(d) *The random variable*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{3.6.9}$$

has a Student t-distribution with $n-1$ degrees of freedom.

Proof: Note that we have proved part (a) in Corollary 3.4.1. Let $\mathbf{X} = (X_1, \ldots, X_n)'$. Because $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ random variables, $\mathbf{X}$ has a multivariate normal distribution $N(\mu\mathbf{1}, \sigma^2\mathbf{I})$, where $\mathbf{1}$ denotes a vector whose components are all 1. Let $\mathbf{v}' = (1/n, \ldots, 1/n) = (1/n)\mathbf{1}'$. Note that $\overline{X} = \mathbf{v}'\mathbf{X}$. Define the random vector $\mathbf{Y}$ by $\mathbf{Y} = (X_1 - \overline{X}, \ldots, X_n - \overline{X})'$. Consider the following transformation:

$$\mathbf{W} = \begin{bmatrix} \overline{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{v}' \\ \mathbf{I} - \mathbf{1}\mathbf{v}' \end{bmatrix} \mathbf{X}. \tag{3.6.10}$$

Because $\mathbf{W}$ is a linear transformation of multivariate normal random vector, by Theorem 3.5.2 it has a multivariate normal distribution with mean

$$E[\mathbf{W}] = \begin{bmatrix} \mathbf{v}' \\ \mathbf{I} - \mathbf{1}\mathbf{v}' \end{bmatrix} \mu\mathbf{1} = \begin{bmatrix} \mu \\ \mathbf{0}_n \end{bmatrix}, \tag{3.6.11}$$

where $\mathbf{0}_n$ denotes a vector whose components are all 0, and covariance matrix

$$\begin{aligned} \mathbf{\Sigma} &= \begin{bmatrix} \mathbf{v}' \\ \mathbf{I} - \mathbf{1}\mathbf{v}' \end{bmatrix} \sigma^2\mathbf{I} \begin{bmatrix} \mathbf{v}' \\ \mathbf{I} - \mathbf{1}\mathbf{v}' \end{bmatrix}' \\ &= \sigma^2 \begin{bmatrix} \frac{1}{n} & \mathbf{0}_n' \\ \mathbf{0}_n & \mathbf{I} - \mathbf{1}\mathbf{v}' \end{bmatrix}. \end{aligned} \tag{3.6.12}$$

Because $\overline{X}$ is the first component of $\mathbf{W}$, we can also obtain part (a) by Theorem 3.5.1. Next, because the covariances are 0, $\overline{X}$ is independent of $\mathbf{Y}$. But $S^2 = (n-1)^{-1}\mathbf{Y}'\mathbf{Y}$. Hence, $\overline{X}$ is independent of S^2, also. Thus part (b) is true.

Consider the random variable

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

Each term in this sum is the square of a $N(0,1)$ random variable and, hence, has a $\chi^2(1)$ distribution (Theorem 3.4.1). Because the summands are independent, it follows from Corollary 3.3.1 that V is a $\chi^2(n)$ random variable. Note the following identity:

$$\begin{aligned} V &= \sum_{i=1}^{n} \left(\frac{(X_i - \overline{X}) + (\overline{X} - \mu)}{\sigma} \right)^2 \\ &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \\ &= \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \right)^2. \end{aligned} \tag{3.6.13}$$

By part (b), the two terms on the right side of the last equation are independent. Further, the second term is the square of a standard normal random variable and, hence, has a $\chi^2(1)$ distribution. Taking mgfs of both sides, we have

$$(1-2t)^{-n/2} = E\left[\exp\{t(n-1)S^2/\sigma^2\}\right](1-2t)^{-1/2}. \tag{3.6.14}$$

Solving for the mgf of $(n-1)S^2/\sigma^2$ on the right side we obtain part (c). Finally, part (d) follows immediately from parts (a)–(c) upon writing T, (3.6.9), as

$$T = \frac{(\overline{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)S^2/(\sigma^2(n-1))}}. \quad \blacksquare$$

EXERCISES

3.6.1. Let T have a t-distribution with 10 degrees of freedom. Find $P(|T| > 2.228)$ from either Table III or by using R.

3.6.2. Let T have a t-distribution with 14 degrees of freedom. Determine b so that $P(-b < T < b) = 0.90$. Use either Table III or by using R.

3.6.3. Let T have a t-distribution with $r > 4$ degrees of freedom. Use expression (3.6.4) to determine the kurtosis of T. See Exercise 1.9.15 for the definition of kurtosis.

3.6.4. Using R, plot the pdfs of the random variables defined in parts (a)–(e) below. Obtain an overlay plot of all five pdfs, also.

(a) X has a standard normal distribution. Use this code:
`x=seq(-6,6,.01); plot(dnorm(x)~x)`.

(b) X has a t-distribution with 1 degree of freedom. Use the code:
`lines(dt(x,1)~x,lty=2)`.

(c) X has a t-distribution with 3 degrees of freedom.

(d) X has a t-distribution with 10 degrees of freedom.

(e) X has a t-distribution with 30 degrees of freedom.

3.6.5. Using R, investigate the probabilities of an "outlier" for a t-random variable and a normal random variable. Specifically, determine the probability of observing the event $\{|X| \geq 2\}$ for the following random variables:

(a) X has a standard normal distribution.

(b) X has a t-distribution with 1 degree of freedom.

(c) X has a t-distribution with 3 degrees of freedom.

(d) X has a t-distribution with 10 degrees of freedom.

(e) X has a t-distribution with 30 degrees of freedom.

3.6.6. In expression (3.4.13), the normal location model was presented. Often real data, though, have more outliers than the normal distribution allows. Based on Exercise 3.6.5, outliers are more probable for t-distributions with small degrees of freedom. Consider a location model of the form

$$X = \mu + e,$$

where e has a t-distribution with 3 degrees of freedom. Determine the standard deviation σ of X and then find $P(|X - \mu| \geq \sigma)$.

3.6.7. Let F have an F-distribution with parameters r_1 and r_2. Assuming that $r_2 > 2k$, continue with Example 3.6.2 and derive the $E(F^k)$.

3.6.8. Let F have an F-distribution with parameters r_1 and r_2. Using the results of the last exercise, determine the kurtosis of F, assuming that $r_2 > 8$.

3.6.9. Let F have an F-distribution with parameters r_1 and r_2. Argue that $1/F$ has an F-distribution with parameters r_2 and r_1.

3.6.10. Suppose F has an F-distribution with parameters $r_1 = 5$ and $r_2 = 10$. Using only 95th percentiles of F-distributions, find a and b so that $P(F \leq a) = 0.05$ and $P(F \leq b) = 0.95$, and, accordingly, $P(a < F < b) = 0.90$.
Hint: Write $P(F \leq a) = P(1/F \geq 1/a) = 1 - P(1/F \leq 1/a)$, and use the result of Exercise 3.6.9 and R.

3.6.11. Let $T = W/\sqrt{V/r}$, where the independent variables W and V are, respectively, normal with mean zero and variance 1 and chi-square with r degrees of freedom. Show that T^2 has an F-distribution with parameters $r_1 = 1$ and $r_2 = r$.
Hint: What is the distribution of the numerator of T^2?

3.6.12. Show that the t-distribution with $r = 1$ degree of freedom and the Cauchy distribution are the same.

3.6.13. Let F have an F-distribution with $2r$ and $2s$ degrees of freedom. Since the support of F is $(0, \infty)$, the F-distribution is often used to model time until failure (lifetime). In this case, $Y = \log F$ is used to model the log of lifetime. The log F family is a rich family of distributions consisting of left- and right-skewed distributions as well as symmetric distributions; see, for example, Chapter 4 of Hettmansperger and McKean (2011). In this exercise, consider the subfamily where $Y = \log F$ and F has 2 and $2s$ degrees of freedom.

(a) Obtain the pdf and cdf of Y.

(b) Using R, obtain a page of plots of these distributions for $s = .4, .6, 1.0, 2.0, 4.0, 8$. Comment on the shape of each pdf.

(c) For $s = 1$, this distribution is called the **logistic** distribution. Show that the pdf is symmetric about 0.

3.6.14. Show that

$$Y = \frac{1}{1 + (r_1/r_2)W},$$

where W has an F-distribution with parameters r_1 and r_2, has a beta distribution.

3.6.15. Let X_1, X_2 be iid with common distribution having the pdf $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Show that $Z = X_1/X_2$ has an F-distribution.

3.6.16. Let X_1, X_2, and X_3 be three independent chi-square variables with r_1, r_2, and r_3 degrees of freedom, respectively.

(a) Show that $Y_1 = X_1/X_2$ and $Y_2 = X_1 + X_2$ are independent and that Y_2 is $\chi^2(r_1 + r_2)$.

(b) Deduce that

$$\frac{X_1/r_1}{X_2/r_2} \quad \text{and} \quad \frac{X_3/r_3}{(X_1 + X_2)/(r_1 + r_2)}$$

are independent F-variables.

3.7 *Mixture Distributions

Recall the discussion on the contaminated normal distribution given in Section 3.4.1. This was an example of a mixture of normal distributions. In this section, we extend this to mixtures of distributions in general. Generally, we use continuous-type notation for the discussion, but discrete pmfs can be handled the same way.

Suppose that we have k distributions with respective pdfs $f_1(x), f_2(x), \ldots, f_k(x)$, with supports $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$, means $\mu_1, \mu_2, \ldots, \mu_k$, and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$, with positive mixing probabilities $p_1, p_2, \ldots, p_k$, where $p_1 + p_2 + \cdots + p_k = 1$. Let $\mathcal{S} = \cup_{i=1}^k \mathcal{S}_i$ and consider the function

$$f(x) = p_1 f_1(x) + p_2 f_2(x) + \cdots + p_k f_k(x) = \sum_{i=1}^k p_i f_i(x), \quad x \in \mathcal{S}. \tag{3.7.1}$$

Note that $f(x)$ is nonnegative and it is easy to see that it integrates to one over $(-\infty, \infty)$; hence, $f(x)$ is a pdf for some continuous-type random variable X. Integrating term-by-term, it follows that the cdf of X is:

$$F(x) = \sum_{i=1}^k p_i F_i(x), \quad x \in \mathcal{S}, \tag{3.7.2}$$

where $F_i(x)$ is the cdf corresponding to the pdf $f_i(x)$. The mean of X is given by

$$E(X) = \sum_{i=1}^k p_i \int_{-\infty}^{\infty} x f_i(x)\, dx = \sum_{i=1}^k p_i \mu_i = \overline{\mu}, \tag{3.7.3}$$

a weighted average of $\mu_1, \mu_2, \ldots, \mu_k$, and the variance equals

$$\begin{aligned}
\text{var}(X) &= \sum_{i=1}^{k} p_i \int_{-\infty}^{\infty} (x-\overline{\mu})^2 f_i(x)\,dx \\
&= \sum_{i=1}^{k} p_i \int_{-\infty}^{\infty} [(x-\mu_i)+(\mu_i-\overline{\mu})]^2 f_i(x)\,dx \\
&= \sum_{i=1}^{k} p_i \int_{-\infty}^{\infty} (x-\mu_i)^2 f_i(x)\,dx + \sum_{i=1}^{k} p_i(\mu_i-\overline{\mu})^2 \int_{-\infty}^{\infty} f_i(x)\,dx,
\end{aligned}$$

because the cross-product terms integrate to zero. That is,

$$\text{var}(X) = \sum_{i=1}^{k} p_i\sigma_i^2 + \sum_{i=1}^{k} p_i(\mu_i-\overline{\mu})^2. \tag{3.7.4}$$

Note that the variance is not simply the weighted average of the k variances, but it also includes a positive term involving the weighted variance of the means.

Remark 3.7.1. It is extremely important to note these characteristics are associated with a mixture of k distributions and have nothing to do with a linear combination, say $\sum a_i X_i$, of k random variables. ■

For the next example, we need the following distribution. We say that X has a **loggamma** pdf with parameters $\alpha > 0$ and $\beta > 0$ if it has pdf

$$f_1(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{-(1+\beta)/\beta} (\log x)^{\alpha-1} & x > 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{3.7.5}$$

The derivation of this pdf is given in Exercise 3.7.1, where its mean and variance are also derived. We denote this distribution of X by $\log\Gamma(\alpha,\beta)$.

Example 3.7.1. Actuaries have found that a mixture of the loggamma and gamma distributions is an important model for claim distributions. Suppose, then, that X_1 is $\log\Gamma(\alpha_1,\beta_1)$, X_2 is $\Gamma(\alpha_2,\beta_2)$, and the mixing probabilities are p and $(1-p)$. Then the pdf of the mixture distribution is

$$f(x) = \begin{cases} \frac{1-p}{\beta_2^{\alpha_2}\Gamma(\alpha_2)} x^{\alpha_2-1} e^{-x/\beta_2} & 0 < x \le 1 \\ \frac{p}{\beta_1^{\alpha_1}\Gamma(\alpha_1)} (\log x)^{\alpha_1-1} x^{-(\beta_1+1)/\beta_1} + \frac{1-p}{\beta_2^{\alpha_2}\Gamma(\alpha_2)} x^{\alpha_2-1} e^{-x/\beta_2} & 1 < x \\ 0 & \text{elsewhere.} \end{cases} \tag{3.7.6}$$

Provided $\beta_1 < 2^{-1}$, the mean and the variance of this mixture distribution are

$$\mu = p(1-\beta_1)^{-\alpha_1} + (1-p)\alpha_2\beta_2 \tag{3.7.7}$$

$$\begin{aligned}
\sigma^2 &= p[(1-2\beta_1)^{-\alpha_1} - (1-\beta_1)^{-2\alpha_1}] \\
&\quad + (1-p)\alpha_2\beta_2^2 + p(1-p)[(1-\beta_1)^{-\alpha_1} - \alpha_2\beta_2]^2;
\end{aligned} \tag{3.7.8}$$

see Exercise 3.7.3. ■

The mixture of distributions is sometimes called **compounding**. Moreover, it does not need to be restricted to a finite number of distributions. As demonstrated in the following example, a continuous weighting function, which is of course a pdf, can replace $p_1, p_2, \ldots, p_k$; i.e., integration replaces summation.

Example 3.7.2. Let X_θ be a Poisson random variable with parameter θ. We want to mix an infinite number of Poisson distributions, each with a different value of θ. We let the weighting function be a pdf of θ, namely, a gamma with parameters α and β. For $x = 0, 1, 2, \ldots$, the pmf of the compound distribution is

$$\begin{aligned} p(x) &= \int_0^\infty \left[\frac{1}{\beta^\alpha \Gamma(\alpha)}\theta^{\alpha-1}e^{-\theta/\beta}\right]\left[\frac{\theta^x e^{-\theta}}{x!}\right] d\theta \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha x!}\int_0^\infty \theta^{\alpha+x-1}e^{-\theta(1+\beta)/\beta}\, d\theta \\ &= \frac{\Gamma(\alpha+x)\beta^x}{\Gamma(\alpha)x!(1+\beta)^{\alpha+x}}, \end{aligned}$$

where the third line follows from the change of variable $t = \theta(1+\beta)/\beta$ to solve the integral of the second line.

An interesting case of this compound occurs when $\alpha = r$, a positive integer, and $\beta = (1-p)/p$, where $0 < p < 1$. In this case the pmf becomes

$$p(x) = \frac{(r+x-1)!}{(r-1)!}\frac{p^r(1-p)^x}{x!}, \quad x = 0, 1, 2, \ldots.$$

That is, this compound distribution is the same as that of the number of excess trials needed to obtain r successes in a sequence of independent trials, each with probability p of success; this is one form of the **negative binomial distribution**. The negative binomial distribution has been used successfully as a model for the number of accidents (see Weber, 1971). ■

In compounding, we can think of the original distribution of X as being a conditional distribution given θ, whose pdf is denoted by $f(x|\theta)$. Then the weighting function is treated as a pdf for θ, say $g(\theta)$. Accordingly, the joint pdf is $f(x|\theta)g(\theta)$, and the compound pdf can be thought of as the marginal (unconditional) pdf of X,

$$h(x) = \int_\theta g(\theta)f(x|\theta)\, d\theta,$$

where a summation replaces integration in case θ has a discrete distribution. For illustration, suppose we know that the mean of the normal distribution is zero but the variance σ^2 equals $1/\theta > 0$, where θ has been selected from some random model. For convenience, say this latter is a gamma distribution with parameters α and β. Thus, given that θ, X is conditionally $N(0, 1/\theta)$ so that the joint distribution of X and θ is

$$f(x|\theta)g(\theta) = \left[\frac{\sqrt{\theta}}{\sqrt{2\pi}}\exp\left(\frac{-\theta x^2}{2}\right)\right]\left[\frac{1}{\beta^\alpha\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\theta/\beta)\right],$$

for $-\infty < x < \infty, \quad 0 < \theta < \infty$. Therefore, the marginal (unconditional) pdf $h(x)$ of X is found by integrating out θ; that is,

$$h(x) = \int_0^\infty \frac{\theta^{\alpha+1/2-1}}{\beta^\alpha \sqrt{2\pi}\Gamma(\alpha)} \exp\left[-\theta\left(\frac{x^2}{2} + \frac{1}{\beta}\right)\right] d\theta.$$

By comparing this integrand with a gamma pdf with parameters $\alpha + \frac{1}{2}$ and $[(1/\beta) + (x^2/2)]^{-1}$, we see that the integral equals

$$h(x) = \frac{\Gamma(\alpha + \frac{1}{2})}{\beta^\alpha \sqrt{2\pi}\Gamma(\alpha)} \left(\frac{2\beta}{2 + \beta x^2}\right)^{\alpha+1/2}, \quad -\infty < x < \infty.$$

It is interesting to note that if $\alpha = r/2$ and $\beta = 2/r$, where r is a positive integer, then X has an unconditional distribution, which is Student's t, with r degrees of freedom. That is, we have developed a generalization of Student's distribution through this type of mixing or compounding. We note that the resulting distribution (a generalization of Student's t) has much thicker tails than those of the conditional normal with which we started.

The next two examples offer two additional illustrations of this type of compounding.

Example 3.7.3. Suppose that we have a binomial distribution, but we are not certain about the probability p of success on a given trial. Suppose p has been selected first by some random process that has a beta pdf with parameters α and β. Thus X, the number of successes on n independent trials, has a conditional binomial distribution so that the joint pdf of X and p is

$$p(x|p)g(p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1},$$

for $x = 0, 1, \ldots, n, \quad 0 < p < 1$. Therefore, the unconditional pmf of X is given by the integral

$$\begin{aligned} h(x) &= \int_0^1 \frac{n!\Gamma(\alpha+\beta)}{x!(n-x)!\Gamma(\alpha)\Gamma(\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1}\, dp \\ &= \frac{n!\Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n-x+\beta)}{x!(n-x)!\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}, \quad x = 0, 1, 2, \ldots, n. \end{aligned}$$

Now suppose α and β are positive integers; since $\Gamma(k) = (k-1)!$, this unconditional (marginal or compound) pdf can be written

$$h(x) = \frac{n!(\alpha+\beta-1)!(x+\alpha-1)!(n-x+\beta-1)!}{x!(n-x)!(\alpha-1)!(\beta-1)!(n+\alpha+\beta-1)!}, \quad x = 0, 1, 2, \ldots, n.$$

Because the conditional mean $E(X|p) = np$, the unconditional mean is $n\alpha/(\alpha+\beta)$ since $E(p)$ equals the mean $\alpha/(\alpha+\beta)$ of the beta distribution. ■

Example 3.7.4. In this example, we develop by compounding a heavy-tailed skewed distribution. Assume X has a conditional gamma pdf with parameters k and θ^{-1}. The weighting function for θ is a gamma pdf with parameters α and β. Thus the unconditional (marginal or compounded) pdf of X is

$$\begin{aligned} h(x) &= \int_0^\infty \left[\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}\right]\left[\frac{\theta^k x^{k-1}e^{-\theta x}}{\Gamma(k)}\right] d\theta \\ &= \int_0^\infty \frac{x^{k-1}\theta^{\alpha+k-1}}{\beta^\alpha\Gamma(\alpha)\Gamma(k)} e^{-\theta(1+\beta x)/\beta}\, d\theta. \end{aligned}$$

Comparing this integrand to the gamma pdf with parameters $\alpha+k$ and $\beta/(1+\beta x)$, we see that

$$h(x) = \frac{\Gamma(\alpha+k)\beta^k x^{k-1}}{\Gamma(\alpha)\Gamma(k)(1+\beta x)^{\alpha+k}}, \quad 0 < x < \infty,$$

which is the pdf of the **generalized Pareto distribution** (and a generalization of the F distribution). Of course, when $k = 1$ (so that X has a conditional exponential distribution), the pdf is

$$h(x) = \alpha\beta(1+\beta x)^{-(\alpha+1)}, \quad 0 < x < \infty,$$

which is the **Pareto pdf**. Both of these compound pdfs have thicker tails than the original (conditional) gamma distribution.

While the cdf of the generalized Pareto distribution cannot be expressed in a simple closed form, that of the Pareto distribution is

$$H(x) = \int_0^x \alpha\beta(1+\beta t)^{-(\alpha+1)}\, dt = 1 - (1+\beta x)^{-\alpha}, \quad 0 \le x < \infty.$$

From this, we can create another useful long-tailed distribution by letting $X = Y^\tau$, $0 < \tau$. Thus Y has the cdf

$$G(y) = P(Y \le y) = P[X^{1/\tau} \le y] = P[X \le y^\tau].$$

Hence, this probability is equal to

$$G(y) = H(y^\tau) = 1 - (1+\beta y^\tau)^{-\alpha}, \quad 0 < y < \infty,$$

with corresponding pdf

$$G'(y) = g(y) = \frac{\alpha\beta\tau y^{\tau-1}}{(1+\beta y^\tau)^{\alpha+1}}, \quad 0 < y < \infty.$$

We call the associated distribution the **transformed Pareto distribution** or the **Burr distribution** (Burr, 1942), and it has proved to be a useful one in modeling thicker-tailed distributions. ■

EXERCISES

3.7.1. Suppose Y has a $\Gamma(\alpha, \beta)$ distribution. Let $X = e^Y$. Show that the pdf of X is given by expression (3.7.5). Determine the cdf of X in terms of the cdf of a Γ-distribution. Derive the mean and variance of X.

3.7.2. Write R functions for the pdf and cdf of the random variable in Exercise 3.7.1.

3.7.3. In Example 3.7.1, derive the pdf of the mixture distribution given in expression (3.7.6), then obtain its mean and variance as given in expressions (3.7.7) and (3.7.8).

3.7.4. Using the R function for the pdf in Exercise 3.7.2 and `dgamma`, write an R function for the mixture pdf (3.7.6). For $\alpha = \beta = 2$, obtain a page of plots of this density for $p = 0.05, 0.10, 0.15$ and 0.20.

3.7.5. Consider the mixture distribution $(9/10)N(0, 1) + (1/10)N(0, 9)$. Show that its kurtosis is 8.34.

3.7.6. Let X have the conditional geometric pmf $\theta(1-\theta)^{x-1}$, $x = 1, 2, \ldots$, where θ is a value of a random variable having a beta pdf with parameters α and β. Show that the marginal (unconditional) pmf of X is

$$\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+1)\Gamma(\beta+x-1)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+x)}, \quad x = 1, 2, \ldots.$$

If $\alpha = 1$, we obtain

$$\frac{\beta}{(\beta+x)(\beta+x-1)}, \quad x = 1, 2, \ldots,$$

which is one form of **Zipf's law**.

3.7.7. Repeat Exercise 3.7.6, letting X have a conditional negative binomial distribution instead of the geometric one.

3.7.8. Let X have a generalized Pareto distribution with parameters k, α, and β. Show, by change of variables, that $Y = \beta X/(1+\beta X)$ has a beta distribution.

3.7.9. Show that the failure rate (hazard function) of the Pareto distribution is

$$\frac{h(x)}{1-H(x)} = \frac{\alpha}{\beta^{-1}+x}.$$

Find the failure rate (hazard function) of the Burr distribution with cdf

$$G(y) = 1 - \left(\frac{1}{1+\beta y^\tau}\right)^\alpha, \quad 0 \le y < \infty.$$

In each of these two failure rates, note what happens as the value of the variable increases.

3.7.10. For the Burr distribution, show that

$$E(X^k) = \frac{1}{\beta^{k/\tau}}\,\Gamma\left(\alpha - \frac{k}{\tau}\right)\Gamma\left(\frac{k}{\tau}+1\right)\Big/\,\Gamma(\alpha),$$

provided $k < \alpha\tau$.

3.7.11. Let the number X of accidents have a Poisson distribution with mean $\lambda\theta$. Suppose λ, the liability to have an accident, has, given θ, a gamma pdf with parameters $\alpha = h$ and $\beta = h^{-1}$; and θ, an accident proneness factor, has a generalized Pareto pdf with parameters α, $\lambda = h$, and k. Show that the unconditional pdf of X is

$$\frac{\Gamma(\alpha+k)\Gamma(\alpha+h)\Gamma(\alpha+h+k)\Gamma(h+k)\Gamma(k+x)}{\Gamma(\alpha)\Gamma(\alpha+k+h)\Gamma(h)\Gamma(k)\Gamma(\alpha+h+k+x)x!}, \quad x = 0, 1, 2, \ldots,$$

sometimes called the **generalized Waring** pmf.

3.7.12. Let X have a conditional Burr distribution with fixed parameters β and τ, given parameter α.

(a) If α has the geometric pmf $p(1-p)^\alpha$, $\alpha = 0, 1, 2, \ldots$, show that the unconditional distribution of X is a Burr distribution.

(b) If α has the exponential pdf $\beta^{-1}e^{-\alpha/\beta}$, $\alpha > 0$, find the unconditional pdf of X.

3.7.13. Let X have the conditional Weibull pdf

$$f(x|\theta) = \theta\tau x^{\tau-1}e^{-\theta x^\tau}, \quad 0 < x < \infty,$$

and let the pdf (weighting function) $g(\theta)$ be gamma with parameters α and β. Show that the compound (marginal) pdf of X is that of Burr.

3.7.14. If X has a Pareto distribution with parameters α and β and if c is a positive constant, show that $Y = cX$ has a Pareto distribution with parameters α and β/c.

Chapter 4

Some Elementary Statistical Inferences

4.1 Sampling and Statistics

In Chapter 2, we introduced the concepts of samples and statistics. We continue with this development in this chapter while introducing the main tools of inference: confidence intervals and tests of hypotheses.

In a typical statistical problem, we have a random variable X of interest, but its pdf $f(x)$ or pmf $p(x)$ is not known. Our ignorance about $f(x)$ or $p(x)$ can roughly be classified in one of two ways:

1. $f(x)$ or $p(x)$ is completely unknown.
2. The form of $f(x)$ or $p(x)$ is known down to a parameter θ, where θ may be a vector.

For now, we consider the second classification, although some of our discussion pertains to the first classification also. Some examples are the following:

(a) X has an exponential distribution, Exp(θ), (3.3.6), where θ is unknown.

(b) X has a binomial distribution $b(n, p)$, (3.1.2), where n is known but p is unknown.

(c) X has a gamma distribution $\Gamma(\alpha, \beta)$, (3.3.2), where both α and β are unknown.

(d) X has a normal distribution $N(\mu, \sigma^2)$, (3.4.6), where both the mean μ and the variance σ^2 of X are unknown.

We often denote this problem by saying that the random variable X has a density or mass function of the form $f(x; \theta)$ or $p(x; \theta)$, where $\theta \in \Omega$ for a specified set Ω. For example, in (a) above, $\Omega = \{\theta \,|\, \theta > 0\}$. We call θ a parameter of the distribution. Because θ is unknown, we want to estimate it.

In this process, our information about the unknown distribution of X or the unknown parameters of the distribution of X comes from a sample on X. The sample observations have the same distribution as X, and we denote them as the random variables $X_1, X_2, \ldots, X_n$, where n denotes the **sample size**. When the sample is actually drawn, we use lower case letters $x_1, x_2, \ldots, x_n$ as the values or **realizations** of the sample. Often we assume that the sample observations $X_1, X_2, \ldots, X_n$ are also mutually independent, in which case we call the sample a random sample, which we now formally define:

Definition 4.1.1. *If the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid), then these random variables constitute a* **random sample** *of size n from the common distribution.*

Often, functions of the sample are used to summarize the information in a sample. These are called statistics, which we define as:

Definition 4.1.2. *Let $X_1, X_2, \ldots, X_n$ denote a sample on a random variable X. Let $T = T(X_1, X_2, \ldots, X_n)$ be a function of the sample. Then T is called a* **statistic**.

Once the sample is drawn, then t is called the realization of T, where $t = T(x_1, x_2, \ldots, x_n)$ and $x_1, x_2, \ldots, x_n$ is the realization of the sample.

4.1.1 Point Estimators

Using the above terminology, the problem we discuss in this chapter is phrased as: Let $X_1, X_2, \ldots, X_n$ denote a random sample on a random variable X with a density or mass function of the form $f(x;\theta)$ or $p(x;\theta)$, where $\theta \in \Omega$ for a specified set Ω. In this situation, it makes sense to consider a statistic T, which is an **estimator** of θ. More formally, T is called a **point estimator** of θ. While we call T an estimator of θ, we call its realization t an **estimate** of θ.

There are several properties of point estimators that we discuss in this book. We begin with a simple one, unbiasedness.

Definition 4.1.3 (Unbiasedness)**.** *Let $X_1, X_2, \ldots, X_n$ denote a sample on a random variable X with pdf $f(x;\theta)$, $\theta \in \Omega$. Let $T = T(X_1, X_2, \ldots, X_n)$ be a statistic. We say that T is an* **unbiased** *estimator of θ if $E(T) = \theta$.*

In Chapters 6 and 7, we discuss several theories of estimation in general. The purpose of this chapter, though, is an introduction to inference, so we briefly discuss the **maximum likelihood estimator (mle)** and then use it to obtain point estimators for some of the examples cited above. We expand on this theory in Chapter 6. Our discussion is for the continuous case. For the discrete case, simply replace the pdf with the pmf.

In our problem, the information in the sample and the parameter θ are involved in the joint distribution of the random sample; i.e., $\prod_{i=1}^{n} f(x_i;\theta)$. We want to view this as a function of θ, so we write it as

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta). \tag{4.1.1}$$

This is called the **likelihood function** of the random sample. As an estimate of θ, a measure of the center of $L(\theta)$ seems appropriate. An often-used estimate is the value of θ that provides a maximum of $L(\theta)$. If it is unique, this is called the **maximum likelihood estimator** (mle), and we denote it as $\widehat{\theta}$; i.e.,

$$\widehat{\theta} = \text{Argmax}\, L(\theta). \tag{4.1.2}$$

In practice, it is often much easier to work with the log of the likelihood, that is, the function $l(\theta) = \log L(\theta)$. Because the log is a strictly increasing function, the value that maximizes $l(\theta)$ is the same as the value that maximizes $L(\theta)$. Furthermore, for most of the models discussed in this book, the pdf (or pmf) is a differentiable function of θ, and frequently $\widehat{\theta}$ solves the equation

$$\frac{\partial l(\theta)}{\partial \theta} = 0. \tag{4.1.3}$$

If θ is a vector of parameters, this results in a system of equations to be solved simultaneously; see Example 4.1.3. These equations are often referred to as the mle **estimating equations**, (EE).

As we show in Chapter 6, under general conditions, mles have some good properties. One property that we need at the moment concerns the situation where, besides the parameter θ, we are also interested in the parameter $\eta = g(\theta)$ for a specified function g. Then, as Theorem 6.1.2 of Chapter 6 shows, the mle of η is $\widehat{\eta} = g(\widehat{\theta})$, where $\widehat{\theta}$ is the mle of θ. We now proceed with some examples, including data realizations.

Example 4.1.1 (Exponential Distribution)**.** Suppose the common pdf of the random sample $X_1, X_2, \ldots, X_n$ is the $\Gamma(1, \theta)$ density $f(x) = \theta^{-1} \exp\{-x/\theta\}$ with support $0 < x < \infty$; see expression (3.3.2). This gamma distribution is often called the exponential distribution. The log of the likelihood function is given by

$$l(\theta) = \log \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = -n \log \theta - \theta^{-1} \sum_{i=1}^{n} x_i.$$

The first partial of the log-likelihood with respect to θ is

$$\frac{\partial l(\theta)}{\partial \theta} = -n\theta^{-1} + \theta^{-2} \sum_{i=1}^{n} x_i.$$

Setting this partial to 0 and solving for θ, we obtain the solution $\overline{x}$. There is only one critical value and, furthermore, the second partial of the log-likelihood evaluated at $\overline{x}$ is strictly negative, verifying that it provides a maximum. Hence, for this example, the statistic $\widehat{\theta} = \overline{X}$ is the mle of θ. Because $E(X) = \theta$, we have that $E(\overline{X}) = \theta$ and, hence, $\widehat{\theta}$ is an unbiased estimator of θ.

Rasmussen (1992), page 92, presents a data set where the variable of interest X is the number of operating hours until the first failure of air-conditioning units for Boeing 720 airplanes. A random sample of size $n = 13$ was obtained and its

realized values are:

```
359 413  25 130  90  50  50 487 102 194  55  74  97
```

For instance, 359 hours is the realization of the random variable X_1. The data range from 25 to 487 hours. Assuming an exponential model, the point estimate of θ discussed above is the arithmetic average of this data. Assuming that the data set is stored in the R vector `ophrs`, this average is computed in R by

```
mean(ophrs);  163.5385
```

Hence our point estimate of θ, the mean of X, is 163.54 hours. How close is 163.54 hours to the true θ? We provide an answer to this question in the next section. ■

Example 4.1.2 (Binomial Distribution)**.** Let X be one or zero if, respectively, the outcome of a Bernoulli experiment is success or failure. Let θ, $0 < \theta < 1$, denote the probability of success. Then by (3.1.1), the pmf of X is

$$p(x;\theta) = \theta^x(1-\theta)^{1-x}, \quad x = 0 \text{ or } 1.$$

If $X_1, X_2, \ldots, X_n$ is a random sample on X, then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} p(x_i;\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}, \quad x_i = 0 \text{ or } 1.$$

Taking logs, we have

$$l(\theta) = \sum_{i=1}^{n} x_i \log\theta + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta), \quad x_i = 0 \text{ or } 1.$$

The partial derivative of $l(\theta)$ is

$$\frac{\partial l(\theta)}{\partial\theta} = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta}.$$

Setting this to 0 and solving for θ, we obtain $\widehat{\theta} = n^{-1}\sum_{i=1}^n X_i = \overline{X}$; i.e., the mle is the proportion of successes in the n trials. Because $E(X) = \theta$, $\widehat{\theta}$ is an unbiased estimator of θ.

Devore (2012) discusses a study involving ceramic hip replacements which for some patients can be squeaky; see, also, page 30 of Kloke and McKean (2014). In this study, 28 out of 143 hip replacements squeaked. In terms of the above discussion, we have a realization of a sample of size $n = 143$ from a binomial distribution where success is a hip replacement that squeaks and failure is one that does not squeak. Let θ denote the probability of success. Then our estimate of θ based on this sample is $\widehat{\theta} = 28/143 = 0.1958$. This is straightforward to calculate but, for later use, the R code `prop.test(28,143)` calculates this proportion. ■

Example 4.1.3 (Normal Distribution)**.** Let X have a $N(\mu,\sigma^2)$ distribution with the pdf given in expression (3.4.6). In this case, $\boldsymbol{\theta}$ is the vector $\boldsymbol{\theta} = (\mu,\sigma)$. If $X_1, X_2, \ldots, X_n$ is a random sample on X, then the log of the likelihood function simplifies to

$$l(\mu,\sigma) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2. \tag{4.1.4}$$

The two partial derivatives simplify to

$$\frac{\partial l(\mu,\sigma)}{\partial \mu} = -\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)\left(-\frac{1}{\sigma}\right) \tag{4.1.5}$$

$$\frac{\partial l(\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{4.1.6}$$

Setting these to 0 and solving simultaneously, we see that the mles are

$$\widehat{\mu} = \overline{X} \tag{4.1.7}$$

$$\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i-\overline{X})^2. \tag{4.1.8}$$

Notice that we have used the property that the mle of $\widehat{\sigma}^2$ is the mle of σ squared. As we have shown in Chapter 2, (2.8.6), the estimator $\overline{X}$ is an unbiased estimator for μ. Further, from Example 2.8.7 of Section 2.8 we know that the following statistic

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{4.1.9}$$

is an unbiased estimator of σ^2. Thus for the mle of σ^2, $E(\widehat{\sigma}^2) = [n/(n-1)]\sigma^2$. Hence, the mle is a biased estimator of σ^2. Note, though, that the bias of $\widehat{\sigma}^2$ is $E(\widehat{\sigma}^2-\sigma^2) = -\sigma^2/n$, which converges to 0 as $n\to\infty$. In practice, however, S^2 is the preferred estimator of σ^2.

Rasmussen (1991), page 65, discusses a study to measure the concentration of sulfur dioxide in a damaged Bavarian forest. The following data set is the realization of a random sample of size $n = 24$ measurements (micro grams per cubic meter) of this sulfur dioxide concentration:

```
33.4 38.6 41.7 43.9 44.4 45.3 46.1 47.6 50.0 52.4 52.7 53.9
54.3 55.1 56.4 56.5 60.7 61.8 62.2 63.4 65.5 66.6 70.0 71.5.
```

These data are also in the R data file `sulfurdio.rda` at the site listed in the Preface. Assuming these data are in the R vector `sulfurdioxide`, the following R segment obtains the estimates of the true mean and variance (both s^2 and $\widehat{\sigma}^2$ are computed):

```
mean(sulfurdioxide);var(sulfurdioxide);(23/24)*var(sulfurdioxide)
53.91667      101.4797      97.25139.
```

Hence, we estimate the true mean concentration of sulfur dioxide in this damaged Bavarian forest to be 53.92 micro grams per cubic meter. The realization of the statistic S^2 is $s^2 = 101.48$, while the biased estimate of σ^2 is 97.25. Rasmussen notes that the average concentration of sulfur dioxide in undamaged areas of Bavaria is 20 micro grams per cubic meter. This value appears to be quite distant from the sample values. This will be discussed statistically in later sections. ■

In all three of these examples, standard differential calculus methods led us to the solution. For the next example, the support of the random variable involves θ and, hence, it is not surprising that for this case differential calculus is not useful.

Example 4.1.4 (Uniform Distribution)**.** Let $X_1, \ldots, X_n$ be iid with the uniform $(0, \theta)$ density; i.e., $f(x) = 1/\theta$ for $0 < x < \theta$, 0 elsewhere. Because θ is in the support, differentiation is not helpful here. The likelihood function can be written as

$$L(\theta) = \theta^{-n} I(\max\{x_i\}, \theta),$$

where $I(a, b)$ is 1 or 0 if $a \leq b$ or $a > b$, respectively. The function $L(\theta)$ is a decreasing function of θ for all $\theta \geq \max\{x_i\}$ and is 0 otherwise [sketch the graph of $L(\theta)$]. So the maximum occurs at the smallest value that θ can assume; i.e., the mle is $\widehat{\theta} = \max\{X_i\}$. ■

4.1.2 Histogram Estimates of pmfs and pdfs

Let $X_1, \ldots, X_n$ be a random sample on a random variable X with cdf $F(x)$. In this section, we briefly discuss a histogram of the sample, which is an estimate of the pmf, $p(x)$, or the pdf, $f(x)$, of X depending on whether X is discrete or continuous. Other than X being a discrete or continuous random variable, we make no assumptions on the form of the distribution of X. In particular, we do not assume a parametric form of the distribution as we did for the above discussion on maximum likelihood estimates; hence, the histogram that we present is often called a **nonparametric** estimator. See Chapter 10 for a general discussion of nonparametric inference. We discuss the discrete situation first.

The Distribution of X Is Discrete

Assume that X is a discrete random variable with pmf $p(x)$. Let $X_1, \ldots, X_n$ be a random sample on X. First, suppose that the space of X is finite, say, $\mathcal{D} = \{a_1, \ldots, a_m\}$. An intuitive estimate of $p(a_j)$ is the relative frequency of a_j in the sample. We express this more formally as follows. For $j = 1, 2, \ldots, m$, define the statistics

$$I_j(X_i) = \begin{cases} 1 & X_i = a_j \\ 0 & X_i \neq a_j. \end{cases}$$

Then our intuitive estimate of $p(a_j)$ can be expressed by the sample average

$$\widehat{p}(a_j) = \frac{1}{n} \sum_{i=1}^{n} I_j(X_i). \tag{4.1.10}$$

These estimators $\{\widehat{p}(a_1), \ldots, \widehat{p}(a_m)\}$ constitute the nonparametric estimate of the pmf $p(x)$. Note that $I_j(X_i)$ has a Bernoulli distribution with probability of success $p(a_j)$. Because

$$E[\widehat{p}(a_j)] = \frac{1}{n} \sum_{i=1}^{n} E[I_j(X_i)] = \frac{1}{n} \sum_{i=1}^{n} p(a_j) = p(a_j), \tag{4.1.11}$$

$\widehat{p}(a_j)$ is an unbiased estimator of $p(a_j)$.

Next, suppose that the space of X is infinite, say, $\mathcal{D} = \{a_1, a_2, \ldots\}$. In practice, we select a value, say, a_m, and make the groupings

$$\{a_1\}, \{a_2\}, \ldots, \{a_m\}, \tilde{a}_{m+1} = \{a_{m+1}, a_{m+2}, \ldots\}. \tag{4.1.12}$$

Let $\widehat{p}(\tilde{a}_{m+1})$ be the proportion of sample items that are greater than or equal to a_{m+1}. Then the estimates $\{\widehat{p}(a_1), \ldots, \widehat{p}(a_m), \widehat{p}(\tilde{a}_{m+1})\}$ form our estimate of $p(x)$. For the merging of groups, a rule of thumb is to select m so that the frequency of the category a_m exceeds twice the combined frequencies of the categories $a_{m+1}, a_{m+2}, \ldots$.

A histogram is a **barplot** of $\widehat{p}(a_j)$ versus a_j. There are two cases to consider. For the first case, suppose the values a_j represent qualitative categories, for example, hair colors of a population of people. In this case, there is no ordinal information in the a_js. The usual histogram for such data consists of nonabutting bars with heights $\widehat{p}(a_j)$ that are plotted in decreasing order of the $\widehat{p}(a_1)$s. Such histograms are usually called **bar charts**. An example is helpful here.

Example 4.1.5 (Hair Color of Scottish School Children)**.** Kendall and Sturat (1979) present data on the eye and hair color of Scottish schoolchildren in the early 1900s. The data are also in the file `scotteyehair.rda` at the site listed in the Preface. In this example, we consider hair color. The discrete random variable is the hair color of a Scottish child with categories fair, red, medium, dark, and black. The results that Kendall and Sturat present are based on a sample of $n = 22{,}361$ Scottish school children. The frequency distribution of this sample and the estimate of the pmf are

	Fair	Red	Medium	Dark	Black
Count	5789	1319	9418	5678	157
$\widehat{p}(a_j)$	0.259	0.059	0.421	0.254	0.007

The bar chart of this sample is shown in Figure 4.1.1. Assume that the counts (second row of the table) are in the R vector `vec`. Then the following R segment computes this bar chart:

```
n=sum(vec); vecs = sort(vec,decreasing=T)/n
nms = c("Medium","Fair","Dark","Red","Black")
barplot(vecs,beside=TRUE,names.arg=nms,ylab="",xlab="Haircolor")
```

■

For the second case, assume that the values in the space $\mathcal{D}$ are **ordinal** in nature; i.e., the natural ordering of the a_js is numerically meaningful. In this case, the usual histogram is an abutting bar chart with heights $\widehat{p}(a_j)$ that are plotted in the natural order of the a_js, as in the following example.

Example 4.1.6 (Simulated Poisson Variates)**.** The following 30 data points are simulated values drawn from a Poisson distribution with mean $\lambda = 2$; see Example 4.8.2 for the generation of Poisson variates.

2	1	1	1	1	5	1	1	3	0	2	1	1	3	4
2	1	2	2	6	5	2	3	2	4	1	3	1	3	0

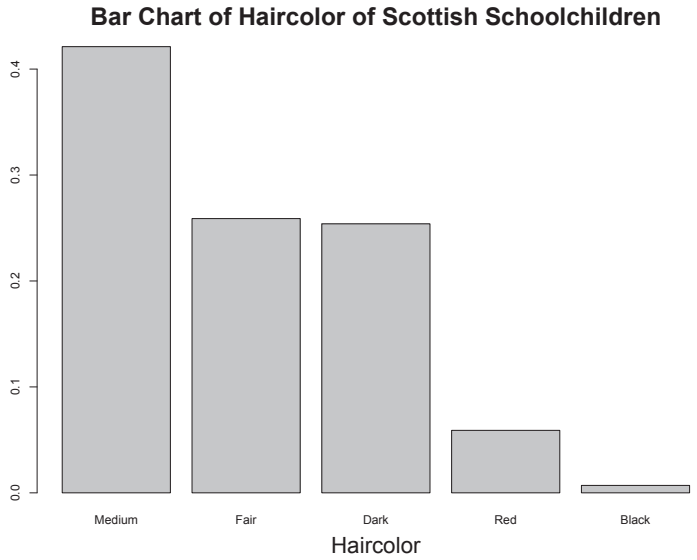

Figure 4.1.1: Bar chart of the Scottish hair color data discussed in Example 4.1.5.

The nonparametric estimate of the pmf is

j	0	1	2	3	4	5	≥ 6
$\widehat{p}(j)$	0.067	0.367	0.233	0.167	0.067	0.067	0.033

The histogram for this data set is given in Figure 4.1.2. Note that counts are used for the vertical axis. If the R vector `x` contains the 30 data points, then the following R code computes this histogram:

```
brs=seq(-.5,6.5,1);hist(x,breaks=brs,xlab="Number of events",ylab="")
```

■

The Distribution of X Is Continuous

For this section, assume that the random sample $X_1, \ldots, X_n$ is from a continuous random variable X with continuous pdf $f(t)$. We first sketch an estimate for this pdf at a specified value of x. Then we use this estimate to develop a histogram estimate of the pdf. For an arbitrary but fixed point x and a given $h > 0$, consider the interval $(x - h, x + h)$. By the mean value theorem for integrals, we have for some ξ, $|x - \xi| < h$, that

$$P(x - h < X < x + h) = \int_{x-h}^{x+h} f(t)\, dt = f(\xi)2h \approx f(x)2h.$$

The nonparametric estimate of the leftside is the proportion of the sample items that fall in the interval $(x - h, x + h)$. This suggests the following nonparametric

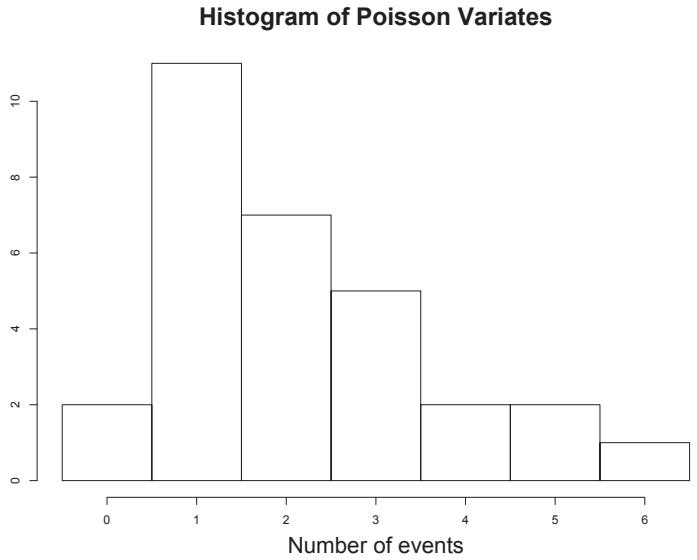

Figure 4.1.2: Histogram of the Poisson variates of Example 4.1.6.

estimate of $f(x)$ at a given x:

$$\widehat{f}(x) = \frac{\#\{x-h < X_i < x+h\}}{2hn}. \tag{4.1.13}$$

To write this more formally, consider the indicator statistic

$$I_i(x) = \begin{cases} 1 & x-h < X_i < x+h \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n.$$

Then a nonparametric estimator of $f(x)$ is

$$\widehat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} I_i(x). \tag{4.1.14}$$

Since the sample items are identically distributed,

$$E[\widehat{f}(x)] = \frac{1}{2hn} nf(\xi)2h = f(\xi) \to f(x),$$

as $h \to 0$. Hence $\widehat{f}(x)$ is approximately an unbiased estimator of the density $f(x)$. In density estimation terminology, the indicator function I_i is called a **rectangular kernel** with **bandwidth** $2h$. See Sheather and Jones (1991) and Chapter 6 of Lehmann (1999) for discussions of density estimation. The R function `density` provides a density estimator with several options. For the examples in the text, we use the default option as in Example 4.1.7.

The histogram provides a somewhat crude but often used estimator of the pdf, so a few remarks on it are pertinent. Let $x_1, \ldots, x_n$ be the realized values of the random sample on a continuous random variable X with pdf $f(x)$. Our histogram estimate of $f(x)$ is obtained as follows. While for the discrete case, there are natural classes for the histogram, for the continuous case these classes must be chosen. One way of doing this is to select a positive integer m, an $h > 0$, and a value a such that $a < \min x_i$, so that the m intervals

$$(a-h, a+h], (a+h, a+3h], (a+3h, a+5h], \ldots, (a+(2m-3)h, a+(2m-1)h] \quad (4.1.15)$$

cover the range of the sample $[\min x_i, \max x_i]$. These intervals form our classes. Let $A_j = (a + (2j-3)h, a + (2j-1)h]$ for $j = 1, \ldots m$.

Let $\widehat{f}_h(x)$ denote our histogram estimate. If $x \leq a - h$ or $x > a + (2m-1)h$ then define $\widehat{f}_h(x) = 0$. For $a - h < x \leq a + (2m-1)h$, x is in one, and only one, A_j. For $x \in A_j$, define $\widehat{f}_h(x)$ to be:

$$\widehat{f}_h(x) = \frac{\#\{x_i \in A_j\}}{2hn}. \quad (4.1.16)$$

Note that $\widehat{f}_h(x) \geq 0$ and that

$$\begin{aligned} \int_{-\infty}^{\infty} \widehat{f}_h(x)\, dx &= \int_{a-h}^{a+(2m-1)h} \widehat{f}_h(x)\, dx = \sum_{j=1}^{m} \int_{A_j} \frac{\#\{x_i \in A_j\}}{2hn}\, dx \\ &= \frac{1}{2hn} \sum_{j=1}^{m} \#\{x_i \in A_j\}[h(2j-1-2j+3)] = \frac{2h}{2hn} n = 1; \end{aligned}$$

so, $\widehat{f}_h(x)$ satisfies the properties of a pdf.

For the discrete case, except when classes are merged, the histogram is unique. For the continuous case, though, the histogram depends on the classes chosen. The resulting picture can be quite different if the classes are changed. Unless there is a compelling reason for the class selection, we recommend using the default classes selected by the computational algorithm. The histogram algorithms in most statistical packages such as R are current on recent research for selection of classes. The histogram in the following example is based on default classes.

Example 4.1.7. In Example 4.1.3, we presented a data set involving sulfur dioxide concentrations in a damaged Bavarian forest. The histogram of this data set is found in Figure 4.1.3. There are only 24 data points in the sample which are far too few for density estimation. With this in mind, although the distribution of data is mound shaped, the center appears to be too flat for normality. We have overlaid the histogram with the default R density estimate (solid line) which confirms some caution on normality. Recall that sample mean and standard deviations for this data are 53.91667 and 10.07371, respectively. So we also plotted the normal pdf with this mean and standard deviation (dashed line). The R code assumes that the data are in the R vector `sulfurdioxide`.

```
hist(sulfurdioxide,xlab="Sulfurdioxide",ylab=" ",pr=T,ylim=c(0,.04))
```

```
lines(density(sulfurdioxide))
y=dnorm(sulfurdioxide,53.91667,10.07371);lines(y~sulfurdioxide,lty=2)
```

The normal density plot seems to be a poor fit. ■

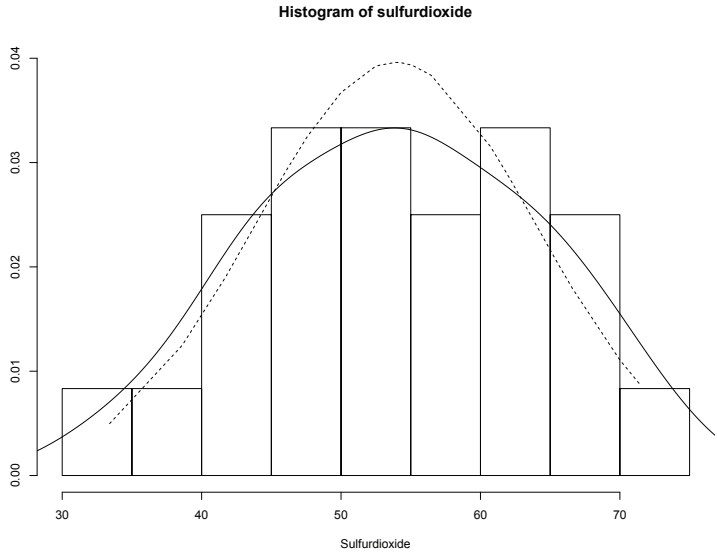

Figure 4.1.3: Histogram of the sulfur dioxide concentrations in a damaged Bavarian forest overlaid with a density estimate (solid line) and a normal pdf (dashed line) with mean and variance replaced by the sample mean and standard deviations, respectively. Data are given in Example 4.1.3.

EXERCISES

4.1.1. Twenty motors were put on test under a high-temperature setting. The lifetimes in hours of the motors under these conditions are given below. Also, the data are in the file `lifetimemotor.rda` at the site listed in the Preface. Suppose we assume that the lifetime of a motor under these conditions, X, has a $\Gamma(1, \theta)$ distribution.

1	4	5	21	22	28	40	42	51	53
58	67	95	124	124	160	202	260	303	363

(a) Obtain a histogram of the data and overlay it with a density estimate, using the code `hist(x,pr=T); lines(density(x))` where the R vector `x` contains the data. Based on this plot, do you think that the $\Gamma(1, \theta)$ model is credible?

(b) Assuming a $\Gamma(1, \theta)$ model, obtain the maximum likelihood estimate $\widehat{\theta}$ of θ and locate it on your histogram. Next overlay the pdf of a $\Gamma(1, \widehat{\theta})$ distribution on

the histogram. Use the R function `dgamma(x,shape=1,scale=`$\hat{\theta}$`)` to evaluate the pdf.

(c) Obtain the sample median of the data, which is an estimate of the median lifetime of a motor. What parameter is it estimating (i.e., determine the median of X)?

(d) Based on the mle, what is another estimate of the median of X?

4.1.2. Here are the weights of 26 professional baseball pitchers; [see page 76 of Hettmansperger and McKean (2011) for the complete data set]. The data are in R file `bb.rda`. Suppose we assume that the weight of a professional baseball pitcher is normally distributed with mean μ and variance σ^2.

160	175	180	185	185	185	190	190	195	195	195	200	200
200	200	205	205	210	210	218	219	220	222	225	225	232

(a) Obtain a histogram of the data. Based on this plot, is a normal probability model credible?

(b) Obtain the maximum likelihood estimates of μ, σ^2, σ, and μ/σ. Locate your estimate of μ on your plot in part (a). Then overlay the normal pdf with these estimates on your histogram in Part (a).

(c) Using the binomial model, obtain the maximum likelihood estimate of the proportion p of professional baseball pitchers who weigh over 215 pounds.

(d) Determine the mle of p assuming that the weight of a professional baseball player follows the normal probability model $N(\mu, \sigma^2)$ with μ and σ unknown.

4.1.3. Suppose the number of customers X that enter a store between the hours 9:00 a.m. and 10:00 a.m. follows a Poisson distribution with parameter θ. Suppose a random sample of the number of customers that enter the store between 9:00 a.m. and 10:00 a.m. for 10 days results in the values

9	7	9	15	10	13	11	7	2	12

(a) Determine the maximum likelihood estimate of θ. Show that it is an unbiased estimator.

(b) Based on these data, obtain the realization of your estimator in part (a). Explain the meaning of this estimate in terms of the number of customers.

4.1.4. For Example 4.1.3, verify equations (4.1.4)–(4.1.8).

4.1.5. Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous-type distribution.

(a) Find $P(X_1 \leq X_2), P(X_1 \leq X_2, X_1 \leq X_3), \ldots, P(X_1 \leq X_i, i = 2, 3, \ldots, n)$.

(b) Suppose the sampling continues until X_1 is no longer the smallest observation (i.e., $X_j < X_1 \leq X_i, i = 2, 3, \ldots, j-1$). Let Y equal the number of trials, not including X_1, until X_1 is no longer the smallest observation (i.e., $Y = j - 1$). Show that the distribution of Y is

$$P(Y = y) = \frac{1}{y(y+1)}, \quad y = 1, 2, 3, \ldots.$$

(c) Compute the mean and variance of Y if they exist.

4.1.6. Consider the estimator of the pmf in expression (4.1.10). In equation (4.1.11), we showed that this estimator is unbiased. Find the variance of the estimator and its mgf.

4.1.7. The data set on Scottish schoolchildren discussed in Example 4.1.5 included the eye colors of the children also. The frequencies of their eye colors are

Blue	Light	Medium	Dark
2978	6697	7511	5175

Use these frequencies to obtain a bar chart and an estimate of the associated pmf.

4.1.8. Recall that for the parameter $\eta = g(\theta)$, the mle of η is $g(\widehat{\theta})$, where $\widehat{\theta}$ is the mle of θ. Assuming that the data in Example 4.1.6 were drawn from a Poisson distribution with mean λ, obtain the mle of λ and then use it to obtain the mle of the pmf. Compare the mle of the pmf to the nonparametric estimate. Note: For the domain value 6, obtain the mle of $P(X \geq 6)$.

4.1.9. Consider the nonparametric estimator, (4.1.14), of a pdf.

(a) Obtain its mean and determine the bias of the estimator.

(b) Obtain the variance of the estimator.

4.1.10. This data set was downloaded from the site http://lib.stat.cmu.edu/DASL/ at Carnegie-Melon university. The original source is Willerman et al. (1991). The data consist of a sample of brain information recorded on 40 college students. The variables include gender, height, weight, three IQ measurements, and Magnetic Resonance Imaging (MRI) counts, as a determination of brain size. The data are in the rda file `braindata.rda` at the sites referenced in the Preface. For this exercise, consider the MRI counts.

(a) Load the rda file `braindata.rda` and print the MRI data, using the code: `mri <- braindata[,7]; print(mri)`.

(b) Obtain a histogram of the data, `hist(mri,pr=T)`. Comment on the shape.

(c) Overlay the default density estimator, `lines(density(mri))`. Comment on the shape.

(d) Obtain the sample mean and standard deviation and on the histogram overlay the normal pdf with these estimates as parameters, using `mris=sort(mri)` and `lines(dnorm(mris,mean(mris),sd(mris))~mris,lty=2)`. Comment on the fit.

(e) Determine the proportions of the data within 1 and 2 standard deviations of the sample mean and compare these with the empirical rule.

4.1.11. This is a famous data set on the speed of light recorded by the scientist Simon Newcomb. The data set was obtained at the Carnegie Melon site given in Exercise 4.1.10 and it can also be found in the rda file `speedlight.rda` at the sites referenced in the Preface. Stigler (1977) presents an informative discussion of this data set.

(a) Load the rda file and type the command `print(speed)`. As Stigler notes, the data values $\times 10^{-3} + 24.8$ are Newcomb's data values; hence, negative items can occur. Also, in the unit of the data the "true value" is 33.02. Discuss the data.

(b) Obtain a histogram of the data. Comment on the shape.

(c) On the histogram overlay the default density estimator. Comment on the shape.

(d) Obtain the sample mean and standard deviation and on the histogram overlay the normal pdf with these estimates as parameters. Comment on the fit.

(e) Determine the proportions of the data within 1 and 2 standard deviations of the sample mean and compare these with the empirical rule.

4.2 Confidence Intervals

Let us continue with the statistical problem that we were discussing in Section 4.1. Recall that the random variable of interest X has density $f(x;\theta), \theta \in \Omega$, where θ is unknown. In that section, we discussed estimating θ by a statistic $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ is a sample from the distribution of X. When the sample is drawn, it is unlikely that the value of $\widehat{\theta}$ is the true value of the parameter. In fact, if $\widehat{\theta}$ has a continuous distribution, then $P_\theta(\widehat{\theta} = \theta) = 0$, where the notation P_θ denotes that the probability is computed when θ is the true parameter. What is needed is an estimate of the error of the estimation; i.e., by how much did $\widehat{\theta}$ miss θ? In this section, we embody this estimate of error in terms of a confidence interval, which we now formally define:

Definition 4.2.1 (Confidence Interval)**.** *Let $X_1, X_2, \ldots, X_n$ be a sample on a random variable X, where X has pdf $f(x;\theta)$, $\theta \in \Omega$. Let $0 < \alpha < 1$ be specified. Let $L = L(X_1, X_2, \ldots, X_n)$ and $U = U(X_1, X_2, \ldots, X_n)$ be two statistics. We say that the interval (L, U) is a $(1-\alpha)100\%$* **confidence interval** *for θ if*

$$1 - \alpha = P_\theta[\theta \in (L, U)]. \tag{4.2.1}$$

That is, the probability that the interval includes θ is $1-\alpha$, which is called the **confidence coefficient** *or the* **confidence level** *of the interval.*

Once the sample is drawn, the realized value of the confidence interval is (l, u), an interval of real numbers. Either the interval (l, u) traps θ or it does not. One way of thinking of a confidence interval is in terms of a Bernoulli trial with probability of success $1-\alpha$. If one makes, say, M independent $(1-\alpha)100\%$ confidence intervals over a period of time, then one would expect to have $(1-\alpha)M$ successful confidence intervals (those that trap θ) over this period of time. Hence one feels $(1-\alpha)100\%$ confident that the true value of θ lies in the interval (l, u).

A measure of efficiency based on a confidence interval is its expected length. Suppose (L_1, U_1) and (L_2, U_2) are two confidence intervals for θ that have the same confidence coefficient. Then we say that (L_1, U_1) is more efficient than (L_2, U_2) if $E_\theta(U_1 - L_1) \leq E_\theta(U_2 - L_2)$ for all $\theta \in \Omega$.

There are several procedures for obtaining confidence intervals. We explore one of them in this section. It is based on a pivot random variable. The pivot is usually a function of an estimator of θ and the parameter and, further, the distribution of the pivot is known. Using this information, an algebraic derivation can often be used to obtain a confidence interval. The next several examples illustrate the pivot method. A second way to obtain a confidence interval involves distribution free techniques, as used in Section 4.4.2 to determine confidence intervals for quantiles.

Example 4.2.1 (Confidence Interval for μ Under Normality)**.** Suppose the random variables $X_1, \ldots, X_n$ are a random sample from a $N(\mu, \sigma^2)$ distribution. Let $\overline{X}$ and S^2 denote the sample mean and sample variance, respectively. Recall from the last section that $\overline{X}$ is the mle of μ and $[(n-1)/n]S^2$ is the mle of σ^2. By part (d) of Theorem 3.6.1, the random variable $T = (\overline{X} - \mu)/(S/\sqrt{n})$ has a t-distribution with $n-1$ degrees of freedom. The random variable T is our pivot variable.

For $0 < \alpha < 1$, define $t_{\alpha/2,n-1}$ to be the upper $\alpha/2$ critical point of a t-distribution with $n-1$ degrees of freedom; i.e., $\alpha/2 = P(T > t_{\alpha/2,n-1})$. Using a simple algebraic derivation, we obtain

$$\begin{aligned} 1-\alpha &= P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) \\ &= P_\mu\left(-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) \\ &= P_\mu\left(-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \overline{X}-\mu < t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) \\ &= P_\mu\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right). \end{aligned} \tag{4.2.2}$$

Once the sample is drawn, let $\overline{x}$ and s denote the realized values of the statistics $\overline{X}$ and S, respectively. Then a $(1-\alpha)100\%$ confidence interval for μ is given by

$$(\overline{x} - t_{\alpha/2,n-1}s/\sqrt{n}, \overline{x} + t_{\alpha/2,n-1}s/\sqrt{n}). \tag{4.2.3}$$

This interval is often referred to as the $(1-\alpha)100\%$ ***t*-interval** for μ. The estimate of the standard deviation of $\overline{X}$, $s/\sqrt{n}$, is referred to as the **standard error** of $\overline{X}$.

In Example 4.1.3, we presented a data set on sulfur dioxide concentrations in a damaged Bavarian forest. Let μ denote the true mean sulfur dioxide concentration. Recall, based on the data, that our estimate of μ is $\overline{x} = 53.92$ with sample standard deviation $s = \sqrt{101.48} = 10.07$. Since the sample size is $n = 24$, for a 99% confidence interval the t-critical value is $t_{0.005,23}$ =`qt(.995,23) = 2.807`. Based on these values, the confidence interval in expression (4.2.3) can be calculated. Assuming that the R vector `sulfurdioxide` contains the sample, the R code to compute this interval is `t.test(sulfurdioxide,conf.level=0.99)`, which results in the 99% confidence interval $(48.14, 59.69)$. Many scientists write this interval as 53.92 ± 5.78. In this way, we can see our estimate of μ and the margin of error. ■

The distribution of the pivot random variable $T = (\overline{X} - \mu)/(s/\sqrt{n})$ of the last example depends on the normality of the sampled items; however, this is approximately true even if the sampled items are not drawn from a normal distribution. The **Central Limit Theorem** (CLT) shows that the distribution of T is approximately $N(0, 1)$. In order to use this result now, we state the CLT now, leaving its proof to Chapter 5; see Theorem 5.3.1.

Theorem 4.2.1 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ denote the observations of a random sample from a distribution that has mean μ and finite variance σ^2. Then the distribution function of the random variable $W_n = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ converges to Φ, the distribution function of the $N(0, 1)$ distribution, as $n \to \infty$.*

As we further show in Chapter 5, the result stays the same if we replace σ by the sample standard deviation S; that is, under the assumptions of Theorem 4.2.1, the distribution of

$$Z_n = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{4.2.4}$$

is approximately $N(0, 1)$. For the nonnormal case, as the next example shows, with this result we can obtain an approximate confidence interval for μ.

Example 4.2.2 (Large Sample Confidence Interval for the Mean μ)**.** Suppose $X_1, X_2, \ldots, X_n$ is a random sample on a random variable X with mean μ and variance σ^2, but, unlike the last example, the distribution of X is not normal. However, from the above discussion we know that the distribution of Z_n, (4.2.4), is approximately $N(0, 1)$. Hence

$$1 - \alpha \approx P_\mu \left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < z_{\alpha/2} \right).$$

Using the same algebraic derivation as in the last example, we obtain

$$1 - \alpha \approx P_\mu \left(\overline{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right). \tag{4.2.5}$$

Again, letting $\overline{x}$ and s denote the realized values of the statistics $\overline{X}$ and S, respectively, after the sample is drawn, an approximate $(1 - \alpha)100\%$ confidence interval for μ is given by

$$(\overline{x} - z_{\alpha/2} s/\sqrt{n}, \overline{x} + z_{\alpha/2} s/\sqrt{n}). \tag{4.2.6}$$

This is called a **large sample** confidence interval for μ. ■

In practice, we often do not know if the population is normal. Which confidence interval should we use? Generally, for the same α, the intervals based on $t_{\alpha/2,n-1}$ are larger than those based on $z_{\alpha/2}$. Hence the interval (4.2.3) is generally more conservative than the interval (4.2.6). So in practice, statisticians generally prefer the interval (4.2.3).

Occasionally in practice, the standard deviation σ is assumed known. In this case, the confidence interval generally used for μ is (4.2.6) with s replaced by σ.

Example 4.2.3 (Large Sample Confidence Interval for p)**.** Let X be a Bernoulli random variable with probability of success p, where X is 1 or 0 if the outcome is success or failure, respectively. Suppose $X_1, \ldots, X_n$ is a random sample from the distribution of X. Let $\widehat{p} = \overline{X}$ be the sample proportion of successes. Note that $\widehat{p} = n^{-1}\sum_{i=1}^{n} X_i$ is a sample average and that $\text{Var}(\widehat{p}) = p(1-p)/n$. It follows immediately from the CLT that the distribution of $Z = (\widehat{p} - p)/\sqrt{p(1-p)/n}$ is approximately $N(0,1)$. Referring to Example 5.1.1 of Chapter 5, we replace $p(1-p)$ with its estimate $\widehat{p}(1-\widehat{p})$. Then proceeding as in the last example, an approximate $(1-\alpha)100\%$ confidence interval for p is given by

$$(\widehat{p} - z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}, \widehat{p} + z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}), \tag{4.2.7}$$

where $\sqrt{\widehat{p}(1-\widehat{p})/n}$ is called the standard error of $\widehat{p}$.

In Example 4.1.2 we discussed a data set involving hip replacements, some of which were squeaky. The outcomes of a hip replacement were squeaky and non-squeaky which we labeled as success or failure, respectively. In the sample there were 28 successes out of 143 replacements. Using R, the 99% confidence interval for p, the probability of a squeaky hip replacement, is computed by `prop.test(28,143,conf.level=.99)`, which results in the interval $(0.122, 0.298)$. So with 99% confidence, we estimate the probability of a squeaky hip replacement to be between 0.122 and 0.298. ■

4.2.1 Confidence Intervals for Difference in Means

A practical problem of interest is the comparison of two distributions, that is, comparing the distributions of two random variables, say X and Y. In this section, we compare the means of X and Y. Denote the means of X and Y by μ_1 and μ_2, respectively. In particular, we obtain confidence intervals for the difference $\Delta = \mu_1 - \mu_2$. Assume that the variances of X and Y are finite and denote them as $\sigma_1^2 = \text{Var}(X)$ and $\sigma_2^2 = \text{Var}(Y)$. Let $X_1, \ldots, X_{n_1}$ be a random sample from the distribution of X and let $Y_1, \ldots, Y_{n_2}$ be a random sample from the distribution of Y. Assume that the samples were gathered independently of one another. Let $\overline{X} = n_1^{-1}\sum_{i=1}^{n_1} X_i$ and $\overline{Y} = n_2^{-1}\sum_{i=1}^{n_2} Y_i$ be the sample means. Let $\widehat{\Delta} = \overline{X} - \overline{Y}$. The statistic $\widehat{\Delta}$ is an unbiased estimator of Δ. This difference, $\widehat{\Delta} - \Delta$, is the numerator of the pivot random variable. By independence of the samples,

$$\text{Var}(\widehat{\Delta}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Let $S_1^2 = (n_1 - 1)^{-1} \sum_{i=1}^{n_1} (X_i - \overline{X})^2$ and $S_2^2 = (n_2 - 1)^{-1} \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2$ be the sample variances. Then estimating the variances by the sample variances, consider the random variable

$$Z = \frac{\widehat{\Delta} - \Delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \tag{4.2.8}$$

By the independence of the samples and Theorem 4.2.1, this pivot variable has an approximate $N(0,1)$ distribution. This leads to the approximate $(1-\alpha)100\%$ confidence interval for $\Delta = \mu_1 - \mu_2$ given by

$$\left((\overline{x} - \overline{y}) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\overline{x} - \overline{y}) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right), \tag{4.2.9}$$

where $\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$ is the standard error of $\overline{X} - \overline{Y}$. This is a large sample $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$.

The above confidence interval is approximate. In this situation we can obtain exact confidence intervals if we assume that the distributions of X and Y are normal with the same variance; i.e., $\sigma_1^2 = \sigma_2^2$. Thus the distributions can differ only in location, i.e., a **location model**. Assume then that X is distributed $N(\mu_1, \sigma^2)$ and Y is distributed $N(\mu_2, \sigma^2)$, where σ^2 is the common variance of X and Y. As above, let $X_1, \ldots, X_{n_1}$ be a random sample from the distribution of X, let $Y_1, \ldots, Y_{n_2}$ be a random sample from the distribution of Y, and assume that the samples are independent of one another. Let $n = n_1 + n_2$ be the total sample size. Our estimator of Δ is $\overline{X} - \overline{Y}$. Our goal is to show that a pivot random variable, defined below, has a t-distribution, which is defined in Section 3.6.

Because $\overline{X}$ is distributed $N(\mu_1, \sigma^2/n_1)$, $\overline{Y}$ is distributed $N(\mu_2, \sigma^2/n_2)$, and $\overline{X}$ and $\overline{Y}$ are independent, we have the result

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ has a } N(0,1) \text{ distribution.} \tag{4.2.10}$$

This serves as the numerator of our T-statistic.

Let

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{4.2.11}$$

Note that S_p^2 is a weighted average of S_1^2 and S_2^2. It is easy to see that S_p^2 is an unbiased estimator of σ^2. It is called the **pooled estimator** of σ^2. Also, because $(n_1 - 1)S_1^2/\sigma^2$ has a $\chi^2(n_1 - 1)$ distribution, $(n_2 - 1)S_2^2/\sigma^2$ has a $\chi^2(n_2 - 1)$ distribution, and S_1^2 and S_2^2 are independent, we have that $(n-2)S_p^2/\sigma^2$ has a $\chi^2(n-2)$ distribution; see Corollary 3.3.1. Finally, because S_1^2 is independent of $\overline{X}$ and S_2^2 is independent of $\overline{Y}$, and the random samples are independent of each other, it follows that S_p^2 is independent of expression (4.2.10). Therefore, from the

result of Section 3.6.1 concerning Student's t-distribution, we have that

$$\begin{aligned} T &= \frac{[(\overline{X}-\overline{Y})-(\mu_1-\mu_2)]/\sigma\sqrt{n_1^{-1}+n_2^{-1}}}{\sqrt{(n-2)S_p^2/(n-2)\sigma^2}} \\ &= \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \end{aligned} \tag{4.2.12}$$

has a t-distribution with $n-2$ degrees of freedom. From this last result, it is easy to see that the following interval is an exact $(1-\alpha)100\%$ confidence interval for $\Delta = \mu_1 - \mu_2$:

$$\left((\overline{x}-\overline{y}) - t_{(\alpha/2,n-2)}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, (\overline{x}-\overline{y}) + t_{(\alpha/2,n-2)}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right). \tag{4.2.13}$$

A consideration of the difficulty encountered when the unknown variances of the two normal distributions are not equal is assigned to one of the exercises.

Example 4.2.4. To illustrate the pooled t-confidence interval, consider the baseball data presented in Hettmansperger and McKean (2011). It consists of 6 variables recorded on 59 professional baseball players of which 33 are hitters and 26 are pitchers. The data can be found in the file `bb.rda` located at the site listed in Chapter 1. The height in inches of a player is one of these measurements and in this example we consider the difference in heights between pitchers and hitters. Denote the true mean heights of the pitchers and hitters by μ_p and μ_h, respectively, and let $\Delta = \mu_p - \mu_h$. The sample averages of the heights are 75.19 and 72.67 inches for the pitchers and hitters, respectively. Hence, our point estimate of Δ is 2.53 inches. Assuming the file `bb.rda` has been loaded in R, the following R segment computes the 95% confidence interval for Δ:

```
hitht=height[hitpitind==1]; pitht=height[hitpitind==0]
t.test(pitht,hitht,var.equal=T)
```

The confidence interval computes to $(1.42, 3.63)$. Note that all values in the confidence interval are positive, indicating that on the average pitchers are taller than hitters. ■

Remark 4.2.1. Suppose X and Y are not normally distributed but that their distributions differ only in location. As we show in Chapter 5, the above interval, (4.2.13), is then approximate and not exact. ■

4.2.2 Confidence Interval for Difference in Proportions

Let X and Y be two independent random variables with Bernoulli distributions $b(1, p_1)$ and $b(1, p_2)$, respectively. Let us now turn to the problem of finding a confidence interval for the difference $p_1 - p_2$. Let $X_1, \ldots, X_{n_1}$ be a random sample from the distribution of X and let $Y_1, \ldots, Y_{n_2}$ be a random sample from the distribution of Y. As above, assume that the samples are independent of one another and let

$n = n_1 + n_2$ be the total sample size. Our estimator of $p_1 - p_2$ is the difference in sample proportions, which, of course, is given by $\overline{X} - \overline{Y}$. We use the traditional notation and write $\hat{p}_1$ and $\hat{p}_2$ instead of $\overline{X}$ and $\overline{Y}$, respectively. Hence, from the above discussion, an interval such as (4.2.9) serves as an approximate confidence interval for $p_1 - p_2$. Here, $\sigma_1^2 = p_1(1-p_1)$ and $\sigma_2^2 = p_2(1-p_2)$. In the interval, we estimate these by $\hat{p}_1(1-\hat{p}_1)$ and $\hat{p}_2(1-\hat{p}_2)$, respectively. Thus our approximate $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}. \tag{4.2.14}$$

Example 4.2.5. Kloke and McKean (2014), page 33, discuss a data set from the original clinical study of the Salk polio vaccine in 1954. At random, one group of children (Treated) received the vaccine while the other group (Control) received a placebo. Let p_c and p_T denote the true proportions of polio cases for control and treated populations, respectively. The tabled results are:

Group	No. Children	No. Polio Cases	Sample Proportion
Treated	200,745	57	0.000284
Control	201,229	199	0.000706

Note that $\hat{p}_C > \hat{p}_T$. The following R segment computes the 95% confidence interval for $p_c - p_T$:

```
prop.test(c(199,57),c(201229,200745))
```

The confidence interval is $(0.00054, 0.00087)$. All values in this interval are positive, indicating that the vaccine is effective in reducing the incidence of polio. ■

EXERCISES

4.2.1. Let the observed value of the mean $\overline{X}$ and of the sample variance of a random sample of size 20 from a distribution that is $N(\mu, \sigma^2)$ be 81.2 and 26.5, respectively. Find respectively 90%, 95% and 99% confidence intervals for μ. Note how the lengths of the confidence intervals increase as the confidence increases.

4.2.2. Consider the data on the lifetimes of motors given in Exercise 4.1.1. Obtain a large sample 95% confidence interval for the mean lifetime of a motor.

4.2.3. Suppose we assume that $X_1, X_2, \ldots, X_n$ is a random sample from a $\Gamma(1, \theta)$ distribution.

(a) Show that the random variable $(2/\theta)\sum_{i=1}^{n} X_i$ has a χ^2-distribution with $2n$ degrees of freedom.

(b) Using the random variable in part (a) as a pivot random variable, find a $(1-\alpha)100\%$ confidence interval for θ.

(c) Obtain the confidence interval in part (b) for the data of Exercise 4.1.1 and compare it with the interval you obtained in Exercise 4.2.2.

4.2.4. In Example 4.2.4, for the baseball data, we found a confidence interval for the mean difference in heights between the pitchers and hitters. In this exercise, find the pooled t 95% confidence interval for the mean difference in weights between the pitchers and hitters.

4.2.5. In the baseball data set discussed in the last exercise, it was found that out of the 59 baseball players, 15 were left-handed. Is this odd, since the proportion of left-handed males in America is about 11%? Answer by using (4.2.7) to construct a 95% approximate confidence interval for p, the proportion of left-handed professional baseball players.

4.2.6. Let $\overline{X}$ be the mean of a random sample of size n from a distribution that is $N(\mu, 9)$. Find n such that $P(\overline{X} - 1 < \mu < \overline{X} + 1) = 0.90$, approximately.

4.2.7. Let a random sample of size 17 from the normal distribution $N(\mu, \sigma^2)$ yield $\overline{x} = 4.7$ and $s^2 = 5.76$. Determine a 90% confidence interval for μ.

4.2.8. Let $\overline{X}$ denote the mean of a random sample of size n from a distribution that has mean μ and variance $\sigma^2 = 10$. Find n so that the probability is approximately 0.954 that the random interval $(\overline{X} - \frac{1}{2},\ \overline{X} + \frac{1}{2})$ includes μ.

4.2.9. Let $X_1, X_2, \ldots, X_9$ be a random sample of size 9 from a distribution that is $N(\mu, \sigma^2)$.

(a) If σ is known, find the length of a 95% confidence interval for μ if this interval is based on the random variable $\sqrt{9}(\overline{X} - \mu)/\sigma$.

(b) If σ is unknown, find the expected value of the length of a 95% confidence interval for μ if this interval is based on the random variable $\sqrt{9}(\overline{X} - \mu)/S$. *Hint:* Write $E(S) = (\sigma/\sqrt{n-1})E[((n-1)S^2/\sigma^2)^{1/2}]$.

(c) Compare these two answers.

4.2.10. Let $X_1, X_2, \ldots, X_n, X_{n+1}$ be a random sample of size $n+1$, $n > 1$, from a distribution that is $N(\mu, \sigma^2)$. Let $\overline{X} = \sum_1^n X_i/n$ and $S^2 = \sum_1^n (X_i - \overline{X})^2/(n-1)$. Find the constant c so that the statistic $c(\overline{X} - X_{n+1})/S$ has a t-distribution. If $n = 8$, determine k such that $P(\overline{X} - kS < X_9 < \overline{X} + kS) = 0.80$. The observed interval $(\overline{x} - ks,\ \overline{x} + ks)$ is often called an 80% **prediction interval** for X_9.

4.2.11. Let $X_1, \ldots, X_n$ be a random sample from a $N(0, 1)$ distribution. Then the probability that the random interval $\overline{X} \pm t_{\alpha/2, n-1}(s/\sqrt{n})$ traps $\mu = 0$ is $(1 - \alpha)$. To verify this empirically, in this exercise, we simulate m such intervals and calculate the proportion that trap 0, which should be "close" to $(1 - \alpha)$.

(a) Set $n = 10$ and $m = 50$. Run the R code `mat=matrix(rnorm(m*n),ncol=n)` which generates m samples of size n from the $N(0, 1)$ distribution. Each row of the matrix `mat` contains a sample. For this matrix of samples, the function below computes the $(1 - \alpha)100\%$ confidence intervals, returning them in a $m \times 2$ matrix. Run this function on your generated matrix `mat`. What is the proportion of successful confidence intervals?

```
getcis <- function(mat,cc=.90){
numb <- length(mat[,1]); ci <- c()
for(j in 1:numb)
{ci<-rbind(ci,t.test(mat[j,],conf.level=cc)$conf.int)}
return(ci)}
```

This function is also at the site discussed in Section 1.1.

(b) Run the following code which plots the intervals. Label the successful intervals. Comment on the variability of the lengths of the confidence intervals.

```
cis<-getcis(mat); x<-1:m
plot(c(cis[,1],cis[,2])~c(x,x),pch="",xlab="Sample",ylab="CI")
points(cis[,1]~x,pch="L");points(cis[,2]~x,pch="U"); abline(h=0)
```

4.2.12. In Exercise 4.2.11, the sampling was from the $N(0,1)$ distribution. Show, however, that setting $\mu = 0$ and $\sigma = 1$ is without loss of generality.
Hint: First, $X_1, \ldots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ if and only if $Z_1, \ldots, Z_n$ is a random sample from the $N(0,1)$, where $Z_i = (X_i - \mu)/\sigma$. Then show the confidence interval based on the Z_i's contains 0 if and only if the confidence interval based on the X_i's contains μ.

4.2.13. Change the code in the R function `getcis` so that it also returns the vector, `ind`, where `ind[i] = 1` if the ith confidence interval is successful and 0 otherwise. Show that the empirical confidence level is `mean(ind)`.

(a) Run 10,000 simulations for the normal setup in Exercise 4.2.11 and compute the empirical confidence level.

(b) Run 10,000 simulations when the sampling is from the Cauchy distribution, (1.8.8), and compute the empirical confidence level. Does it differ from (a)? Note that the R code `rcauchy(k)` returns a sample of size k from this Cauchy distribution.

(c) Note that these empirical confidence levels are proportions from samples that are independent. Hence, use the 95% confidence interval given in expression (4.2.14) to statistically investigate whether or not the true confidence levels differ. Comment.

4.2.14. Let $\overline{X}$ denote the mean of a random sample of size 25 from a gamma-type distribution with $\alpha = 4$ and $\beta > 0$. Use the Central Limit Theorem to find an approximate 0.954 confidence interval for μ, the mean of the gamma distribution.
Hint: Use the random variable $(\overline{X} - 4\beta)/(4\beta^2/25)^{1/2} = 5\overline{X}/2\beta - 10$.

4.2.15. Let $\overline{x}$ be the observed mean of a random sample of size n from a distribution having mean μ and known variance σ^2. Find n so that $\overline{x} - \sigma/4$ to $\overline{x} + \sigma/4$ is an approximate 95% confidence interval for μ.

4.2.16. Assume a binomial model for a certain random variable. If we desire a 90% confidence interval for p that is at most 0.02 in length, find n.
Hint: Note that $\sqrt{(y/n)(1-y/n)} \leq \sqrt{(\frac{1}{2})(1-\frac{1}{2})}$.

4.2.17. It is known that a random variable X has a Poisson distribution with parameter μ. A sample of 200 observations from this distribution has a mean equal to 3.4. Construct an approximate 90% confidence interval for μ.

4.2.18. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, where both parameters μ and σ^2 are unknown. A **confidence interval** for σ^2 can be found as follows. We know that $(n-1)S^2/\sigma^2$ is a random variable with a $\chi^2(n-1)$ distribution. Thus we can find constants a and b so that $P((n-1)S^2/\sigma^2 < b) = 0.975$ and $P(a < (n-1)S^2/\sigma^2 < b) = 0.95$. In R, b = qchisq(0.975,n-1), while a = qchisq(0.025,n-1).

(a) Show that this second probability statement can be written as

$$P((n-1)S^2/b < \sigma^2 < (n-1)S^2/a) = 0.95.$$

(b) If $n = 9$ and $s^2 = 7.93$, find a 95% confidence interval for σ^2.

(c) If μ is known, how would you modify the preceding procedure for finding a confidence interval for σ^2?

4.2.19. Let $X_1, X_2, \ldots, X_n$ be a random sample from a gamma distribution with known parameter $\alpha = 3$ and unknown $\beta > 0$. In Exercise 4.2.14, we obtained an approximate confidence interval for β based on the Central Limit Theorem. In this exercise obtain an exact confidence interval by first obtaining the distribution of $2\sum_1^n X_i/\beta$.
Hint: Follow the procedure outlined in Exercise 4.2.18.

4.2.20. When 100 tacks were thrown on a table, 60 of them landed point up. Obtain a 95% confidence interval for the probability that a tack of this type lands point up. Assume independence.

4.2.21. Let two independent random samples, each of size 10, from two normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ yield $\overline{x} = 4.8$, $s_1^2 = 8.64$, $\overline{y} = 5.6$, $s_2^2 = 7.88$. Find a 95% confidence interval for $\mu_1 - \mu_2$.

4.2.22. Let two independent random variables, Y_1 and Y_2, with binomial distributions that have parameters $n_1 = n_2 = 100$, p_1, and p_2, respectively, be observed to be equal to $y_1 = 50$ and $y_2 = 40$. Determine an approximate 90% confidence interval for $p_1 - p_2$.

4.2.23. Discuss the problem of finding a confidence interval for the difference $\mu_1 - \mu_2$ between the two means of two normal distributions if the variances σ_1^2 and σ_2^2 are known but not necessarily equal.

4.2.24. Discuss Exercise 4.2.23 when it is assumed that the variances are unknown and unequal. This is a very difficult problem, and the discussion should point out exactly where the difficulty lies. If, however, the variances are unknown but their ratio σ_1^2/σ_2^2 is a known constant k, then a statistic that is a T random variable can again be used. Why?

4.2.25. To illustrate Exercise 4.2.24, let $X_1, X_2, \ldots, X_9$ and $Y_1, Y_2, \ldots, Y_{12}$ represent two independent random samples from the respective normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. It is given that $\sigma_1^2 = 3\sigma_2^2$, but σ_2^2 is unknown. Define a random variable that has a t-distribution that can be used to find a 95% confidence interval for $\mu_1 - \mu_2$.

4.2.26. Let $\overline{X}$ and $\overline{Y}$ be the means of two independent random samples, each of size n, from the respective distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, where the common variance is known. Find n such that

$$P(\overline{X} - \overline{Y} - \sigma/5 < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + \sigma/5) = 0.90.$$

4.2.27. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two independent random samples from the respective normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, where the four parameters are unknown. To construct a *confidence interval for the ratio*, σ_1^2/σ_2^2, of the variances, form the quotient of the two independent χ^2 variables, each divided by its degrees of freedom, namely,

$$F = \frac{\frac{(m-1)S_2^2}{\sigma_2^2}/(m-1)}{\frac{(n-1)S_1^2}{\sigma_1^2}/(n-1)} = \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2},$$

where S_1^2 and S_2^2 are the respective sample variances.

(a) What kind of distribution does F have?

(b) Critical values a and b can be found so that $P(F < b) = 0.975$ and $P(a < F < b) = 0.95$. In R, b = qf(0.975,m-1,n-1), while a = qf(0.025,m-1,n-1).

(c) Rewrite the second probability statement as

$$P\left[a\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < b\frac{S_1^2}{S_2^2}\right] = 0.95.$$

The observed values, s_1^2 and s_2^2, can be inserted in these inequalities to provide a 95% confidence interval for σ_1^2/σ_2^2.

We caution the reader on the use of this confidence interval. This interval does depend on the normality of the distributions. If the distributions of X and Y are not normal then the true confidence coefficient may be far from the nominal confidence coefficient; see, for example, page 142 of Hettmansperger and McKean (2011) for discussion.

4.3 *Confidence Intervals for Parameters of Discrete Distributions

In this section, we outline a procedure that can be used to obtain exact confidence intervals for the parameters of discrete random variables. Let $X_1, X_2, \ldots, X_n$ be a

random sample on a discrete random variable X with pmf $p(x;\theta)$, $\theta \in \Omega$, where Ω is an interval of real numbers. Let $T = T(X_1, X_2, \ldots, X_n)$ be an estimator of θ with cdf $F_T(t;\theta)$. Assume that $F_T(t;\theta)$ is a nonincreasing and continuous function of θ for every t in the support of T. For a given realization of the sample, let t be the realized value of the statistic T. Let $\alpha_1 > 0$ and $\alpha_2 > 0$ be given such that $\alpha = \alpha_1 + \alpha_2 < 0.50$. Let $\underline{\theta}$ and $\overline{\theta}$ be the solutions of the equations

$$F_T(t-;\underline{\theta}) = 1 - \alpha_2 \text{ and } F_T(t;\overline{\theta}) = \alpha_1, \tag{4.3.1}$$

where $T-$ is the statistic whose support lags by one value of T's support. For instance, if $t_i < t_{i+1}$ are consecutive support values of T, then $T = t_{i+1}$ if and only if $T- = t_i$. Under these conditions, the interval $(\underline{\theta}, \overline{\theta})$ is a confidence interval for θ with confidence coefficient of at least $1 - \alpha$. We sketch a proof of this at the end of this section.

Before proceeding with discrete examples, we provide an example in the continuous case where the solution of equations (4.3.1) produces a familiar confidence interval.

Example 4.3.1. Assume $X_1, \ldots, X_n$ is a random sample from a $N(\theta, \sigma^2)$ distribution, where σ^2 is known. Let $\overline{X}$ be the sample mean and let $\overline{x}$ be its value for a given realization of the sample. Recall, from expression (4.2.6), that $\overline{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ is a $(1-\alpha)100\%$ confidence interval for θ. Assuming θ is the true mean, the cdf of $\overline{X}$ is $F_{\overline{X};\theta}(t) = \Phi[(t-\theta)/(\sigma/\sqrt{n})]$, where $\Phi(z)$ is the cdf of a standard normal distribution. Note for the continuous case that $\overline{X}-$ has the same distribution as $\overline{X}$. Then the first equation of (4.3.1) yields

$$\Phi[(\overline{x} - \theta)/(\sigma/\sqrt{n})] = 1 - (\alpha/2);$$

i.e.,

$$(\overline{x} - \theta)/(\sigma/\sqrt{n}) = \Phi^{-1}[1 - (\alpha/2)] = z_{\alpha/2}.$$

Solving for θ, we obtain the lower bound of the confidence interval $\overline{x} - z_{\alpha/2}(\sigma/\sqrt{n})$. Similarly, the solution of the second equation is the upper bound of the confidence interval. ■

For the discrete case, generally iterative algorithms are used to solve equations (4.3.1). In practice, the function $F_T(T;\overline{\theta})$ is often strictly decreasing and continuous in θ, so a simple algorithm often suffices. We illustrate the examples below by using the simple **bisection algorithm**, which we now briefly discuss.

Remark 4.3.1 (Bisection Algorithm)**.** Suppose we want to solve the equation $g(x) = d$, where $g(x)$ is strictly decreasing. Assume on a given step of the algorithm that $a < b$ bracket the solution; i.e., $g(a) > d > g(b)$. Let $c = (a+b)/2$. Then on the next step of the algorithm, the new bracket values a and b are determined by

$$\begin{aligned} &\text{if}(g(c) > d) \quad \text{then} \quad \{a \leftarrow c \text{ and } b \leftarrow b\} \\ &\text{if}(g(c) < d) \quad \text{then} \quad \{a \leftarrow a \text{ and } b \leftarrow c\}. \end{aligned}$$

The algorithm continues until $|a - b| < \epsilon$, where $\epsilon > 0$ is a specified tolerance. ■

Example 4.3.2 (Confidence Interval for a Bernoulli Proportion)**.** Let X have a Bernoulli distribution with θ as the probability of success. Let $\Omega = (0, 1)$. Suppose $X_1, X_2, \ldots, X_n$ is a random sample on X. As our point estimator of θ, we consider $\overline{X}$, which is the sample proportion of successes. The cdf of $n\overline{X}$ is binomial$(n.\theta)$. Thus

$$\begin{aligned} F_{\overline{X}}(\overline{x};\theta) &= P(n\overline{X} \leq n\overline{x}) \\ &= \sum_{j=0}^{n\overline{x}} \binom{n}{j} \theta^j (1-\theta)^{n-j} \\ &= 1 - \sum_{j=n\overline{x}+1}^{n} \binom{n}{j} \theta^j (1-\theta)^{n-j} \\ &= 1 - \int_0^\theta \frac{n!}{(n\overline{x})![n-(n\overline{x}+1)]!} z^{n\overline{x}} (1-z)^{n-(n\overline{x}+1)}\, dz, \end{aligned} \quad (4.3.2)$$

where the last equality, involving the incomplete β-function, follows from Exercise 4.3.6. By the fundamental theorem of calculus and expression (4.3.2),

$$\frac{d}{d\theta} F_{\overline{X}}(\overline{x};\theta) = -\frac{n!}{(n\overline{x})![n-(n\overline{x}+1]!} \theta^{n\overline{x}} (1-\theta)^{n-(n\overline{x}+1)} < 0;$$

hence, $F_{\overline{X}}(\overline{x};\theta)$ is a strictly decreasing function of θ, for each $\overline{x}$. Next, let $\alpha_1, \alpha_2 > 0$ be specified constants such that $\alpha_1 + \alpha_2 < 1/2$ and let $\underline{\theta}$ and $\overline{\theta}$ solve the equations

$$F_{\overline{X}}(\overline{x}-;\underline{\theta}) = 1 - \alpha_2 \text{ and } F_{\overline{x}}(\overline{X};\overline{\theta}) = \alpha_1. \quad (4.3.3)$$

Then $(\underline{\theta}, \overline{\theta})$ is a confidence interval for θ with confidence coefficient at least $1 - \alpha$, where $\alpha = \alpha_1 + \alpha_2$. These equations can be solved iteratively, as discussed in the following numerical illustration.

Numerical Illustration. Suppose $n = 30$ and the realization of the sample mean is $\overline{x} = 0.60$, i.e., the sample produced $n\overline{x} = 18$ successes. Take $\alpha_1 = \alpha_2 = 0.05$. Because the support of the binomial consists of integers and $n\overline{x} = 18$, we can write equations (4.3.3) as

$$\textstyle\sum_{j=0}^{17} \binom{n}{j} \underline{\theta}^j (1-\underline{\theta})^{n-j} = 0.95 \text{ and } \sum_{j=0}^{18} \binom{n}{j} \overline{\theta}^j (1-\overline{\theta})^{n-j} = 0.05\,. \quad (4.3.4)$$

Let $bin(n, p)$ denote a random variable with binomial distribution with parameters n and p. Because $P(bin(30, 0.4) \leq 17) =$ pbinom(17,30,.4) $= 0.9787$ and because $P(bin(30, 0.45) \leq 17) =$ pbinom(17,30,.45) $= 0.9286$, the values 0.4 and 0.45 bracket the solution to the first equation. We use these bracket values as input to the R function[1] `binomci.r` which iteratively solves the equation. The call and its output are:

```
> binomci(17,30,.4,.45,.95);    $solution  0.4339417
```

[1]Download this function at the site given in the preface.

So the solution to the first equation is $\underline{\theta} = 0.434$. In the same way, because $P(bin(30, 0.7) \leq 18) = 0.1593$ and $P(bin(30, 0.8) \leq 18) = 0.0094$, the values 0.7 and 0.8 bracket the solution to the second equation. The R segment for the solution is:

```
> binomci(18,30,.7,.8,.05);    $solution  0.75047
```

Thus the confidence interval is $(0.434, 0.750)$, with a confidence of at least 90%. For comparison, the asymptotic 90% confidence interval of expression (4.2.7) is $(0.453, 0.747)$; see Exercise 4.3.2. ■

Example 4.3.3 (Confidence Interval for the Mean of a Poisson Distribution)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample on a random variable X that has a Poisson distribution with mean θ. Let $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ be our point estimator of θ. As with the Bernoulli confidence interval in the last example, we can work with $n\overline{X}$, which, in this case, has a Poisson distribution with mean $n\theta$. The cdf of $\overline{X}$ is

$$\begin{aligned} F_{\overline{X}}(\overline{x};\theta) &= \sum_{j=0}^{n\overline{x}} e^{-n\theta}\frac{(n\theta)^j}{j!} \\ &= \frac{1}{\Gamma(n\overline{x}+1)}\int_{n\theta}^{\infty} x^{n\overline{x}}e^{-x}\,dx, \end{aligned} \tag{4.3.5}$$

where the integral equation is obtained in Exercise 4.3.7. From expression (4.3.5), we immediately have

$$\frac{d}{d\theta}F_{\overline{X}}(\overline{x};\theta) = \frac{-n}{\Gamma(n\overline{x}+1)}(n\theta)^{n\overline{x}}e^{-n\theta} < 0.$$

Therefore, $F_{\overline{X}}(\overline{x};\theta)$ is a strictly decreasing function of θ for every fixed $\overline{x}$. For a given sample, let $\overline{x}$ be the realization of the statistic $\overline{X}$. Hence, as discussed above, for $\alpha_1, \alpha_2 > 0$ such that $\alpha_1 + \alpha_2 < 1/2$, the confidence interval is given by $(\underline{\theta}, \overline{\theta})$, where

$$\textstyle\sum_{j=0}^{n\overline{x}-1} e^{-n\underline{\theta}}\frac{(n\underline{\theta})^j}{j!} = 1-\alpha_2 \;\text{ and }\; \sum_{j=0}^{n\overline{x}} e^{-n\overline{\theta}}\frac{(n\overline{\theta})^j}{j!} = \alpha_1. \tag{4.3.6}$$

The confidence coefficient of the interval $(\underline{\theta}, \overline{\theta})$ is at least $1-\alpha = 1-(\alpha_1+\alpha_2)$. As with the Bernoulli proportion, these equations can be solved iteratively.

Numerical Illustration. Suppose $n = 25$ and the realized value of $\overline{X}$ is $\overline{x} = 5$; hence, $n\overline{x} = 125$ events have occurred. We select $\alpha_1 = \alpha_2 = 0.05$. Then, by (4.3.7), our confidence interval solves the equations

$$\textstyle\sum_{j=0}^{124} e^{-n\underline{\theta}}\frac{(n\underline{\theta})^j}{j!} = 0.95 \;\text{ and }\; \sum_{j=0}^{125} e^{-n\overline{\theta}}\frac{(n\overline{\theta})^j}{j!} = 0.05. \tag{4.3.7}$$

Our R function[2] `poissonci.r` uses the bisection algorithm to solve these equations. Since $\text{ppois}(124, 25*4) = 0.9932$ and $\text{ppois}(124, 25*4.4) = 0.9145$, for the first equation, 4.0 and 4.4 bracket the solution. Here is the call to `poissonci.r` along with the solution (the lower bound of the confidence interval):

[2]Download this function at the site given in the Preface.

```
> poissonci(124,25,4,4.4,.95);    $solution  4.287836
```

Since ppois$(125, 25 * 5.5) = 0.1528$ and ppois$(125, 25 * 6.0) = 0.0204$, for the second equation, 5.5 and 6.0 bracket the solution. Hence, the computation of the lower bound of the confidence interval is:

```
> poissonci(125,25,5.5,6,.05);   $solution  5.800575
```

So the confidence interval is $(4.287, 5.8)$, with confidence at least 90%. Note that the confidence interval is right-skewed, similar to the Poisson distribution. ■

A brief sketch of the theory behind this confidence interval follows. Consider the general setup in the first paragraph of this section, where T is an estimator of the unknown parameter θ and $F_T(t;\theta)$ is the cdf of T. Define

$$\overline{\theta} = \sup\{\theta : F_T(T;\theta) \geq \alpha_1\} \tag{4.3.8}$$

$$\underline{\theta} = \inf\{\theta : F_T(T-;\theta) \leq 1-\alpha_2\}. \tag{4.3.9}$$

Hence, we have

$$\begin{aligned} \theta > \overline{\theta} &\Rightarrow F_T(T;\theta) \leq \alpha_1 \\ \theta < \underline{\theta} &\Rightarrow F_T(T-;\theta) \geq 1-\alpha_2. \end{aligned}$$

These implications lead to

$$\begin{aligned} P[\underline{\theta} < \theta < \overline{\theta}] &= 1 - P[\{\theta < \underline{\theta}\} \cup \{\theta > \overline{\theta}\}] \\ &= 1 - P[\theta < \underline{\theta}] - P[\theta > \overline{\theta}] \\ &\geq 1 - P[F_T(T-;\theta) \geq 1-\alpha_2] - P[F_T(T;\theta) \leq \alpha_1] \\ &\geq 1 - \alpha_1 - \alpha_2, \end{aligned}$$

where the last inequality is evident from equations (4.3.8) and (4.3.9). A rigorous proof can be based on Exercise 4.8.13; see page 425 of Shao (1998) for details.

EXERCISES

4.3.1. Recall For the baseball data (`bb.rda`), 15 out of 59 ballplayers are left-handed. Let p be the probability that a professional baseball player is left-handed. Determine an exact 90% confidence interval for p. Show first that the equations to be solved are:

$$\textstyle\sum_{j=0}^{14} \binom{n}{j} \underline{\theta}^j (1-\underline{\theta})^{n-j} = 0.95 \text{ and } \sum_{j=0}^{15} \binom{n}{j} \overline{\theta}^j (1-\overline{\theta})^{n-j} = 0.05\,.$$

Then do the following steps to obtain the confidence interval.

(a) Show that 0.10 and 0.17 bracket the solution to the first equation.

(b) Show that 0.34 and 0.38 bracket the solution to the second equation.

(c) Then use the R function `binomci.r` to solve the equations.

4.3.2. In Example 4.3.2, verify the result for the asymptotic confidence interval for θ.

4.3.3. In Exercise 4.2.20, the large sample confidence interval was obtained for the probability that a tack tossed on a table lands point up. Find the discrete exact confidence interval for this proportion.

4.3.4. Suppose $X_1, X_2, \ldots, X_{10}$ is a random sample on a random variable X that has a Poisson distribution with mean θ. Suppose the realized value of the sample mean is 0.5; i.e., $n\overline{x} = 5$ events occurred. Suppose we want to compute the exact 90% confidence interval for θ, as determined by equations (4.3.7).

(a) Show that 0.19 and 0.20 bracket the solution to the first equation.

(b) Show that 1.0 and 1.1 bracket the solution to the second equation.

(c) Then use the R function `poissonci.r` to solve the equations.

4.3.5. Consider the same setup as in Example 4.3.1 except now assume that σ^2 is unknown. Using the distribution of $(\overline{X} - \theta)/(S/\sqrt{n})$, where S is the sample standard deviation, set up the equations and derive the t-interval, (4.2.3), for θ.

4.3.6. Using Exercise 3.3.22, show that

$$\int_0^p \frac{n!}{(k-1)!(n-k)!} z^{k-1}(1-z)^{n-k}\, dz = \sum_{w=k}^{n} \binom{n}{w} p^w (1-p)^{n-w},$$

where $0 < p < 1$, and k and n are positive integers such that $k \leq n$.
Hint: Differentiate both sides with respect to p. The derivative of the right side is a sum of differences. Show it simplifies to the derivative of the left side. Hence, the sides differ by a constant. Finally, show that the constant is 0.

4.3.7. This exercise obtains a useful identity for the cdf of a Poisson cdf.

(a) Use Exercise 3.3.5 to show that this identity is true:

$$\frac{\lambda^n}{\Gamma(n)} \int_1^\infty x^{n-1} e^{-x\lambda}\, dx = \sum_{j=0}^{n-1} e^{-\lambda} \frac{\lambda^j}{j!},$$

for $\lambda > 0$ and n a positive integer.

Hint: Just consider a Poisson process on the unit interval with mean λ. Let W_n be the waiting time until the nth event. Then the left side is $P(W_n > 1)$. Why?

(b) Obtain the identity used in Example 4.3.3, by making the transformation $z = \lambda x$ in the above integral.

4.4 Order Statistics

In this section the notion of an **order statistic** is defined and some of its simple properties are investigated. These statistics have in recent times come to play an

important role in statistical inference partly because some of their properties do not depend upon the distribution from which the random sample is obtained.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution of the *continuous type* having a pdf $f(x)$ that has support $\mathcal{S} = (a, b)$, where $-\infty \leq a < b \leq \infty$. Let Y_1 be the smallest of these X_i, Y_2 the next X_i in order of magnitude, $\ldots$, and Y_n the largest of X_i. That is, $Y_1 < Y_2 < \cdots < Y_n$ represent $X_1, X_2, \ldots, X_n$ when the latter are arranged in ascending order of magnitude. We call $Y_i,\ i = 1, 2, \ldots, n$, the ith **order statistic** of the random sample $X_1, X_2, \ldots, X_n$. Then the joint pdf of $Y_1, Y_2, \ldots, Y_n$ is given in the following theorem.

Theorem 4.4.1. *Using the above notation, let $Y_1 < Y_2 < \cdots < Y_n$ denote the n order statistics based on the random sample $X_1, X_2, \ldots, X_n$ from a continuous distribution with pdf $f(x)$ and support (a, b). Then the joint pdf of $Y_1, Y_2, \ldots, Y_n$ is given by*

$$g(y_1, y_2, \ldots, y_n) = \begin{cases} n! f(y_1) f(y_2) \cdots f(y_n) & a < y_1 < y_2 < \cdots < y_n < b \\ 0 & \text{elsewhere.} \end{cases} \tag{4.4.1}$$

Proof: Note that the support of $X_1, X_2, \ldots, X_n$ can be partitioned into $n!$ mutually disjoint sets that map onto the support of $Y_1, Y_2, \ldots, Y_n$, namely, $\{(y_1, y_2, \ldots, y_n) : a < y_1 < y_2 < \cdots < y_n < b\}$. One of these $n!$ sets is $a < x_1 < x_2 < \cdots < x_n < b$, and the others can be found by permuting the n xs in all possible ways. The transformation associated with the one listed is $x_1 = y_1, x_2 = y_2, \ldots, x_n = y_n$, which has a Jacobian equal to 1. However, the Jacobian of each of the other transformations is either ± 1. Thus

$$\begin{aligned} g(y_1, y_2, \ldots, y_n) &= \sum_{i=1}^{n!} |J_i| f(y_1) f(y_2) \cdots f(y_n) \\ &= \begin{cases} n! f(y_1) f(y_2) \cdots f(y_n) & a < y_1 < y_2 < \cdots < y_n < b \\ 0 & \text{elsewhere,} \end{cases} \end{aligned}$$

as was to be proved. ■

Example 4.4.1. Let X denote a random variable of the continuous type with a pdf $f(x)$ that is positive and continuous, with support $\mathcal{S} = (a, b)$, $-\infty \leq a < b \leq \infty$. The distribution function $F(x)$ of X may be written

$$F(x) = \int_a^x f(w)\, dw, \quad a < x < b.$$

If $x \leq a,\ F(x) = 0$; and if $b \leq x,\ F(x) = 1$. Thus there is a unique median m of the distribution with $F(m) = \frac{1}{2}$. Let $X_1,\ X_2,\ X_3$ denote a random sample from this distribution and let $Y_1 < Y_2 < Y_3$ denote the order statistics of the sample. Note that Y_2 is the sample median. We compute the probability that $Y_2 \leq m$. The joint pdf of the three order statistics is

$$g(y_1, y_2, y_3) = \begin{cases} 6 f(y_1) f(y_2) f(y_3) & a < y_1 < y_2 < y_3 < b \\ 0 & \text{elsewhere.} \end{cases}$$

The pdf of Y_2 is then

$$\begin{aligned} h(y_2) &= 6f(y_2)\int_{y_2}^{b}\int_{a}^{y_2} f(y_1)f(y_3)\,dy_1 dy_3 \\ &= \begin{cases} 6f(y_2)F(y_2)[1-F(y_2)] & a < y_2 < b \\ 0 & \text{elsewhere.} \end{cases} \end{aligned}$$

Accordingly,

$$\begin{aligned} P(Y_2 \le m) &= 6\int_a^m \{F(y_2)f(y_2) - [F(y_2)]^2 f(y_2)\}\,dy_2 \\ &= 6\left\{\frac{[F(y_2)]^2}{2} - \frac{[F(y_2)]^3}{3}\right\}_a^m = \frac{1}{2}. \end{aligned}$$

Hence, for this situation, the median of the sample median Y_2 is the population median m. ■

Once it is observed that

$$\int_a^x [F(w)]^{\alpha-1} f(w)\,dw = \frac{[F(x)]^\alpha}{\alpha}, \quad \alpha > 0,$$

and that

$$\int_y^b [1-F(w)]^{\beta-1} f(w)\,dw = \frac{[1-F(y)]^\beta}{\beta}, \quad \beta > 0,$$

it is easy to express the marginal pdf of any order statistic, say Y_k, in terms of $F(x)$ and $f(x)$. This is done by evaluating the integral

$$g_k(y_k) = \int_a^{y_k}\cdots\int_a^{y_2}\int_{y_k}^{b}\cdots\int_{y_{n-1}}^{b} n!f(y_1)f(y_2)\cdots f(y_n)\,dy_n\cdots dy_{k+1}\,dy_1\cdots dy_{k-1}.$$

The result is

$$g_k(y_k) = \begin{cases} \frac{n!}{(k-1)!(n-k)!}[F(y_k)]^{k-1}[1-F(y_k)]^{n-k}f(y_k) & a < y_k < b \\ 0 & \text{elsewhere.} \end{cases} \tag{4.4.2}$$

Example 4.4.2. Let $Y_1 < Y_2 < Y_3 < Y_4$ denote the order statistics of a random sample of size 4 from a distribution having pdf

$$f(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

We express the pdf of Y_3 in terms of $f(x)$ and $F(x)$ and then compute $P(\frac{1}{2} < Y_3)$. Here $F(x) = x^2$, provided that $0 < x < 1$, so that

$$g_3(y_3) = \begin{cases} \frac{4!}{2!\,1!}(y_3^2)^2(1-y_3^2)(2y_3) & 0 < y_3 < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Thus

$$\begin{aligned} P(\tfrac{1}{2} < Y_3) &= \int_{1/2}^{\infty} g_3(y_3)\,dy_3 \\ &= \int_{1/2}^{1} 24(y_3^5 - y_3^7)\,dy_3 = \frac{243}{256}. \end{aligned}$$

Finally, the joint pdf of any two order statistics, say $Y_i < Y_j$, is easily expressed in terms of $F(x)$ and $f(x)$. We have

$$\begin{aligned} g_{ij}(y_i, y_j) = \int_a^{y_i} \cdots \int_a^{y_2} \int_{y_i}^{y_j} \cdots \int_{y_{j-2}}^{y_j} \int_{y_j}^{b} \cdots \int_{y_{n-1}}^{b} n! f(y_1) \times \cdots \\ \times f(y_n)\,dy_n \cdots dy_{j+1} dy_{j-1} \cdots dy_{i+1} dy_1 \cdots dy_{i-1}. \end{aligned}$$

Since, for $\gamma > 0$,

$$\begin{aligned} \int_x^y [F(y) - F(w)]^{\gamma-1} f(w)\,dw &= -\frac{[F(y) - F(w)]^{\gamma}}{\gamma}\Bigg|_x^y \\ &= \frac{[F(y) - F(x)]^{\gamma}}{\gamma}, \end{aligned}$$

it is found that

$$g_{ij}(y_i, y_j) = \begin{cases} \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(y_i)]^{i-1}[F(y_j) - F(y_i)]^{j-i-1} \\ \quad \times [1 - F(y_j)]^{n-j} f(y_i) f(y_j) & a < y_i < y_j < b \\ 0 & \text{elsewhere.} \end{cases} \tag{4.4.3}$$

■

Remark 4.4.1 (Heuristic Derivation)**.** There is an easy method of remembering the pdf of a vector of order statistics such as the one given in formula (4.4.3). The probability $P(y_i < Y_i < y_i + \Delta_i,\ y_j < Y_j < y_j + \Delta_j)$, where Δ_i and Δ_j are small, can be approximated by the following multinomial probability. In n independent trials, $i-1$ outcomes must be less than y_i [an event that has probability $p_1 = F(y_i)$ on each trial]; $j - i - 1$ outcomes must be between $y_i + \Delta_i$ and y_j [an event with approximate probability $p_2 = F(y_j) - F(y_i)$ on each trial]; $n - j$ outcomes must be greater than $y_j + \Delta_j$ [an event with approximate probability $p_3 = 1 - F(y_j)$ on each trial]; one outcome must be between y_i and $y_i + \Delta_i$ [an event with approximate probability $p_4 = f(y_i)\Delta_i$ on each trial]; and, finally, one outcome must be between y_j and $y_j + \Delta_j$ [an event with approximate probability $p_5 = f(y_j)\Delta_j$ on each trial]. This multinomial probability is

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!\ 1!\ 1!} p_1^{i-1} p_2^{j-i-1} p_3^{n-j} p_4 p_5,$$

which is $g_{i,j}(y_i, y_j)\Delta_i\Delta_j$, where $g_{i,j}(y_i, y_j)$ is given in expression (4.4.3). ■

Certain functions of the order statistics $Y_1,\ Y_2, \ldots, Y_n$ are important statistics themselves. The **sample range** of the random sample is given by $Y_n - Y_1$ and the **sample midrange** is given by $(Y_1 + Y_n)/2$, which is called the **midrange** of the random sample. The **sample median** of the random sample is defined by

$$Q_2 = \begin{cases} Y_{(n+1)/2} & \text{if } n \text{ is odd} \\ (Y_{n/2} + Y_{(n/2)+1})/2 & \text{if } n \text{ is even.} \end{cases} \tag{4.4.4}$$

Example 4.4.3. Let $Y_1,\ Y_2,\ Y_3$ be the order statistics of a random sample of size 3 from a distribution having pdf

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

We seek the pdf of the sample range $Z_1 = Y_3 - Y_1$. Since $F(x) = x,\ 0 < x < 1$, the joint pdf of Y_1 and Y_3 is

$$g_{13}(y_1, y_3) = \begin{cases} 6(y_3 - y_1) & 0 < y_1 < y_3 < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

In addition to $Z_1 = Y_3 - Y_1$, let $Z_2 = Y_3$. The functions $z_1 = y_3 - y_1,\ z_2 = y_3$ have respective inverses $y_1 = z_2 - z_1,\ y_3 = z_2$, so that the corresponding Jacobian of the one-to-one transformation is

$$J = \begin{vmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} \\ \frac{\partial y_3}{\partial z_1} & \frac{\partial y_3}{\partial z_2} \end{vmatrix} = \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = -1.$$

Thus the joint pdf of Z_1 and Z_2 is

$$h(z_1, z_2) = \begin{cases} |-1|6z_1 = 6z_1 & 0 < z_1 < z_2 < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Accordingly, the pdf of the range $Z_1 = Y_3 - Y_1$ of the random sample of size 3 is

$$h_1(z_1) = \begin{cases} \int_{z_1}^{1} 6z_1\, dz_2 = 6z_1(1 - z_1) & 0 < z_1 < 1 \\ 0 & \text{elsewhere.} \end{cases}$$ ■

4.4.1 Quantiles

Let X be a random variable with a continuous cdf $F(x)$. For $0 < p < 1$, define the p**th quantile** of X to be $\xi_p = F^{-1}(p)$. For example, $\xi_{0.5}$, the median of X, is the 0.5 quantile. Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution of X and let $Y_1 < Y_2 < \cdots < Y_n$ be the corresponding order statistics. Let k be the greatest integer less than or equal to $[p(n + 1)]$. We next define an estimator of ξ_p after making the following observation. The area under the pdf $f(x)$ to the left of Y_k is $F(Y_k)$. The expected value of this area is

$$E(F(Y_k)) = \int_a^b F(y_k) g_k(y_k)\, dy_k,$$

where $g_k(y_k)$ is the pdf of Y_k given in expression (4.4.2). If, in this integral, we make a change of variables through the transformation $z = F(y_k)$, we have

$$E(F(Y_k)) = \int_0^1 \frac{n!}{(k-1)!(n-k)!} z^k (1-z)^{n-k}\, dz.$$

Comparing this to the integral of a beta pdf, we see that it is equal to

$$E(F(Y_k)) = \frac{n!k!(n-k)!}{(k-1)!(n-k)!(n+1)!} = \frac{k}{n+1}.$$

On the average, there is $k/(n+1)$ of the total area to the left of Y_k. Because $p \doteq k/(n+1)$, it seems reasonable to take Y_k as an estimator of the quantile ξ_p. Hence, we call Y_k the pth **sample quantile** It is also called the $100p$th **percentile of the sample**.

Remark 4.4.2. Some statisticians define sample quantiles slightly differently from what we have. For one modification with $1/(n+1) < p < n/(n+1)$, if $(n+1)/p$ is not equal to an integer, then the pth quantile of the sample may be defined as follows. Write $(n+1)p = k + r$, where $k = [(n+1)p]$ and r is a proper fraction, using the weighted average. Then the pth quantile of the sample is the weighted average

$$(1-r)Y_k + rY_{k+1}, \tag{4.4.5}$$

which is an estimator of the pth quantile. As n becomes large, however, all these modified definitions are essentially the same. For R code, let the R vector `x` contain the realization of the sample. Then the call `quantile(x,p)` computes a pth quantile of form (4.4.5). ■

Sample quantiles are useful descriptive statistics. For instance, if y_k is the pth quantile of the realized sample, then we know that approximately $p100\%$ of the data are less than or equal to y_k and approximately $(1-p)100\%$ of the data are greater than or equal to y_k. Next we discuss two statistical applications of quantiles.

A **five-number** summary of the data consists of the following five sample quantiles: the minimum (Y_1), the first quartile ($Y_{.25(n+1)}$), the median defined in expression (4.4.4), the third quartile ($Y_{.75(n+1)}$), and the maximum (Y_n). For this section, we use the notation Q_1, Q_2, and Q_3 to denote, respectively, the first quartile, median, and third quartile of the sample.

The five-number summary divides the data into their quartiles, offering a simple and easily interpretable description of the data. Five-number summaries were made popular by the work of the late Professor John Tukey [see Tukey (1977) and Mosteller and Tukey (1977)]. Tukey used the median of the lower half of the data (from minimum to median) and the median of the upper half of the data instead of the first and third quartiles. He referred to these quantities as the **hinges** of the data. The R function `fivenum(x)` returns the hinges along with the minimum, median, and maximum of the data.

Example 4.4.4. The following data are the ordered realizations of a random sample of size 15 on a random variable X.

56	70	89	94	96	101	102	102
102	105	106	108	110	113	116	

For these data, since $n + 1 = 16$, the realizations of the five-number summary are $y_1 = 56$, $Q_1 = y_4 = 94$, $Q_2 = y_8 = 102$, $Q_3 = y_{12} = 108$, and $y_{15} = 116$. Hence, based on the five-number summary, the data range from 56 to 116; the middle 50% of the data range from 94 to 108; and the middle of the data occurred at 102. The data are in the file `eg4.4.4data.rda`. ■

The five-number summary is the basis for a useful and quick plot of the data. This is called a **boxplot** of the data. The box encloses the middle 50% of the data and a line segment is usually used to indicate the median. The extreme order statistics, however, are very sensitive to outlying points. So care must be used in placing these on the plot. We make use of the **box and whisker** plots defined by John Tukey. In order to define this plot, we need to define a potential outlier. Let $h = 1.5(Q_3 - Q_1)$ and define the **lower fence** (LF) and the **upper fence** (UF) by

$$LF = Q_1 - h \text{ and } UF = Q_3 + h. \tag{4.4.6}$$

Points that lie outside the fences, i.e., outside the interval (LF, UF), are called **potential outliers** and they are denoted by the symbol "0" on the boxplot. The whiskers then protrude from the sides of the box to what are called the **adjacent points**, which are the points within the fences but closest to the fences. Exercise 4.4.2 shows that the probability of an observation from a normal distribution being a potential outlier is 0.006977.

Example 4.4.5 (Example 4.4.4, Continued)**.** Consider the data given in Example 4.4.4. For these data, $h = 1.5(108 - 94) = 21$, $LF = 73$, and $UF = 129$. Hence the observations 56 and 70 are potential outliers. There are no outliers on the high side of the data. The lower adjacent point is 89. The boxplot of the data set is given in Panel A of Figure 4.4.1, which was computed by the R segment `boxplot(x)` where the R vector `x` contains the data.

Note that the point 56 is over $2h$ from Q_1. Some statisticians call such a point an "outlier" and label it with a symbol other than "O," but we do not make this distinction. ■

In practice, we often assume that the data follow a certain distribution. For example, we may assume that $X_1, \ldots, X_n$ are a random sample from a normal distribution with unknown mean and variance. Thus the form of the distribution of X is known, but the specific parameters are not. Such an assumption needs to be checked and there are many statistical tests which do so; see D'Agostino and Stephens (1986) for a thorough discussion of such tests. As our second statistical application of quantiles, we discuss one such diagnostic plot in this regard.

We consider the location and scale family. Suppose X is a random variable with cdf $F((x - a)/b)$, where $F(x)$ is known but a and $b > 0$ may not be. Let $Z = (X - a)/b$; then Z has cdf $F(z)$. Let $0 < p < 1$ and let $\xi_{X,p}$ be the pth quantile of X. Let $\xi_{Z,p}$ be the pth quantile of $Z = (X - a)/b$. Because $F(z)$ is known, $\xi_{Z,p}$ is known. But

$$p = P[X \leq \xi_{X,p}] = P\left[Z \leq \frac{\xi_{X,p} - a}{b}\right],$$

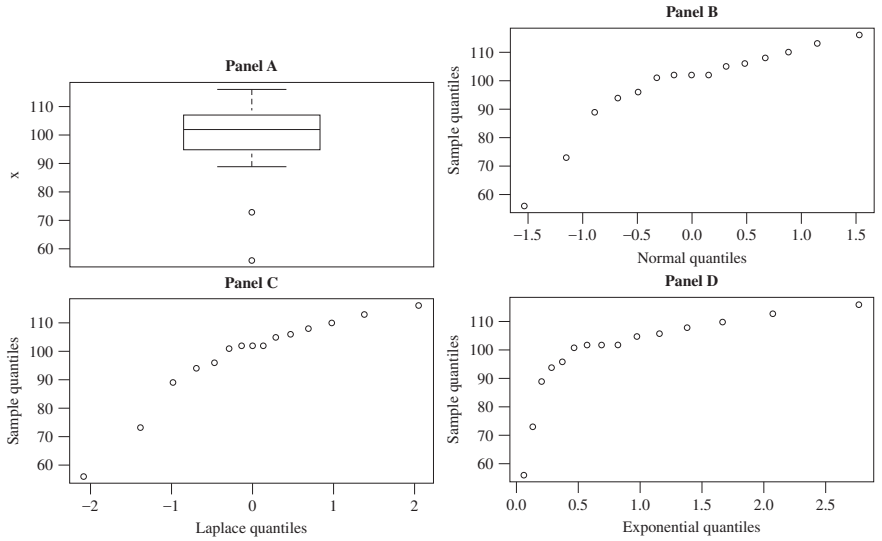

Figure 4.4.1: Boxplot and quantile plots for the data of Example 4.4.4.

from which we have the linear relationship

$$\xi_{X,p} = b\xi_{Z,p} + a. \tag{4.4.7}$$

Thus, if X has a cdf of the form of $F((x-a)/b)$, then the quantiles of X are linearly related to the quantiles of Z. Of course, in practice, we do not know the quantiles of X, but we can estimate them. Let $X_1, \ldots, X_n$ be a random sample from the distribution of X and let $Y_1 < \cdots < Y_n$ be the order statistics. For $k = 1, \ldots, n$, let $p_k = k/(n+1)$. Then Y_k is an estimator of ξ_{X,p_k}. Denote the corresponding quantiles of the cdf $F(z)$ by $\xi_{Z,p_k} = F^{-1}(p_k)$. Let y_k denote the realized value of Y_k. The plot of y_k versus ξ_{Z,p_k} is called a **q−q plot**, as it plots one set of quantiles from the sample against another set from the theoretical cdf $F(z)$. Based on the above discussion, the linearity of such a plot indicates that the cdf of X is of the form $F((x-a)/b)$.

Example 4.4.6 (Example 4.4.5, Continued)**.** Panels B, C, and D of Figure 4.4.1 contain $q-q$ plots of the data of Example 4.4.4 for three different distributions. The quantiles of a standard normal random variable are used for the plot in Panel B. Hence, as described above, this is the plot of y_k versus $\Phi^{-1}(k/(n+1))$, for $k = 1, 2, \ldots, n$. For Panel C, the population quantiles of the standard **Laplace** distribution are used; that is, the density of Z is $f(z) = (1/2)e^{-|z|}$, $-\infty < z < \infty$. For Panel D, the quantiles were generated from an exponential distribution with density $f(z) = e^{-z}$, $0 < z < \infty$, zero elsewhere. The generation of these quantiles is discussed in Exercise 4.4.1.

The plot farthest from linearity is that of Panel D. Note that this plot gives an indication of a more correct distribution. For the points to lie on a line, the

lower quantiles of Z must be spread out as are the higher quantiles; i.e., symmetric distributions may be more appropriate. The plots in Panels B and C are more linear than that of Panel D, but they still contain some curvature. Of the two, Panel C appears to be more linear. Actually, the data were generated from a Laplace distribution, so one would expect that Panel C would be the most linear of the three plots.

Many computer packages have commands to obtain the population quantiles used in this example. The R function `qqplotc4s2.r`, at the site listed in Chapter 1, obtains the normal, Laplace, and exponential quantiles used for Figure 4.4.1 and the plot. The call is `qqplotc4s2(x)` where the R vector `x` contains the data. ■

The $q-q$ plot using normal quantiles is often called a **normal** $q-q$ plot. If the data are in the R vector `x`, the plot is obtained by the call `qqnorm(x)`.

4.4.2 Confidence Intervals for Quantiles

Let X be a continuous random variable with cdf $F(x)$. For $0 < p < 1$, define the $100p$th distribution percentile to be ξ_p, where $F(\xi_p) = p$. For a sample of size n on X, let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics. Let $k = [(n+1)p]$. Then the $100p$th sample percentile Y_k is a point estimate of ξ_p.

We now derive a **distribution free** confidence interval for ξ_p, meaning it is a confidence interval for ξ_p which is free of any assumptions about $F(x)$ other than it is of the continuous type. Let $i < [(n+1)p] < j$, and consider the order statistics $Y_i < Y_j$ and the event $Y_i < \xi_p < Y_j$. For the ith order statistic Y_i to be less than ξ_p, it must be true that at least i of the X values are less than ξ_p. Moreover, for the jth order statistic to be greater than ξ_p, fewer than j of the X values are less than ξ_p. To put this in the context of a binomial distribution, the probability of success is $P(X < \xi_p) = F(\xi_p) = p$. Further, the event $Y_i < \xi_p < Y_j$ is equivalent to obtaining between i (inclusive) and j (exclusive) successes in n independent trials. Thus, taking probabilities, we have

$$P(Y_i < \xi_p < Y_j) = \sum_{w=i}^{j-1} \binom{n}{w} p^w (1-p)^{n-w}. \tag{4.4.8}$$

When particular values of n, i, and j are specified, this probability can be computed. By this procedure, suppose that it has been found that $\gamma = P(Y_i < \xi_p < Y_j)$. Then the probability is γ that the random interval (Y_i, Y_j) includes the quantile of order p. If the experimental values of Y_i and Y_j are, respectively, y_i and y_j, the interval (y_i, y_j) serves as a $100\gamma\%$ confidence interval for ξ_p, the quantile of order p. We use this in the next example to find a confidence interval for the median.

Example 4.4.7 (Confidence Interval for the Median)**.** Let X be a continuous random variable with cdf $F(x)$. Let $\xi_{1/2}$ denote the median of $F(x)$; i.e., $\xi_{1/2}$ solves $F(\xi_{1/2}) = 1/2$. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from the distribution of X with corresponding order statistics $Y_1 < Y_2 < \cdots < Y_n$. As before, let Q_2 denote the sample median, which is a point estimator of $\xi_{1/2}$. Select α, so that $0 < \alpha < 1$. Take $c_{\alpha/2}$ to be the $\alpha/2$th quantile of a binomial $b(n, 1/2)$ distribution;

that is, $P[S \leq c_{\alpha/2}] = \alpha/2$, where S is distributed $b(n, 1/2)$. Then note also that $P[S \geq n - c_{\alpha/2}] = \alpha/2$. (Because of the discreteness of the binomial distribution, either take a value of α for which these probabilities are correct or change the equalities to approximations.) Thus it follows from expression (4.4.8) that

$$P[Y_{c_{\alpha/2}+1} < \xi_{1/2} < Y_{n-c_{\alpha/2}}] = 1 - \alpha. \tag{4.4.9}$$

Hence, when the sample is drawn, if $y_{c_{\alpha/2}+1}$ and $y_{n-c_{\alpha/2}}$ are the realized values of the order statistics $Y_{c_{\alpha/2}+1}$ and $Y_{n-c_{\alpha/2}}$, then the interval

$$(y_{c_{\alpha/2}+1}, y_{n-c_{\alpha/2}}) \tag{4.4.10}$$

is a $(1 - \alpha)100\%$ confidence interval for $\xi_{1/2}$.

To illustrate this confidence interval, consider the data of Example 4.4.4. Suppose we want an 88% confidence interval for $\xi_{1/2}$. Then $\alpha/2 = 0.060$. Then $c_{\alpha/2} = 4$ because $P[S \leq 4] =$`pbinom(4,15,.5)`$= 0.059$, where the distribution of S is binomial with $n = 15$ and $p = 0.5$. Therefore, an 88% confidence interval for $\xi_{1/2}$ is $(y_5, y_{11}) = (96, 106)$.

The R function `onesampsgn(x)` computes a confidence interval for the median. For the data in Example 4.4.4, the code `onesampsgn(x,alpha=.12)` computes the confidence interval $(96, 106)$ for the median. ■

Note that because of the discreteness of the binomial distribution, only certain confidence levels are possible for this confidence interval for the median. If we further assume that $f(x)$ is symmetric about ξ, Chapter 10 presents other distribution free confidence intervals where this discreteness is much less of a problem.

EXERCISES

4.4.1. Obtain closed-form expressions for the distribution quantiles based on the exponential and Laplace distributions as discussed in Example 4.4.6.

4.4.2. Suppose the pdf $f(x)$ is symmetric about 0 with cdf $F(x)$. Show that the probability of a potential outlier from this distribution is $2F(4q_1)$, where $F^{-1}(0.25) = q_1$ Use this to obtain the probability that an observation is a potential outlier for the following distributions.

(a) The underlying distribution is normal. Use the $N(0, 1)$ distribution.

(b) The underlying distribution is **logistic**; that is, the pdf is given by

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}, \quad -\infty < x < \infty. \tag{4.4.11}$$

(c) The underlying distribution is Laplace, with the pdf

$$f(x) = \frac{1}{2}e^{-|x|}, \quad -\infty < x < \infty. \tag{4.4.12}$$

4.4.3. Consider the sample of data (data are in the file `ex4.4.3data.rda`):

13	5	202	15	99	4	67	83	36	11	301
23	213	40	66	106	78	69	166	84	64	

(a) Obtain the five-number summary of these data.

(b) Determine if there are any outliers.

(c) Boxplot the data. Comment on the plot.

4.4.4. Consider the data in Exercise 4.4.3. Obtain the normal $q-q$ plot for these data. Does the plot suggest that the underlying distribution is normal? If not, use the plot to determine a more appropriate distribution. Confirm your choice with a $q-q$ based on the quantiles using your chosen distribution.

4.4.5. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size 4 from the distribution having pdf $f(x) = e^{-x},\ 0 < x < \infty$, zero elsewhere. Find $P(Y_4 \geq 3)$.

4.4.6. Let $X_1,\ X_2,\ X_3$ be a random sample from a distribution of the continuous type having pdf $f(x) = 2x,\ 0 < x < 1$, zero elsewhere.

(a) Compute the probability that the smallest of X_1, X_2, X_3 exceeds the median of the distribution.

(b) If $Y_1 < Y_2 < Y_3$ are the order statistics, find the correlation between Y_2 and Y_3.

4.4.7. Let $f(x) = \frac{1}{6},\ x = 1, 2, 3, 4, 5, 6$, zero elsewhere, be the pmf of a distribution of the discrete type. Show that the pmf of the smallest observation of a random sample of size 5 from this distribution is

$$g_1(y_1) = \left(\frac{7 - y_1}{6}\right)^5 - \left(\frac{6 - y_1}{6}\right)^5, \quad y_1 = 1, 2, \ldots, 6,$$

zero elsewhere. Note that in this exercise the random sample is from a distribution of the discrete type. All formulas in the text were derived under the assumption that the random sample is from a distribution of the continuous type and are not applicable. Why?

4.4.8. Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ denote the order statistics of a random sample of size 5 from a distribution having pdf $f(x) = e^{-x},\ 0 < x < \infty$, zero elsewhere. Show that $Z_1 = Y_2$ and $Z_2 = Y_4 - Y_2$ are independent.
Hint: First find the joint pdf of Y_2 and Y_4.

4.4.9. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from a distribution with pdf $f(x) = 1,\ 0 < x < 1$, zero elsewhere. Show that the kth order statistic Y_k has a beta pdf with parameters $\alpha = k$ and $\beta = n - k + 1$.

4.4.10. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics from a Weibull distribution, Exercise 3.3.26. Find the distribution function and pdf of Y_1.

4.4.11. Find the probability that the range of a random sample of size 4 from the uniform distribution having the pdf $f(x) = 1,\ 0 < x < 1$, zero elsewhere, is less than $\frac{1}{2}$.

4.4.12. Let $Y_1 < Y_2 < Y_3$ be the order statistics of a random sample of size 3 from a distribution having the pdf $f(x) = 2x,\ 0 < x < 1$, zero elsewhere. Show that $Z_1 = Y_1/Y_2,\ Z_2 = Y_2/Y_3$, and $Z_3 = Y_3$ are mutually independent.

4.4.13. Suppose a random sample of size 2 is obtained from a distribution that has pdf $f(x) = 2(1-x),\ 0 < x < 1$, zero elsewhere. Compute the probability that one sample observation is at least twice as large as the other.

4.4.14. Let $Y_1 < Y_2 < Y_3$ denote the order statistics of a random sample of size 3 from a distribution with pdf $f(x) = 1,\ 0 < x < 1$, zero elsewhere. Let $Z = (Y_1 + Y_3)/2$ be the midrange of the sample. Find the pdf of Z.

4.4.15. Let $Y_1 < Y_2$ denote the order statistics of a random sample of size 2 from $N(0, \sigma^2)$.

(a) Show that $E(Y_1) = -\sigma/\sqrt{\pi}$.
Hint: Evaluate $E(Y_1)$ by using the joint pdf of Y_1 and Y_2 and first integrating on y_1.

(b) Find the covariance of Y_1 and Y_2.

4.4.16. Let $Y_1 < Y_2$ be the order statistics of a random sample of size 2 from a distribution of the continuous type which has pdf $f(x)$ such that $f(x) > 0$, provided that $x \geq 0$, and $f(x) = 0$ elsewhere. Show that the independence of $Z_1 = Y_1$ and $Z_2 = Y_2 - Y_1$ characterizes the gamma pdf $f(x)$, which has parameters $\alpha = 1$ and $\beta > 0$. That is, show that Y_1 and Y_2 are independent if and only if $f(x)$ is the pdf of a $\Gamma(1, \beta)$ distribution.
Hint: Use the change-of-variable technique to find the joint pdf of Z_1 and Z_2 from that of Y_1 and Y_2. Accept the fact that the functional equation $h(0)h(x+y) \equiv h(x)h(y)$ has the solution $h(x) = c_1 e^{c_2 x}$, where c_1 and c_2 are constants.

4.4.17. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size $n = 4$ from a distribution with pdf $f(x) = 2x,\ 0 < x < 1$, zero elsewhere.

(a) Find the joint pdf of Y_3 and Y_4.

(b) Find the conditional pdf of Y_3, given $Y_4 = y_4$.

(c) Evaluate $E(Y_3|y_4)$.

4.4.18. Two numbers are selected at random from the interval $(0, 1)$. If these values are uniformly and independently distributed, by cutting the interval at these numbers, compute the probability that the three resulting line segments can form a triangle.

4.4.19. Let X and Y denote independent random variables with respective probability density functions $f(x) = 2x,\ 0 < x < 1$, zero elsewhere, and $g(y) = 3y^2,\ 0 < y < 1$, zero elsewhere. Let $U = \min(X, Y)$ and $V = \max(X, Y)$. Find the joint pdf of U and V.
Hint: Here the two inverse transformations are given by $x = u,\ y = v$ and $x = v,\ y = u$.

4.4.20. Let the joint pdf of X and Y be $f(x, y) = \frac{12}{7}x(x+y),\ 0 < x < 1,\ 0 < y < 1$, zero elsewhere. Let $U = \min(X, Y)$ and $V = \max(X, Y)$. Find the joint pdf of U and V.

4.4.21. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution of either type. A measure of spread is *Gini's mean difference*

$$G = \sum_{j=2}^{n} \sum_{i=1}^{j-1} |X_i - X_j| / \binom{n}{2}. \tag{4.4.13}$$

(a) If $n = 10$, find $a_1, a_2, \ldots, a_{10}$ so that $G = \sum_{i=1}^{10} a_i Y_i$, where $Y_1, Y_2, \ldots, Y_{10}$ are the order statistics of the sample.

(b) Show that $E(G) = 2\sigma/\sqrt{\pi}$ if the sample arises from the normal distribution $N(\mu, \sigma^2)$.

4.4.22. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from the exponential distribution with pdf $f(x) = e^{-x},\ 0 < x < \infty$, zero elsewhere.

(a) Show that $Z_1 = nY_1,\ Z_2 = (n-1)(Y_2 - Y_1),\ Z_3 = (n-2)(Y_3 - Y_2), \ldots, Z_n = Y_n - Y_{n-1}$ are independent and that each Z_i has the exponential distribution.

(b) Demonstrate that all linear functions of $Y_1, Y_2, \ldots, Y_n$, such as $\sum_1^n a_i Y_i$, can be expressed as linear functions of independent random variables.

4.4.23. In the Program Evaluation and Review Technique (PERT), we are interested in the total time to complete a project that is comprised of a large number of subprojects. For illustration, let $X_1,\ X_2,\ X_3$ be three independent random times for three subprojects. If these subprojects are in series (the first one must be completed before the second starts, etc.), then we are interested in the sum $Y = X_1 + X_2 + X_3$. If these are in parallel (can be worked on simultaneously), then we are interested in $Z = \max(X_1,\ X_2,\ X_3)$. In the case each of these random variables has the uniform distribution with pdf $f(x) = 1,\ 0 < x < 1$, zero elsewhere, find (a) the pdf of Y and (b) the pdf of Z.

4.4.24. Let Y_n denote the nth order statistic of a random sample of size n from a distribution of the continuous type. Find the smallest value of n for which the inequality $P(\xi_{0.9} < Y_n) \geq 0.75$ is true.

4.4.25. Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ denote the order statistics of a random sample of size 5 from a distribution of the continuous type. Compute:

(a) $P(Y_1 < \xi_{0.5} < Y_5)$.

(b) $P(Y_1 < \xi_{0.25} < Y_3)$.

(c) $P(Y_4 < \xi_{0.80} < Y_5)$.

4.4.26. Compute $P(Y_3 < \xi_{0.5} < Y_7)$ if $Y_1 < \cdots < Y_9$ are the order statistics of a random sample of size 9 from a distribution of the continuous type.

4.4.27. Find the smallest value of n for which $P(Y_1 < \xi_{0.5} < Y_n) \geq 0.99$, where $Y_1 < \cdots < Y_n$ are the order statistics of a random sample of size n from a distribution of the continuous type.

4.4.28. Let $Y_1 < Y_2$ denote the order statistics of a random sample of size 2 from a distribution that is $N(\mu, \sigma^2)$, where σ^2 is known.

(a) Show that $P(Y_1 < \mu < Y_2) = \frac{1}{2}$ and compute the expected value of the random length $Y_2 - Y_1$.

(b) If $\overline{X}$ is the mean of this sample, find the constant c that solves the equation $P(\overline{X} - c\sigma < \mu < \overline{X} + c\sigma) = \frac{1}{2}$, and compare the length of this random interval with the expected value of that of part (a).

4.4.29. Let $y_1 < y_2 < y_3$ be the observed values of the order statistics of a random sample of size $n = 3$ from a continuous type distribution. Without knowing these values, a statistician is given these values in a random order, and she wants to select the largest; but once she refuses an observation, she cannot go back. Clearly, if she selects the first one, her probability of getting the largest is 1/3. Instead, she decides to use the following algorithm: She looks at the first but refuses it and then takes the second if it is larger than the first, or else she takes the third. Show that this algorithm has probability of 1/2 of selecting the largest.

4.4.30. Refer to Exercise 4.1.1. Using expression (4.4.10), obtain a confidence interval (with confidence close to 90%) for the median lifetime of a motor. What does the interval mean?

4.4.31. Let $Y_1 < Y_2 < \cdots < Y_n$ denote the order statistics of a random sample of size n from a distribution that has pdf $f(x) = 3x^2/\theta^3$, $0 < x < \theta$, zero elsewhere.

(a) Show that $P(c < Y_n/\theta < 1) = 1 - c^{3n}$, where $0 < c < 1$.

(b) If n is 4 and if the observed value of Y_4 is 2.3, what is a 95% confidence interval for θ?

4.4.32. Reconsider the weight of professional baseball players in the data file `bb.rda`. Obtain comparison boxplots of the weights of the hitters and pitchers (use the R code `boxplot(x,y)` where `x` and `y` contain the weights of the hitters and pitchers, respectively). Then obtain 95% confidence intervals for the median weights of the hitters and pitchers (use the R function `onesampsgn`). Comment.

4.5 Introduction to Hypothesis Testing

Point estimation and confidence intervals are useful statistical inference procedures. Another type of inference that is frequently used concerns tests of hypotheses. As in Sections 4.1 through 4.3, suppose our interest centers on a random variable X that has density function $f(x;\theta)$, where $\theta \in \Omega$. Suppose we think, due to theory or a preliminary experiment, that $\theta \in \omega_0$ or $\theta \in \omega_1$, where ω_0 and ω_1 are disjoint subsets of Ω and $\omega_0 \cup \omega_1 = \Omega$. We label these hypotheses as

$$H_0 : \theta \in \omega_0 \text{ versus } H_1 : \theta \in \omega_1. \tag{4.5.1}$$

The hypothesis H_0 is referred to as the **null hypothesis**, while H_1 is referred to as the **alternative hypothesis**. Often the null hypothesis represents no change or no difference from the past, while the alternative represents change or difference. The alternative is often referred to as the research worker's hypothesis. The decision rule to take H_0 or H_1 is based on a sample $X_1, \ldots, X_n$ from the distribution of X and, hence, the decision could be wrong. For instance, we could decide that $\theta \in \omega_1$ when really $\theta \in \omega_0$ or we could decide that $\theta \in \omega_0$ when, in fact, $\theta \in \omega_1$. We label these errors Type I and Type II errors, respectively, later in this section. As we show in Chapter 8, a careful analysis of these errors can lead in certain situations to optimal decision rules. In this section, though, we simply want to introduce the elements of hypothesis testing. To set ideas, consider the following example.

Example 4.5.1 (*Zea mays* Data)**.** In 1878 Charles Darwin recorded some data on the heights of *Zea mays* plants to determine what effect cross-fertilization or self-fertilization had on the height of *Zea mays*. The experiment was to select one cross-fertilized plant and one self-fertilized plant, grow them in the same pot, and then later measure their heights. An interesting hypothesis for this example would be that the cross-fertilized plants are generally taller than the self-fertilized plants. This is the alternative hypothesis, i.e., the research worker's hypothesis. The null hypothesis is that the plants generally grow to the same height regardless of whether they were self- or cross-fertilized. Data for 15 pots were recorded.

We represent the data as $(Y_1, Z_1), \ldots, (Y_{15}, Z_{15})$, where Y_i and Z_i are the heights of the cross-fertilized and self-fertilized plants, respectively, in the ith pot. Let $X_i = Y_i - Z_i$. Due to growing in the same pot, Y_i and Z_i may be dependent random variables, but it seems appropriate to assume independence between pots, i.e., independence between the paired random vectors. So we assume that $X_1, \ldots, X_{15}$ form a random sample. As a tentative model, consider the location model

$$X_i = \mu + e_i, \quad i = 1, \ldots, 15,$$

where the random variables e_i are iid with continuous density $f(x)$. For this model, there is no loss in generality in assuming that the mean of e_i is 0, for, otherwise, we can simply redefine μ. Hence, $E(X_i) = \mu$. Further, the density of X_i is $f_X(x;\mu) = f(x-\mu)$. In practice, the goodness of the model is always a concern and diagnostics based on the data would be run to confirm the quality of the model.

If $\mu = E(X_i) = 0$, then $E(Y_i) = E(Z_i)$; i.e., on average, the cross-fertilized plants grow to the same height as the self-fertilized plants. While, if $\mu > 0$ then

Table 4.5.1: 2×2 Decision Table for a Hypothesis Test

	True State of Nature	
Decision	H_0 is True	H_1 is True
Reject H_0	Type I Error	Correct Decision
Accept H_0	Correct Decision	Type II Error

$E(Y_i) > E(Z_i)$; i.e., on average the cross-fertilized plants are taller than the self-fertilized plants. Under this model, our hypotheses are

$$H_0 : \mu = 0 \text{ versus } H_1 : \mu > 0. \tag{4.5.2}$$

Hence, $\omega_0 = \{0\}$ represents no difference in the treatments, while $\omega_1 = (0, \infty)$ represents that the mean height of cross-fertilized *Zea mays* exceeds the mean height of self-fertilized *Zea mays*. ■

To complete the testing structure for the general problem described at the beginning of this section, we need to discuss decision rules. Recall that $X_1, \ldots, X_n$ is a random sample from the distribution of a random variable X that has density $f(x; \theta)$, where $\theta \in \Omega$. Consider testing the hypotheses $H_0 : \theta \in \omega_0$ versus $H_1 : \theta \in \omega_1$, where $\omega_0 \cup \omega_1 = \Omega$. Denote the space of the sample by $\mathcal{D}$; that is, $\mathcal{D} = \text{space}\{(X_1, \ldots, X_n)\}$. A **test** of H_0 versus H_1 is based on a subset C of $\mathcal{D}$. This set C is called the **critical region** and its corresponding decision rule (test) is

$$\begin{aligned} &\text{Reject } H_0 \text{ (Accept } H_1) && \text{if } (X_1, \ldots, X_n) \in C \\ &\text{Retain } H_0 \text{ (Reject } H_1) && \text{if } (X_1, \ldots, X_n) \in C^c. \end{aligned} \tag{4.5.3}$$

For a given critical region, the 2×2 decision table as shown in Table 4.5.1, summarizes the results of the hypothesis test in terms of the true state of nature. Besides the correct decisions, two errors can occur. A **Type I** error occurs if H_0 is rejected when it is true, while a **Type II** error occurs if H_0 is accepted when H_1 is true.

The goal, of course, is to select a critical region from all possible critical regions which minimizes the probabilities of these errors. In general, this is not possible. The probabilities of these errors often have a seesaw effect. This can be seen immediately in an extreme case. Simply let $C = \phi$. With this critical region, we would never reject H_0, so the probability of Type I error would be 0, but the probability of Type II error is 1. Often we consider Type I error to be the worse of the two errors. We then proceed by selecting critical regions that bound the probability of Type I error and then among these critical regions we try to select one that minimizes the probability of Type II error.

Definition 4.5.1. *We say a critical region C is of* **size** α *if*

$$\alpha = \max_{\theta \in \omega_0} P_\theta[(X_1, \ldots, X_n) \in C]. \tag{4.5.4}$$

Over all critical regions of size α, we want to consider critical regions that have lower probabilities of Type II error. We also can look at the complement of a Type II error, namely, rejecting H_0 when H_1 is true, which is a correct decision, as marked in Table 4.5.1. Since we desire to maximize the probability of this latter decision, we want the probability of it to be as large as possible. That is, for $\theta \in \omega_1$, we want to maximize

$$1 - P_\theta[\text{Type II Error}] = P_\theta[(X_1, \ldots, X_n) \in C].$$

The probability on the right side of this equation is called the **power** of the test at θ. It is the probability that the test detects the alternative θ when $\theta \in \omega_1$ is the true parameter. So minimizing the probability of Type II error is equivalent to maximizing power.

We define the **power function** of a critical region to be

$$\gamma_C(\theta) = P_\theta[(X_1, \ldots, X_n) \in C]; \quad \theta \in \omega_1. \tag{4.5.5}$$

Hence, given two critical regions C_1 and C_2, which are both of size α, C_1 is better than C_2 if $\gamma_{C_1}(\theta) \geq \gamma_{C_2}(\theta)$ for all $\theta \in \omega_1$. In Chapter 8, we obtain optimal critical regions for specific situations. In this section, we want to illustrate these concepts of hypothesis testing with several examples.

Example 4.5.2 (Test for a Binomial Proportion of Success)**.** Let X be a Bernoulli random variable with probability of success p. Suppose we want to test, at size α,

$$H_0 : \ p = p_0 \text{ versus } H_1 : \ p < p_0, \tag{4.5.6}$$

where p_0 is specified. As an illustration, suppose "success" is dying from a certain disease and p_0 is the probability of dying with some standard treatment. A new treatment is used on several (randomly chosen) patients, and it is hoped that the probability of dying under this new treatment is less than p_0. Let $X_1, \ldots, X_n$ be a random sample from the distribution of X and let $S = \sum_{i=1}^n X_i$ be the total number of successes in the sample. An intuitive decision rule (critical region) is

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } S \leq k, \tag{4.5.7}$$

where k is such that $\alpha = P_{H_0}[S \leq k]$. Since S has a $b(n, p_0)$ distribution under H_0, k is determined by $\alpha = P_{p_0}[S \leq k]$. Because the binomial distribution is discrete, however, it is likely that there is no integer k that solves this equation. For example, suppose $n = 20$, $p_0 = 0.7$, and $\alpha = 0.15$. Then under H_0, S has a binomial $b(20, 0.7)$ distribution. Hence, computationally, $P_{H_0}[S \leq 11] =$`pbinom(11,20,0.7)`$= 0.1133$ and $P_{H_0}[S \leq 12] =$`pbinom(12,20,0.7)`$= 0.2277$. Hence, erring on the conservative side, we would probably choose k to be 11 and $\alpha = 0.1133$. As n increases, this is less of a problem; see, also, the later discussion on p-values. In general, the power of the test for the hypotheses (4.5.6) is

$$\gamma(p) = P_p[S \leq k] \,, \quad p < p_0. \tag{4.5.8}$$

The curve labeled Test 1 in Figure 4.5.1 is the power function for the case $n = 20$, $p_0 = 0.7$, and $\alpha = 0.1133$. Notice that the power function is decreasing. The

power is higher to detect the alternative $p = 0.2$ than $p = 0.6$. In Section 8.2, we prove in general the monotonicity of the power function for binomial tests of these hypotheses. Using this monotonicity, we extend our test to the more general null hypothesis $H_0 : p \geq p_0$ rather than simply $H_0 : p = p_0$. Using the same decision rule as we used for the hypotheses (4.5.6), the definition of the size of a test (4.5.4), and the monotonicity of the power curve, we have

$$\max_{p \geq p_0} P_p[S \leq k] = P_{p_0}[S \leq k] = \alpha,$$

i.e., the same size as for the original null hypothesis.

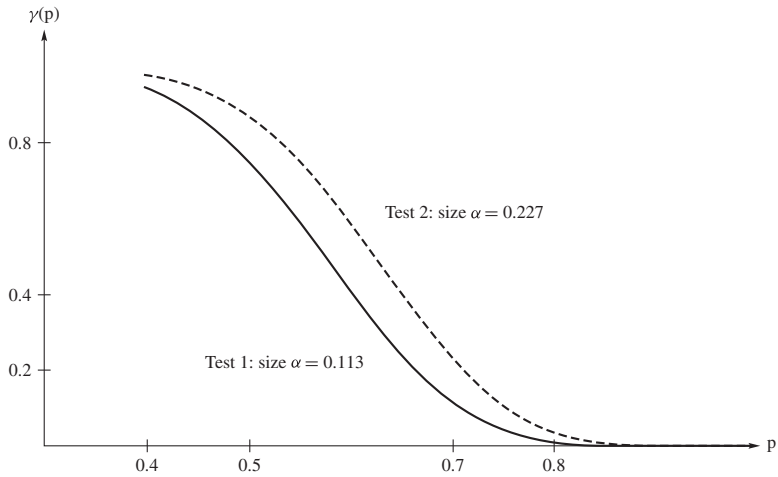

Figure 4.5.1: Power curves for tests 1 and 2; see Example 4.5.2.

Denote by Test 1 the test for the situation with $n = 20$, $p_0 = 0.70$, and size $\alpha = 0.1133$. Suppose we have a second test (Test 2) with an increased size. How does the power function of Test 2 compare to Test 1? As an example, suppose for Test 2, we select $\alpha = 0.2277$. Hence, for Test 2, we reject H_0 if $S \leq 12$. Figure 4.5.1 displays the resulting power function. Note that while Test 2 has a higher probability of committing a Type I error, it also has a higher power at each alternative $p < 0.7$. Exercise 4.5.7 shows that this is true for these binomial tests. It is true in general; that is, if the size of the test increases, power does too. For this example, the R function `binpower.r`, found at the site listed in the Preface, produces a version of Figure 4.5.1. ■

Remark 4.5.1 (Nomenclature)**.** Since in Example 4.5.2, the first null hypothesis $H_0 : p = p_0$ completely specifies the underlying distribution, it is called a **simple** hypothesis. Most hypotheses, such as $H_1 : p < p_0$, are **composite** hypotheses, because they are composed of many simple hypotheses and, hence, do not completely specify the distribution.

As we study more and more statistics, we discover that often other names are used for the size, α, of the critical region. Frequently, α is also called the **signifi-**

cance level of the test associated with that critical region. Moreover, sometimes α is called the "maximum of probabilities of committing an error of Type I" and the "maximum of the power of the test when H_0 is true." It is disconcerting to the student to discover that there are so many names for the same thing. However, all of them are used in the statistical literature, and we feel obligated to point out this fact. ■

The test in the last example is based on the exact distribution of its test statistic, i.e., the binomial distribution. Often we cannot obtain the distribution of the test statistic in closed form. As with approximate confidence intervals, however, we can frequently appeal to the Central Limit Theorem to obtain an approximate test; see Theorem 4.2.1. Such is the case for the next example.

Example 4.5.3 (Large Sample Test for the Mean)**.** Let X be a random variable with mean μ and finite variance σ^2. We want to test the hypotheses

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu > \mu_0, \tag{4.5.9}$$

where μ_0 is specified. To illustrate, suppose μ_0 is the mean level on a standardized test of students who have been taught a course by a standard method of teaching. Suppose it is hoped that a new method that incorporates computers has a mean level $\mu > \mu_0$, where $\mu = E(X)$ and X is the score of a student taught by the new method. This conjecture is tested by having n students (randomly selected) taught under this new method.

Let $X_1, \ldots, X_n$ be a random sample from the distribution of X and denote the sample mean and variance by $\overline{X}$ and S^2, respectively. Because $\overline{X}$ is an unbiased estimate of μ, an intuitive decision rule is given by

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } \overline{X} \text{ is much larger than } \mu_0. \tag{4.5.10}$$

In general, the distribution of the sample mean cannot be obtained in closed form. In Example 4.5.4, under the strong assumption of normality for the distribution of X, we obtain an exact test. For now, the Central Limit Theorem (Theorem 4.2.1) shows that the distribution of $(\overline{X} - \mu)/(S/\sqrt{n})$ is approximately $N(0, 1)$. Using this, we obtain a test with an approximate size α, with the decision rule

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } \tfrac{\overline{X}-\mu_0}{S/\sqrt{n}} \geq z_\alpha. \tag{4.5.11}$$

The test is intuitive. To reject H_0, $\overline{X}$ must exceed μ_0 by at least $z_\alpha S/\sqrt{n}$. To approximate the power function of the test, we use the Central Limit Theorem. Upon substituting σ for S, it readily follows that the approximate power function is

$$\begin{aligned}
\gamma(\mu) &= P_\mu(\overline{X} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}) \\
&= P_\mu\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) \\
&\approx 1 - \Phi\left(z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) \\
&= \Phi\left(-z_\alpha - \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right).
\end{aligned} \tag{4.5.12}$$

So if we have some reasonable idea of what σ equals, we can compute the approximate power function. As Exercise 4.5.1 shows, this approximate power function is strictly increasing in μ, so as in the last example, we can change the null hypotheses to

$$H_0 : \mu \leq \mu_0 \text{ versus } H_1 : \mu > \mu_0. \tag{4.5.13}$$

Our asymptotic test has approximate size α for these hypotheses. ■

Example 4.5.4 (Test for μ Under Normality)**.** Let X have a $N(\mu, \sigma^2)$ distribution. As in Example 4.5.3, consider the hypotheses

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu > \mu_0, \tag{4.5.14}$$

where μ_0 is specified. Assume that the desired size of the test is α, for $0 < \alpha < 1$, Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution. Let $\overline{X}$ and S^2 denote the sample mean and variance, respectively. Our intuitive rejection rule is to reject H_0 in favor of H_1 if $\overline{X}$ is much larger than μ_0. Unlike Example 4.5.3, we now know the distribution of the statistic $\overline{X}$. In particular, by Part (d) of Theorem 3.6.1, under H_0 the statistic $T = (\overline{X} - \mu_0)/(S/\sqrt{n})$ has a t-distribution with $n-1$ degrees of freedom. Using the distribution of T, it follows that this rejection rule has exact level α:

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq t_{\alpha, n-1}, \tag{4.5.15}$$

where $t_{\alpha,n-1}$ is the upper α critical point of a t-distribution with $n-1$ degrees of freedom; i.e., $\alpha = P(T > t_{\alpha,n-1})$. This is often called the ***t*-test** of $H_0 : \mu = \mu_0$.

Note the differences between this rejection rule and the large sample rule, (4.5.11). The large sample rule has approximate level α, while this has exact level α. Of course, we now have to assume that X has a normal distribution. In practice, we may not be willing to assume that the population is normal. Usually t-critical values are larger than z-critical values; hence, the t-test is conservative relative to the large sample test. So, in practice, many statisticians often use the t-test.

The R code `t.test(x,mu=mu0,alt="greater")` computes the t-test for the hypotheses (4.5.14), where the R vector `x` contains the sample. ■

Example 4.5.5 (Example 4.5.1, Continued)**.** The data for Darwin's experiment on *Zea mays* are recorded in Table 4.5.2 and are, also, in the file `darwin.rda`. A boxplot and a normal $q-q$ plot of the 15 differences, $x_i = y_i - z_i$, are found in Figure 4.5.2. Based on these plots, we can see that there seem to be two outliers, Pots 2 and 15. In these two pots, the self-fertilized *Zea mays* are much taller than their cross-fertilized pairs. Except for these two outliers, the differences, $y_i - z_i$, are positive, indicating that the cross-fertilization leads to taller plants. We proceed to conduct a test of hypotheses (4.5.2), as discussed in Example 4.5.4. We use the decision rule given by (4.5.15) with $\alpha = 0.05$. As Exercise 4.5.2 shows, the values of the sample mean and standard deviation for the differences, x_i, are $\overline{x} = 2.62$ and $s_x = 4.72$. Hence the t-test statistic is 2.15, which exceeds the t-critical value, $t_{.05,14}$ =`qt(0.95,14)`= 1.76. Thus we reject H_0 and conclude that cross-fertilized *Zea mays* are on the average taller than self-fertilized *Zea mays*. Because of the

Table 4.5.2: Plant Growth

Pot	1	2	3	4	5	6	7	8
Cross	23.500	12.000	21.000	22.000	19.125	21.500	22.125	20.375
Self	17.375	20.375	20.000	20.000	18.375	18.625	18.625	15.250
Pot	9	10	11	12	13	14	15	
Cross	18.250	21.625	23.250	21.000	22.125	23.000	12.000	
Self	16.500	18.000	16.250	18.000	12.750	15.500	18.000	

outliers, normality of the error distribution is somewhat dubious, and we use the test in a conservative manner, as discussed at the end of Example 4.5.4.

Assuming that the rda file `darwin.rda` has been loaded in R, the code for the above t-test is `t.test(cross-self,mu=0,alt="greater")` which evaluates the t-test statistic to be 2.1506. ■

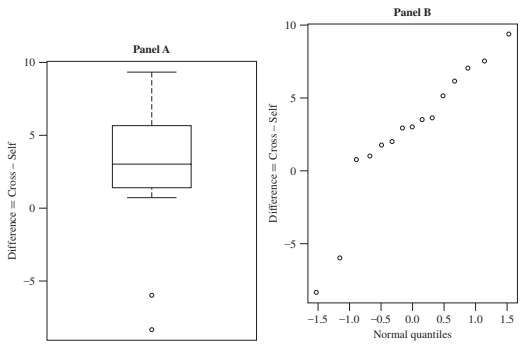

Figure 4.5.2: Boxplot and normal $q-q$ plot for the data of Example 4.5.5.

EXERCISES

In many of these exercises, use R or another statistical package for computations and graphs of power functions.

4.5.1. Show that the approximate power function given in expression (4.5.12) of Example 4.5.3 is a strictly increasing function of μ. Show then that the test discussed in this example has approximate size α for testing

$$H_0 : \mu \leq \mu_0 \text{ versus } H_1 : \mu > \mu_0.$$

4.5.2. For the Darwin data tabled in Example 4.5.5, verify that the Student t-test statistic is 2.15.

4.5.3. Let X have a pdf of the form $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, zero elsewhere, where $\theta \in \{\theta : \theta = 1, 2\}$. To test the simple hypothesis $H_0 : \theta = 1$ against the alternative simple hypothesis $H_1 : \theta = 2$, use a random sample X_1, X_2 of size $n = 2$

and define the critical region to be $C = \{(x_1, x_2) : \frac{3}{4} \leq x_1 x_2\}$. Find the power function of the test.

4.5.4. Let X have a binomial distribution with the number of trials $n = 10$ and with p either $1/4$ or $1/2$. The simple hypothesis $H_0 : p = \frac{1}{2}$ is rejected, and the alternative simple hypothesis $H_1 : p = \frac{1}{4}$ is accepted, if the observed value of X_1, a random sample of size 1, is less than or equal to 3. Find the significance level and the power of the test.

4.5.5. Let X_1, X_2 be a random sample of size $n = 2$ from the distribution having pdf $f(x;\theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, zero elsewhere. We reject $H_0 : \theta = 2$ and accept $H_1 : \theta = 1$ if the observed values of X_1, X_2, say x_1, x_2, are such that

$$\frac{f(x_1;2)f(x_2;2)}{f(x_1;1)f(x_2;1)} \leq \frac{1}{2}.$$

Here $\Omega = \{\theta : \theta = 1, 2\}$. Find the significance level of the test and the power of the test when H_0 is false.

4.5.6. Consider the tests Test 1 and Test 2 for the situation discussed in Example 4.5.2. Consider the test that rejects H_0 if $S \leq 10$. Find the level of significance for this test and sketch its power curve as in Figure 4.5.1.

4.5.7. Consider the situation described in Example 4.5.2. Suppose we have two tests A and B defined as follows. For Test A, H_0 is rejected if $S \leq k_A$, while for Test B, H_0 is rejected if $S \leq k_B$. If Test A has a higher level of significance than Test B, show that Test A has higher power than Test B at each alternative.

4.5.8. Let us say the life of a tire in miles, say X, is normally distributed with mean θ and standard deviation 5000. Past experience indicates that $\theta = 30{,}000$. The manufacturer claims that the tires made by a new process have mean $\theta > 30{,}000$. It is possible that $\theta = 35{,}000$. Check his claim by testing $H_0 : \theta = 30{,}000$ against $H_1 : \theta > 30{,}000$. We observe n independent values of X, say $x_1, \ldots, x_n$, and we reject H_0 (thus accept H_1) if and only if $\overline{x} \geq c$. Determine n and c so that the power function $\gamma(\theta)$ of the test has the values $\gamma(30{,}000) = 0.01$ and $\gamma(35{,}000) = 0.98$.

4.5.9. Let X have a Poisson distribution with mean θ. Consider the simple hypothesis $H_0 : \theta = \frac{1}{2}$ and the alternative composite hypothesis $H_1 : \theta < \frac{1}{2}$. Thus $\Omega = \{\theta : 0 < \theta \leq \frac{1}{2}\}$. Let $X_1, \ldots, X_{12}$ denote a random sample of size 12 from this distribution. We reject H_0 if and only if the observed value of $Y = X_1 + \cdots + X_{12} \leq 2$. Show that the following R code graphs the power function of this test:

```
theta=seq(.1,.5,.05); gam=ppois(2,theta*12)
plot(gam~theta,pch=" ",xlab=expression(theta),ylab=expression(gamma))
lines(gam~theta)
```

Run the code. Determine the significance level from the plot.

4.5.10. Let Y have a binomial distribution with parameters n and p. We reject $H_0 : p = \frac{1}{2}$ and accept $H_1 : p > \frac{1}{2}$ if $Y \geq c$. Find n and c to give a power function $\gamma(p)$ which is such that $\gamma(\frac{1}{2}) = 0.10$ and $\gamma(\frac{2}{3}) = 0.95$, approximately.

4.5.11. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size $n = 4$ from a distribution with pdf $f(x;\theta) = 1/\theta,\ 0 < x < \theta$, zero elsewhere, where $0 < \theta$. The hypothesis $H_0 : \theta = 1$ is rejected and $H_1 : \theta > 1$ is accepted if the observed $Y_4 \geq c$.

(a) Find the constant c so that the significance level is $\alpha = 0.05$.

(b) Determine the power function of the test.

4.5.12. Let $X_1, X_2, \ldots, X_8$ be a random sample of size $n = 8$ from a Poisson distribution with mean μ. Reject the simple null hypothesis $H_0 : \mu = 0.5$ and accept $H_1 : \mu > 0.5$ if the observed sum $\sum_{i=1}^{8} x_i \geq 8$.

(a) Show that the significance level is `1-ppois(7,8*.5)`.

(b) Use R to determine $\gamma(0.75),\ \gamma(1)$, and $\gamma(1.25)$.

(c) Modify the code in Exercise 4.5.9 to obtain a plot of the power function.

4.5.13. Let p denote the probability that, for a particular tennis player, the first serve is good. Since $p = 0.40$, this player decided to take lessons in order to increase p. When the lessons are completed, the hypothesis $H_0 : p = 0.40$ is tested against $H_1 : p > 0.40$ based on $n = 25$ trials. Let Y equal the number of first serves that are good, and let the critical region be defined by $C = \{Y : Y \geq 13\}$.

(a) Show that α is computed by $\alpha =$`1-pbinom(12,25,.4)`.

(b) Find $\beta = P(Y < 13)$ when $p = 0.60$; that is, $\beta = P(Y \leq 12;\ p = 0.60)$ so that $1 - \beta$ is the power at $p = 0.60$.

4.5.14. Let S denote the number of success in $n = 40$ Bernoulli trials with probability of success p. Consider the hypotheses: $H_0 : p \leq 0.3$ versus $H_1 : p > 0.3$. Consider the two tests: (1) Reject H_0 if $S \geq 16$ and (2) Reject H_0 if $S \geq 17$. Determine the level of these tests. The R function `binpower.r` produces a version of Figure 4.5.1. For this exercise, write a similar R function that graphs the power functions of the above two tests.

4.6 Additional Comments About Statistical Tests

All of the alternative hypotheses considered in Section 4.5 were **one-sided hypotheses**. For illustration, in Exercise 4.5.8 we tested $H_0 : \mu = 30{,}000$ against the one-sided alternative $H_1 : \mu > 30{,}000$, where μ is the mean of a normal distribution having standard deviation $\sigma = 5000$. Perhaps in this situation, though, we think the manufacturer's process has changed but are unsure of the direction. That is, we are interested in the alternative $H_1 :\ \mu \neq 30{,}000$. In this section, we further explore hypotheses testing and we begin with the construction of a test for a two-sided alternative.

Example 4.6.1 (Large Sample Two-Sided Test for the Mean)**.** In order to see how to construct a test for a two-sided alternative, reconsider Example 4.5.3, where we constructed a large sample one-sided test for the mean of a random variable. As in Example 4.5.3, let X be a random variable with mean μ and finite variance σ^2. Here, though, we want to test

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0, \tag{4.6.1}$$

where μ_0 is specified. Let $X_1, \ldots, X_n$ be a random sample from the distribution of X and denote the sample mean and variance by $\overline{X}$ and S^2, respectively. For the one-sided test, we rejected H_0 if $\overline{X}$ was too large; hence, for the hypotheses (4.6.1), we use the decision rule

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } \overline{X} \leq h \text{ or } \overline{X} \geq k, \tag{4.6.2}$$

where h and k are such that $\alpha = P_{H_0}[\overline{X} \leq h \text{ or } \overline{X} \geq k]$. Clearly, $h < k$; hence, we have

$$\alpha = P_{H_0}[\overline{X} \leq h \text{ or } \overline{X} \geq k] = P_{H_0}[\overline{X} \leq h] + P_{H_0}[\overline{X} \geq k].$$

Since, at least for large samples, the distribution of $\overline{X}$ is symmetrically distributed about μ_0, under H_0, an intuitive rule is to divide α equally between the two terms on the right side of the above expression; that is, h and k are chosen by

$$P_{H_0}[\overline{X} \leq h] = \alpha/2 \text{ and } P_{H_0}[\overline{X} \geq k] = \alpha/2. \tag{4.6.3}$$

From Theorem 4.2.1, it follows that $(\overline{X} - \mu_0)/(S/\sqrt{n})$ is approximately $N(0,1)$. This and (4.6.3) lead to the approximate decision rule

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } \left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right| \geq z_{\alpha/2}. \tag{4.6.4}$$

Upon substituting σ for S, it readily follows that the approximate power function is

$$\begin{aligned} \gamma(\mu) &= P_\mu(\overline{X} \leq \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}) + P_\mu(\overline{X} \geq \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}) \\ &= \Phi\left(\frac{\sqrt{n}(\mu_0-\mu)}{\sigma} - z_{\alpha/2}\right) + 1 - \Phi\left(\frac{\sqrt{n}(\mu_0-\mu)}{\sigma} + z_{\alpha/2}\right), \end{aligned} \tag{4.6.5}$$

where $\Phi(z)$ is the cdf of a standard normal random variable; see (3.4.9). So if we have some reasonable idea of what σ equals, we can compute the approximate power function. Note that the derivative of the power function is

$$\gamma'(\mu) = \frac{\sqrt{n}}{\sigma}\left[\phi\left(\frac{\sqrt{n}(\mu_0-\mu)}{\sigma} + z_{\alpha/2}\right) - \phi\left(\frac{\sqrt{n}(\mu_0-\mu)}{\sigma} - z_{\alpha/2}\right)\right], \tag{4.6.6}$$

where $\phi(z)$ is the pdf of a standard normal random variable. Then we can show that $\gamma(\mu)$ has a critical value at μ_0 which is the minimum; see Exercise 4.6.2. Further, $\gamma(\mu)$ is strictly decreasing for $\mu < \mu_0$ and strictly increasing for $\mu > \mu_0$. ■

Consider again the situation at the beginning of this section. Suppose we want to test

$$H_0 : \mu = 30,000 \text{ versus } H_1 : \mu \neq 30,000. \tag{4.6.7}$$

Suppose $n = 20$ and $\alpha = 0.01$. Then the rejection rule (4.6.4) becomes

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } \left| \frac{\overline{X} - 30,000}{S/\sqrt{20}} \right| \geq 2.575. \tag{4.6.8}$$

Figure 4.6.1 displays the power curve for this test when $\sigma = 5000$ is substituted in for S. For comparison, the power curve for the test with level $\alpha = 0.05$ is also shown. The R function `zpower` computes a version of this figure.

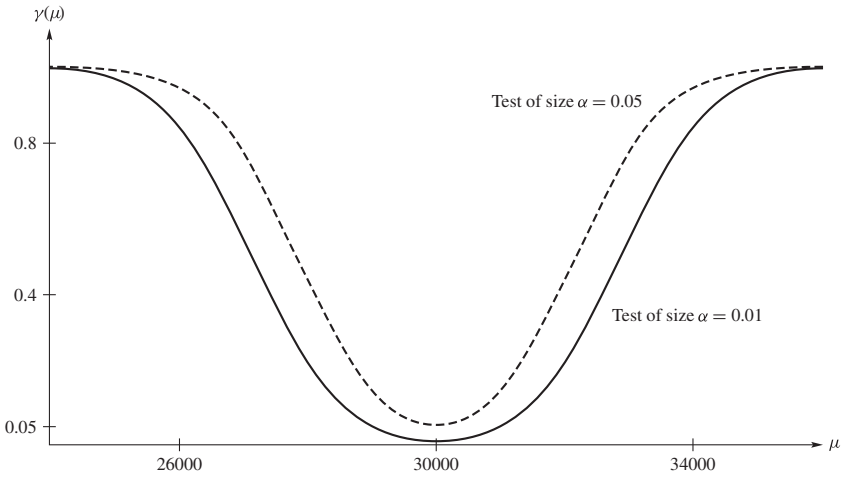

Figure 4.6.1: Power curves for the tests of the hypotheses (4.6.7).

This two-sided test for the mean is approximate. If we assume that X has a normal distribution, then, as Exercise 4.6.3 shows, the following test has exact size α for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$:

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } \left| \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \right| \geq t_{\alpha/2,n-1}. \tag{4.6.9}$$

It too has a bowl-shaped power curve similar to Figure 4.6.1, although it is not as easy to show; see Lehmann (1986).

For computation in R, the code `t.test(x,mu=mu0)` obtains the two-sided t-test of hypotheses (4.6.1), when the R vector `x` contains the sample.

There exists a relationship between two-sided tests and confidence intervals. Consider the two-sided t-test (4.6.9). Here, we use the rejection rule with "if and only if" replacing "if." Hence, in terms of acceptance, we have

$$\text{Accept } H_0 \text{ if and only if } \mu_0 - t_{\alpha/2,n-1} S/\sqrt{n} < \overline{X} < \mu_0 + t_{\alpha/2,n-1} S/\sqrt{n}.$$

But this is easily shown to be

$$\text{Accept } H_0 \text{ if and only if } \mu_0 \in (\overline{X} - t_{\alpha/2,n-1} S/\sqrt{n}, \overline{X} + t_{\alpha/2,n-1} S/\sqrt{n}); \tag{4.6.10}$$

that is, we accept H_0 at significance level α if and only if μ_0 is in the $(1-\alpha)100\%$ confidence interval for μ. Equivalently, we reject H_0 at significance level α if and only if μ_0 is not in the $(1-\alpha)100\%$ confidence interval for μ. This is true for all the two-sided tests and hypotheses discussed in this text. There is also a similar relationship between one-sided tests and one-sided confidence intervals.

Once we recognize this relationship between confidence intervals and tests of hypothesis, we can use all those statistics that we used to construct confidence intervals to test hypotheses, not only against two-sided alternatives but one-sided ones as well. Without listing all of these in a table, we present enough of them so that the principle can be understood.

Example 4.6.2. Let independent random samples be taken from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. Say these have the respective sample characteristics n_1, $\overline{X}$, S_1^2 and n_2, $\overline{Y}$, S_2^2. Let $n = n_1 + n_2$ denote the combined sample size and let $S_p^2 = [(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n-2)$, (4.2.11), be the pooled estimator of the common variance. At $\alpha = 0.05$, reject $H_0 : \mu_1 = \mu_2$ and accept the one-sided alternative $H_1 : \mu_1 > \mu_2$ if

$$T = \frac{\overline{X} - \overline{Y} - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{.05,n-2},$$

because, under $H_0 : \mu_1 = \mu_2$, T has a $t(n-2)$-distribution. A rigorous development of this test is given in Example 8.3.1. ■

Example 4.6.3. Say X is $b(1, p)$. Consider testing $H_0 : p = p_0$ against $H_1 : p < p_0$. Let $X_1 \ldots, X_n$ be a random sample from the distribution of X and let $\widehat{p} = \overline{X}$. To test H_0 versus H_1, we use either

$$Z_1 = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \leq c \quad \text{or} \quad Z_2 = \frac{\widehat{p} - p_0}{\sqrt{\widehat{p}(1-\widehat{p})/n}} \leq c.$$

If n is large, both Z_1 and Z_2 have approximate standard normal distributions provided that $H_0 : p = p_0$ is true. Hence, if c is set at -1.645, then the approximate significance level is $\alpha = 0.05$. Some statisticians use Z_1 and others Z_2. We do not have strong preferences one way or the other because the two methods provide about the same numerical results. As one might suspect, using Z_1 provides better probabilities for power calculations if the true p is close to p_0, while Z_2 is better if H_0 is clearly false. However, with a two-sided alternative hypothesis, Z_2 does provide a better relationship with the confidence interval for p. That is, $|Z_2| < z_{\alpha/2}$ is equivalent to p_0 being in the interval from

$$\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \quad \text{to} \quad \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

which is the interval that provides a $(1-\alpha)100\%$ approximate confidence interval for p as considered in Section 4.2. ■

In closing this section, we introduce the concept of **randomized tests**.

Example 4.6.4. Let $X_1, X_2, \ldots, X_{10}$ be a random sample of size $n = 10$ from a Poisson distribution with mean θ. A critical region for testing $H_0 : \theta = 0.1$ against $H_1 : \theta > 0.1$ is given by $Y = \sum_1^{10} X_i \geq 3$. The statistic Y has a Poisson distribution with mean 10θ. Thus, with $\theta = 0.1$ so that the mean of Y is 1, the significance level of the test is

$$P(Y \geq 3) = 1 - P(Y \leq 2) = 1 - \text{ppois(2,1)} = 1 - 0.920 = 0.080.$$

If, on the other hand, the critical region defined by $\sum_1^{10} x_i \geq 4$ is used, the significance level is

$$\alpha = P(Y \geq 4) = 1 - P(Y \leq 3) = 1 - \text{ppois(3,1)} = 1 - 0.981 = 0.019.$$

For instance, if a significance level of about $\alpha = 0.05$, say, is desired, most statisticians would use one of these tests; that is, they would adjust the significance level to that of one of these convenient tests. However, a significance level of $\alpha = 0.05$ can be achieved in the following way. Let W have a Bernoulli distribution with probability of success equal to

$$P(W = 1) = \frac{0.050 - 0.019}{0.080 - 0.019} = \frac{31}{61}.$$

Assume that W is selected independently of the sample. Consider the rejection rule

$$\text{Reject } H_0 \text{ if } \sum_1^{10} x_i \geq 4 \text{ or if } \sum_1^{10} x_i = 3 \text{ and } W = 1.$$

The significance level of this rule is

$$\begin{aligned} P_{H_0}(Y \geq 4) + P_{H_0}(\{Y = 3\} \cap \{W = 1\}) &= P_{H_0}(Y \geq 4) \\ &\quad + P_{H_0}(Y = 3)P(W = 1) \\ &= 0.019 + 0.061\frac{31}{61} = 0.05; \end{aligned}$$

hence, the decision rule has exactly level 0.05. The process of performing the auxiliary experiment to decide whether to reject or not when $Y = 3$ is sometimes referred to as a **randomized test**. ■

4.6.1 Observed Significance Level, p-value

Not many statisticians like randomized tests in practice, because the use of them means that two statisticians could make the same assumptions, observe the same data, apply the same test, and yet make different decisions. Hence, they usually adjust their significance level so as not to randomize. As a matter of fact, many statisticians report what are commonly called **observed significance levels** or p**-values** (for *probability values*).

A general example suffices to explain observed significance levels. Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where both μ and σ^2 are unknown.

Consider, first, the one-sided hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$, where μ_0 is specified. Write the rejection rule as

$$\text{Reject } H_0 \text{ in favor of } H_1, \text{ if } \overline{X} \geq k, \tag{4.6.11}$$

where $\overline{X}$ is the sample mean. Previously we have specified a level and then solved for k. In practice, though, the level is not specified. Instead, once the sample is observed, the realized value $\overline{x}$ of $\overline{X}$ is computed and we ask the question: Is $\overline{x}$ sufficiently large to reject H_0 in favor of H_1? To answer this we calculate the p-value which is the probability,

$$p\text{-value} = P_{H_0}(\overline{X} \geq \overline{x}). \tag{4.6.12}$$

Note that this is a data-based "significance level" and we call it the **observed significance level** or the p-value. The hypothesis H_0 is rejected at all levels greater than or equal to the p-value. For example, if the p-value is 0.048, and the nominal α level is 0.05 then H_0 would be rejected; however, if the nominal α level is 0.01, then H_0 would not be rejected. In summary, the experimenter sets the hypotheses; the statistician selects the test statistic and rejection rule; the data are observed and the statistician reports the p-value to the experimenter; and the experimenter decides whether the p-value is sufficiently small to warrant rejection of H_0 in favor of H_1. The following example provides a numerical illustration.

Example 4.6.5. Recall the Darwin data discussed in Example 4.5.5. It was a paired design on the heights of cross and self-fertilized *Zea mays* plants. In each of 15 pots, one cross-fertilized and one self-fertilized were grown. The data of interest are the 15 paired differences, (cross − self). As in Example 4.5.5, let X_i denote the paired difference for the ith pot. Let μ be the true mean difference. The hypotheses of interest are $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. The standardized rejection rule is

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } T \geq k,$$

where $T = \overline{X}/(S/\sqrt{15})$, where $\overline{X}$ and S are respectively the sample mean and standard deviation of the differences. The alternative hypothesis states that on the average cross-fertilized plants are taller than self-fertilized plants. From Example 4.5.5 the t-test statistic has the value 2.15. Letting $t(14)$ denote a random variable with the t-distribution with 14 degrees of freedom, and using R the p-value for the experiment is

$$P[t(14) > 2.15] = 1 - \text{ pt}(2.15,14) = 1 - 0.9752 = 0.0248. \tag{4.6.13}$$

In practice, with this p-value, H_0 would be rejected at all levels greater than or equal to 0.0248. This observed significance level is also part of the output from the R call `t.test(cross-self,mu=0,alt="greater")`. ■

Returning to the discussion above, suppose the hypotheses are $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$. Obviously, the observed significance level in this case is p-value $= P_{H_0}(\overline{X} \leq \overline{x})$. For the two-sided hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, our "unspecified" rejection rule is

$$\text{Reject } H_0 \text{ in favor of } H_1, \text{ if } \overline{X} \leq l \text{ or } \overline{X} \geq k. \tag{4.6.14}$$

For the p-value, compute each of the one-sided p-values, take the smaller p-value, and double it. For an illustration, in the Darwin example, suppose the the hypotheses are $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Then the p-value is $2(0.0248) = 0.0496$. As a final note on p-values for two-sided hypotheses, suppose the test statistic can be expressed in terms of a t-test statistic. In this case the p-value can be found equivalently as follows. If d is the realized value of the t-test statistic then the p-value is

$$p\text{-value} = P_{H_0}[|t| \geq |d|], \tag{4.6.15}$$

where, under H_0, t has a t-distribution with $n - 1$ degrees of freedom.

In this discussion on p-values, keep in mind that good science dictates that the hypotheses should be known before the data are drawn.

EXERCISES

4.6.1. The R function `zpower`, found at the site listed in the Preface, computes the plot in Figure 4.6.1. Consider the two-sided test for proportions discussed in Example 4.6.3 based on the test statistic Z_1. Specifically consider the hypotheses $H_0 : p = .0.6$ versus $H_1 : p \neq 0.6$. Using the sample size $n = 50$ and the level $\alpha = 0.05$, write a R program, similar to `zpower`, which computes a plot of the power curve for this test on a proportion.

4.6.2. Consider the power function $\gamma(\mu)$ and its derivative $\gamma'(\mu)$ given by (4.6.5) and (4.6.6). Show that $\gamma'(\mu)$ is strictly negative for $\mu < \mu_0$ and strictly positive for $\mu > \mu_0$.

4.6.3. Show that the test defined by 4.6.9 has exact size α for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

4.6.4. Consider the one-sided t-test for $H_0 : \mu = \mu_0$ versus $H_{A1} : \mu > \mu_0$ constructed in Example 4.5.4 and the two-sided t-test for t-test for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ given in (4.6.9). Assume that both tests are of size α. Show that for $\mu > \mu_0$, the power function of the one-sided test is larger than the power function of the two-sided test.

4.6.5. On page 373 Rasmussen (1992) discussed a paired design. A baseball coach paired 20 members of his team by their speed; i.e., each member of the pair has about the same speed. Then for each pair, he randomly chose one member of the pair and told him that if could beat his best time in circling the bases he would give him an award (call this response the time of the "self" member). For the other member of the pair the coach's instruction was an award if he could beat the time of the other member of the pair (call this response the time of the "rival" member). Each member of the pair knew who his rival was. The data are given below, but are also in the file `selfrival.rda`. Let μ_d be the true difference in times (rival minus self) for a pair. The hypotheses of interest are $H_0 : \mu_d = 0$ versus $H_1 : \mu_d < 0$. The data are in order by pairs, so do not mix the order.

```
self:  16.20 16.78 17.38 17.59 17.37 17.49 18.18 18.16 18.36 18.53
```

```
15.92 16.58 17.57 16.75 17.28 17.32 17.51 17.58 18.26 17.87

rival: 15.95 16.15 17.05 16.99 17.34 17.53 17.34 17.51 18.10 18.19
16.04 16.80 17.24 16.81 17.11 17.22 17.33 17.82 18.19 17.88
```

(a) Obtain comparison boxplots of the data. Comment on the comparison plots. Are there any outliers?

(b) Compute the paired t-test and obtain the p-value. Are the data significant at the 5% level of significance?

(c) Obtain a point estimate of μ_d and a 95% confidence interval for it.

(d) Conclude in terms of the problem.

4.6.6. Verzani (2014), page 323, presented a data set concerning the effect that different dosages of the drug AZT have on patients with HIV. The responses we consider are the p24 antigen levels of HIV patients after their treatment with AZT. Of the 20 HIV patients in the study, 10 were randomly assign the dosage of 300 mg of AZT while the other 10 were assigned 600 mg. The hypotheses of interest are $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$ where $\Delta = \mu_{600} - \mu_{300}$ and μ_{600} and μ_{300} are the true mean p24 antigen levels under dosages of 600 mg and 300 mg of AZT, respectively. The data are given below but are also available in the file `aztdoses.rda`.

300 mg	284	279	289	292	287	295	285	279	306	298
600 mg	298	307	297	279	291	335	299	300	306	291

(a) Obtain comparison boxplots of the data. Identify outliers by patient. Comment on the comparison plots.

(b) Compute the two-sample t-test and obtain the p-value. Are the data significant at the 5% level of significance?

(c) Obtain a point estimate of Δ and a 95% confidence interval for it.

(d) Conclude in terms of the problem.

4.6.7. Among the data collected for the World Health Organization air quality monitoring project is a measure of suspended particles in $\mu g/m^3$. Let X and Y equal the concentration of suspended particles in $\mu g/m^3$ in the city center (commercial district) for Melbourne and Houston, respectively. Using $n = 13$ observations of X and $m = 16$ observations of Y, we test $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X < \mu_Y$.

(a) Define the test statistic and critical region, assuming that the unknown variances are equal. Let $\alpha = 0.05$.

(b) If $\overline{x} = 72.9$, $s_x = 25.6$, $\overline{y} = 81.7$, and $s_y = 28.3$, calculate the value of the test statistic and state your conclusion.

4.6.8. Let p equal the proportion of drivers who use a seat belt in a country that does not have a mandatory seat belt law. It was claimed that $p = 0.14$. An advertising campaign was conducted to increase this proportion. Two months after the campaign, $y = 104$ out of a random sample of $n = 590$ drivers were wearing their seat belts. Was the campaign successful?

(a) Define the null and alternative hypotheses.

(b) Define a critical region with an $\alpha = 0.01$ significance level.

(c) Determine the approximate p-value and state your conclusion.

4.6.9. In Exercise 4.2.18 we found a confidence interval for the variance σ^2 using the variance S^2 of a random sample of size n arising from $N(\mu, \sigma^2)$, where the mean μ is unknown. In testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 > \sigma_0^2$, use the critical region defined by $(n-1)S^2/\sigma_0^2 \geq c$. That is, reject H_0 and accept H_1 if $S^2 \geq c\sigma_0^2/(n-1)$. If $n = 13$ and the significance level $\alpha = 0.025$, determine c.

4.6.10. In Exercise 4.2.27, in finding a confidence interval for the ratio of the variances of two normal distributions, we used a statistic S_1^2/S_2^2, which has an F-distribution when those two variances are equal. If we denote that statistic by F, we can test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$ using the critical region $F \geq c$. If $n = 13,\ m = 11$, and $\alpha = 0.05$, find c.

4.7 Chi-Square Tests

In this section we introduce tests of statistical hypotheses called **chi-square tests**. A test of this sort was originally proposed by Karl Pearson in 1900, and it provided one of the earlier methods of statistical inference.

Let the random variable X_i be $N(\mu_i, \sigma_i^2),\ i = 1, 2, \ldots, n$, and let $X_1, X_2, \ldots, X_n$ be mutually independent. Thus the joint pdf of these variables is

$$\frac{1}{\sigma_1\sigma_2\cdots\sigma_n(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_1^n \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right], \quad -\infty < x_i < \infty.$$

The random variable that is defined by the exponent (apart from the coefficient $-\frac{1}{2}$) is $\sum_1^n [(X_i - \mu_i)/\sigma_i]^2$, and this random variable has a $\chi^2(n)$ distribution. In Section 3.5 we generalized this joint normal distribution of probability to n random variables that are *dependent* and we called the distribution a *multivariate normal distribution*. Theorem 3.5.1 shows a similar result holds for the exponent in the multivariate normal case, also.

Let us now discuss some random variables that have approximate chi-square distributions. Let X_1 be $b(n, p_1)$. Consider the random variable

$$Y = \frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}},$$

which has, as $n \to \infty$, an approximate $N(0,1)$ distribution (see Theorem 4.2.1). Furthermore, as discussed in Example 5.3.6, the distribution of Y^2 is approximately $\chi^2(1)$. Let $X_2 = n - X_1$ and let $p_2 = 1 - p_1$. Let $Q_1 = Y^2$. Then Q_1 may be written as

$$\begin{aligned} Q_1 = \frac{(X_1 - np_1)^2}{np_1(1-p_1)} &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \end{aligned} \tag{4.7.1}$$

because $(X_1 - np_1)^2 = (n - X_2 - n + np_2)^2 = (X_2 - np_2)^2$. This result can be generalized as follows.

Let $X_1, X_2, \ldots, X_{k-1}$ have a multinomial distribution with the parameters n and $p_1, \ldots, p_{k-1}$, as in Section 3.1. Let $X_k = n - (X_1 + \cdots + X_{k-1})$ and let $p_k = 1 - (p_1 + \cdots + p_{k-1})$. Define Q_{k-1} by

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}.$$

It is proved in a more advanced course that, as $n \to \infty$, Q_{k-1} has an approximate $\chi^2(k-1)$ distribution. Some writers caution the user of this approximation to be certain that n is large enough so that each np_i, $i = 1, 2, \ldots, k$, is at least equal to 5. In any case, it is important to realize that Q_{k-1} does not have a chi-square distribution, only an approximate chi-square distribution.

The random variable Q_{k-1} may serve as the basis of the tests of certain statistical hypotheses which we now discuss. Let the sample space $\mathcal{A}$ of a random experiment be the union of a finite number k of mutually disjoint sets $A_1, A_2, \ldots, A_k$. Furthermore, let $P(A_i) = p_i$, $i = 1, 2, \ldots, k$, where $p_k = 1 - p_1 - \cdots - p_{k-1}$, so that p_i is the probability that the outcome of the random experiment is an element of the set A_i. The random experiment is to be repeated n independent times and X_i represents the number of times the outcome is an element of set A_i. That is, $X_1, X_2, \ldots, X_k = n - X_1 - \cdots - X_{k-1}$ are the frequencies with which the outcome is, respectively, an element of $A_1, A_2, \ldots, A_k$. Then the joint pmf of $X_1, X_2, \ldots, X_{k-1}$ is the multinomial pmf with the parameters n, $p_1, \ldots, p_{k-1}$. Consider the simple hypothesis (concerning this multinomial pmf) $H_0 : p_1 = p_{10}$, $p_2 = p_{20}, \ldots, p_{k-1} = p_{k-1,0}$ ($p_k = p_{k0} = 1 - p_{10} - \cdots - p_{k-1,0}$), where $p_{10}, \ldots, p_{k-1,0}$ are specified numbers. It is desired to test H_0 against all alternatives.

If the hypothesis H_0 is true, the random variable

$$Q_{k-1} = \sum_{1}^{k} \frac{(X_i - np_{i0})^2}{np_{i0}}$$

has an approximate chi-square distribution with $k-1$ degrees of freedom. Since, when H_0 is true, np_{i0} is the expected value of X_i, one would feel intuitively that observed values of Q_{k-1} should not be too large if H_0 is true. Our test is then

to reject H_0 if $Q_{k-1} \geq c$. To determine a test with level of significance α, we can use tables of the χ^2-distribution or a computer package. Using R, we compute the critical value c by `qchisq(1 − α,k-1)`. If, then, the hypothesis H_0 is rejected when the observed value of Q_{k-1} is at least as great as c, the test of H_0 has a significance level that is approximately equal to α. Also if q is the realized value of the test statistic Q_{k-1} then the observed significance level of the test is computed in R by `1-pchisq(q,k-1)`. This is frequently called a **goodness-of-fit test**. Some illustrative examples follow.

Example 4.7.1. One of the first six positive integers is to be chosen by a random experiment (perhaps by the cast of a die). Let $A_i = \{x : x = i\},\ i = 1, 2, \ldots, 6$. The hypothesis $H_0 : P(A_i) = p_{i0} = \frac{1}{6},\ i = 1, 2, \ldots, 6$, is tested, at the approximate 5% significance level, against all alternatives. To make the test, the random experiment is repeated under the same conditions, 60 independent times. In this example, $k = 6$ and $np_{i0} = 60(\frac{1}{6}) = 10,\ i = 1, 2, \ldots, 6$. Let X_i denote the frequency with which the random experiment terminates with the outcome in $A_i,\ i = 1, 2, \ldots, 6$, and let $Q_5 = \sum_1^6 (X_i - 10)^2/10$. Since there are $6 - 1 = 5$ degrees of freedom, the critical value for a level $\alpha = 0.05$ test is `qchisq(0.95,5) = 11.0705`. Now suppose that the experimental frequencies of $A_1, A_2, \ldots, A_6$ are, respectively, 13, 19, 11, 8, 5, and 4. The observed value of Q_5 is

$$\frac{(13-10)^2}{10} + \frac{(19-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(4-10)^2}{10} = 15.6.$$

Since $15.6 > 11.0705$, the hypothesis $P(A_i) = \frac{1}{6},\ i = 1, 2, \ldots, 6$, is rejected at the (approximate) 5% significance level.

The following R segment computes this test, returning the test statistic and the p-value as shown:

```
ps=rep(1/6,6); x=c(13,19,11,8,5,4); chisq.test(x,p=ps)
X-squared = 15.6, df = 5, p-value = 0.008084.
```

■

Example 4.7.2. A point is to be selected from the unit interval $\{x : 0 < x < 1\}$ by a random process. Let $A_1 = \{x : 0 < x \leq \frac{1}{4}\},\ A_2 = \{x : \frac{1}{4} < x \leq \frac{1}{2}\},\ A_3 = \{x : \frac{1}{2} < x \leq \frac{3}{4}\}$, and $A_4 = \{x : \frac{3}{4} < x < 1\}$. Let the probabilities $p_i,\ i = 1, 2, 3, 4$, assigned to these sets under the hypothesis be determined by the pdf $2x,\ 0 < x < 1$, zero elsewhere. Then these probabilities are, respectively,

$$p_{10} = \int_0^{1/4} 2x\,dx = \tfrac{1}{16}, \quad p_{20} = \tfrac{3}{16}, \quad p_{30} = \tfrac{5}{16}, \quad p_{40} = \tfrac{7}{16}.$$

Thus the hypothesis to be tested is that p_1, p_2, p_3, and $p_4 = 1 - p_1 - p_2 - p_3$ have the preceding values in a multinomial distribution with $k = 4$. This hypothesis is to be tested at an approximate 0.025 significance level by repeating the random experiment $n = 80$ independent times under the same conditions. Here the np_{i0} for $i = 1, 2, 3, 4$, are, respectively, 5, 15, 25, and 35. Suppose the observed frequencies of A_1, A_2, A_3, and A_4 are 6, 18, 20, and 36, respectively. Then the observed value

of $Q_3 = \sum_1^4 (X_i - np_{i0})^2/(np_{i0})$ is

$$\frac{(6-5)^2}{5} + \frac{(18-15)^2}{15} + \frac{(20-25)^2}{25} + \frac{(36-35)^2}{35} = \frac{64}{35} = 1.83.$$

The following R segment calculates the test and p-value:

```
x=c(6,18,20,36); ps=c(1,3,5,7)/16; chisq.test(x,p=ps)
X-squared = 1.8286, df = 3, p-value = 0.6087
```

Hence, we fail to reject H_0 at level 0.0250. ■

Thus far we have used the chi-square test when the hypothesis H_0 is a simple hypothesis. More often we encounter hypotheses H_0 in which the multinomial probabilities $p_1, p_2, \ldots, p_k$ are not completely specified by the hypothesis H_0. That is, under H_0, these probabilities are functions of unknown parameters. For an illustration, suppose that a certain random variable Y can take on any real value. Let us partition the space $\{y : -\infty < y < \infty\}$ into k mutually disjoint sets $A_1, A_2, \ldots, A_k$ so that the events $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive. Let H_0 be the hypothesis that Y is $N(\mu, \sigma^2)$ with μ and σ^2 unspecified. Then each

$$p_i = \int_{A_i} \frac{1}{\sqrt{2\pi}\sigma} \exp[-(y-\mu)^2/2\sigma^2]\, dy, \quad i = 1, 2, \ldots, k,$$

is a function of the unknown parameters μ and σ^2. Suppose that we take a random sample $Y_1, \ldots, Y_n$ of size n from this distribution. If we let X_i denote the frequency of A_i, $i = 1, 2, \ldots, k$, so that $X_1 + X_2 + \cdots + X_k = n$, the random variable

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

cannot be computed once $X_1, \ldots, X_k$ have been observed, since each p_i, and hence Q_{k-1}, is a function of μ and σ^2. Accordingly, choose the values of μ and σ^2 that minimize Q_{k-1}. These values depend upon the observed $X_1 = x_1, \ldots, X_k = x_k$ and are called **minimum chi-square estimates** of μ and σ^2. These point estimates of μ and σ^2 enable us to compute numerically the estimates of each p_i. Accordingly, if these values are used, Q_{k-1} can be computed once $Y_1, Y_2, \ldots, Y_n$, and hence $X_1, X_2, \ldots, X_k$, are observed. However, a very important aspect of the fact, which we accept without proof, is that now Q_{k-1} is approximately $\chi^2(k-3)$. That is, the number of degrees of freedom of the approximate chi-square distribution of Q_{k-1} is reduced by one for each parameter estimated by the observed data. This statement applies not only to the problem at hand but also to more general situations. Two examples are now be given. The first of these examples deals with the test of the hypothesis that two multinomial distributions are the same.

Remark 4.7.1. In many cases, such as that involving the mean μ and the variance σ^2 of a normal distribution, minimum chi-square estimates are difficult to compute. Other estimates, such as the maximum likelihood estimates of Example 4.1.3, $\hat{\mu} = \overline{Y}$ and $\hat{\sigma^2} = (n-1)S^2/n$, are used to evaluate p_i and Q_{k-1}. In general, Q_{k-1} is not minimized by maximum likelihood estimates, and thus its computed value

is somewhat greater than it would be if minimum chi-square estimates are used. Hence, when comparing it to a critical value listed in the chi-square table with $k-3$ degrees of freedom, there is a greater chance of rejection than there would be if the actual minimum of Q_{k-1} is used. Accordingly, the approximate significance level of such a test may be higher than the p-value as calculated in the χ^2-analysis. This modification should be kept in mind and, if at all possible, each p_i should be estimated using the frequencies $X_1, \ldots, X_k$ rather than directly using the observations $Y_1, Y_2, \ldots, Y_n$ of the random sample. ■

Example 4.7.3. In this example, we consider two multinomial distributions with parameters $n_j, p_{1j}, p_{2j}, \ldots, p_{kj}$ and $j = 1, 2$, respectively. Let $X_{ij},\ i = 1, 2, \ldots, k$, $j = 1, 2$, represent the corresponding frequencies. If n_1 and n_2 are large and the observations from one distribution are independent of those from the other, the random variable

$$\sum_{j=1}^{2}\sum_{i=1}^{k}\frac{(X_{ij}-n_jp_{ij})^2}{n_jp_{ij}}$$

is the sum of two independent random variables each of which we treat as though it were $\chi^2(k-1)$; that is, the random variable is approximately $\chi^2(2k-2)$. Consider the hypothesis

$$H_0 : p_{11} = p_{12}, p_{21} = p_{22}, \ldots, p_{k1} = p_{k2},$$

where each $p_{i1} = p_{i2},\ i = 1, 2, \ldots, k$, is unspecified. Thus we need point estimates of these parameters. The maximum likelihood estimator of $p_{i1} = p_{i2}$, based upon the frequencies X_{ij}, is $(X_{i1} + X_{i2})/(n_1 + n_2),\ i = 1, 2, \ldots, k$. Note that we need only $k-1$ point estimates, because we have a point estimate of $p_{k1} = p_{k2}$ once we have point estimates of the first $k-1$ probabilities. In accordance with the fact that has been stated, the random variable

$$Q_{k-1} = \sum_{j=1}^{2}\sum_{i=1}^{k}\frac{\{X_{ij}-n_j[(X_{i1}+X_{i2})/(n_1+n_2)]\}^2}{n_j[(X_{i1}+X_{i2})/(n_1+n_2)]}$$

has an approximate χ^2 distribution with $2k-2-(k-1) = k-1$ degrees of freedom. Thus we are able to test the hypothesis that two multinomial distributions are the same. For a specified level α, the hypothesis H_0 is rejected when the computed value of Q_{k-1} exceeds the $1-\alpha$ quantile of a χ^2-distribution with $k-1$ degrees of freedom. This test is often called the chi-square test for **homogeneity** (the null is equivalent to **homogeneous distributions**). ■

The second example deals with the subject of **contingency tables**.

Example 4.7.4. Let the result of a random experiment be classified by two attributes (such as the color of the hair and the color of the eyes). That is, one attribute of the outcome is one and only one of certain mutually exclusive and exhaustive events, say $A_1, A_2, \ldots, A_a$; and the other attribute of the outcome is also one and only one of certain mutually exclusive and exhaustive events, say $B_1, B_2, \ldots, B_b$. Let $p_{ij} = P(A_i \cap B_j),\ i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b$. The random

experiment is repeated n independent times and X_{ij} denotes the frequency of the event $A_i \cap B_j$. Since there are $k = ab$ such events as $A_i \cap B_j$, the random variable

$$Q_{ab-1} = \sum_{j=1}^{b} \sum_{i=1}^{a} \frac{(X_{ij} - np_{ij})^2}{np_{ij}}$$

has an approximate chi-square distribution with $ab-1$ degrees of freedom, provided that n is large. Suppose that we wish to test the independence of the A and the B attributes, i.e., the hypothesis $H_0 : P(A_i \cap B_j) = P(A_i)P(B_j),\ i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b$. Let us denote $P(A_i)$ by $p_{i.}$ and $P(B_j)$ by $p_{.j}$. It follows that

$$p_{i.} = \sum_{j=1}^{b} p_{ij}, \quad p_{.j} = \sum_{i=1}^{a} p_{ij}, \text{ and } 1 = \sum_{j=1}^{b}\sum_{i=1}^{a} p_{ij} = \sum_{j=1}^{b} p_{.j} = \sum_{i=1}^{a} p_{i.}.$$

Then the hypothesis can be formulated as $H_0 : p_{ij} = p_{i.}p_{.j},\ i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b$. To test H_0, we can use Q_{ab-1} with p_{ij} replaced by $p_{i.}p_{.j}$. But if $p_{i.},\ i = 1, 2, \ldots, a$, and $p_{.j},\ j = 1, 2, \ldots, b$, are unknown, as they frequently are in applications, we cannot compute Q_{ab-1} once the frequencies are observed. In such a case, we estimate these unknown parameters by

$$\hat{p}_{i\cdot} = \tfrac{X_{i.}}{n}, \text{ where } X_{i.} = \sum_{j=1}^{b} X_{ij}, \text{ for } i = 1, 2, \ldots, a,$$

and

$$\hat{p}_{\cdot j} = \tfrac{X_{.j}}{n}, \text{ where } X_{.j} = \sum_{i=1}^{a} X_{ij}, \text{ for } j = 1, 2, \ldots, b.$$

Since $\sum_i p_{i.} = \sum_j p_{.j} = 1$, we have estimated only $a-1+b-1 = a+b-2$ parameters. So if these estimates are used in Q_{ab-1}, with $p_{ij} = p_{i.}p_{.j}$, then, according to the rule that has been stated in this section, the random variable

$$\sum_{j=1}^{b} \sum_{i=1}^{a} \frac{[X_{ij} - n(X_{i.}/n)(X_{.j}/n)]^2}{n(X_{i.}/n)(X_{.j}/n)} \tag{4.7.2}$$

has an approximate chi-square distribution with $ab-1-(a+b-2) = (a-1)(b-1)$ degrees of freedom provided that H_0 is true. For a specified level α, the hypothesis H_0 is then rejected if the computed value of this statistic exceeds the $1-\alpha$ quantile of a χ^2-distribution with $(a-1)(b-1)$ degrees of freedom. This is the χ^2-**test** for **independence**.

For an illustration, reconsider Example 4.1.5 in which we presented data on hair color of Scottish children. The eye colors of the children were also recorded. The complete data are in the following contingency table (with additionally the marginal sums). The contingency table is also in the file `scotteyehair.rda`.

	Fair	Red	Medium	Dark	Black	Margin
Blue	1368	170	1041	398	1	2978
Light	2577	474	2703	932	11	6697
Medium	1390	420	3826	1842	33	7511
Dark	454	255	1848	2506	112	5175
Margin	5789	1319	9418	5678	157	22361

The table indicates that hair and eye color are dependent random variables. For example, the observed frequency of children with blue eyes and black hair is 1 while the expected frequency under independence is $2978 \times 157/22361 = 20.9$. The contribution to the test statistic from this one cell is $(1 - 20.9)^2/20.9 = 19.95$ that nearly exceeds the test statistic's χ^2 critical value at level 0.05, which is `qchisq(.95,12) = 21.026`. The χ^2-test statistic for independence is tedious to compute and the reader is advised to use a statistical package. For R, assume that the contingency table without margin sums is in the matrix `scotteyehair`. Then the code `chisq.test(scotteyehair)` returns the χ^2 test statistic and the p-value as: `X-squared = 3683.9, df = 12, p-value < 2.2e-16`. Thus the result is highly significant. Based on this study, hair color and eye color of Scottish children are dependent on one another. To investigate where the dependence is the strongest in a contingency table, we recommend considering the table of expected frequencies and the table of **Pearson residuals**. The later are the square roots (with the sign of the numerators) of the summands in expression (4.7.2) defining the test statistic. The sum of the squared Pearson residuals equals the χ^2-test statistic. In R, the following code obtains both of these items:

```
fit = chisq.test(scotteyehair); fit$expected; fit$residual
```

Based on running this code, the largest residual is 32.8 for the cell dark hair and dark eyes. The observed frequency is 2506 while the expected frequency under independence is 1314. ■

In each of the four examples of this section, we have indicated that the statistic used to test the hypothesis H_0 has an approximate chi-square distribution, provided that n is sufficiently large and H_0 is true. To compute the power of any of these tests for values of the parameters not described by H_0, we need the distribution of the statistic when H_0 is not true. In each of these cases, the statistic has an approximate distribution called a **noncentral chi-square distribution**. The noncentral chi-square distribution is discussed later in Section 9.3.

EXERCISES

4.7.1. Consider Example 4.7.2. Suppose the observed frequencies of $A_1, \ldots, A_4$ are 20, 30, 92, and 105, respectively. Modify the R code given in the example to calculate the test for these new frequencies. Report the p-value.

4.7.2. A number is to be selected from the interval $\{x : 0 < x < 2\}$ by a random process. Let $A_i = \{x : (i-1)/2 < x \leq i/2\}$, $i = 1, 2, 3$, and let $A_4 = \{x : \frac{3}{2} < x < 2\}$. For $i = 1, 2, 3, 4$, suppose a certain hypothesis assigns probabilities p_{i0} to these sets in accordance with $p_{i0} = \int_{A_i} (\frac{1}{2})(2 - x)\, dx$, $i = 1, 2, 3, 4$. This

hypothesis (concerning the multinomial pdf with $k = 4$) is to be tested at the 5% level of significance by a chi-square test. If the observed frequencies of the sets $A_i,\ i = 1, 2, 3, 4$, are respectively, 30, 30, 10, 10, would H_0 be accepted at the (approximate) 5% level of significance? Use R code similar to that of Example 4.7.2 for the computation.

4.7.3. Define the sets $A_1 = \{x : -\infty < x \leq 0\}$, $A_i = \{x : i - 2 < x \leq i - 1\}$, $i = 2, \ldots, 7$, and $A_8 = \{x : 6 < x < \infty\}$. A certain hypothesis assigns probabilities p_{i0} to these sets A_i in accordance with

$$p_{i0} = \int_{A_i} \frac{1}{2\sqrt{2\pi}} \exp\left[-\frac{(x-3)^2}{2(4)}\right] dx, \quad i = 1, 2, \ldots, 7, 8.$$

This hypothesis (concerning the multinomial pdf with $k = 8$) is to be tested, at the 5% level of significance, by a chi-square test. If the observed frequencies of the sets $A_i,\ i = 1, 2, \ldots, 8$, are, respectively, 60, 96, 140, 210, 172, 160, 88, and 74, would H_0 be accepted at the (approximate) 5% level of significance? Use R code similar to that discussed in Example 4.7.2. The probabilities are easily computed in R; for example, p_{30} = pnorm(2,3,2) − pnorm(1,3,2).

4.7.4. A die was cast $n = 120$ independent times and the following data resulted:

Spots Up	1	2	3	4	5	6
Frequency	b	20	20	20	20	$40 - b$

If we use a chi-square test, for what values of b would the hypothesis that the die is unbiased be rejected at the 0.025 significance level?

4.7.5. Consider the problem from genetics of crossing two types of peas. The Mendelian theory states that the probabilities of the classifications (a) round and yellow, (b) wrinkled and yellow, (c) round and green, and (d) wrinkled and green are $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$, and $\frac{1}{16}$, respectively. If, from 160 independent observations, the observed frequencies of these respective classifications are 86, 35, 26, and 13, are these data consistent with the Mendelian theory? That is, test, with $\alpha = 0.01$, the hypothesis that the respective probabilities are $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$, and $\frac{1}{16}$.

4.7.6. Two different teaching procedures were used on two different groups of students. Each group contained 100 students of about the same ability. At the end of the term, an evaluating team assigned a letter grade to each student. The results were tabulated as follows.

	Grade					
Group	A	B	C	D	F	Total
I	15	25	32	17	11	100
II	9	18	29	28	16	100

If we consider these data to be independent observations from two respective multinomial distributions with $k = 5$, test at the 5% significance level the hypothesis