1 (Sample Variance)**.** Let $\mathbf{X}' = (X_1, \ldots, X_n)$ be an n-dimensional vector of random variables. Let $\mathbf{1}' = (1, \ldots, 1)$ be the n-dimensional vector whose components are 1. Let $\mathbf{I}$ be the $n \times n$ identity matrix. Consider the quadratic form $Q = \mathbf{X}'(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{X}$, where $\mathbf{J} = \mathbf{1}\mathbf{1}'$; i.e., $\mathbf{J}$ is an $n \times n$ matrix with all entries equal to 1. Note that the off-diagonal entries of $(\mathbf{I} - \frac{1}{n}, \mathbf{J})$ are $-n^{-1}$ while the diagonal entries are $1 - n^{-1}$; hence, by (9.8.4), Q simplifies to

$$\begin{aligned} Q &= \sum_{i=1}^{n} X_i^2\left(1 - \frac{1}{n}\right) + \sum\sum_{i\neq j}\left(-\frac{1}{n}\right)X_iX_j \\ &= \sum_{i=1}^{n} X_i^2\left(1 - \frac{1}{n}\right) - \frac{1}{n}\sum_{i=1}^{n} X_i\sum_{j=1}^{n} X_j + \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = (n-1)S^2, \end{aligned} \quad (9.8.7)$$

where $\overline{X}$ and S^2 denote the sample mean and variance of $X_1, \ldots, X_n$.

Suppose we further assume that $X_1, \ldots, X_n$ are iid random variables with common mean μ and variance σ^2. Using Theorem 9.8.1, we can obtain yet another

proof that S^2 is an unbiased estimate of σ^2. Note that the mean of the random vector $\mathbf{X}$ is $\mu\mathbf{1}$ and that its variance–covariance matrix is $\sigma^2\mathbf{I}$. Based on Theorem 9.8.1, we find immediately that

$$E(S^2) = \frac{1}{n-1}\left\{\text{tr}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\sigma^2\mathbf{I} + \mu^2\left(\mathbf{1}'\mathbf{1} - \frac{1}{n}\mathbf{1}'\mathbf{1}\mathbf{1}'\mathbf{1}\right)\right\} = \sigma^2 \ . \quad \blacksquare$$

The spectral decomposition of symmetric matrices proves quite useful in this part of the chapter. As discussed around expression (3.5.8), a real symmetric matrix $\mathbf{A}$ can be diagonalized as

$$\mathbf{A} = \mathbf{\Gamma}'\mathbf{\Lambda}\mathbf{\Gamma}, \tag{9.8.8}$$

where $\mathbf{\Lambda}$ is the diagonal matrix $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_1 \geq \cdots \geq \lambda_n$ are the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{\Gamma}' = [\mathbf{v}_1 \cdots \mathbf{v}_n]$ are the corresponding orthonormal eigenvectors (i.e., $\mathbf{\Gamma}$ is an orthogonal matrix). Recall from linear algebra that the rank of $\mathbf{A}$ is the number of nonzero eigenvalues. Further, because $\mathbf{\Lambda}$ is diagonal, we can write this expression as

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i'. \tag{9.8.9}$$

The R command to compute the spectral decomposition of $\mathbf{A}$ is `sdc=eigen(amat)`, where `amat` is the R matrix for $\mathbf{A}$. The eigenvalues and eigenvectors are in the respective attributes `sdc$values` and `sdc$vectors`. For normal random variables, we make use of equation (9.8.9) to obtain the mgf of the quadratic form Q in the next theorem, Theorem 9.8.2.

Theorem 9.8.2. *Let $\mathbf{X}' = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are iid $N(0, \sigma^2)$. Consider the quadratic form $Q = \sigma^{-2}\mathbf{X}'\mathbf{A}\mathbf{X}$ for a symmetric matrix $\mathbf{A}$ of rank $r \leq n$. Then Q has the moment generating function*

$$M(t) = \prod_{i=1}^{r}(1 - 2t\lambda_i)^{-1/2} = |\mathbf{I} - 2t\mathbf{A}|^{-1/2}, \tag{9.8.10}$$

where $\lambda_1, \ldots, \lambda_r$ are the nonzero eigenvalues of $\mathbf{A}$, $|t| < 1/(2\lambda^*)$, and the value of λ^* is given by $\lambda^* = \max_{1 \leq i \leq r} |\lambda_i|$.

Proof: Write the spectral decomposition of $\mathbf{A}$ as in expression (9.8.9). Since the rank of $\mathbf{A}$ is r, exactly r of the eigenvalues are not 0. Denote the nonzero eigenvalues by $\lambda_1, \ldots, \lambda_r$. Then we can write Q as

$$Q = \sum_{i=1}^{r} \lambda_i (\sigma^{-1}\mathbf{v}_i'\mathbf{X})^2. \tag{9.8.11}$$

Let $\mathbf{\Gamma}_1' = [\mathbf{v}_1 \cdots \mathbf{v}_r]$ and define the r-dimensional random vector $\mathbf{W}$ by $\mathbf{W} = \sigma^{-1}\mathbf{\Gamma}_1\mathbf{X}$. Since $\mathbf{X}$ is $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and $\mathbf{\Gamma}_1'\mathbf{\Gamma}_1 = \mathbf{I}_r$, Theorem 3.5.2 shows that $\mathbf{W}$ has a $N_r(\mathbf{0}, \mathbf{I}_r)$ distribution. In terms of the W_i, we can write (9.8.11) as

$$Q = \sum_{i=1}^{r} \lambda_i W_i^2. \tag{9.8.12}$$

Because $W_1, \ldots, W_r$ are independent $N(0,1)$ random variables, $W_1^2, \ldots, W_r^2$ are independent $\chi^2(1)$ random variables. Thus the mgf of Q is

$$\begin{aligned} E[\exp\{tQ\}] &= E\left[\exp\left\{\sum_{i=1}^{r} t\lambda_i W_i^2\right\}\right] \\ &= \prod_{i=1}^{r} E[\exp\{t\lambda_i W_i^2\}] = \prod_{i=1}^{r}(1-2t\lambda_i)^{-1/2}. \end{aligned} \tag{9.8.13}$$

The last equality holds if we assume that $|t| < 1/(2\lambda^*)$, where $\lambda^* = \max_{1\le i\le r}|\lambda_i|$; see Exercise 9.8.6. To obtain the second form in (9.8.10), recall that the determinant of an orthogonal matrix is 1. The result then follows from

$$\begin{aligned} |\mathbf{I}-2t\mathbf{A}| = |\boldsymbol{\Gamma}'\boldsymbol{\Gamma} - 2t\boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}| &= |\boldsymbol{\Gamma}'(\mathbf{I}-2t\boldsymbol{\Lambda})\boldsymbol{\Gamma}| \\ &= |\mathbf{I}-2t\boldsymbol{\Lambda}| = \left\{\prod_{i=1}^{r}(1-2t\lambda_i)^{-1/2}\right\}^{-2}. \quad \blacksquare \end{aligned}$$

Example 9.8.2. To illustrate this theorem, suppose $X_i,\ i = 1, 2, \ldots, n$, are independent random variables with X_i distributed as $N(\mu_i, \sigma_i^2),\ i = 1, 2, \ldots, n$, respectively. Let $Z_i = (X_i - \mu_i)/\sigma_i$. We know that $\sum_{i=1}^{n} Z_i^2$ has a χ^2 distribution with n degrees of freedom. To illustrate Theorem 9.8.2, let $\mathbf{Z}' = (Z_1, \ldots, Z_n)$. Let $Q = \mathbf{Z}'\mathbf{I}\mathbf{Z}$. Hence the symmetric matrix associated with Q is the identity matrix $\mathbf{I}$, which has n eigenvalues, all of value 1; i.e., $\lambda_i \equiv 1$. By Theorem 9.8.2, the mgf of Q is $(1-2t)^{-n/2}$; i.e., Q is distributed χ^2 with n degrees of freedom. $\blacksquare$

In general, from Theorem 9.8.2, note how close the mgf of the quadratic form Q is to the mgf of a χ^2 distribution. The next two theorems give conditions where this is true.

Theorem 9.8.3. *Let $\mathbf{X}' = (X_1, X_2, \ldots, X_n)$ have a $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\Sigma}$ is positive definite. Then $Q = (\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$ has a $\chi^2(n)$ distribution.*

Proof: Write the spectral decomposition of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\boldsymbol{\Sigma}$. Because $\boldsymbol{\Sigma}$ is positive definite, all $\lambda_i > 0$. Hence we can write

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma},$$

where $\boldsymbol{\Lambda}^{-1/2} = \text{diag}\{\lambda_1^{-1/2}, \ldots, \lambda_n^{-1/2}\}$. Thus we have

$$Q = \left\{\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma}(\mathbf{X}-\boldsymbol{\mu})\right\}' \mathbf{I} \left\{\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma}(\mathbf{X}-\boldsymbol{\mu})\right\}.$$

But by Theorem 3.5.2, it is easy to show that the random vector $\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma}(\mathbf{X}-\boldsymbol{\mu})$ has a $N_n(\mathbf{0}, \mathbf{I})$ distribution; hence, Q has a $\chi^2(n)$ distribution. $\blacksquare$

The remarkable fact that the random variable Q in the last theorem is $\chi^2(n)$ stimulates a number of questions about quadratic forms in normally distributed

variables. We would like to treat this problem generally, but limitations of space forbid this, and we find it necessary to restrict ourselves to some special cases; see, for instance, Stapleton (2009) for discussion.

Recall from linear algebra that a symmetric matrix $\mathbf{A}$ is **idempotent** if $\mathbf{A}^2 = \mathbf{A}$. In Section 9.1, we have already met some idempotent matrices. For example, the matrix $\mathbf{I} - \frac{1}{n}\mathbf{J}$ of Example 9.8.1 is idempotent. Idempotent matrices possess some important characteristics. Suppose λ is an eigenvalue of an idempotent matrix $\mathbf{A}$ with corresponding eigenvector $\mathbf{v}$. Then the following identity is true:

$$\lambda\mathbf{v} = \mathbf{A}\mathbf{v} = \mathbf{A}^2\mathbf{v} = \lambda\mathbf{A}\mathbf{v} = \lambda^2\mathbf{v}.$$

Hence $\lambda(\lambda - 1)\mathbf{v} = \mathbf{0}$. Since $\mathbf{v} \neq \mathbf{0}$, $\lambda = 0$ or 1. Conversely, if the eigenvalues of a real symmetric matrix are only 0s and 1s then it is idempotent; see Exercise 9.8.10. Thus the rank of an idempotent matrix $\mathbf{A}$ is the number of its eigenvalues which are 1. Denote the spectral decomposition of $\mathbf{A}$ by $\mathbf{A} = \mathbf{\Gamma}'\mathbf{\Lambda}\mathbf{\Gamma}$, where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues and $\mathbf{\Gamma}$ is an orthogonal matrix whose columns are the corresponding orthonormal eigenvectors. Because the diagonal entries of $\mathbf{\Lambda}$ are 0 or 1 and $\mathbf{\Gamma}$ is orthogonal, we have

$$\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{\Lambda}\mathbf{\Gamma}\mathbf{\Gamma}' = \operatorname{tr}\mathbf{\Lambda} = \operatorname{rank}(\mathbf{A});$$

i.e., the rank of an idempotent matrix is equal to its trace.

Theorem 9.8.4. *Let $\mathbf{X}' = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are iid $N(0, \sigma^2)$. Let $Q = \sigma^{-2}\mathbf{X}'\mathbf{A}\mathbf{X}$ for a symmetric matrix $\mathbf{A}$ with rank r. Then Q has a $\chi^2(r)$ distribution if and only if $\mathbf{A}$ is idempotent.*

Proof: By Theorem 9.8.2, the mgf of Q is

$$M_Q(t) = \prod_{i=1}^{r}(1 - 2t\lambda_i)^{-1/2}, \tag{9.8.14}$$

where $\lambda_1, \ldots, \lambda_r$ are the r nonzero eigenvalues of $\mathbf{A}$. Suppose, first, that $\mathbf{A}$ is idempotent. Then $\lambda_1 = \cdots = \lambda_r = 1$ and the mgf of Q is $M_Q(t) = (1 - 2t)^{-r/2}$; i.e., Q has a $\chi^2(r)$ distribution. Next, suppose Q has a $\chi^2(r)$ distribution. Then for t in a neighborhood of 0, we have the identity

$$\prod_{i=1}^{r}(1 - 2t\lambda_i)^{-1/2} = (1 - 2t)^{-r/2},$$

which, upon squaring both sides, leads to

$$\prod_{i=1}^{r}(1 - 2t\lambda_i) = (1 - 2t)^{r},$$

By the uniqueness of the factorization of polynomials, $\lambda_1 = \cdots = \lambda_r = 1$. Hence $\mathbf{A}$ is idempotent. ■

Example 9.8.3. Based on this last theorem, we can obtain quickly the distribution of the sample variance when sampling from a normal distribution. Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$. Then $\mathbf{X}$ has a $N_n(\mu\mathbf{1}, \sigma^2\mathbf{I})$ distribution, where $\mathbf{1}$ denotes a $n \times 1$ vector with all components equal to 1. Let $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Then by Example 9.8.1, we can write

$$\frac{(n-1)S^2}{\sigma^2} = \sigma^{-2}\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{X} = \sigma^{-2}(\mathbf{X} - \mu\mathbf{1})'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)(\mathbf{X} - \mu\mathbf{1}),$$

where the last equality holds because $\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{1} = \mathbf{0}$. Because the matrix $\mathbf{I} - \frac{1}{n}\mathbf{J}$ is idempotent, $\text{tr}\,(\mathbf{I} - \frac{1}{n}\mathbf{J}) = n - 1$, and $\mathbf{X} - \mu\mathbf{1}$ is $N_n(\mathbf{0}, \sigma^2\mathbf{I})$, it follows from Theorem 9.8.4 that $(n-1)S^2/\sigma^2$ has a $\chi^2(n-1)$ distribution. ■

Remark 9.8.3. If the normal distribution in Theorem 9.8.4 is $N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, the condition $\mathbf{A}^2 = \mathbf{A}$ remains a necessary and sufficient condition that Q/σ^2 have a chi-square distribution. In general, however, Q/σ^2 is not central $\chi^2(r)$ but instead, Q/σ^2 has a noncentral chi-square distribution if $\mathbf{A}^2 = \mathbf{A}$. The number of degrees of freedom is r, the rank of $\mathbf{A}$, and the noncentrality parameter is $\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}/\sigma^2$. If $\boldsymbol{\mu} = \mu\mathbf{1}$, then $\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = \mu^2\sum_{i,j} a_{ij}$, where $\mathbf{A} = [a_{ij}]$. Then, if $\mu \neq 0$, the conditions $\mathbf{A}^2 = \mathbf{A}$ and $\sum_{i,j} a_{ij} = 0$ are necessary and sufficient conditions that Q/σ^2 be central $\chi^2(r)$. Moreover, the theorem may be extended to a quadratic form in random variables which have a multivariate normal distribution with positive definite covariance matrix $\boldsymbol{\Sigma}$; here the necessary and sufficient condition that Q have a chi-square distribution is $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}$. See Exercise 9.8.9. ■

EXERCISES

9.8.1. Let $Q = X_1X_2 - X_3X_4$, where X_1, X_2, X_3, X_4 is a random sample of size 4 from a distribution that is $N(0, \sigma^2)$. Show that Q/σ^2 does not have a chi-square distribution. Find the mgf of Q/σ^2.

9.8.2. Let $\mathbf{X}' = [X_1, X_2]$ be bivariate normal with matrix of means $\boldsymbol{\mu}' = [\mu_1, \mu_2]$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. Let

$$Q_1 = \frac{X_1^2}{\sigma_1^2(1-\rho^2)} - 2\rho\frac{X_1X_2}{\sigma_1\sigma_2(1-\rho^2)} + \frac{X_2^2}{\sigma_2^2(1-\rho^2)}.$$

Show that Q_1 is $\chi^2(r, \theta)$ and find r and θ. When and only when does Q_1 have a central chi-square distribution?

9.8.3. Let $\mathbf{X}' = [X_1, X_2, X_3]$ denote a random sample of size 3 from a distribution that is $N(4, 8)$ and let

$$\mathbf{A} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

Let $Q = \mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$.

(a) Use Theorem 9.8.1 to find the $E(Q)$.

(b) Justify the assertion that Q is $\chi^2(2, 6)$.

9.8.4. Suppose $X_1, \ldots, X_n$ are independent random variables with the common mean μ but with unequal variances $\sigma_i^2 = \text{Var}(X_i)$.

(a) Determine the variance of $\overline{X}$.

(b) Determine the constant K so that $Q = K\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimate of the variance of $\overline{X}$. (*Hint:* Proceed as in Example 9.8.3.)

9.8.5. Suppose $X_1, \ldots, X_n$ are correlated random variables, with common mean μ and variance σ^2 but with correlations ρ (all correlations are the same).

(a) Determine the variance of $\overline{X}$.

(b) Determine the constant K so that $Q = K\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimate of the variance of $\overline{X}$. (*Hint:* Proceed as in Example 9.8.3.)

9.8.6. Fill in the details for expression (9.8.13).

9.8.7. For the trace operator defined in expression (9.8.1), prove the following properties are true.

(a) If $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ matrices and a and b are scalars, then

$$\text{tr}\,(a\mathbf{A} + b\mathbf{B}) = a\,\text{tr}\,\mathbf{A} + b\,\text{tr}\,\mathbf{B}.$$

(b) If $\mathbf{A}$ is an $n \times m$ matrix, $\mathbf{B}$ is an $m \times k$ matrix, and $\mathbf{C}$ is a $k \times n$ matrix, then

$$\text{tr}\,(\mathbf{ABC}) = \text{tr}\,(\mathbf{BCA}) = \text{tr}\,(\mathbf{CAB}).$$

(c) If $\mathbf{A}$ is a square matrix and $\mathbf{\Gamma}$ is an orthogonal matrix, use the result of part (a) to show that $\text{tr}(\mathbf{\Gamma}'\mathbf{A}\mathbf{\Gamma}) = \text{tr}\mathbf{A}$.

(d) If $\mathbf{A}$ is a real symmetric idempotent matrix, use the result of part (b) to prove that the rank of $\mathbf{A}$ is equal to $\text{tr}\mathbf{A}$.

9.8.8. Let $\mathbf{A} = [a_{ij}]$ be a real symmetric matrix. Prove that $\sum_i \sum_j a_{ij}^2$ is equal to the sum of the squares of the eigenvalues of $\mathbf{A}$.
Hint: If $\mathbf{\Gamma}$ is an orthogonal matrix, show that $\sum_j \sum_i a_{ij}^2 = \text{tr}(\mathbf{A}^2) = \text{tr}(\mathbf{\Gamma}'\mathbf{A}^2\mathbf{\Gamma}) = \text{tr}[(\mathbf{\Gamma}'\mathbf{A}\mathbf{\Gamma})(\mathbf{\Gamma}'\mathbf{A}\mathbf{\Gamma})]$.

9.8.9. Suppose $\mathbf{X}$ has a $N_n(0, \mathbf{\Sigma})$ distribution, where $\mathbf{\Sigma}$ is positive definite. Let $Q = \mathbf{X}'\mathbf{A}\mathbf{X}$ for a symmetric matrix $\mathbf{A}$ with rank r. Prove Q has a $\chi^2(r)$ distribution if and only if $\mathbf{A}\mathbf{\Sigma}\mathbf{A} = \mathbf{A}$.
Hint: Write Q as

$$Q = (\mathbf{\Sigma}^{-1/2}\mathbf{X})'\mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}(\mathbf{\Sigma}^{-1/2}\mathbf{X}),$$

where $\mathbf{\Sigma}^{1/2} = \mathbf{\Gamma}'\mathbf{\Lambda}^{1/2}\mathbf{\Gamma}$ and $\mathbf{\Sigma} = \mathbf{\Gamma}'\mathbf{\Lambda}\mathbf{\Gamma}$ is the spectral decomposition of $\mathbf{\Sigma}$. Then use Theorem 9.8.4.

9.8.10. Suppose $\mathbf{A}$ is a real symmetric matrix. If the eigenvalues of $\mathbf{A}$ are only 0s and 1s then prove that $\mathbf{A}$ is idempotent.

9.9 The Independence of Certain Quadratic Forms

We have previously investigated the independence of linear functions of normally distributed variables. In this section we shall prove some theorems about the independence of quadratic forms. We shall confine our attention to normally distributed variables that constitute a random sample of size n from a distribution that is $N(0, \sigma^2)$.

Remark 9.9.1. In the proof of the next theorem, we use the fact that if $\mathbf{A}$ is an $m \times n$ matrix of rank n (i.e., $\mathbf{A}$ has full column rank), then the matrix $\mathbf{A}'\mathbf{A}$ is nonsingular. A proof of this linear algebra fact is sketched in Exercises 9.9.12 and 9.9.13. ■

Theorem 9.9.1 (Craig)**.** *Let* $\mathbf{X}' = (X_1, \ldots, X_n)$*, where* $X_1, \ldots, X_n$ *are iid* $N(0, \sigma^2)$ *random variables. For real symmetric matrices* $\mathbf{A}$ *and* $\mathbf{B}$*, let* $Q_1 = \sigma^{-2}\mathbf{X}'\mathbf{A}\mathbf{X}$ *and* $Q_2 = \sigma^{-2}\mathbf{X}'\mathbf{B}\mathbf{X}$ *denote quadratic forms in* $\mathbf{X}$*. The random variables* Q_1 *and* Q_2 *are independent if and only if* $\mathbf{A}\mathbf{B} = \mathbf{0}$.

Proof: First, we obtain some preliminary results. Based on these results, the proof follows immediately. Assume the ranks of the matrices $\mathbf{A}$ and $\mathbf{B}$ are r and s, respectively. Let $\mathbf{\Gamma}_1'\mathbf{\Lambda}_1\mathbf{\Gamma}_1$ denote the spectral decomposition of $\mathbf{A}$. Denote the r nonzero eigenvalues of $\mathbf{A}$ by $\lambda_1, \ldots, \lambda_r$. Without loss of generality, assume that these nonzero eigenvalues of $\mathbf{A}$ are the first r elements on the main diagonal of $\mathbf{\Lambda}_1$ and let $\mathbf{\Gamma}_{11}'$ be the $n \times r$ matrix whose columns are the corresponding eigenvectors. Finally, let $\mathbf{\Lambda}_{11} = \text{diag}\{\lambda_1, \ldots, \lambda_r\}$. Then we can write the spectral decomposition of $\mathbf{A}$ in either of the two ways

$$\mathbf{A} = \mathbf{\Gamma}_1'\mathbf{\Lambda}_1\mathbf{\Gamma}_1 = \mathbf{\Gamma}_{11}'\mathbf{\Lambda}_{11}\mathbf{\Gamma}_{11}. \tag{9.9.1}$$

Note that we can write Q_1 as

$$Q_1 = \sigma^{-2}\mathbf{X}'\mathbf{\Gamma}_{11}'\mathbf{\Lambda}_{11}\mathbf{\Gamma}_{11}\mathbf{X} = \sigma^{-2}(\mathbf{\Gamma}_{11}\mathbf{X})'\mathbf{\Lambda}_{11}(\mathbf{\Gamma}_{11}\mathbf{X}) = \mathbf{W}_1'\mathbf{\Lambda}_{11}\mathbf{W}_1, \tag{9.9.2}$$

where $\mathbf{W}_1 = \sigma^{-1}\mathbf{\Gamma}_{11}\mathbf{X}$. Next, obtain a similar representation based on the s nonzero eigenvalues $\gamma_1, \ldots, \gamma_s$ of $\mathbf{B}$. Let $\mathbf{\Lambda}_{22} = \text{diag}\{\gamma_1, \ldots, \gamma_s\}$ denote the $s \times s$ diagonal matrix of nonzero eigenvalues and form the $n \times s$ matrix $\mathbf{\Gamma}_{21}' = [\mathbf{u}_1 \cdots \mathbf{u}_s]$ of corresponding eigenvectors. Then we can write the spectral decomposition of $\mathbf{B}$ as

$$\mathbf{B} = \mathbf{\Gamma}_{21}'\mathbf{\Lambda}_{22}\mathbf{\Gamma}_{21}. \tag{9.9.3}$$

Also, we can write Q_2 as

$$Q_2 = \mathbf{W}_2'\mathbf{\Lambda}_{22}\mathbf{W}_2, \tag{9.9.4}$$

where $\mathbf{W}_2 = \sigma^{-1}\mathbf{\Gamma}_{21}\mathbf{X}$. Letting $\mathbf{W}' = (\mathbf{W}_1', \mathbf{W}_2')$, we have

$$\mathbf{W} = \sigma^{-1}\begin{bmatrix} \mathbf{\Gamma}_{11} \\ \mathbf{\Gamma}_{21} \end{bmatrix}\mathbf{X}.$$

Because $\mathbf{X}$ has a $N_n(\mathbf{0}, \sigma^2\mathbf{I})$ distribution, Theorem 3.5.2 shows that $\mathbf{W}$ has an $(r + s)$–dimensional multivariate normal distribution with mean $\mathbf{0}$ and variance–covariance matrix

$$\text{Var}(\mathbf{W}) = \begin{bmatrix} \mathbf{I}_r & \mathbf{\Gamma}_{11}\mathbf{\Gamma}_{21}' \\ \mathbf{\Gamma}_{21}\mathbf{\Gamma}_{11}' & \mathbf{I}_s \end{bmatrix}. \tag{9.9.5}$$

Finally, using (9.9.1) and (9.9.3), we have the identity

$$\mathbf{AB} = \{\mathbf{\Gamma}'_{11}\mathbf{\Lambda}_{11}\}\mathbf{\Gamma}_{11}\mathbf{\Gamma}'_{21}\{\mathbf{\Lambda}_{22}\mathbf{\Gamma}_{21}\}. \tag{9.9.6}$$

Let $\mathbf{U}$ denote the matrix in the first set of braces. Note that $\mathbf{U}$ has full column rank, so its kernel is null; i.e., its kernel consists of the vector $\mathbf{0}$. Let $\mathbf{V}$ denote the matrix in the second set of braces. Note that $\mathbf{V}$ has full row rank, hence the kernel of $\mathbf{V}'$ is null.

For the proof then, suppose $\mathbf{AB} = \mathbf{0}$. Then

$$\mathbf{U}\left[\mathbf{\Gamma}_{11}\mathbf{\Gamma}'_{21}\mathbf{V}\right] = \mathbf{0}.$$

Because the kernel of $\mathbf{U}$ is null this implies each column of the matrix in the brackets is the vector $\mathbf{0}$; i.e., the matrix in the brackets is the matrix $\mathbf{0}$. This implies that

$$\mathbf{V}'\left[\mathbf{\Gamma}_{21}\mathbf{\Gamma}'_{11}\right] = \mathbf{0}.$$

In the same way, because the kernel of $\mathbf{V}'$ is null, we have $\mathbf{\Gamma}_{11}\mathbf{\Gamma}'_{21} = \mathbf{0}$. Hence, by (9.9.5), the random vectors $\mathbf{W}_1$ and $\mathbf{W}_2$ are independent. Therefore, by (9.9.2) and (9.9.4), Q_1 and Q_2 are independent.

Conversely, if Q_1 and Q_2 are independent, then

$$\{E[\exp\{t_1Q_1 + t_2Q_2\}]\}^{-2} = \{E[\exp\{t_1Q_1\}]\}^{-2}\{E[\exp\{t_2Q_2\}]\}^{-2}, \tag{9.9.7}$$

for (t_1, t_2) in an open neighborhood of $(0, 0)$. Note that $t_1Q_1 + t_2Q_2$ is a quadratic form in $\mathbf{X}$ with symmetric matrix $t_1\mathbf{A}+t_2\mathbf{B}$. Recall that the matrix $\mathbf{\Gamma}_1$ is orthogonal and hence has determinant ± 1. Using this and Theorem 9.8.2, we can write the left side of (9.9.7) as

$$\begin{aligned} E^{-2}[\exp\{t_1Q_1 + t_2Q_2\}] &= |\mathbf{I}_n - 2t_1\mathbf{A} - 2t_2\mathbf{B}| \\ &= |\mathbf{\Gamma}'_1\mathbf{\Gamma}_1 - 2t_1\mathbf{\Gamma}'_1\mathbf{\Lambda}_1\mathbf{\Gamma}_1 - 2t_2\mathbf{\Gamma}'_1(\mathbf{\Gamma}_1\mathbf{B}\mathbf{\Gamma}'_1)\mathbf{\Gamma}_1| \\ &= |\mathbf{I}_n - 2t_1\mathbf{\Lambda}_1 - 2t_2\mathbf{D}|, \end{aligned} \tag{9.9.8}$$

where the matrix $\mathbf{D}$ is given by

$$\mathbf{D} = \mathbf{\Gamma}_1\mathbf{B}\mathbf{\Gamma}'_1 = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{bmatrix}, \tag{9.9.9}$$

and $\mathbf{D}_{11}$ is $r \times r$. By (9.9.2), (9.9.3), and Theorem 9.8.2, the right side of (9.9.7) can be written as

$$\{E[\exp\{t_1Q_1\}]\}^{-2}\{E[\exp\{t_2Q_2\}]\}^{-2} = \left\{\prod_{i=1}^{r}(1 - 2t_1\lambda_i)\right\}|\mathbf{I}_n - 2t_2\mathbf{D}|. \tag{9.9.10}$$

This leads to the identity

$$|\mathbf{I}_n - 2t_1\mathbf{\Lambda}_1 - 2t_2\mathbf{D}| = \left\{\prod_{i=1}^{r}(1 - 2t_1\lambda_i)\right\}|\mathbf{I}_n - 2t_2\mathbf{D}|, \tag{9.9.11}$$

for (t_1, t_2) in an open neighborhood of $(0, 0)$.

The coefficient of $(-2t_1)^r$ on the right side of (9.9.11) is $\lambda_1 \cdots \lambda_r |\mathbf{I} - 2t_2\mathbf{D}|$. It is not so easy to find the coefficient of $(-2t_1)^r$ in the left side of the equation (9.9.11). Conceive of expanding this determinant in terms of minors of order r formed from the first r columns. One term in this expansion is the product of the minor of order r in the upper left-hand corner, namely, $|\mathbf{I}_r - 2t_1\mathbf{\Lambda}_{11} - 2t_2\mathbf{D}_{11}|$, and the minor of order $n - r$ in the lower right-hand corner, namely, $|\mathbf{I}_{n-r} - 2t_2\mathbf{D}_{22}|$. Moreover, this product is the only term in the expansion of the determinant that involves $(-2t_1)^r$. Thus the coefficient of $(-2t_1)^r$ in the left-hand member of Equation (9.9.11) is $\lambda_1 \cdots \lambda_r |\mathbf{I}_{n-r} - 2t_2\mathbf{D}_{22}|$. If we equate these coefficients of $(-2t_1)^r$, we have

$$|\mathbf{I} - 2t_2\mathbf{D}| = |\mathbf{I}_{n-r} - 2t_2\mathbf{D}_{22}|, \tag{9.9.12}$$

for t_2 in an open neighborhood of 0. Equation (9.9.12) implies that the nonzero eigenvalues of the matrices $\mathbf{D}$ and $\mathbf{D}_{22}$ are the same (see Exercise 9.9.8). Recall that the sum of the squares of the eigenvalues of a symmetric matrix is equal to the sum of the squares of the elements of that matrix (see Exercise 9.8.8). Thus the sum of the squares of the elements of matrix $\mathbf{D}$ is equal to the sum of the squares of the elements of $\mathbf{D}_{22}$. Since the elements of the matrix $\mathbf{D}$ are real, it follows that each of the elements of $\mathbf{D}_{11}, \mathbf{D}_{12}$, and $\mathbf{D}_{21}$ is zero. Hence we can write

$$\mathbf{0} = \mathbf{\Lambda}_1\mathbf{D} = \mathbf{\Gamma}_1\mathbf{A}\mathbf{\Gamma}_1'\mathbf{\Gamma}_1\mathbf{B}\mathbf{\Gamma}_1'$$

because $\mathbf{\Gamma}_1$ is an orthogonal matrix, $\mathbf{AB} = \mathbf{0}$. ■

Remark 9.9.2. Theorem 9.9.1 remains valid if the random sample is from a distribution that is $N(\mu, \sigma^2)$, whatever the real value of μ. Moreover, Theorem 9.9.1 may be extended to quadratic forms in random variables that have a joint multivariate normal distribution with a positive definite covariance matrix $\mathbf{\Sigma}$. The necessary and sufficient condition for the independence of two such quadratic forms with symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ then becomes $\boldsymbol{A\Sigma B} = \mathbf{0}$. In our Theorem 9.9.1, we have $\mathbf{\Sigma} = \sigma^2\boldsymbol{I}$, so that $\boldsymbol{A\Sigma B} = \boldsymbol{A}\sigma^2\boldsymbol{IB} = \sigma^2\boldsymbol{AB} = \mathbf{0}$. ■

The following theorem is from Hogg and Craig (1958).

Theorem 9.9.2 (Hogg and Craig)**.** *Define the sum $Q = Q_1 + \cdots + Q_{k-1} + Q_k$, where $Q, Q_1, \ldots, Q_{k-1}, Q_k$ are $k+1$ random variables that are quadratic forms in the observations of a random sample of size n from a distribution that is $N(0, \sigma^2)$. Let Q/σ^2 be $\chi^2(r)$, let Q_i/σ^2 be $\chi^2(r_i)$, $i = 1, 2, \ldots, k-1$, and let Q_k be nonnegative. Then the random variables $Q_1, Q_2, \ldots, Q_k$ are independent and, hence, Q_k/σ^2 is $\chi^2(r_k = r - r_1 - \cdots - r_{k-1})$.*

Proof: Take first the case of $k = 2$ and let the real symmetric matrices Q, Q_1, and Q_2 be denoted, respectively, by $\boldsymbol{A}, \boldsymbol{A}_1, \boldsymbol{A}_2$. We are given that $Q = Q_1 + Q_2$ or, equivalently, that $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2$. We are also given that Q/σ^2 is $\chi^2(r)$ and that Q_1/σ^2 is $\chi^2(r_1)$. In accordance with Theorem 9.8.4, we have $\boldsymbol{A}^2 = \boldsymbol{A}$ and $\boldsymbol{A}_1^2 = \boldsymbol{A}$.

Since $Q_2 \geq 0$, each of the matrices $\boldsymbol{A}, \boldsymbol{A}_1$, and $\boldsymbol{A}_2$ is positive semidefinite. Because $\boldsymbol{A}^2 = \boldsymbol{A}$, we can find an orthogonal matrix Γ such that

$$\boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{\Gamma} = \begin{bmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}.$$

If we multiply both members of $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2$ on the left by $\boldsymbol{\Gamma}'$ and on the right by $\boldsymbol{\Gamma}$, we have

$$\begin{bmatrix} \boldsymbol{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma} + \boldsymbol{\Gamma}'\boldsymbol{A}_2\boldsymbol{\Gamma}.$$

Now each of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, and hence each of $\boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}\prime\boldsymbol{A}_2\boldsymbol{\Gamma}$ is positive semidefinite. Recall that if a real symmetric matrix is positive semidefinite, each element on the principal diagonal is positive or zero. Moreover, if an element on the principal diagonal is zero, then all elements in that row and all elements in that column are zero. Thus $\boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma} + \boldsymbol{\Gamma}'\boldsymbol{A}_2\boldsymbol{\Gamma}$ can be written as

$$\begin{bmatrix} \boldsymbol{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{H}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{9.9.13}$$

Since $\boldsymbol{A}_1^2 = \boldsymbol{A}_1$, we have

$$(\boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma})^2 = \boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{G}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

If we multiply both members of Equation (9.9.13) on the left by the matrix $\boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma}$, we see that

$$\begin{bmatrix} \boldsymbol{G}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{G}_r\boldsymbol{H}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

or, equivalently, $\boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma} = \boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma} + (\boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma})(\boldsymbol{\Gamma}'\boldsymbol{A}_2\boldsymbol{\Gamma})$. Thus $(\boldsymbol{\Gamma}'\boldsymbol{A}_1\boldsymbol{\Gamma}) \times (\boldsymbol{\Gamma}'\boldsymbol{A}_2\boldsymbol{\Gamma}) = \mathbf{0}$ and $\boldsymbol{A}_1\boldsymbol{A}_2 = \mathbf{0}$. In accordance with Theorem 9.9.1, Q_1 and Q_2 are independent. This independence immediately implies that Q_2/σ^2 is $\chi^2(r_2 = r - r_1)$. This completes the proof when $k = 2$. For $k > 2$, the proof may be made by induction. We shall merely indicate how this can be done by using $k = 3$. Take $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2 + \boldsymbol{A}_3$, where $\boldsymbol{A}^2 = \boldsymbol{A}$, $\boldsymbol{A}_1^2 = \boldsymbol{A}_1$, $\boldsymbol{A}_2^2 = \boldsymbol{A}_2$, and $\boldsymbol{A}_3$ is positive semidefinite. Write $\boldsymbol{A} = \boldsymbol{A}_1 + (\boldsymbol{A}_2 + \boldsymbol{A}_3) = \boldsymbol{A}_1 + \boldsymbol{B}_1$, say. Now $\boldsymbol{A}^2 = \boldsymbol{A}$, $\boldsymbol{A}_1^2 = \boldsymbol{A}_1$, and $\boldsymbol{B}_1$ is positive semidefinite. In accordance with the case of $k = 2$, we have $\boldsymbol{A}_1\boldsymbol{B}_1 = \mathbf{0}$, so that $\boldsymbol{B}_1^2 = \boldsymbol{B}_1$. With $\boldsymbol{B}_1 = \boldsymbol{A}_2 + \boldsymbol{A}_3$, where $\boldsymbol{B}_1^2 = \boldsymbol{B}_1$, $\boldsymbol{A}_2^2 = \boldsymbol{A}_2$, it follows from the case of $k = 2$ that $\boldsymbol{A}_2\boldsymbol{A}_3 = \mathbf{0}$ and $\boldsymbol{A}_3^2 = \boldsymbol{A}_3$. If we regroup by writing $\boldsymbol{A} = \boldsymbol{A}_2 + (\boldsymbol{A}_1 + \boldsymbol{A}_3)$, we obtain $\boldsymbol{A}_1\boldsymbol{A}_3 = \mathbf{0}$, and so on. ■

Remark 9.9.3. In our statement of Theorem 9.9.2, we took $X_1, X_2, \ldots, X_n$ to be observations of a random sample from a distribution that is $N(0, \sigma^2)$. We did this because our proof of Theorem 9.9.1 was restricted to that case. In fact, if $Q', Q_1', \ldots, Q_k'$ are quadratic forms in any normal variables (including multivariate normal variables), if $Q' = Q_1' + \cdots + Q_k'$, if $Q', Q_1', \ldots, Q_{k-1}'$ are central or noncentral chi-square, and if Q_k' is nonnegative, then $Q_1', \ldots, Q_k'$ are independent and Q_k' is either central or noncentral chi-square. ■

This section concludes with a proof of a frequently quoted theorem due to Cochran.

Theorem 9.9.3 (Cochran)**.** *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $N(0, \sigma^2)$. Let the sum of the squares of these observations be written in the form*

$$\sum_1^n X_i^2 = Q_1 + Q_2 + \cdots + Q_k,$$

where Q_j is a quadratic form in $X_1, X_2, \ldots, X_n$, with matrix $\boldsymbol{A}_j$ that has rank r_j, $j = 1, 2, \ldots, k$. The random variables $Q_1, Q_2, \ldots, Q_k$ are independent and Q_j/σ^2 is $\chi^2(r_j)$, $j = 1, 2, \ldots, k$, if and only if $\sum_1^k r_j = n$.

Proof. First assume the two conditions $\sum_1^k r_j = n$ and $\sum_1^n X_i^2 = \sum_1^k Q_j$ to be satisfied. The latter equation implies that $\boldsymbol{I} = \boldsymbol{A}_1 + \boldsymbol{A}_2 + \cdots + \boldsymbol{A}_k$. Let $\boldsymbol{B}_i = \boldsymbol{I} - \boldsymbol{A}_i$; that is, $\boldsymbol{B}_i$ is the sum of the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$ exclusive of $\boldsymbol{A}_i$. Let R_i denote the rank of $\boldsymbol{B}_i$. Since the rank of the sum of several matrices is less than or equal to the sum of the ranks, we have $R_i \leq \sum_1^k r_j - r_i = n - r_i$. However, $\boldsymbol{I} = \boldsymbol{A}_i + \boldsymbol{B}_i$, so that $n \leq r_i + R_i$ and $n - r_i \leq R_i$. Hence $R_i = n - r_i$. The eigenvalues of $\boldsymbol{B}_i$ are the roots of the equation $|\boldsymbol{B}_i - \lambda \boldsymbol{I}| = 0$. Since $\boldsymbol{B}_i = \boldsymbol{I} - \boldsymbol{A}_i$, this equation can be written as $|\boldsymbol{I} - \boldsymbol{A}_i - \lambda \boldsymbol{I}| = 0$. Thus we have $|\boldsymbol{A}_i - (1 - \lambda)\boldsymbol{I}| = 0$. But each root of the last equation is 1 minus an eigenvalue of $\boldsymbol{A}_i$. Since $\boldsymbol{B}_i$ has exactly $n - R_i = r_i$ eigenvalues that are zero, then $\boldsymbol{A}_i$ has exactly r_i eigenvalues that are equal to 1. However, r_i is the rank of $\boldsymbol{A}_i$. Thus each of the r_i nonzero eigenvalues of $\boldsymbol{A}_i$ is 1. That is, $\boldsymbol{A}_i^2 = \boldsymbol{A}_i$ and thus Q_i/σ^2 has a $\chi^2(r_i)$, for $i = 1, 2, \ldots, k$. In accordance with Theorem 9.9.2, the random variables $Q_1, Q_2, \ldots, Q_k$ are independent.

To complete the proof of Theorem 9.9.3, take

$$\sum_1^n X_i^2 = Q_1 + Q_2 + \cdots + Q_k,$$

let $Q_1, Q_2, \ldots, Q_k$ be independent, and let Q_j/σ^2 be $\chi^2(r_j)$, $j = 1, 2, \ldots, k$. Then $\sum_1^k Q_j/\sigma^2$ is $\chi^2(\sum_1^k r_j)$. But $\sum_1^k Q_j/\sigma^2 = \sum_1^n X_i^2/\sigma^2$ is $\chi^2(n)$. Thus $\sum_1^k r_j = n$ and the proof is complete. ■

EXERCISES

9.9.1. Let X_1, X_2, X_3 be a random sample from the normal distribution $N(0, \sigma^2)$. Are the quadratic forms $X_1^2 + 3X_1X_2 + X_2^2 + X_1X_3 + X_3^2$ and $X_1^2 - 2X_1X_2 + \frac{2}{3}X_2^2 - 2X_1X_2 - X_3^2$ independent or dependent?

9.9.2. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution that is $N(0, \sigma^2)$. Prove that $\sum_1^n X_i^2$ and every quadratic form, that is nonidentically zero in $X_1, X_2, \ldots, X_n$, are dependent.

9.9.3. Let X_1, X_2, X_3, X_4 denote a random sample of size 4 from a distribution that is $N(0, \sigma^2)$. Let $Y = \sum_1^4 a_i X_i$, where a_1, a_2, a_3, and a_4 are real constants. If Y^2 and $Q = X_1X_2 - X_3X_4$ are independent, determine a_1, a_2, a_3, and a_4.

9.9.4. Let $\boldsymbol{A}$ be the real symmetric matrix of a quadratic form Q in the observations of a random sample of size n from a distribution that is $N(0, \sigma^2)$. Given that Q and the mean $\overline{X}$ of the sample are independent, what can be said of the elements of each row (column) of $\boldsymbol{A}$?
Hint: Are Q and $\overline{X}^2$ independent?

9.9.5. Let $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_k$ be the matrices of $k > 2$ quadratic forms $Q_1, Q_2, \ldots, Q_k$ in the observations of a random sample of size n from a distribution that is $N(0, \sigma^2)$. Prove that the pairwise independence of these forms implies that they are mutually independent.
Hint: Show that $\boldsymbol{A}_i\boldsymbol{A}_j = \boldsymbol{0}$, $i \neq j$, permits $E[\exp(t_1Q_1 + t_2Q_2 + \cdots + t_kQ_k)]$ to be written as a product of the mgfs of $Q_1, Q_2, \ldots, Q_k$.

9.9.6. Let $\boldsymbol{X}' = [X_1, X_2, \ldots, X_n]$, where $X_1, X_2, \ldots, X_n$ are observations of a random sample from a distribution that is $N(0, \sigma^2)$. Let $\boldsymbol{b}' = [b_1, b_2, \ldots, b_n]$ be a real nonzero vector, and let $\boldsymbol{A}$ be a real symmetric matrix of order n. Prove that the linear form $\boldsymbol{b}'\boldsymbol{X}$ and the quadratic form $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ are independent if and only if $\boldsymbol{b}'\boldsymbol{A} = \boldsymbol{0}$. Use this fact to prove that $\boldsymbol{b}'\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ are independent if and only if the two quadratic forms $(\boldsymbol{b}'\boldsymbol{X})^2 = \boldsymbol{X}'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ are independent.

9.9.7. Let Q_1 and Q_2 be two nonnegative quadratic forms in the observations of a random sample from a distribution that is $N(0, \sigma^2)$. Show that another quadratic form Q is independent of $Q_1 + Q_2$ if and only if Q is independent of each of Q_1 and Q_2.
Hint: Consider the orthogonal transformation that diagonalizes the matrix of $Q_1 + Q_2$. After this transformation, what are the forms of the matrices Q, Q_1 and Q_2 if Q and $Q_1 + Q_2$ are independent?

9.9.8. Prove that Equation (9.9.12) of this section implies that the nonzero eigenvalues of the matrices $\boldsymbol{D}$ and $\boldsymbol{D}_{22}$ are the same.
Hint: Let $\lambda = 1/(2t_2)$, $t_2 \neq 0$, and show that Equation (9.9.12) is equivalent to $|\boldsymbol{D} - \lambda\boldsymbol{I}| = (-\lambda)^r|\boldsymbol{D}_{22} - \lambda\boldsymbol{I}_{n-r}|$.

9.9.9. Here Q_1 and Q_2 are quadratic forms in observations of a random sample from $N(0, 1)$. If Q_1 and Q_2 are independent and if $Q_1 + Q_2$ has a chi-square distribution, prove that Q_1 and Q_2 are chi-square variables.

9.9.10. Often in regression the mean of the random variable Y is a linear function of p-values $x_1, x_2, \ldots, x_p$, say $\beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$, where $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ are the *regression coefficients*. Suppose that n values, $\boldsymbol{Y}' = (Y_1, Y_2, \ldots, Y_n)$, are observed for the x-values in $\boldsymbol{X} = [x_{ij}]$, where $\boldsymbol{X}$ is an $n \times p$ *design matrix* and its ith row is associated with Y_i, $i = 1, 2, \ldots, n$. Assume that $\boldsymbol{Y}$ is multivariate normal with mean $\boldsymbol{X}\boldsymbol{\beta}$ and variance–covariance matrix $\sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix.

(a) Note that $Y_1, Y_2, \ldots, Y_n$ are independent. Why?

(b) Since $\boldsymbol{Y}$ should approximately equal its mean $\boldsymbol{X}\boldsymbol{\beta}$, we estimate $\boldsymbol{\beta}$ by solving the *normal equations* $\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$ for $\boldsymbol{\beta}$. Assuming that $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular, solve the equations to get $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. Show that $\hat{\boldsymbol{\beta}}$ has a

multivariate normal distribution with mean $\boldsymbol{\beta}$ and variance–covariance matrix $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

(c) Show that

$$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

For the remainder of the exercise, let Q denote the quadratic form on the left side of this expression and Q_1 and Q_2 denote the respective quadratic forms on the right side. Hence, $Q = Q_1 + Q_2$.

(d) Show that Q_1/σ^2 is $\chi^2(p)$.

(e) Show that Q_1 and Q_2 are independent.

(f) Argue that Q_2/σ^2 is $\chi^2(n-p)$.

(g) Find c so that cQ_1/Q_2 has an F-distribution.

(h) The fact that a value d can be found so that $P(cQ_1/Q_2 \leq d) = 1-\alpha$ could be used to find a $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\beta}$. Explain.

9.9.11. Say that G.P.A. (Y) is thought to be a linear function of a "coded" high school rank (x_2) and a "coded" American College Testing score (x_3), namely, $\beta_1 + \beta_2 x_2 + \beta_3 x_3$. Note that all x_1 values equal 1. We observe the following five points:

x_1	x_2	x_3	Y
1	1	2	3
1	4	3	6
1	2	2	4
1	4	2	4
1	3	2	4

(a) Compute $\boldsymbol{X}'\boldsymbol{X}$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

(b) Compute a 95% confidence ellipsoid for $\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_3)$. See part (h) of Exercise 9.9.10.

9.9.12. Assume that $\mathbf{X}$ is an $n \times p$ matrix. Then the kernel of $\mathbf{X}$ is defined to be the space $\ker(\mathbf{X}) = \{\mathbf{b} : \mathbf{X}\mathbf{b} = \mathbf{0}\}$.

(a) Show that $\ker(\mathbf{X})$ is a subspace of R^p.

(b) The dimension of $\ker(\mathbf{X})$ is called the **nullity** of $\mathbf{X}$ and is denoted by $\nu(\mathbf{X})$. Let $\rho(\mathbf{X})$ denote the rank of $\mathbf{X}$. A fundamental theorem of linear algebra says that $\rho(\mathbf{X}) + \nu(\mathbf{X}) = p$. Use this to show that if $\mathbf{X}$ has full column rank, then $\ker(\mathbf{X}) = \{\mathbf{0}\}$.

9.9.13. Suppose $\mathbf{X}$ is an $n \times p$ matrix with rank p.

(a) Show that $\ker(\mathbf{X}'\mathbf{X}) = \ker(\mathbf{X})$.

(b) Use part (a) and the last exercise to show that if $\mathbf{X}$ has full column rank, then $\mathbf{X}'\mathbf{X}$ is nonsingular.

Chapter 10

Nonparametric and Robust Statistics

10.1 Location Models

In this chapter, we present some nonparametric procedures for the simple location problems. As we shall show, the test procedures associated with these methods are distribution-free under null hypotheses. We also obtain point estimators and confidence intervals associated with these tests. The distributions of the estimators are not distribution-free; hence, we use the term **rank-based** to refer collectively to these procedures. The asymptotic relative efficiencies of these procedures are easily obtained, thus facilitating comparisons among them and procedures that we have discussed in earlier chapters. We also obtain estimators that are asymptotically efficient; that is, they achieve asymptotically the Rao–Cramér bound.

Our purpose is not a rigorous development of these concepts, and at times we simply sketch the theory. A rigorous treatment can be found in several advanced texts, such as Randles and Wolfe (1979) or Hettmansperger and McKean (2011). For an applied discussion using R, see Kloke and McKean (2014).

In this and the following section, we consider the one-sample problem. For the most part, we consider continuous random variables X with cdf and pdf $F_X(x)$ and $f_X(x)$, respectively. We assume that $f_X(x) > 0$ on the support of X; so, in particular, $F_X(x)$ is strictly increasing on the support. In this and the succeeding chapters, we want to identify classes of parameters. Think of a parameter as a function of the cdf (or pdf) of a given random variable. For example, consider the mean μ of X. We can write it as $\mu_X = T(F_X)$ if T is defined as

$$T(F_X) = E(X).$$

As another example, recall that the median of a random variable X is a parameter ξ such that $F_X(\xi) = 1/2$; i.e., $\xi = F_X^{-1}(1/2)$. Hence, in this notation, we say that the parameter ξ is defined by the function $T(F_X) = F_X^{-1}(1/2)$. Note that these Ts are functions of the cdfs (or pdfs). We shall call them **functionals**.

Remark 10.1.1 (Natural Nonparametric Estimators)**.** Functionals induce nonparametric estimators naturally. Let $X_1, X_2, \ldots, X_n$ denote a random sample from some distribution with cdf $F(x)$ and let $T(F)$ be a functional. Let $x_1, x_2, \ldots, x_n$ be a realization of this sample. Recall that the empirical distribution function of the sample is given by

$$\widehat{F}_n(x) = n^{-1}[\#\{x_i \leq x\}], \quad -\infty < x < \infty. \tag{10.1.1}$$

Hence, F_n is a discrete cdf that puts mass (probability) $1/n$ at each x_i. Because $\widehat{F}_n(x)$ is a cdf, $T(\widehat{F}_n)$ is well defined. Furthermore, $T(\widehat{F}_n)$ depends only on the sample; hence, it is a statistic. We call $T(\widehat{F}_n)$ the **induced estimator** of $T(F)$. For example, if $T(F)$ is the mean of the distribution, then it is easy to see that $T(\widehat{F}_n) = \overline{x}$; see Exercise 10.1.3.

For another example, consider the median. Note that $\hat{F}_n$ is a discrete cdf; hence, we use the general definition of a median of a distribution that is given in Definition 1.7.2 of Chapter 1. Let $\hat{\theta}$ denote the usual sample median which is defined in expression (4.4.4); that is, $\hat{\theta} = x_{((n+1)/2)}$ if n is odd while $\hat{\theta} = [x_{(n/2)} + x_{((n/2)+1)}]/2$ if n is even. To show that $\hat{\theta}$ satisfies Definition 1.7.2, note that:

- If n is even, then $\hat{F}_n(\hat{\theta}) = 1/2$.
- If n is odd then

$$n^{-1}\#\{x_i < \hat{\theta}\} = \tfrac{1}{2} - \tfrac{1}{n} \leq 1/2 \text{ and } F_n(\hat{\theta}) \geq 1/2.$$

Thus in either case, by Definition 1.7.2, $\hat{\theta}$ is a median of $\hat{F}_n$. Note that when n is even any point in the interval $(X_{(n/2)}, X_{((n/2)+1)})$ satisfies the definition of a median. ■

We begin with the definition of a location functional.

Definition 10.1.1. *Let X be a continuous random variable with cdf $F_X(x)$ and pdf $f_X(x)$. We say that $T(F_X)$ is a* **location functional** *if it satisfies*

$$\text{If } Y = X + a, \text{ then } T(F_Y) = T(F_X) + a, \text{ for all } a \in R, \tag{10.1.2}$$
$$\text{If } Y = aX; \text{ then } T(F_Y) = aT(F_X), \text{ for all } a \neq 0. \tag{10.1.3}$$

For example, suppose T is the mean functional; i.e., $T(F_X) = E(X)$. Let $Y = X + a$; then $E(Y) = E(X + a) = E(X) + a$. Secondly, if $Y = aX$, then $E(Y) = aE(X)$. Hence the mean is a location functional. The next example shows that the median is a location functional.

Example 10.1.1. Let $F(x)$ be the cdf of X and let $T(F_X) = F_X^{-1}(1/2)$ be the median functional of X. Note that another way to state this is $F_X(T(F_X)) = 1/2$. Let $Y = X + a$. It then follows that the cdf of Y is $F_Y(y) = F_X(y - a)$. The following identity shows that $T(F_Y) = T(F_X) + a$:

$$F_Y(T(F_X) + a) = F_X(T(F_X) + a - a) = F_X(T(F_X)) = 1/2.$$

Next, suppose $Y = aX$. If $a > 0$, then $F_Y(y) = F_X(y/a)$ and, hence,

$$F_Y(aT(F_X)) = F_X(aT(F_X)/a) = F_X(T(F_X)) = 1/2.$$

Thus $T(F_Y) = aT(F_X)$ when $a > 0$. On the other hand, if $a < 0$, then $F_Y(y) = 1 - F_X(y/a)$. Hence

$$F_Y(aT(F_X)) = 1 - F_X(aT(F_X)/a) = 1 - F_X(T(F_X)) = 1 - \frac{1}{2} = \frac{1}{2}.$$

Therefore, (10.1.3) holds for all $a \neq 0$. Thus the median is a location functional.

Recall that the median is a percentile, namely, the 50th percentile of a distribution. As Exercise 10.1.1 shows, the median is the only percentile that is a location functional. ■

We often continue to use parameter notation to denote functionals. For example, $\theta_X = T(F_X)$.

In Chapters 4 and 6, we wrote the location model for specified pdfs. In this chapter, we write it for a general pdf in terms of a specified location functional. Let X be a random variable with cdf $F_X(x)$ and pdf $f_X(x)$. Let $\theta_X = T(F_X)$ be a location functional. Define the random variable ε to be $\varepsilon = X - T(F_X)$. Then by (10.1.2), $T(F_\varepsilon) = 0$; i.e., ε has location 0, according to T. Further, the pdf of X can be written as $f_X(x) = f(x - T(F_X))$, where $f(x)$ is the pdf of ε.

Definition 10.1.2 (Location Model)**.** *Let $\theta_X = T(F_X)$ be a location functional. We say that the observations $X_1, X_2, \ldots, X_n$ follow a* **location model** *with functional $\theta_X = T(F_X)$ if*

$$X_i = \theta_X + \varepsilon_i, \tag{10.1.4}$$

where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are iid random variables with pdf $f(x)$ and $T(F_\varepsilon) = 0$. Hence, from the above discussion, $X_1, X_2, \ldots, X_n$ are iid with pdf $f_X(x) = f(x - T(F_X))$.

Example 10.1.2. Let ε be a random variable with cdf $F(x)$, such that $F(0) = 1/2$. Assume that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are iid with cdf $F(x)$. Let $\theta \in R$ and define

$$X_i = \theta + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Then $X_1, X_2, \ldots, X_n$ follow the location model with the locational functional θ, which is the median of X_i. ■

Note that the location model very much depends on the functional. It forces one to state clearly which location functional is being used in order to write the model statement. For the class of symmetric densities, though, all location functionals are the same.

Theorem 10.1.1. *Let X be a random variable with cdf $F_X(x)$ and pdf $f_X(x)$ such that the distribution of X is symmetric about a. Let $T(F_X)$ be any location functional. Then $T(F_X) = a$.*

Proof: By (10.1.2), we have

$$T(F_{X-a}) = T(F_X) - a. \tag{10.1.5}$$

Since the distribution of X is symmetric about a, it is easy to show that $X - a$ and $-(X - a)$ have the same distribution; see Exercise 10.1.2. Hence, using (10.1.2) and (10.1.3), we have

$$T(F_{X-a}) = T(F_{-(X-a)}) = -(T(F_X) - a) = -T(F_X) + a. \tag{10.1.6}$$

Putting (10.1.5) and (10.1.6) together gives the result. ■

The assumption of symmetry is very appealing, because the concept of "center" is unique when it is true.

EXERCISES

10.1.1. Let X be a continuous random variable with cdf $F(x)$. For $0 < p < 1$, let ξ_p be the pth quantile; i.e., $F(\xi_p) = p$. If $p \neq 1/2$, show that while property (10.1.2) holds, property (10.1.3) does not. Thus ξ_p is not a location parameter.

10.1.2. Let X be a continuous random variable with pdf $f(x)$. Suppose $f(x)$ is symmetric about a; i.e., $f(x - a) = f(-(x - a))$. Show that the random variables $X - a$ and $-(X - a)$ have the same pdf.

10.1.3. Let $\widehat{F}_n(x)$ denote the empirical cdf of the sample $X_1, X_2, \ldots, X_n$. The distribution of $\widehat{F}_n(x)$ puts mass $1/n$ at each sample item X_i. Show that its mean is $\overline{X}$. If $T(F) = F^{-1}(1/2)$ is the median, show that $T(\widehat{F}_n) = Q_2$, the sample median.

10.1.4. Let X be a random variable with cdf $F(x)$ and let $T(F)$ be a functional. We say that $T(F)$ is a **scale functional** if it satisfies the three properties

$$\begin{array}{lrcll} (i) & T(F_{aX}) & = & aT(F_X), & \text{for } a > 0 \\ (ii) & T(F_{X+b}) & = & T(F_X), & \text{for all } b \\ (iii) & T(F_{-X}) & = & T(F_X). & \end{array}$$

Show that the following functionals are scale functionals.

(a) The standard deviation, $T(F_X) = (\text{Var}(X))^{1/2}$.

(b) The interquartile range, $T(F_X) = F_X^{-1}(3/4) - F_X^{-1}(1/4)$.

10.2 Sample Median and the Sign Test

In this section, we consider inference for the median of a distribution using the sample median. Fundamental to this discussion is the sign test statistic, which we present first.

Let $X_1, X_2, \ldots, X_n$ be a random sample that follows the location model

$$X_i = \theta + \varepsilon_i, \tag{10.2.1}$$

where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are iid with cdf $F(x)$, pdf $f(x)$, and median 0. Note that in terms of Section 10.1, the location functional is the median and, hence, θ is the median of X_i. We begin with a test for the one-sided hypotheses

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta > \theta_0. \tag{10.2.2}$$

Consider the statistic

$$S(\theta_0) = \#\{X_i > \theta_0\}, \tag{10.2.3}$$

which is called the **sign statistic** because it counts the number of positive signs in the differences $X_i - \theta_0$, $i = 1, 2, \ldots, n$. If we define $I(x > a)$ to be 1 or 0 depending on whether $x > a$ or $x \leq a$, then we can express $S(\theta_0)$ as

$$S(\theta_0) = \sum_{i=1}^{n} I(X_i > \theta_0). \tag{10.2.4}$$

Note that if H_0 is true, then we expect one half of the observations to exceed θ_0, while if H_1 is true, we expect more than half of the observations to exceed θ_0. Consider then the test of the hypotheses (10.2.2) given by

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } S(\theta_0) \geq c. \tag{10.2.5}$$

Under the null hypothesis, the random variables $I(X_i > \theta_0)$ are iid with a Bernoulli $b(1, 1/2)$ distribution. Hence the null distribution of $S(\theta_0)$ is $b(n, 1/2)$ with mean $n/2$ and variance $n/4$. Note that under H_0, the sign test does not depend on the distribution of X_i. In general, we call such a test a **distribution free** test.

For a level α test, select c to be c_α, where c_α is the upper α critical point of a binomial $b(n, 1/2)$ distribution. The test statistic, though, has a discrete distribution, so for an exact test there are only a finite number of levels α available. The values of c_α are easily found by most computer packages. For instance, the R command `pbinom(0:15,15,.5)` returns the cdf of a binomial distribution with $n = 15$ and $p = 0.5$, from which all possible levels can be seen.

For a given data set, the p-value associated with the sign test is given by $\widehat{p} = P_{H_0}(S(\theta_0) \geq s)$, where s is the realized value of $S(\theta_0)$ based on the sample. For computation, the R command `1 - pbinom(s-1,n,.5)` computes $\widehat{p}$.

It is convenient at times to use a large sample test based on the asymptotic distribution of the test statistic. By the Central Limit Theorem, under H_0 the standardized statistic $[S(\theta_0) - (n/2)]/\sqrt{n}/2$ is asymptotically normal, $N(0, 1)$. Hence the large sample test rejects H_0 if

$$\frac{S(\theta_0) - (n/2)}{\sqrt{n}/2} \geq z_\alpha; \tag{10.2.6}$$

see Exercise 10.2.2.

We briefly touch on the two-sided hypotheses given by

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0. \tag{10.2.7}$$

The following symmetric decision rule seems appropriate:

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } S(\theta_0) \leq c_1 \text{ or if } S(\theta_0) \geq n - c_1. \tag{10.2.8}$$

For a level α test, c_1 would be chosen such that $\alpha/2 = P_{H_0}(S(\theta_0) \leq c_1)$. Recall that the p-value is given by $\widehat{p} = 2\min\{P_{H_0}(S(\theta_0) \leq s), P_{H_0}(S(\theta_0) \geq s)\}$, where s is the realized value of $S(\theta_0)$ based on the sample.

Example 10.2.1 (Shoshoni Rectangles)**.** A golden rectangle is a rectangle in which the ratio of the width (w) to the length (l) is the golden ratio, which is approximately 0.618. It can be characterized in various ways. For example, $w/l = l/(w+l)$ characterizes the golden rectangle. It is considered to be an aesthetic standard in Western civilization and appears in art and architecture going back to the ancient Greeks. It now appears in such items as credit and business cards. In a cultural anthropology study, DuBois (1960) reports on a study of the Shoshoni beaded baskets. These baskets contain beaded rectangles, and the question was whether the Shoshonis use the same aesthetic standard as the West. Let X denote the ratio of the width to the length of a Shoshoni beaded basket. Let θ be the median of X. The hypotheses of interest are

$$H_0 : \theta = 0.618 \text{ versus } H_1 : \theta \neq 0.618.$$

These are two-sided hypotheses. It follows from the above discussion that the sign test rejects H_0 in favor of H_1 if $S(0.618) \leq c$ or $S(0.618) \geq n - c$.

A sample of 20 width to length (ordered) ratios from Shoshoni baskets resulted in the data

Width-to-Length Ratios of Rectangles

0.553	0.570	0.576	0.601	0.606	0.606	0.609	0.611	0.615	0.628
0.654	0.662	0.668	0.670	0.672	0.690	0.693	0.749	0.844	0.933

The data can be found in the file `shoshoni.rda`. For these data, the sign test statistic is $S(0.618) = 11$. Using R the p-value is: `2*(1-pbinom(10,20,.5))= 0.8238`. Thus there is no evidence to reject H_0 based on these data.

A boxplot and a normal $q-q$ plot of the data are given in Figure 10.2.1. Notice that the data contain two, possibly three, potential outliers. The data do not appear to be drawn from a normal distribution. ■

We next obtain several useful results concerning the power function of the sign test for the hypotheses (10.2.2). The following function proves useful here and in the associated estimation and confidence intervals described below. Define

$$S(\theta) = \#\{X_i > \theta\}. \tag{10.2.9}$$

The sign test statistic is given by $S(\theta_0)$. We can easily describe the function $S(\theta)$. First, note that we can write it in terms of the order statistics $Y_1 < \cdots < Y_n$ of $X_1, \ldots, X_n$ because $\#\{Y_i > \theta\} = \#\{X_i > \theta\}$. Now if $\theta < Y_1$, then all the Y_is are larger than θ and, hence $S(\theta) = n$. Next, if $Y_1 \leq \theta < Y_2$ then $S(\theta) = n - 1$. Continuing this way, we see that $S(\theta)$ is a decreasing step function of θ, which steps

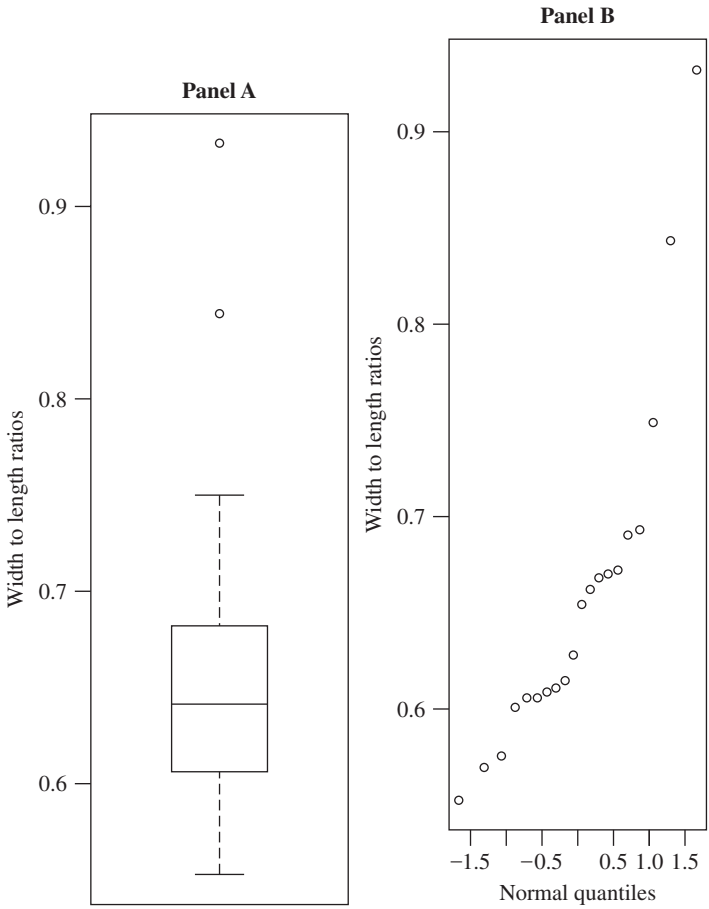

Figure 10.2.1: Boxplot (Panel A) and normal $q-q$ plot (Panel B) of the Shoshoni data.

down one unit at each order statistic Y_i, attaining its maximum and minimum values n and 0 at Y_1 and Y_n, respectively. Figure 10.2.2 depicts this function.

We need the following translation property. Because we can always subtract θ_0 from each X_i, we can assume without loss of generality that $\theta_0 = 0$.

Lemma 10.2.1. *For every k,*

$$P_\theta[S(0) \geq k] = P_0[S(-\theta) \geq k]. \tag{10.2.10}$$

Proof: Note that the left side of equation (10.2.10) concerns the probability of the event $\#\{X_i > 0\}$, where X_i has median θ. The right side concerns the probability of the event $\#\{(X_i + \theta) > 0\}$, where the random variable $X_i + \theta$ has median θ (because under $\theta = 0$, X_i has median 0). Hence the left and right sides give the same probability. ■

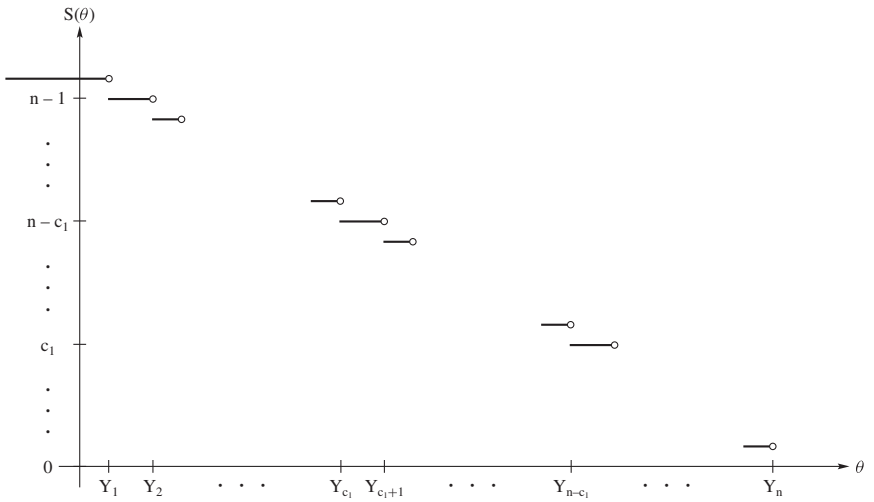

Figure 10.2.2: The sketch shows the graph of the decreasing step function $S(\theta)$. The function drops one unit at each order statistic Y_i.

Based on this lemma, it is easy to show that the power function of the sign test is monotone for one-sided tests.

Theorem 10.2.1. *Suppose Model (10.2.1) is true. Let $\gamma(\theta)$ be the power function of the sign test of level α for the one-sided hypotheses (10.2.2). Then $\gamma(\theta)$ is a nondecreasing function of θ.*

Proof: Let c_α denote the $b(n, 1/2)$ upper critical value as defined after expression (10.2.8). Without loss of generality, assume that $\theta_0 = 0$. The power function of the sign test is

$$\gamma(\theta) = P_\theta[S(0) \geq c_\alpha], \quad \text{for } -\infty < \theta < \infty.$$

Suppose $\theta_1 < \theta_2$. Then $-\theta_1 > -\theta_2$ and hence, since $S(\theta)$ is nonincreasing, $S(-\theta_1) \leq S(-\theta_2)$. This and Lemma 10.2.1 yield the desired result; i.e.,

$$\begin{aligned} \gamma(\theta_1) &= P_{\theta_1}[S(0) \geq c_\alpha] \\ &= P_0[S(-\theta_1) \geq c_\alpha] \\ &\leq P_0[S(-\theta_2) \geq c_\alpha] \\ &= P_{\theta_2}[S(0) \geq c_\alpha] \\ &= \gamma(\theta_2). \quad \blacksquare \end{aligned}$$

This is a very desirable property for any test. Because the monotonicity of the power function of the sign test holds for all θ, $-\infty < \theta < \infty$, we can extend the simple null hypothesis of (10.2.2) to the composite null hypothesis

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0. \tag{10.2.11}$$

Recall from Definition 4.5.4 of Chapter 4 that the size of the test for a composite null hypothesis is given by $\max_{\theta \leq \theta_0} \gamma(\theta)$. Because $\gamma(\theta)$ is nondecreasing, the size of the sign test is α for this extended null hypothesis. As a second result, it follows immediately that the sign test is an unbiased test; see Section 8.3. As Exercise 10.2.8 shows, the power function of the sign test for the other one-sided alternative, $H_1 : \theta < \theta_0$, is nonincreasing.

Under an alternative, say $\theta = \theta_1$, the test statistic $S(\theta_0)$ has the binomial distribution $b(n, p_1)$, where p_1 is given by

$$p_1 = P_{\theta_1}(X > 0) = 1 - F(-\theta_1), \tag{10.2.12}$$

where $F(x)$ is the cdf of ε in Model (10.2.1). Hence $S(\theta_0)$ is not distribution free under alternative hypotheses. As in Exercise 10.2.3, we can determine the power of the test for specified θ_1 and $F(x)$. We want to compare the power of the sign test to other size α tests, in particular the test based on the sample mean. However, for these comparison purposes, we need more general results, some of which are obtained in the next subsection.

10.2.1 Asymptotic Relative Efficiency

One solution to this problem is to consider the behavior of a test under a sequence of local alternatives. In this section, we often take $\theta_0 = 0$ in hypotheses (10.2.2). As noted before Lemma 10.2.1, this is without loss of generality. For the hypotheses (10.2.2), consider the sequence of alternatives

$$H_0 : \theta = 0 \text{ versus } H_{1n} : \theta_n = \frac{\delta}{\sqrt{n}}, \tag{10.2.13}$$

where $\delta > 0$. Note that this sequence of alternatives converges to the null hypothesis as $n \to \infty$. We often call such a sequence of alternatives **local alternatives**. The idea is to consider how the power function of a test behaves relative to the power functions of other tests under this sequence of alternatives. We only sketch this development. For more details, the reader can consult the more advanced books cited in Section 10.1. As a first step in that direction, we obtain the asymptotic power lemma for the sign test.

Consider the large sample size α test given by (10.2.6). Under the alternative

θ_n, we can approximate the mean of this test as follows:

$$\begin{aligned} E_{\theta_n}\left[\frac{1}{\sqrt{n}}\left(S(0)-\frac{n}{2}\right)\right] &= E_0\left[\frac{1}{\sqrt{n}}\left(S(-\theta_n)-\frac{n}{2}\right)\right] \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E_0[I(X_i > -\theta_n)] - \frac{\sqrt{n}}{2} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} P_0(X_i > -\theta_n) - \frac{\sqrt{n}}{2} \\ &= \sqrt{n}\left(1 - F(-\theta_n) - \frac{1}{2}\right) \\ &= \sqrt{n}\left(\frac{1}{2} - F(-\theta_n)\right) \\ &\approx \sqrt{n}\theta_n f(0) = \delta f(0), \end{aligned} \tag{10.2.14}$$

where the step to the last line is due to the mean value theorem. It can be shown in more advanced texts that the variance of $[S(0) - (n/2)]/(\sqrt{n}/2)$ converges to 1 under θ_n, just as under H_0, and that, furthermore, $[S(0)-(n/2)-\sqrt{n}\delta f(0)]/(\sqrt{n}/2)$ has a limiting standard normal distribution. This leads to the **asymptotic power lemma**, which we state in the form of a theorem.

Theorem 10.2.2 (Asymptotic Power Lemma)**.** *Consider the sequence of hypotheses (10.2.13). The limit of the power function of the large sample, size α, sign test is*

$$\lim_{n\to\infty} \gamma(\theta_n) = 1 - \Phi(z_\alpha - \delta\tau_S^{-1}), \tag{10.2.15}$$

where $\tau_S = 1/[2f(0)]$ and $\Phi(z)$ is the cdf of a standard normal random variable.

Proof: Using expression (10.2.14) and the discussion that followed its derivation, we have

$$\begin{aligned} \gamma(\theta_n) &= P_{\theta_n}\left[\frac{n^{-1/2}[S(0)-(n/2)]}{1/2} \geq z_\alpha\right] \\ &= P_{\theta_n}\left[\frac{n^{-1/2}[S(0)-(n/2)-\sqrt{n}\delta f(0)]}{1/2} \geq z_\alpha - \delta 2f(0)\right] \\ &\to 1 - \Phi(z_\alpha - \delta 2f(0)), \end{aligned}$$

which was to be shown. ■

As shown in Exercise 10.2.5, the parameter $\tau_S = 1/[2f(0)]$ is a scale parameter (functional) as defined in Exercise 10.1.4 of the last section. We later show that $\tau_S/\sqrt{n}$ is the asymptotic standard deviation of the sample median.

Note that there were several approximations used in the proof of Theorem 10.2.2. A rigorous proof can be found in more advanced texts, such as those cited in Section 10.1. It is quite helpful for the next sections to reconsider the approximation of the

mean given in (10.2.14) in terms of another concept called **efficacy**. Consider another standardization of the test statistic given by

$$\overline{S}(0) = \frac{1}{n}\sum_{i=1}^{n} I(X_i > 0), \tag{10.2.16}$$

where the bar notation is used to signify that $\overline{S}(0)$ is an average of $I(X_i > 0)$ and, in this case under H_0, converges in probability to $\frac{1}{2}$. Let $\mu(\theta) = E_\theta(\overline{S}(0) - \frac{1}{2})$. Then, by expression (10.2.14), we have

$$\mu(\theta_n) = E_{\theta_n}\left(\overline{S}(0) - \frac{1}{2}\right) = \frac{1}{2} - F(-\theta_n). \tag{10.2.17}$$

Let $\sigma^2_{\overline{S}} = \text{Var}(\overline{S}(0)) = \frac{1}{4n}$. Finally, define the **efficacy** of the sign test to be

$$c_S = \lim_{n\to\infty} \frac{\mu'(0)}{\sqrt{n}\sigma_{\overline{S}}}. \tag{10.2.18}$$

That is, the efficacy is the rate of change of the mean of the test statistic at the null divided by the product of $\sqrt{n}$ and the standard deviation of the test statistic at the null. So the efficacy increases with an increase in this rate, as it should. We use this formulation of efficacy throughout this chapter.

Hence, by expression (10.2.14), the efficacy of the sign test is

$$c_S = \frac{f(0)}{1/2} = 2f(0) = \tau_S^{-1}, \tag{10.2.19}$$

the reciprocal of the scale parameter τ_S. In terms of efficacy, we can write the conclusion of the Asymptotic Power Lemma as

$$\lim_{n\to\infty} \gamma(\theta_n) = 1 - \Phi(z_\alpha - \delta c_S). \tag{10.2.20}$$

This is not a coincidence, and it is true for the procedures we consider in the next section.

Remark 10.2.1. In this chapter, we compare nonparametric procedures with traditional parametric procedures. For instance, we compare the sign test with the test based on the sample mean. Traditionally, tests based on sample means are referred to as t-tests. Even though our comparisons are asymptotic and we could use the terminology of z-tests, we instead use the traditional terminology of t-tests. ■

As a second illustration of efficacy, we determine the efficacy of the t-test for the mean. Assume that the random variables ε_i in Model (10.2.1) are symmetrically distributed about 0 and their mean exists. Hence the parameter θ is the location parameter. In particular, $\theta = E(X_i) = \text{med}(X_i)$. Denote the variance of X_i by σ^2. This allows us to easily compare the sign and t-tests. Recall for hypotheses (10.2.2) that the t-test rejects H_0 in favor of H_1 if $\overline{X} \geq c$. The form of the test statistic is then $\overline{X}$. Furthermore, we have

$$\mu_{\overline{X}}(\theta) = E_\theta(\overline{X}) = \theta \tag{10.2.21}$$

and

$$\sigma^2_{\overline{X}}(0) = V_0(\overline{X}) = \frac{\sigma^2}{n}. \tag{10.2.22}$$

Thus, by (10.2.21) and (10.2.22), the efficacy of the t-test is

$$c_t = \lim_{n\to\infty} \frac{\mu'_{\overline{X}}(0)}{\sqrt{n}(\sigma/\sqrt{n})} = \frac{1}{\sigma}. \tag{10.2.23}$$

As confirmed in Exercise 10.2.9, the asymptotic power of the large sample level α, t-test under the sequence of alternatives (10.2.13) is $1 - \Phi(z_\alpha - \delta c_t)$. Thus we can compare the sign and t-tests by comparing their efficacies. We do this from the perspective of sample size determination.

Assume without loss of generality that $H_0 : \theta = 0$. Now suppose we want to determine the sample size so that a level α sign test can detect the alternative $\theta^* > 0$ with (approximate) probability γ^*. That is, find n so that

$$\gamma^* = \gamma(\theta^*) = P_{\theta^*}\left[\frac{S(0) - (n/2)}{\sqrt{n}/2} \geq z_\alpha\right]. \tag{10.2.24}$$

Write $\theta^* = \sqrt{n}\theta^*/\sqrt{n}$. Then, using the asymptotic power lemma, we have

$$\gamma^* = \gamma(\sqrt{n}\theta^*/\sqrt{n}) \approx 1 - \Phi(z_\alpha - \sqrt{n}\theta^*\tau_S^{-1}).$$

Now denote z_{γ^*} to be the upper $1-\gamma^*$ quantile of the standard normal distribution. Then, from this last equation, we have

$$z_{\gamma^*} = z_\alpha - \sqrt{n}\theta^*\tau_S^{-1}.$$

Solving for n, we get

$$n_S = \left(\frac{(z_\alpha - z_{\gamma^*})\tau_S}{\theta^*}\right)^2. \tag{10.2.25}$$

As outlined in Exercise 10.2.9, for this situation the sample size determination for the test based on the sample mean is

$$n_{\overline{X}} = \left(\frac{(z_\alpha - z_{\gamma^*})\sigma}{\theta^*}\right)^2, \tag{10.2.26}$$

where $\sigma^2 = \mathrm{Var}(\varepsilon)$.

Suppose we have two tests of the same level for which the asymptotic power lemma holds and for each we determine the sample size necessary to achieve power γ^* at the alternative θ^*. Then the ratio of these sample sizes is called the **asymptotic relative efficiency** (ARE) between the tests. We show later that this is the same as the ARE defined in Chapter 6 between estimators. Hence the ARE of the sign test to the t-test is

$$\mathrm{ARE}(S,t) = \frac{n_{\overline{X}}}{n_S} = \frac{\sigma^2}{\tau_S^2} = \frac{c_S^2}{c_t^2}. \tag{10.2.27}$$

Note that this is the same relative efficiency that was discussed in Example 6.2.5 when the sample median was compared to the sample mean. In the next two examples we revisit this discussion by examining the AREs when X_i has a normal distribution and then a Laplace (double exponential) distribution.

Example 10.2.2 (ARE(S,t): normal distribution)**.** Suppose $X_1, X_2, \ldots, X_n$ follow the location model (10.1.4), where $f(x)$ is a $N(0,\sigma^2)$ pdf. Then $\tau_S = (2f(0))^{-1} = \sigma\sqrt{\pi/2}$. Hence the ARE$(S,t)$ is given by

$$\text{ARE}(S,t) = \frac{\sigma^2}{\tau_S^2} = \frac{\sigma^2}{(\pi/2)\sigma^2} = \frac{2}{\pi} \approx 0.637. \tag{10.2.28}$$

Hence at the normal distribution the sign test is only 64% as efficient as the t-test. In terms of sample size at the normal distribution, the t-test requires a smaller sample, $0.64n_s$, where n_s is the sample size of the sign test, to achieve the same power as the sign test. A cautionary note is needed here because this is asymptotic efficiency. There have been ample empirical (simulation) studies that give credence to these numbers. ■

Example 10.2.3 (ARE(S,t) at the Laplace distribution)**.** For this example, consider Model (10.1.4), where $f(x)$ is the Laplace pdf $f(x) = (2b)^{-1}\exp\{-|x|/b\}$ for $-\infty < x < \infty$ and $b > 0$. Then $\tau_S = (2f(0))^{-1} = b$, while $\sigma^2 = E(X^2) = 2b^2$. Hence the ARE(S,t) is given by

$$\text{ARE}(S,t) = \frac{\sigma^2}{\tau_S^2} = \frac{2b^2}{b^2} = 2. \tag{10.2.29}$$

So, at the Laplace distribution, the sign test is (asymptotically) twice as efficient as the t-test. In terms of sample size at the Laplace distribution, the t-test requires twice as large a sample as the sign test to achieve the same asymptotic power as the sign test.

Recall from Example 6.3.4 that the sign test is the scores type likelihood ratio test when the true distribution is the Laplace. ■

The normal distribution has much lighter tails than the Laplace distribution, because the two pdfs are proportional to $\exp\{-t^2/2\sigma^2\}$ and $\exp\{-|t|/b\}$, respectively. Based on the last two examples, it seems that the t-test is more efficient for light-tailed distributions while the sign test is more efficient for heavier-tailed distributions. This is true in general and we illustrate this in the next example where we can easily vary the tail weight from light to heavy.

Example 10.2.4 (ARE(S,t) at a family of contaminated normals)**.** Consider the location Model (10.1.4), where the cdf of ε_i is the contaminated normal cdf given in expression (3.4.19). Assume that $\theta_0 = 0$. Recall that for this distribution, $(1-\epsilon)$ proportion of the time the sample is drawn from a $N(0,b^2)$ distribution, while ϵ proportion of the time the sample is drawn from a $N(0,b^2\sigma_c^2)$ distribution. The corresponding pdf is given by

$$f(x) = \frac{1-\epsilon}{b}\phi\left(\frac{x}{b}\right) + \frac{\epsilon}{b\sigma_c}\phi\left(\frac{x}{b\sigma_c}\right), \tag{10.2.30}$$

where $\phi(z)$ is the pdf of a standard normal random variable. As shown in Section 3.4, the variance of ε_i is $b^2(1+\epsilon(\sigma_c^2-1))$. Also, $\tau_s = (b\sqrt{\pi/2})/[1-\epsilon+(\epsilon/\sigma_c)]$. Thus the ARE is

$$\text{ARE}(S,t) = \frac{2}{\pi}[(1+\epsilon(\sigma_c^2-1)][1-\epsilon+(\epsilon/\sigma_c)]^2. \tag{10.2.31}$$

For example, the following table (see Exercise 6.2.6) shows the AREs for various values of ϵ when σ_c is set at 3.0:

ϵ	0	0.01	0.02	0.03	0.05	0.10	0.15	0.25
ARE(S,t)	0.636	0.678	0.718	0.758	0.832	0.998	1.134	1.326

Note: if ϵ increases over the range of values in the table, then the contamination effect becomes larger (generally resulting in a heavier-tailed distribution) and as the table shows, the sign test becomes more efficient relative to the t-test. Increasing σ_c has the same effect. It does take, however, with $\sigma_c = 3$, over 10% contamination before the sign test becomes more efficient than the t-test. ■

10.2.2 Estimating Equations Based on the Sign Test

In practice, we often want to estimate θ, the median of X_i, in Model (10.2.1). The associated point estimate based on the sign test can be described with a simple geometry, which is analogous to the geometry of the sample mean. As Exercise 10.2.6 shows, the sample mean $\overline{X}$ is such that

$$\overline{X} = \text{Argmin}\sqrt{\sum_{i=1}^{n}(X_i-\theta)^2}. \tag{10.2.32}$$

The quantity $\sqrt{\sum_{i=1}^{n}(X_i-\theta)^2}$ is the Euclidean distance between the vector of observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ and the vector $\theta\mathbf{1}$. If we simply interchange the square root and the summation symbols, we go from the Euclidean distance to the L_1 distance. Let

$$\widehat{\theta} = \text{Argmin}\sum_{i=1}^{n}|X_i-\theta|. \tag{10.2.33}$$

To determine $\widehat{\theta}$, simply differentiate the quantity on the right side with respect to θ (as in Chapter 6, define the derivative of $|x|$ to be 0 at $x=0$). We then obtain

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}|X_i-\theta| = -\sum_{i=1}^{n}\text{sgn}(X_i-\theta).$$

Setting this to 0, we obtain the estimating equations (EE)

$$\sum_{i=1}^{n}\text{sgn}(X_i-\theta) = 0, \tag{10.2.34}$$

whose solution is the sample median Q_2, (4.4.4).

Because our observations are continuous random variables, we have the identity

$$\sum_{i=1}^{n} \text{sgn}(X_i - \theta) = 2S(\theta) - n.$$

Hence the sample median also solves $S(\theta) \approx n/2$. Consider again Figure 10.2.2. Imagine $n/2$ on the vertical axis. This is halfway in the total drop of $S(\theta)$, from n to 0. The order statistic on the horizontal axis corresponding to $n/2$ is essentially the sample median (middle order statistic). In terms of testing, this last equation says that, based on the data, the sample median is the "most acceptable" hypothesis, because $n/2$ is the null expected value of the test statistic. We often think of this as estimation by the inversion of a test.

We now sketch the asymptotic distribution of the sample median. Assume without loss of generality that the true median of X_i is 0. Suppose $-\infty < x < \infty$. Using the fact that $S(\theta)$ is nonincreasing and the identity $S(\theta) \approx n/2$, we have the following equivalences:

$$\{\sqrt{n}Q_2 \leq x\} \Leftrightarrow \left\{Q_2 \leq \frac{x}{\sqrt{n}}\right\} \Leftrightarrow \left\{S\left(\frac{x}{\sqrt{n}}\right) \leq \frac{n}{2}\right\}.$$

Hence we have

$$\begin{aligned} P_0(\sqrt{n}Q_2 \leq x) &= P_0\left[S\left(\frac{x}{\sqrt{n}}\right) \leq \frac{n}{2}\right] \\ &= P_{-x/\sqrt{n}}\left[S(0) \leq \frac{n}{2}\right] \\ &= P_{-x/\sqrt{n}}\left[\frac{S(0) - (n/2)}{\sqrt{n}/2} \leq 0\right] \\ &\rightarrow \Phi(0 - x\tau_S^{-1}) = P(\tau_S Z \leq x), \end{aligned}$$

where Z has a standard normal distribution, Notice that the limit was obtained by invoking the Asymptotic Power Lemma with $\alpha = 0.5$ and hence $z_\alpha = 0$. Rearranging the last term earlier, we obtain the asymptotic distribution of the sample median, which we state as a theorem:

Theorem 10.2.3. *For the random sample $X_1, X_2, \ldots, X_n$, assume that Model (10.2.1) holds. Suppose that $f(0) > 0$. Let Q_2 denote the sample median. Then*

$$\sqrt{n}(Q_2 - \theta) \rightarrow N(0, \tau_S^2), \tag{10.2.35}$$

where $\tau_S = (2f(0))^{-1}$.

In Section 6.2 we defined the ARE between two estimators to be the reciprocal of their asymptotic variances. For the sample median and mean, this is the same ratio as that based on sample size determinations of their respective tests given earlier in expression (10.2.27).

10.2.3 Confidence Interval for the Median

In Section 4.4, we obtained a confidence interval for the median. In this section, we derive this confidence interval by inverting the sign test. Based on the monotonicity of $S(\theta)$, the derivation is straightforward, but the technique will prove useful in subsequent sections of this chapter.

Suppose the random sample $X_1, X_2, \ldots, X_n$ follows the location model (10.2.1). In this subsection, we develop a confidence interval for the median θ of X_i. Assuming that θ is the true median, the random variable $S(\theta)$, (10.2.9), has a binomial $b(n, 1/2)$ distribution. For $0 < \alpha < 1$, select c_1 so that $P_\theta[S(\theta) \leq c_1] = \alpha/2$. Hence we have

$$1 - \alpha = P_\theta[c_1 < S(\theta) < n - c_1]. \tag{10.2.36}$$

Recall in our derivation for the t-confidence interval for the mean in Chapter 3, we began with such a statement and then "inverted" the pivot random variable $t = \sqrt{n}(\overline{X} - \mu)/S$ (S in this expression is the sample standard deviation) to obtain an equivalent inequality with μ isolated in the middle. In this case, the function $S(\theta)$ does not have an inverse, but it is a decreasing step function of θ and the inversion can still be performed. As depicted in Figure 10.2.2, $c_1 < S(\theta) < n - c_1$ if and only if $Y_{c_1+1} \leq \theta < Y_{n-c_1}$, where $Y_1 < Y_2 < \cdots < Y_n$ are the order statistics of the sample $X_1, X_2, \ldots, X_n$. Therefore, the interval $[Y_{c_1+1}, Y_{n-c_1})$ is a $(1-\alpha)100\%$ confidence interval for the median θ. Because the order statistics are continuous random variables, the interval (Y_{c_1+1}, Y_{n-c_1}) is an equivalent confidence interval.

If n is large, then there is a large sample approximation to c_1. We know from the Central Limit Theorem that $S(\theta)$ is approximately normal with mean $n/2$ and variance $n/4$. Then, using the continuity correction, we obtain the approximation

$$c_1 \approx \frac{n}{2} - \frac{z_{\alpha/2}\sqrt{n}}{2} - \frac{1}{2}, \tag{10.2.37}$$

where $\Phi(-z_{\alpha/2}) = \alpha/2$; see Exercise 10.2.7. In practice, we use the closest integer to c_1.

Example 10.2.5 (Example 10.2.1, Continued)**.** There are 20 data points in the Shoshoni basket data. The sample median of the width to the length is $0.5(0.628 + 0.654) = 0.641$. Because $0.021 = P_{H_0}(S(0.618) \leq 5)$, a 95.8% confidence interval for θ is the interval $(y_6, y_{15}) = (0.606, 0.672)$, which includes 0.618, the ratio of the width to the length, which characterizes the golden rectangle.

Currently, there is not an intrinsic R function for the one-sample sign analysis. The R function `onesampsgn.R`, which can be downloaded at the site listed in the Preface, computes this analysis. For these data, its default 95% confidence interval is the same as that computed above. ■

EXERCISES

10.2.1. Sketch Figure 10.2.2 for the Shoshoni basket data found in Example 10.2.1. Show the values of the test statistic, the point estimate, and the 95.8% confidence interval of Example 10.2.5 on the sketch.

10.2.2. Show that the test given by (10.2.6) has asymptotically level α; that is, show that under H_0,

$$\frac{S(\theta_0) - (n/2)}{\sqrt{n}/2} \xrightarrow{D} Z,$$

where Z has a $N(0,1)$ distribution.

10.2.3. Let θ denote the median of a random variable X. Consider testing

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta > 0 .$$

Suppose we have a sample of size $n = 25$.

(a) Let $S(0)$ denote the sign test statistic. Determine the level of the test: reject H_0 if $S(0) \geq 16$.

(b) Determine the power of the test in part (a) if X has $N(0.5, 1)$ distribution.

(c) Assuming X has finite mean $\mu = \theta$, consider the asymptotic test of rejecting H_0 if $\overline{X}/(\sigma/\sqrt{n}) \geq k$. Assuming that $\sigma = 1$, determine k so the asymptotic test has the same level as the test in part (a). Then determine the power of this test for the situation in part (b).

10.2.4. To appreciate the importance of setting the location functional, consider the length of rivers data set, as taken from Tukey (1977). This data set contains the lengths of 141 American rivers in miles and it can be found in the file `lengthriver.rda`.

(a) Suppose the location functional is the median. Obtain the estimate and a 95% confidence interval for it. Use the confidence interval discussed in Section 10.2.3. Interpret it in terms of the data. Use the R function `onesampsgn.R` for computation.

(b) Suppose the location functional is the mean. Obtain the estimate and the 95% t-confidence interval for it. Interpret it in terms of the data.

(c) Obtain the boxplot of the data and sketch the estimates and confidence intervals on it. Discuss.

10.2.5. Recall the definition of a scale functional given in Exercise 10.1.4. Show that the parameter τ_S defined in Theorem 10.2.2 is a scale functional.

10.2.6. Show that the sample mean solves Equation (10.2.32).

10.2.7. Derive the approximation (10.2.37).

10.2.8. Show that the power function of the sign test is nonincreasing for the hypotheses

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta < \theta_0. \tag{10.2.38}$$

10.2.9. Let $X_1, X_2, \ldots, X_n$ be a random sample that follows the location model (10.2.1). In this exercise we want to compare the sign tests and t-test of the hypotheses (10.2.2); so we assume the random errors ε_i are symmetrically distributed about 0. Let $\sigma^2 = \text{Var}(\varepsilon_i)$. Hence the mean and the median are the same for this location model. Assume, also, that $\theta_0 = 0$. Consider the large sample version of the t-test, which rejects H_0 in favor of H_1 if $\overline{X}/(\sigma/\sqrt{n}) > z_\alpha$.

(a) Obtain the power function, $\gamma_t(\theta)$, of the large sample version of the t-test.

(b) Show that $\gamma_t(\theta)$ is nondecreasing in θ.

(c) Show that $\gamma_t(\theta_n) \to 1 - \Phi(z_\alpha - \sigma\theta^*)$, under the sequence of local alternatives (10.2.13).

(d) Based on part (c), obtain the sample size determination for the t-test to detect θ^* with approximate power γ^*.

(e) Derive the ARE(S, t) given in (10.2.27).

10.3 Signed-Rank Wilcoxon

Let $X_1, X_2, \ldots, X_n$ be a random sample that follows Model (10.2.1). Inference for θ based on the sign test is simple and requires few assumptions about the underlying distribution of X_i. On the other hand, sign procedures have the low efficiency of 0.64 relative to procedures based on the t-test given an underlying normal distribution. In this section, we discuss a nonparametric procedure that does attain high efficiency relative to the t-test. We make the additional assumption that the pdf $f(x)$ of ε_i in Model (10.2.1) is symmetric; i.e., $f(x) = f(-x)$, for all x such that $-\infty < x < \infty$. Hence X_i is symmetrically distributed about θ. Thus, by Theorem 10.1.1, all location parameters are identical.

First, consider the one-sided hypotheses

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta > 0. \tag{10.3.1}$$

There is no loss of generality in assuming that the null hypothesis is $H_0 : \theta = 0$, for if it were $H_0 : \theta = \theta_0$, we would consider the sample $X_1 - \theta_0, \ldots, X_n - \theta_0$. Under a symmetric pdf, observations X_i that are the same distance from 0 are equilikely and hence should receive the same weight. A test statistic that does this is the **signed-rank Wilcoxon** given by

$$T = \sum_{i=1}^{n} \text{sgn}(X_i) R|X_i|, \tag{10.3.2}$$

where $R|X_i|$ denotes the rank of X_i among $|X_1|, \ldots, |X_n|$, where the rankings are from low to high. Intuitively, under the null hypothesis, we expect half of the X_is to be positive and half to be negative. Further, the ranks are uniformly distributed on the integers $\{1, 2, \ldots, n\}$. Hence values of T around 0 are indicative of H_0. On the

other hand, if H_1 is true, then we expect more than half of the X_is to be positive and further, the positive observations are more likely to receive the higher ranks. Thus an appropriate decision rule is

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } T \geq c, \tag{10.3.3}$$

where c is determined by the level α of the test.

Given α, we need the null distribution of T to determine the critical point c. The set of integers $\{-n(n+1)/2, -[n(n+1)/2]+2, \ldots, n(n+1)/2\}$ form the support of T. Also, from Section 10.2, we know that the signs are iid with support $\{-1, 1\}$ and pmf

$$p(-1) = p(1) = \frac{1}{2}. \tag{10.3.4}$$

A key result is the following lemma:

Lemma 10.3.1. *Under H_0 and symmetry about 0 for the pdf, the random variables $|X_1|, \ldots, |X_n|$ are independent of the random variables $sgn(X_1), \ldots, sgn(X_n)$.*

Proof: Because $X_1, \ldots, X_n$ is a random sample from the cdf $F(x)$, it suffices to show that $P[|X_i| \leq x, \text{sgn}(X_i) = 1] = P[|X_i| \leq x]P[\text{sgn}(X_i) = 1]$. Due to H_0 and the symmetry of $f(x)$, this follows from the following string of equalities

$$\begin{aligned} P[|X_i| \leq x, \text{sgn}(X_i) = 1] &= P[0 < X_i \leq x] = F(x) - \frac{1}{2} \\ &= [2F(x) - 1]\frac{1}{2} = P[|X_i| \leq x]P[\text{sgn}(X_i) = 1]. \quad \blacksquare \end{aligned}$$

Based on this lemma, the ranks of the $|X_i|$s are independent of the signs of the X_is. Note that the ranks are a permutation of the integers $1, 2, \ldots, n$. By the lemma this independence is true for any permutation. In particular, suppose we use the permutation that orders the absolute values. For example, suppose the observations are $-6.1, 4.3, 7.2, 8.0, -2.1$. Then the permutation $5, 2, 1, 3, 4$ orders the absolute values; that is, the fifth observation is the smallest in absolute value, the second observation is the next smallest, etc. This permutation is called the **anti-ranks**, which we denote generally by by $i_1, i_2, \ldots, i_n$. Using the anti-ranks, we can write T as

$$T = \sum_{j=1}^{n} j \, \text{sgn}(X_{i_j}), \tag{10.3.5}$$

where, by Lemma 10.3.1, $\text{sgn}(X_{i_j})$ are iid with support $\{-1, 1\}$ and pmf (10.3.4).

Based on this observation, for s such that $-\infty < s < \infty$, the mgf of T is

$$\begin{aligned} E[\exp\{sT\}] &= E\left[\exp\left\{\sum_{j=1}^{n} sj\,\mathrm{sgn}(X_{i_j})\right\}\right] \\ &= \prod_{j=1}^{n} E[\exp\{sj\,\mathrm{sgn}(X_{i_j})\}] \\ &= \prod_{j=1}^{n}\left(\frac{1}{2}e^{-sj} + \frac{1}{2}e^{sj}\right) \\ &= \frac{1}{2^n}\prod_{j=1}^{n}\left(e^{-sj} + e^{sj}\right). \end{aligned} \tag{10.3.6}$$

Because the mgf does not depend on the underlying symmetric pdf $f(x)$, the test statistic T is distribution free under H_0. Although the pmf of T cannot be obtained in closed form, this mgf can be used to generate the pmf for a specified n; see Exercise 10.3.1.

Because the $\mathrm{sgn}(X_{i_j})$s are mutually independent with mean zero, it follows that $E_{H_0}[T] = 0$. Further, because the variance of $\mathrm{sgn}(X_{i_j})$ is 1, we have

$$\mathrm{Var}_{H_0}(T) = \sum_{j=1}^{n} \mathrm{Var}_{H_0}(j\,\mathrm{sgn}(X_{i_j})) = \sum_{j=1}^{n} j^2 = n(n+1)(2n+1)/6.$$

We summarize these results in the following theorem:

Theorem 10.3.1. *Assume that Model (10.2.1) is true for the random sample $X_1, \ldots, X_n$. Assume also that the pdf $f(x)$ is symmetric about 0. Then under H_0,*

$$T \text{ is distribution free with a symmetric pmf} \tag{10.3.7}$$

$$E_{H_0}[T] = 0 \tag{10.3.8}$$

$$Var_{H_0}(T) = \frac{n(n+1)(2n+1)}{6} \tag{10.3.9}$$

$$\frac{T}{\sqrt{Var_{H_0}(T)}} \text{ has an asymptotically } N(0,1) \text{ distribution.} \tag{10.3.10}$$

Proof: The first part of (10.3.7) and the expressions (10.3.8) and (10.3.9) were derived above. The asymptotic distribution of T certainly is plausible and its proof can be found in more advanced books. To obtain the second part of (10.3.7), we need to show that the distribution of T is symmetric about 0. But by the mgf of Y, (10.3.6), we have

$$E[\exp\{s(-T)\} = E[\exp\{(-s)T\}] = E[\exp\{sT\}].$$

Hence T and $-T$ have the same distribution, so T is symmetrically distributed about 0. ■

Note that the support of T is much denser than that of the sign test, so the normal approximation is good even for a sample size of 10.

There is another formulation of T that is convenient. Let T^+ denote the sum of the ranks of the positive X_is. Then, because the sum of all ranks is $n(n+1)/2$, we have

$$\begin{aligned} T &= \sum_{i=1}^{n} \text{sgn}(X_i) R|X_i| = \sum_{X_i>0} R|X_i| - \sum_{X_i<0} R|X_i| \\ &= 2\sum_{X_i>0} R|X_i| - \frac{n(n+1)}{2} \\ &= 2T^+ - \frac{n(n+1)}{2}. \end{aligned} \tag{10.3.11}$$

Hence T^+ is a linear function of T and thus is an equivalent formulation of the signed-rank test statistic T. For the record, we note the null mean and variance of T^+:

$$E_{H_0}(T^+) = \tfrac{n(n+1)}{4} \text{ and } \text{Var}_{H_0}(T^+) = \tfrac{n(n+1)(2n+1)}{24}. \tag{10.3.12}$$

The intrinsic R function `wilcox.test` computes the signed-rank analysis, returning the test statistic T^+ and the p-value. If the sample is in the R vector `x` then the signed-rank test of the hypotheses (10.3.1) is computed by the R command `wilcox.test(x,alt="greater")`. The arguments for the other one-sided and the two-sided hypotheses are respectively `alt="less"` and `alt="two.sided"`. To compute the signed-rank test of the hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, use the command `wilcox.test(x,alt="two.sided",mu=theta0)`. Also, the R call `psignrank(y,n)` computes the cdf of T^+ at y.

Example 10.3.1 (*Zea mays* Data of Darwin)**.** Reconsider the data set discussed in Example 4.5.1. Recall that W_i is the difference in heights of the cross-fertilized plant minus the self-fertilized plant in pot i, for $i = 1, \ldots, 15$. Let θ be the location parameter and consider the one-sided hypotheses

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta > 0. \tag{10.3.13}$$

Table 10.3.1 displays the data and the signed ranks.

Adding up the ranks of the positive items in column 5 of Table 10.3.1, we obtain $T^+ = 96$. Using the exact distribution, the R command is `1-psignrank(95,15)`), we obtain the p-value, $\widehat{p} = P_{H_0}(T^+ \geq 96) = 0.021$. For comparison, the asymptotic p-value, using the continuity correction is

$$\begin{aligned} P_{H_0}(T^+ \geq 96) = P_{H_0}(T^+ \geq 95.5) &\approx P\left(Z \geq \frac{95.5 - 60}{\sqrt{15 \cdot 16 \cdot 31/24}}\right) \\ &= P(Z \geq 2.016) = 0.022, \end{aligned}$$

which is quite close to the exact value of 0.021.

Suppose the R vector `ds` contains the paired differences between cross and self-fertilized. Then the R command `wilcox.test(ds,alt="greater")` computes the

Table 10.3.1: Signed Ranks for Darwin Data, Example 10.3.1

Pot	Cross-Fertilized	Self-Fertilized	Difference	Signed-Rank
1	23.500	17.375	6.125	11
2	12.000	20.375	−8.375	−14
3	21.000	20.000	1.000	2
4	22.000	20.000	2.000	4
5	19.125	18.375	0.750	1
6	21.550	18.625	2.925	5
7	22.125	18.625	3.500	7
8	20.375	15.250	5.125	9
9	18.250	16.500	1.750	3
10	21.625	18.000	3.625	8
11	23.250	16.250	7.000	12
12	21.000	18.000	3.000	6
13	22.125	12.750	9.375	15
14	23.000	15.500	7.500	13
15	12.000	18.000	−6.000	−10

value of T^+ along with the p-value. The computed values are the same as those computed above. ■

There is another formulation of T^+ which is useful for obtaining the properties of the Wilcoxon signed-rank test and confidence intervals for θ. Let $X_i > 0$ and consider all X_j such that $-X_i < X_j < X_i$. Thus all the averages $(X_i + X_j)/2$, under these restrictions, are positive, including $(X_i + X_i)/2$. From the restriction, though, the number of these positive averages is simply the $R|X_i|$. Doing this for all $X_i > 0$, we obtain

$$T^+ = \#_{i \le j}\{(X_j + X_i)/2 > 0\}. \tag{10.3.14}$$

The pairwise averages $(X_j + X_i)/2$ are often called the *Walsh averages*. Hence the signed-rank Wilcoxon can be obtained by counting the number of positive Walsh averages.

Based on the identity (10.3.14), we obtain the corresponding process. Let

$$T^+(\theta) = \#_{i \le j}\{[(X_j - \theta) + (X_i - \theta)]/2 > 0\} = \#_{i \le j}\{(X_j + X_i)/2 > \theta\}. \tag{10.3.15}$$

The process associated with $T^+(\theta)$ is much like the sign process, (10.2.9). Let $W_1 < W_2 < \cdots < W_{n(n+1)/2}$ denote the $n(n+1)/2$ ordered Walsh averages. Then a graph of $T^+(\theta)$ would appear as in Figure 10.2.2, except the ordered Walsh averages would be on the horizontal axis and the largest value on the vertical would be $n(n+1)/2$. Hence the function $T^+(\theta)$ is a decreasing step function of θ, which steps down one unit at each Walsh average. This observation greatly simplifies the discussion on the properties of the signed-rank Wilcoxon.

Let c_α denote the critical value of a level α test of the hypotheses (10.3.1) based on the signed-rank test statistic T^+; i.e., $\alpha = P_{H_0}(T^+ \geq c_\alpha)$. Let $\gamma_{SW}(\theta) = P_\theta(T^+ \geq c_\alpha)$, for $\theta \geq \theta_0$, denote the power function of the test. The translation property, Lemma 10.2.1, holds for the signed-rank Wilcoxon. Hence, as in Theorem 10.2.1, the power function is a nondecreasing function of θ. In particular, the signed-rank Wilcoxon test is an unbiased test for the one-sided hypotheses (10.3.1).

10.3.1 Asymptotic Relative Efficiency

We investigate the efficiency of the signed-rank Wilcoxon by first determining its efficacy. Without loss of generality, we can assume that $\theta_0 = 0$. Consider the same sequence of local alternatives discussed in the last section; i.e.,

$$H_0 : \theta = 0 \text{ versus } H_{1n} : \theta_n = \tfrac{\delta}{\sqrt{n}}, \tag{10.3.16}$$

where $\delta > 0$. Contemplate the modified statistic, which is the average of $T^+(\theta)$,

$$\overline{T}^+(\theta) = \frac{2}{n(n+1)} T^+(\theta). \tag{10.3.17}$$

Then, by (10.3.12),

$$E_0[\overline{T}^+(0)] = \tfrac{2}{n(n+1)} \tfrac{n(n+1)}{4} = \tfrac{1}{2} \text{ and } \sigma^2_{\overline{T}^+}(0) = \text{Var}_0[\overline{T}^+(0)] = \tfrac{2n+1}{6n(n+1)}. \tag{10.3.18}$$

Let $a_n = 2/n(n+1)$. Note that we can decompose $\overline{T}^+(\theta_n)$ into two parts as

$$\overline{T}^+(\theta_n) = a_n S(\theta_n) + a_n \sum_{i<j} I(X_i + X_j > 2\theta_n) = a_n S(\theta_n) + a_n T^*(\theta_n), \tag{10.3.19}$$

where $S(\theta)$ is the sign process (10.2.9) and

$$T^*(\theta_n) = \sum_{i<j} I(X_i + X_j > 2\theta_n). \tag{10.3.20}$$

To obtain the efficacy, we require the mean

$$\mu_{\overline{T}^+}(\theta_n) = E_{\theta_n}[\overline{T}^+(0)] = E_0[\overline{T}^+(-\theta_n)]. \tag{10.3.21}$$

But by (10.2.14), $a_n E_0(S(-\theta_n)) = a_n n(2^{-1} - F(-\theta_n)) \to 0$. Hence we need only be concerned with the second term in (10.3.19). But note that the Walsh averages in $T^*(\theta)$ are identically distributed. Thus

$$a_n E_0(T^*(-\theta_n)) = a_n \binom{n}{2} P_0(X_1 + X_2 > -2\theta_n). \tag{10.3.22}$$

This latter probability can be expressed as follows:

$$\begin{aligned} P_0(X_1 + X_2 > -2\theta_n) &= E_0[P_0(X_1 > -2\theta_n - X_2|X_2)] = E_0[1 - F(-2\theta_n - X_2)] \\ &= \int_{-\infty}^{\infty} [1 - F(-2\theta_n - x)] f(x)\, dx \\ &= \int_{-\infty}^{\infty} F(2\theta_n + x) f(x)\, dx \\ &\approx \int_{-\infty}^{\infty} [F(x) + 2\theta_n f(x)] f(x)\, dx \\ &= \frac{1}{2} + 2\theta_n \int_{-\infty}^{\infty} f^2(x)\, dx, \end{aligned} \tag{10.3.23}$$

where we have used the facts that X_1 and X_2 are iid and symmetrically distributed about 0, and the mean value theorem. Hence

$$\mu_{\overline{T}^+}(\theta_n) \approx a_n \binom{n}{2} \left(\frac{1}{2} + 2\theta_n \int_{-\infty}^{\infty} f^2(x)\, dx \right). \tag{10.3.24}$$

Putting (10.3.18) and (10.3.24) together, we have the efficacy

$$c_{T^+} = \lim_{n\to\infty} \frac{\mu'_{\overline{T}^+}(0)}{\sqrt{n}\sigma_{\overline{T}^+}(0)} = \sqrt{12} \int_{-\infty}^{\infty} f^2(x)\, dx. \tag{10.3.25}$$

In a more advanced text, this development can be made into a rigorous argument for the following asymptotic power lemma.

Theorem 10.3.2 (Asymptotic Power Lemma)**.** *Consider the sequence of hypotheses (10.3.16). The limit of the power function of the large sample, size α, signed-rank Wilcoxon test is given by*

$$\lim_{n\to\infty} \gamma_{SR}(\theta_n) = 1 - \Phi(z_\alpha - \delta\tau_W^{-1}), \tag{10.3.26}$$

where $\tau_W = 1/[\sqrt{12}\int_{-\infty}^{\infty} f^2(x)\, dx]$ is the reciprocal of the efficacy c_{T^+} and $\Phi(z)$ is the cdf of a standard normal random variable.

As shown in Exercise 10.3.10, the parameter τ_W is a scale functional.

The arguments used in the determination of the sample size in Section 10.2 for the sign test were based on the asymptotic power lemma; hence, these arguments follow almost verbatim for the signed-rank Wilcoxon. In particular, the sample size needed so that a level α signed-rank Wilcoxon test of the hypotheses (10.3.1) can detect the alternative $\theta = \theta_0 + \theta^*$ with approximate probability γ^* is

$$n_W = \left(\frac{(z_\alpha - z_{\gamma^*})\tau_W}{\theta^*} \right)^2. \tag{10.3.27}$$

Using (10.2.26), the ARE between the signed-rank Wilcoxon test and the t-test based on the sample mean is

$$\text{ARE}(T, t) = \frac{n_t}{n_T} = \frac{\sigma^2}{\tau_W^2}. \tag{10.3.28}$$

We now derive some AREs between the Wilcoxon and the t-test. As noted above, the parameter τ_W is a scale functional and, hence, varies directly with scale transformations of the form aX, for $a > 0$. Likewise, the standard deviation σ is also a scale functional. Therefore, because the AREs are ratios of scale functionals, they are scale invariant. Hence, for derivations of AREs, we can select a pdf with a convenient choice of scale. For example, if we are considering an ARE at the normal distribution, we can work with the $N(0,1)$ pdf.

Example 10.3.2 (ARE(W,t) at the normal distribution)**.** If $f(x)$ is a $N(0,1)$ pdf, then

$$\begin{aligned}\tau_W^{-1} &= \sqrt{12}\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\exp\{-x^2/2\}\right)^2 dx \\ &= \frac{\sqrt{12}}{\sqrt{2}\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}(1/\sqrt{2})}\exp\{-2^{-1}(x/(1/\sqrt{2}))^2\}\,dx = \sqrt{\frac{3}{\pi}}\end{aligned}$$

Hence $\tau_W^2 = \pi/3$. Since $\sigma = 1$, we have

$$\text{ARE}(W,t) = \frac{\sigma^2}{\tau_W^2} = \frac{3}{\pi} = 0.955. \tag{10.3.29}$$

As discussed above, this ARE holds for all normal distributions. Hence, at the normal distribution, the Wilcoxon signed-rank test is 95.5% efficient as the t-test. The Wilcoxon is called a **highly efficient** procedure. ■

Example 10.3.3 (ARE(W,t) at a Family of Contaminated Normals)**.** For this example, suppose that $f(x)$ is the pdf of a contaminated normal distribution. For convenience, we use the standardized pdf given in expression (10.2.30) with $b = 1$. Recall that for this distribution, $(1-\epsilon)$ proportion of the time the sample is drawn from a $N(0,1)$ distribution, while ϵ proportion of the time the sample is drawn from a $N(0,\sigma_c^2)$ distribution. Recall that the variance is $\sigma^2 = 1 + \epsilon(\sigma_c^2 - 1)$. Note that the formula for the pdf $f(x)$ is given in expression (3.4.17). In Exercise 10.3.5 it is shown that

$$\int_{-\infty}^{\infty} f^2(x)\,dx = \frac{(1-\epsilon)^2}{2\sqrt{\pi}} + \frac{\epsilon^2}{6\sqrt{\pi}} + \frac{\epsilon(1-\epsilon)}{2\sqrt{\pi}}. \tag{10.3.30}$$

Based on this, an expression for the ARE can be obtained; see Exercise 10.3.5. We used this expression to determine the AREs between the Wilcoxon and the t-tests for the situations with $\sigma_c = 3$ and ϵ varying from 0.00–0.25, displaying them in Table 10.3.2. For convenience, we have also displayed the AREs between the sign test and these two tests.

Note that the signed-rank Wilcoxon is more efficient than the t-test even at 1% contamination and increases to 150% efficiency for 15% contamination. ■

10.3.2 Estimating Equations Based on Signed-Rank Wilcoxon

For the sign procedure, the estimation of θ was based on minimizing the L_1 norm. The estimator associated with the signed-rank test minimizes another norm, which

Table 10.3.2: AREs among the sign, the Signed-Rank Wilcoxon, and the t-Tests for Contaminated Normals with $\sigma_c = 3$ and Proportion of Contamination ϵ

ϵ	0.00	0.01	0.02	0.03	0.05	0.10	0.15	0.25
ARE(W,t)	0.955	1.009	1.060	1.108	1.196	1.373	1.497	1.616
ARE(S,t)	0.637	0.678	0.719	0.758	0.833	0.998	1.134	1.326
ARE(W,S)	1.500	1.487	1.474	1.461	1.436	1.376	1.319	1.218

is discussed in Exercises 10.3.7 and 10.3.8. Recall that we also show that the location estimator based on the sign test could be obtained by inverting the test. Considering this for the Wilcoxon, the estimator $\widehat{\theta}_W$ solves

$$T^+(\widehat{\theta}_W) = \frac{n(n+1)}{4}. \tag{10.3.31}$$

Using the description of the function $T^+(\theta)$ after its definition, (10.3.15), it is easily seen that $\widehat{\theta}_W = \text{median}\{(X_i + X_j)/2\}$; i.e., the median of the Walsh averages. This is often called the Hodges–Lehmann estimator because of several seminal articles by Hodges and Lehmann on the properties of this estimator; see Hodges and Lehmann (1963).

The R function `wilcox.test` computes the Hodges–Lehmann estimate. To illustrate its computation, consider the Darwin data in Example 10.3.1. Let the R vector `ds` contain the paired differences, Cross − Self. The R code segment given by `wilcox.test(ds,conf.int=T)` then computes the Hodges–Lehmann estimate to be 3.1375. So we estimate the difference in heights to be 3.1375 inches.

Once again, we can use practically the same argument that we used for the sign process to obtain the asymptotic distribution of the Hodges–Lehmann estimator. We summarize the result in the next theorem.

Theorem 10.3.3. *Consider a random sample $X_1, X_2, X_3, \ldots, X_n$ which follows Model (10.2.1). Suppose that $f(x)$ is symmetric about 0. Then*

$$\sqrt{n}(\widehat{\theta}_W - \theta) \rightarrow N(0, \tau_W^2), \tag{10.3.32}$$

where $\tau_W = \left(\sqrt{12}\int_{-\infty}^{\infty} f^2(x)\,dx\right)^{-1}$.

Using this theorem, the AREs based on asymptotic variances for the signed-rank Wilcoxon are the same as those defined above.

10.3.3 Confidence Interval for the Median

Because of the similarity between the processes $S(\theta)$ and $T^+(\theta)$, confidence intervals for θ based on the signed-rank Wilcoxon follow the same way as do those based on $S(\theta)$. For a given level α, let c_{W1}, an integer, denote the critical point of the signed-rank Wilcoxon distribution such that $P_\theta[T^+(\theta) \leq c_{W1}] = \alpha/2$. As in Section 10.2.3,