7.6.8, the joint distribution of X_1 and $\overline{X}$ is bivariate normal with mean vector (θ, θ), variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1/n$, and correlation coefficient $\rho = 1/\sqrt{n}$. Thus the conditional pdf of X_1, given $\overline{X} = \overline{x}$, is normal with linear conditional mean

$$\theta + \frac{\rho\sigma_1}{\sigma_2}(\overline{x} - \theta) = \overline{x}$$

and with variance

$$\sigma_1^2(1 - \rho^2) = \frac{n-1}{n}.$$

The conditional expectation of $u(X_1)$, given $\overline{X} = \overline{x}$, is then

$$\begin{aligned} \varphi(\overline{x}) &= \int_{-\infty}^{\infty} u(x_1)\sqrt{\frac{n}{n-1}}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{n(x_1-\overline{x})^2}{2(n-1)}\right]dx_1 \\ &= \int_{-\infty}^{c} \sqrt{\frac{n}{n-1}}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{n(x_1-\overline{x})^2}{2(n-1)}\right]dx_1. \end{aligned}$$

The change of variable $z = \sqrt{n}(x_1 - \overline{x})/\sqrt{n-1}$ enables us to write this conditional expectation as

$$\varphi(\overline{x}) = \int_{-\infty}^{c'} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz = \Phi(c') = \Phi\left[\frac{\sqrt{n}(c-\overline{x})}{\sqrt{n-1}}\right],$$

where $c' = \sqrt{n}(c - \overline{x})/\sqrt{n-1}$. Thus the unique MVUE of $\Phi(c - \theta)$ is, for every fixed constant c, given by $\varphi(\overline{X}) = \Phi[\sqrt{n}(c - \overline{X})/\sqrt{n-1}]$.

In this example the mle of $\Phi(c - \theta)$ is $\Phi(c - \overline{X})$. These two estimators are close because $\sqrt{n/(n-1)} \to 1$, as $n \to \infty$. ■

Remark 7.6.1. We should like to draw the attention of the reader to a rather important fact. This has to do with the adoption of a *principle*, such as the principle of unbiasedness and minimum variance. A principle is not a theorem; and seldom does a principle yield satisfactory results in all cases. So far, this principle has provided quite satisfactory results. To see that this is not always the case, let X have a Poisson distribution with parameter θ, $0 < \theta < \infty$. We may look upon X as a random sample of size 1 from this distribution. Thus X is a complete sufficient statistic for θ. We seek the estimator of $e^{-2\theta}$ that is unbiased and has minimum variance. Consider $Y = (-1)^X$. We have

$$E(Y) = E[(-1)^X] = \sum_{x=0}^{\infty} \frac{(-\theta)^x e^{-\theta}}{x!} = e^{-2\theta}.$$

Accordingly, $(-1)^X$ is the MVUE of $e^{-2\theta}$. Here this estimator leaves much to be desired. We are endeavoring to elicit some information about the number $e^{-2\theta}$, where $0 < e^{-2\theta} < 1$; yet our point estimate is either -1 or $+1$, each of which is a very poor estimate of a number between 0 and 1. We do not wish to leave the reader with the impression that an MVUE is *bad*. That is not the case at all. We merely wish to point out that if one tries hard enough, one can find instances where such a statistic is *not good*. Incidentally, the maximum likelihood estimator of $e^{-2\theta}$ is, in the case where the sample size equals 1, e^{-2X}, which is probably a much better estimator in practice than is the unbiased estimator $(-1)^X$. ■

7.6.1 Bootstrap Standard Errors

Section 6.3 presented the asymptotic theory of maximum likelihood estimators (mles). In many cases, this theory also provides consistent estimators of the asymptotic standard deviation of mles. This allows a simple, but very useful, summary of the estimation process; i.e., $\hat{\theta} \pm \text{SE}(\hat{\theta})$ where $\hat{\theta}$ is the mle of θ and $\text{SE}(\hat{\theta})$ is the corresponding standard error. For example, these summaries can be used descriptively as labels on plots and tables as well as in the formation of asymptotic confidence intervals for inference. Section 4.9 presented percentile confidence intervals for θ based on the bootstrap. The bootstrap, though, can also be used to obtain standard errors for estimates including MVUE's.

Consider a random variable X with pdf $f(x;\theta)$, where $\theta \in \Omega$. Let $X_1, \ldots, X_n$ be a random sample on X. Let $\hat{\theta}$ be an estimator of θ based on the sample. Suppose $x_1, \ldots, x_n$ is a realization of the sample and let $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ be the corresponding estimate of θ. Recall in Section 4.9 that the bootstrap uses the empirical cdf $\hat{F}_n$ of the realization. This is the discrete distribution which places mass $1/n$ at each point x_i. The bootstrap procedure samples, with replacement, from $\hat{F}_n$.

For the bootstrap procedure, we obtain B bootstrap samples. For $i = 1, \ldots, B$, let the vector $\mathbf{x}_i^* = (x_{i,1}^*, \ldots, x_{i,n}^*)'$ denote the ith bootstrap sample. Let $\hat{\theta}_i^* = \hat{\theta}(\mathbf{x}_i^*)$ denote the estimate of θ based on the ith sample. We then have the bootstrap estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$, which we used in Section 4.9 to obtain the bootstrap percentile confidence interval for θ. Suppose instead we consider the standard deviation of these bootstrap estimates; that is,

$$\text{SE}_B = \left[\frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_1^* - \overline{\hat{\theta}^*})^2, \right]^{1/2}, \tag{7.6.3}$$

where $\overline{\hat{\theta}^*} = (1/B) \sum_{i=1}^{B} \hat{\theta}_1^*$. This is the bootstrap estimate of the standard error of $\hat{\theta}$.

Example 7.6.4. For this example, we consider a data set drawn from a normal distribution, $N(\theta, \sigma^2)$. In this case the MVUE of θ is the sample mean $\overline{X}$ and its usual standard error is $s/\sqrt{n}$, where s is the sample standard deviation. The rounded data[1] are:

```
27.5 50.9 71.1 43.1 40.4 44.8 36.6 53.5 65.2 47.7
75.7 55.4 61.1 39.8 33.4 57.6 47.9 60.7 27.8 65.2
```

Assuming the data are in the R vector `x`, the mean and standard error are computed as

```
mean(x);  50.27;  sd(x)/sqrt(n); 3.094461
```

The R function `bootse1.R` runs the bootstrap for standard errors as described above. Using 3,000 bootstraps, our run of this function estimated the standard error by 3.050878. Thus, the estimate and the bootstrap standard error are summarized as 50.27 ± 3.05. ■

[1] The data are in the file sect76data.rda. The true mean and sd are: 50 and 15.

The bootstrap process described above is often called the **nonparametric bootstrap** because it makes no assumptions about the pdf $f(x;\theta)$. In this chapter, though, strong assumptions are made about the model. For instance, in the last example, we assume that the pdf is normal. What if we make use of this information in the bootstrap? This is called the **parametric bootstrap**. For the last example, instead of sampling from the empirical cdf $\widehat{F}_n$, we sample randomly from the normal distribution, using as mean $\overline{x}$ and as standard deviation s, the sample standard deviation. The R function `bootse2.R` performs this parametric bootstrap. For our run on the data set in the example, it computed the standard error as 3.162918. Notice how close the three estimated standard deviations are.

Which bootstrap, nonparametric or parametric, should we use? We recommend the nonparametric bootstrap in general. The strong model assumptions are not needed for its validity. See pages 55–56 of Efron and Tibshirani (1993) for discussion.

EXERCISES

7.6.1. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $N(\theta, 1)$, $-\infty < \theta < \infty$. Find the MVUE of θ^2.
Hint: First determine $E(\overline{X}^2)$.

7.6.2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $N(0, \theta)$. Then $Y = \sum X_i^2$ is a complete sufficient statistic for θ. Find the MVUE of θ^2.

7.6.3. Consider Example 7.6.3 where the parameter of interest is $P(X < c)$ for X distributed $N(\theta, 1)$. Modify the R function `bootse1.R` so that for a specified value of c it returns the MVUE of $P(X < c)$ and the bootstrap standard error of the estimate. Run your function on the data in `ex763data.rda` with $c = 11$ and 3,000 bootstraps. These data are generated from a $N(10, 1)$ distribution. Report (a) the true parameter, (b) the MVUE, and (c) the bootstrap standard error.

7.6.4. For Example 7.6.4, modify the R function `bootse1.R` so that the estimate is the median not the mean. Using 3,000 bootstraps, run your function on the data set discussed in the example and report (a) the estimate and (b) the bootstrap standard error.

7.6.5. Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform $(0, \theta)$ distribution. Continuing with Example 7.6.2, find the MVUEs for the following functions of θ.

(a) $g(\theta) = \frac{\theta^2}{12}$, i.e., the variance of the distribution.

(b) $g(\theta) = \frac{1}{\theta}$, i.e., the pdf of the distribution.

(c) For t real, $g(\theta) = \frac{e^{t\theta}-1}{t\theta}$, i.e., the mgf of the distribution.

7.6.6. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter $\theta > 0$.

(a) Find the MVUE of $P(X \leq 1) = (1+\theta)e^{-\theta}$.

Hint: Let $u(x_1) = 1,\ x_1 \leq 1$, zero elsewhere, and find $E[u(X_1)|Y = y]$, where $Y = \sum_1^n X_i$.

(b) Express the MVUE as a function of the mle of θ.

(c) Determine the asymptotic distribution of the mle of θ.

(d) Obtain the mle of $P(X \leq 1)$. Then use Theorem 5.2.9 to determine its asymptotic distribution.

7.6.7. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson distribution with parameter $\theta > 0$. From Remark 7.6.1, we know that $E[(-1)^{X_1}] = e^{-2\theta}$.

(a) Show that $E[(-1)^{X_1}|Y_1 = y_1] = (1-2/n)^{y_1}$, where $Y_1 = X_1 + X_2 + \cdots + X_n$. *Hint:* First show that the conditional pdf of $X_1, X_2, \ldots, X_{n-1}$, given $Y_1 = y_1$, is multinomial, and hence that of X_1, given $Y_1 = y_1$, is $b(y_1, 1/n)$.

(b) Show that the mle of $e^{-2\theta}$ is $e^{-2\overline{X}}$.

(c) Since $y_1 = n\overline{x}$, show that $(1-2/n)^{y_1}$ is approximately equal to $e^{-2\overline{x}}$ when n is large.

7.6.8. As in Example 7.6.3, let $X_1, X_2, \ldots, X_n$ be a random sample of size $n > 1$ from a distribution that is $N(\theta, 1)$. Show that the joint distribution of X_1 and $\overline{X}$ is bivariate normal with mean vector (θ, θ), variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1/n$, and correlation coefficient $\rho = 1/\sqrt{n}$.

7.6.9. Let a random sample of size n be taken from a distribution that has the pdf $f(x;\theta) = (1/\theta)\exp(-x/\theta)I_{(0,\infty)}(x)$. Find the mle and MVUE of $P(X \leq 2)$.

7.6.10. Let $X_1, X_2, \ldots, X_n$ be a random sample with the common pdf $f(x) = \theta^{-1}e^{-x/\theta}$, for $x > 0$, zero elsewhere; that is, $f(x)$ is a $\Gamma(1, \theta)$ pdf.

(a) Show that the statistic $\overline{X} = n^{-1}\sum_{i=1}^n X_i$ is a complete and sufficient statistic for θ.

(b) Determine the MVUE of θ.

(c) Determine the mle of θ.

(d) Often, though, this pdf is written as $f(x) = \tau e^{-\tau x}$, for $x > 0$, zero elsewhere. Thus $\tau = 1/\theta$. Use Theorem 6.1.2 to determine the mle of τ.

(e) Show that the statistic $\overline{X} = n^{-1}\sum_{i=1}^n X_i$ is a complete and sufficient statistic for τ. Show that $(n-1)/(n\overline{X})$ is the MVUE of $\tau = 1/\theta$. Hence, as usual, the reciprocal of the mle of θ is the mle of $1/\theta$, but, in this situation, the reciprocal of the MVUE of θ is not the MVUE of $1/\theta$.

(f) Compute the variances of each of the unbiased estimators in parts (b) and (e).

7.6.11. Consider the situation of the last exercise, but suppose we have the following two independent random samples: (1) $X_1, X_2, \ldots, X_n$ is a random sample with the common pdf $f_X(x) = \theta^{-1}e^{-x/\theta}$, for $x > 0$, zero elsewhere, and (2) $Y_1, Y_2, \ldots, Y_n$ is a random sample with common pdf $f_Y(y) = \theta e^{-\theta y}$, for $y > 0$, zero elsewhere. The last exercise suggests that, for some constant c, $Z = c\overline{X}/\overline{Y}$ might be an unbiased estimator of θ^2. Find this constant c and the variance of Z.
Hint: Show that $\overline{X}/(\theta^2\overline{Y})$ has an F-distribution.

7.6.12. Obtain the asymptotic distribution of the MVUE in Example 7.6.1 for the case $\theta = 1/2$.

7.7 The Case of Several Parameters

In many of the interesting problems we encounter, the pdf or pmf may not depend upon a single parameter θ, but perhaps upon two (or more) parameters. In general, our parameter space Ω is a subset of R^p, but in many of our examples p is 2.

Definition 7.7.1. *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has pdf or pmf $f(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Omega \subset R^p$. Let $\mathcal{S}$ denote the support of X. Let $\mathbf{Y}$ be an m-dimensional random vector of statistics $\mathbf{Y} = (Y_1, \ldots, Y_m)'$, where $Y_i = u_i(X_1, X_2, \ldots, X_n)$, for $i = 1, \ldots, m$. Denote the pdf or pmf of $\mathbf{Y}$ by $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ for $\mathbf{y} \in R^m$. The random vector of statistics $\mathbf{Y}$ is* **jointly sufficient** *for $\boldsymbol{\theta}$ if and only if*

$$\frac{\prod_{i=1}^n f(x_i;\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})} = H(x_1, x_2, \ldots, x_n), \quad \text{for all } x_i \in \mathcal{S},$$

where $H(x_1, x_2, \ldots, x_n)$ does not depend upon $\boldsymbol{\theta}$.

In general, $m \neq p$, i.e., the number of sufficient statistics does not have to be the same as the number of parameters, but in most of our examples this is the case.

As may be anticipated, the factorization theorem can be extended. In our notation it can be stated in the following manner. The vector of statistics $\mathbf{Y}$ is jointly sufficient for the parameter $\boldsymbol{\theta} \in \Omega$ if and only if we can find two nonnegative functions k_1 and k_2 such that

$$\prod_{i=1}^n f(x_i;\boldsymbol{\theta}) = k_1(\mathbf{y};\boldsymbol{\theta})k_2(x_1, \ldots, x_n), \quad \text{for all } x_i \in \mathcal{S}, \tag{7.7.1}$$

where the function $k_2(x_1, x_2, \ldots, x_n)$ does not depend upon $\boldsymbol{\theta}$.

Example 7.7.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution having pdf

$$f(x;\theta_1,\theta_2) = \begin{cases} \frac{1}{2\theta_2} & \theta_1 - \theta_2 < x < \theta_1 + \theta_2 \\ 0 & \text{elsewhere,} \end{cases}$$

where $-\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics. The joint pdf of Y_1 and Y_n is given by

$$f_{Y_1,Y_2}(y_1, y_n;\theta_1,\theta_2) = \frac{n(n-1)}{(2\theta_2)^n}(y_n - y_1)^{n-2}, \quad \theta_1 - \theta_2 < y_1 < y_n < \theta_1 + \theta_2,$$

and equals zero elsewhere. Accordingly, the joint pdf of $X_1, X_2, \ldots, X_n$ can be written, for all points in its support (all x_i such that $\theta_1 - \theta_2 < x_i < \theta_1 + \theta_2$),

$$\left(\frac{1}{2\theta_2}\right)^n = \frac{n(n-1)[\max(x_i) - \min(x_i)]^{n-2}}{(2\theta_2)^n} \left(\frac{1}{n(n-1)[\max(x_i) - \min(x_i)]^{n-2}}\right).$$

Since $\min(x_i) \leq x_j \leq \max(x_i)$, $j = 1, 2, \ldots, n$, the last factor does not depend upon the parameters. Either the definition or the factorization theorem assures us that Y_1 and Y_n are joint sufficient statistics for θ_1 and θ_2. ■

The concept of a complete family of probability density functions is generalized as follows: Let

$$\{f(v_1, v_2, \ldots, v_k; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega\}$$

denote a family of pdfs of k random variables $V_1, V_2, \ldots, V_k$ that depends upon the p-dimensional vector of parameters $\boldsymbol{\theta} \in \Omega$. Let $u(v_1, v_2, \ldots, v_k)$ be a function of $v_1, v_2, \ldots, v_k$ (but not a function of any or all of the parameters). If

$$E[u(V_1, V_2, \ldots, V_k)] = 0$$

for all $\boldsymbol{\theta} \in \Omega$ implies that $u(v_1, v_2, \ldots, v_k) = 0$ at all points $(v_1, v_2, \ldots, v_k)$, except on a set of points that has probability zero for all members of the family of probability density functions, we shall say that the family of probability density functions is a complete family.

In the case where $\boldsymbol{\theta}$ is a vector, we generally consider best estimators of functions of $\boldsymbol{\theta}$, that is, parameters δ, where $\delta = g(\boldsymbol{\theta})$ for a specified function g. For example, suppose we are sampling from a $N(\theta_1, \theta_2)$ distribution, where θ_2 is the variance. Let $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ and consider the two parameters $\delta_1 = g_1(\boldsymbol{\theta}) = \theta_1$ and $\delta_2 = g_2(\boldsymbol{\theta}) = \sqrt{\theta_2}$. Hence we are interested in best estimates of δ_1 and δ_2.

The Rao–Blackwell, Lehmann–Scheffé theory outlined in Sections 7.3 and 7.4 extends naturally to this vector case. Briefly, suppose $\delta = g(\boldsymbol{\theta})$ is the parameter of interest and $\mathbf{Y}$ is a vector of sufficient and complete statistics for $\boldsymbol{\theta}$. Let T be a statistic that is a function of $\mathbf{Y}$, such as $T = T(\mathbf{Y})$. If $E(T) = \delta$, then T is the unique MVUE of δ.

The remainder of our treatment of the case of several parameters is restricted to probability density functions that represent what we shall call regular cases of the exponential class. Here $m = p$.

Definition 7.7.2. *Let X be a random variable with pdf or pmf $f(x; \boldsymbol{\theta})$, where the vector of parameters $\boldsymbol{\theta} \in \Omega \subset R^m$. Let $\mathcal{S}$ denote the support of X. If X is continuous, assume that $\mathcal{S} = (a, b)$, where a or b may be $-\infty$ or ∞, respectively. If X is discrete, assume that $\mathcal{S} = \{a_1, a_2, \ldots\}$. Suppose $f(x; \boldsymbol{\theta})$ is of the form*

$$f(x; \boldsymbol{\theta}) = \begin{cases} \exp\left[\sum_{j=1}^m p_j(\boldsymbol{\theta}) K_j(x) + H(x) + q(\theta_1, \theta_2, \ldots, \theta_m)\right] & \text{for all } x \in \mathcal{S} \\ 0 & \text{elsewhere.} \end{cases} \tag{7.7.2}$$

Then we say this pdf or pmf is a member of the **exponential class**. *We say it is a* **regular case** *of the exponential family if, in addition,*

1. *the support does not depend on the vector of parameters* $\boldsymbol{\theta}$,
2. *the space* Ω *contains a nonempty, m-dimensional open rectangle,*
3. *the* $p_j(\boldsymbol{\theta})$, $j = 1, \ldots, m$, *are nontrivial, functionally independent, continuous functions of* $\boldsymbol{\theta}$,
4. *and, depending on whether* X *is continuous or discrete, one of the following holds, respectively:*
   (a) *if* X *is a continuous random variable, then the* m *derivatives* $K_j'(x)$, *for* $j = 1, 2, \ldots, m$, *are continuous for* $a < x < b$ *and no one is a linear homogeneous function of the others, and* $H(x)$ *is a continuous function of* x, $a < x < b$.
   (b) *if* X *is discrete, the* $K_j(x)$, $j = 1, 2, \ldots, m$, *are nontrivial functions of* x *on the support* $\mathcal{S}$ *and no one is a linear homogeneous function of the others.*

Let $X_1, \ldots, X_n$ be a random sample on X where the pdf or pmf of X is a regular case of the exponential class with the same notation as in Definition 7.7.2. It follows from (7.7.2) that the joint pdf or pmf of the sample is given by

$$\prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = \exp\left[\sum_{j=1}^{m} p_j(\boldsymbol{\theta}) \sum_{i=1}^{n} K_j(x_i) + nq(\boldsymbol{\theta})\right] \exp\left[\sum_{i=1}^{n} H(x_i)\right], \tag{7.7.3}$$

for all $x_i \in \mathcal{S}$. In accordance with the factorization theorem, the statistics

$$Y_1 = \sum_{i=1}^{n} K_1(x_i), \quad Y_2 = \sum_{i=1}^{n} K_2(x_i), \ldots, Y_m = \sum_{i=1}^{n} K_m(x_i)$$

are joint sufficient statistics for the m-dimensional vector of parameters $\boldsymbol{\theta}$. It is left as an exercise to prove that the joint pdf of $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ is of the form

$$R(\mathbf{y}) \exp\left[\sum_{j=1}^{m} p_j(\boldsymbol{\theta}) y_j + nq(\boldsymbol{\theta})\right], \tag{7.7.4}$$

at points of positive probability density. These points of positive probability density and the function $R(\mathbf{y})$ do not depend upon the vector of parameters $\boldsymbol{\theta}$. Moreover, in accordance with a theorem in analysis, it can be asserted that in a regular case of the exponential class, the family of probability density functions of these joint sufficient statistics $Y_1, Y_2, \ldots, Y_m$ is complete when $n > m$. In accordance with a convention previously adopted, we shall refer to $Y_1, Y_2, \ldots, Y_m$ as **joint complete sufficient statistics** for the vector of parameters $\boldsymbol{\theta}$.

Example 7.7.2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $N(\theta_1, \theta_2)$, $-\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$. Thus the pdf $f(x; \theta_1, \theta_2)$ of the distribution may be written as

$$f(x; \theta_1, \theta_2) = \exp\left(\frac{-1}{2\theta_2} x^2 + \frac{\theta_1}{\theta_2} x - \frac{\theta_1^2}{2\theta_2} - \ln\sqrt{2\pi\theta_2}\right).$$

Therefore, we can take $K_1(x) = x^2$ and $K_2(x) = x$. Consequently, the statistics

$$Y_1 = \sum_1^n X_i^2 \quad \text{and} \quad Y_2 = \sum_1^n X_i$$

are joint complete sufficient statistics for θ_1 and θ_2. Since the relations

$$Z_1 = \frac{Y_2}{n} = \overline{X}, \quad Z_2 = \frac{Y_1 - Y_2^2/n}{n-1} = \frac{\sum (X_i - \overline{X})^2}{n-1}$$

define a one-to-one transformation, Z_1 and Z_2 are also joint complete sufficient statistics for θ_1 and θ_2. Moreover,

$$E(Z_1) = \theta_1 \quad \text{and} \quad E(Z_2) = \theta_2.$$

From completeness, we have that Z_1 and Z_2 are the only functions of Y_1 and Y_2 that are unbiased estimators of θ_1 and θ_2, respectively. Hence Z_1 and Z_2 are the unique minimum variance estimators of θ_1 and θ_2, respectively. The MVUE of the standard deviation $\sqrt{\theta_2}$ is derived in Exercise 7.7.5. ■

In this section we have extended the concepts of sufficiency and completeness to the case where $\boldsymbol{\theta}$ is a p-dimensional vector. We now extend these concepts to the case where $\mathbf{X}$ is a k-dimensional random vector. We only consider the regular exponential class.

Suppose $\mathbf{X}$ is a k-dimensional random vector with pdf or pmf $f(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Omega \subset R^p$. Let $\mathcal{S} \subset R^k$ denote the support of $\mathbf{X}$. Suppose $f(\mathbf{x}; \boldsymbol{\theta})$ is of the form

$$f(\mathbf{x}; \boldsymbol{\theta}) = \begin{cases} \exp\left[\sum_{j=1}^m p_j(\boldsymbol{\theta}) K_j(\mathbf{x}) + H(\mathbf{x}) + q(\boldsymbol{\theta})\right] & \text{for all } \mathbf{x} \in \mathcal{S} \\ 0 & \text{elsewhere.} \end{cases} \tag{7.7.5}$$

Then we say this pdf or pmf is a member of the **exponential class**. If, in addition, $p = m$, the support does not depend on the vector of parameters $\boldsymbol{\theta}$, and conditions similar to those of Definition 7.7.2 hold, then we say this pdf is a **regular case** of the exponential class.

Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ constitute a random sample on $\mathbf{X}$. Then the statistics,

$$Y_j = \sum_{i=1}^n K_j(\mathbf{X}_i), \quad \text{for } j = 1, \ldots, m, \tag{7.7.6}$$

are sufficient and complete statistics for $\boldsymbol{\theta}$. Let $\mathbf{Y} = (Y_1, \ldots, Y_m)'$. Suppose $\delta = g(\boldsymbol{\theta})$ is a parameter of interest. If $T = h(\mathbf{Y})$ for some function h and $E(T) = \delta$ then T is the unique minimum variance unbiased estimator of δ.

Example 7.7.3 (Multinomial)**.** In Example 6.4.5, we consider the mles of the multinomial distribution. In this example we determine the MVUEs of several of the parameters. As in Example 6.4.5, consider a random trial that can result in one, and only one, of k outcomes or categories. Let X_j be 1 or 0 depending on whether the jth outcome does or does not occur, for $j = 1, \ldots, k$. Suppose the probability

that outcome j occurs is p_j; hence, $\sum_{j=1}^{k} p_j = 1$. Let $\mathbf{X} = (X_1, \ldots, X_{k-1})'$ and $\mathbf{p} = (p_1, \ldots, p_{k-1})'$. The distribution of $\mathbf{X}$ is multinomial and can be found in expression (6.4.18), which can be reexpressed as

$$f(\mathbf{x}, \mathbf{p}) = \exp\left\{\sum_{j=1}^{k-1}\left(\log\left[\frac{p_j}{1-\sum_{i\neq k} p_i}\right]\right)x_j + \log\left(1-\sum_{i\neq k} p_i\right)\right\}.$$

Because this a regular case of the exponential family, the following statistics, resulting from a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from the distribution of $\mathbf{X}$, are jointly sufficient and complete for the parameters $\mathbf{p} = (p_1, \ldots, p_{k-1})'$:

$$Y_j = \sum_{i=1}^{n} X_{ij}, \quad \text{for } j = 1, \ldots, k-1.$$

Each random variable X_{ij} is Bernoulli with parameter p_j and the variables X_{ij} are independent for $i = 1, \ldots, n$. Hence the variables Y_j are binomial(n, p_j) for $j = 1, \ldots, k$. Thus the MVUE of p_j is the statistic $n^{-1}Y_j$.

Next, we shall find the MVUE of $p_j p_l$, for $j \neq l$. Exercise 7.7.8 shows that the mle of $p_j p_l$ is $n^{-2}Y_j Y_l$. Recall from Section 3.1 that the conditional distribution of Y_j, given Y_l, is $b[n - Y_l, p_j/(1-p_l)]$. As an initial guess at the MVUE, consider the mle, which, as shown by Exercise 7.7.8, is $n^{-2}Y_j Y_l$. Hence

$$\begin{aligned}
E[n^{-2}Y_j Y_l] &= \frac{1}{n^2}E[E(Y_j Y_l|Y_l)] = \frac{1}{n^2}E[Y_l E(Y_j|Y_l)] \\
&= \frac{1}{n^2}E\left[Y_l(n-Y_l)\frac{p_j}{1-p_l}\right] = \frac{1}{n^2}\frac{p_j}{1-p_l}\{E[nY_l] - E[Y_l^2]\} \\
&= \frac{1}{n^2}\frac{p_j}{1-p_l}\{n^2 p_l - np_l(1-p_l) - n^2 p_l^2\} \\
&= \frac{1}{n^2}\frac{p_j}{1-p_l}np_l(n-1)(1-p_l) = \frac{(n-1)}{n}p_j p_l.
\end{aligned}$$

Hence the MVUE of $p_j p_l$ is $\frac{1}{n(n-1)}Y_j Y_l$. ■

Example 7.7.4 (Multivariate Normal)**.** Let $\mathbf{X}$ have the multivariate normal distribution $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a positive definite $k \times k$ matrix. The pdf of $\mathbf{X}$ is given in expression (3.5.16). In this case $\boldsymbol{\theta}$ is a $\{k + [k(k+1)/2]\}$-dimensional vector whose first k components consist of the mean vector $\boldsymbol{\mu}$ and whose last $\frac{k(k+1)}{2}$ components consist of the componentwise variances σ_i^2 and the covariances σ_{ij}, for $j \geq i$. The density of $\mathbf{X}$ can be written as

$$f_{\mathbf{X}}(\mathbf{x}) = \exp\left\{-\frac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{k}{2}\log 2\pi\right\}, \tag{7.7.7}$$

for $\mathbf{x} \in R^k$. Hence, by (7.7.5), the multivariate normal pdf is a regular case of the exponential class of distributions. We need only identify the functions $K(\mathbf{x})$. The second term in the exponent on the right side of (7.7.7) can be written as $(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1})\mathbf{x}$;

hence, $K_1(\mathbf{x}) = \mathbf{x}$. The first term is easily seen to be a linear combination of the products $x_i x_j$, $i, j = 1, 2, \ldots, k$, which are the entries of the matrix $\mathbf{x}\mathbf{x}'$. Hence we can take $K_2(\mathbf{x}) = \mathbf{x}\mathbf{x}'$. Now, let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample on $\mathbf{X}$. Based on (7.7.7) then, a set of sufficient and complete statistics is given by

$$\mathbf{Y}_1 = \sum_{i=1}^{n} \mathbf{X}_i \text{ and } \mathbf{Y}_2 = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'. \tag{7.7.8}$$

Note that $\mathbf{Y}_1$ is a vector of k statistics and that $\mathbf{Y}_2$ is a $k \times k$ symmetric matrix. Because the matrix is symmetric, we can eliminate the bottom-half [elements (i, j) with $i > j$] of the matrix, which results in $\{k + [k(k+1)]\}$ complete sufficient statistics, i.e., as many complete sufficient statistics as there are parameters.

Based on marginal distributions, it is easy to show that $\overline{X}_j = n^{-1} \sum_{i=1}^{n} X_{ij}$ is the MVUE of μ_j and that $(n-1)^{-1} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2$ is the MVUE of σ_j^2. The MVUEs of the covariance parameters are obtained in Exercise 7.7.9. ■

For our last example, we consider a case where the set of parameters is the cdf.

Example 7.7.5. Let $X_1, X_2, \ldots, X_n$ be a random sample having the common continuous cdf $F(x)$. Let $Y_1 < Y_2 < \cdots < Y_n$ denote the corresponding order statistics. Note that given $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, the conditional distribution of $X_1, X_2, \ldots, X_n$ is discrete with probability $\frac{1}{n!}$ on each of the $n!$ permutations of the vector $(y_1, y_2, \ldots, y_n)$, [because $F(x)$ is continuous, we can assume that each of the values $y_1, y_2, \ldots, y_n$ is distinct]. That is, the conditional distribution does not depend on $F(x)$. Hence, by the definition of sufficiency, the order statistics are sufficient for $F(x)$. Furthermore, while the proof is beyond the scope of this book, it can be shown that the order statistics are also complete; see page 72 of Lehmann and Casella (1998).

Let $T = T(x_1, x_2, \ldots, x_n)$ be any statistic that is *symmetric in its arguments*; i.e., $T(x_1, x_2, \ldots, x_n) = T(x_{i_1}, x_{i_2}, \ldots, x_{i_n})$ for any permutation $(x_{i_1}, x_{i_2}, \ldots, x_{i_n})$ of $(x_1, x_2, \ldots, x_n)$. Then T is a function of the order statistics. This is useful in determining MVUEs for this situation; see Exercises 7.7.12 and 7.7.13. ■

EXERCISES

7.7.1. Let $Y_1 < Y_2 < Y_3$ be the order statistics of a random sample of size 3 from the distribution with pdf

$$f(x; \theta_1, \theta_2) = \begin{cases} \frac{1}{\theta_2} \exp\left(-\frac{x-\theta_1}{\theta_2}\right) & \theta_1 < x < \infty, \ -\infty < \theta_1 < \infty, \ 0 < \theta_2 < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

Find the joint pdf of $Z_1 = Y_1, Z_2 = Y_2$, and $Z_3 = Y_1 + Y_2 + Y_3$. The corresponding transformation maps the space $\{(y_1, y_2, y_3) : \theta_1 < y_1 < y_2 < y_3 < \infty\}$ onto the space

$$\{(z_1, z_2, z_3) : \theta_1 < z_1 < z_2 < (z_3 - z_1)/2 < \infty\}.$$

Show that Z_1 and Z_3 are joint sufficient statistics for θ_1 and θ_2.

7.7.2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that has a pdf of the form (7.7.2) of this section. Show that $Y_1 = \sum_{i=1}^{n} K_1(X_i), \ldots, Y_m = \sum_{i=1}^{m} K_m(X_i)$ have a joint pdf of the form (7.7.4) of this section.

7.7.3. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ denote a random sample of size n from a bivariate normal distribution with means μ_1 and μ_2, positive variances σ_1^2 and σ_2^2, and correlation coefficient ρ. Show that $\sum_1^n X_i$, $\sum_1^n Y_i$, $\sum_1^n X_i^2$, $\sum_1^n Y_i^2$, and $\sum_1^n X_i Y_i$ are joint complete sufficient statistics for the five parameters. Are $\overline{X} = \sum_1^n X_i/n$, $\overline{Y} = \sum_1^n Y_i/n$, $S_1^2 = \sum_1^n (X_i - \overline{X})^2/(n-1)$, $S_2^2 = \sum_1^n (Y_i - \overline{Y})^2/(n-1)$, and $\sum_1^n (X_i - \overline{X})(Y_i - \overline{Y})/(n-1)S_1 S_2$ also joint complete sufficient statistics for these parameters?

7.7.4. Let the pdf $f(x; \theta_1, \theta_2)$ be of the form

$$\exp[p_1(\theta_1, \theta_2)K_1(x) + p_2(\theta_1, \theta_2)K_2(x) + H(x) + q_1(\theta_1, \theta_2)], \quad a < x < b,$$

zero elsewhere. Suppose that $K_1'(x) = cK_2'(x)$. Show that $f(x; \theta_1, \theta_2)$ can be written in the form

$$\exp[p(\theta_1, \theta_2)K_2(x) + H(x) + q(\theta_1, \theta_2)], \quad a < x < b,$$

zero elsewhere. This is the reason why it is required that no one $K_j'(x)$ be a linear homogeneous function of the others, that is, so that the number of sufficient statistics equals the number of parameters.

7.7.5. In Example 7.7.2:

(a) Find the MVUE of the standard deviation $\sqrt{\theta_2}$.

(b) Modify the R function `bootse1.R` so that it returns the estimate in (a) and its bootstrap standard error. Run it on the Bavarian forest data discussed in Example 4.1.3, where the response is the concentration of sulfur dioxide. Using 3,000 bootstraps, report the estimate and its bootstrap standard error.

7.7.6. Let $X_1, X_2, \ldots, X_n$ be a random sample from the uniform distribution with pdf $f(x; \theta_1, \theta_2) = 1/(2\theta_2)$, $\theta_1 - \theta_2 < x < \theta_1 + \theta_2$, where $-\infty < \theta_1 < \infty$ and $\theta_2 > 0$, and the pdf is equal to zero elsewhere.

(a) Show that $Y_1 = \min(X_i)$ and $Y_n = \max(X_i)$, the joint sufficient statistics for θ_1 and θ_2, are complete.

(b) Find the MVUEs of θ_1 and θ_2.

7.7.7. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\theta_1, \theta_2)$.

(a) If the constant b is defined by the equation $P(X \leq b) = p$ where p is specified, find the mle and the MVUE of b.

(b) Modify the R function `bootse1.R` so that it returns the MVUE of Part (a) and its bootstrap standard error.

(c) Run your function in Part (b) on the data set discussed in Example 7.6.4 for $p = 0.75$ and 3,000 bootstraps.

7.7.8. In the notation of Example 7.7.3, show that the mle of $p_j p_l$ is $n^{-2}Y_jY_l$.

7.7.9. Refer to Example 7.7.4 on sufficiency for the multivariate normal model.

(a) Determine the MVUE of the covariance parameters σ_{ij}.

(b) Let $g = \sum_{i=1}^{k} a_i\mu_i$, where $a_1, \ldots, a_k$ are specified constants. Find the MVUE for g.

7.7.10. In a personal communication, LeRoy Folks noted that the inverse Gaussian pdf

$$f(x;\theta_1,\theta_2) = \left(\frac{\theta_2}{2\pi x^3}\right)^{1/2} \exp\left[\frac{-\theta_2(x-\theta_1)^2}{2\theta_1^2 x}\right], \quad 0 < x < \infty, \tag{7.7.9}$$

where $\theta_1 > 0$ and $\theta_2 > 0$, is often used to model lifetimes. Find the complete sufficient statistics for (θ_1, θ_2) if $X_1, X_2, \ldots, X_n$ is a random sample from the distribution having this pdf.

7.7.11. Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\theta_1, \theta_2)$ distribution.

(a) Show that $E[(X_1 - \theta_1)^4] = 3\theta_2^2$.

(b) Find the MVUE of $3\theta_2^2$.

7.7.12. Let $X_1, \ldots, X_n$ be a random sample from a distribution of the continuous type with cdf $F(x)$. Suppose the mean, $\mu = E(X_1)$, exists. Using Example 7.7.5, show that the sample mean, $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$, is the MVUE of μ.

7.7.13. Let $X_1, \ldots, X_n$ be a random sample from a distribution of the continuous type with cdf $F(x)$. Let $\theta = P(X_1 \leq a) = F(a)$, where a is known. Show that the proportion $n^{-1}\#\{X_i \leq a\}$ is the MVUE of θ.

7.8 Minimal Sufficiency and Ancillary Statistics

In the study of statistics, it is clear that we want to reduce the data contained in the entire sample as much as possible without losing relevant information about the important characteristics of the underlying distribution. That is, a large collection of numbers in the sample is not as meaningful as a few good summary statistics of those data. Sufficient statistics, if they exist, are valuable because we know that the statisticians with those summary measures have as much information as the statistician with the entire sample. Sometimes, however, there are several sets of joint sufficient statistics, and thus we would like to find the simplest one of these sets. For illustration, in a sense, the observations $X_1, X_2, \ldots, X_n,\ n > 2$, of a random sample from $N(\theta_1, \theta_2)$ could be thought of as joint sufficient statistics for θ_1 and θ_2. We know, however, that we can use $\overline{X}$ and S^2 as joint sufficient statistics for those parameters, which is a great simplification over using $X_1, X_2, \ldots, X_n$, particularly if n is large.

In most instances in this chapter, we have been able to find a single sufficient statistic for one parameter or two joint sufficient statistics for two parameters. Possibly the most complicated cases considered so far are given in Example 7.7.3, in

which we find $k+k(k+1)/2$ joint sufficient statistics for $k+k(k+1)/2$ parameters; or the multivariate normal distribution given in Example 7.7.4; or in the use the order statistics of a random sample for some completely unknown distribution of the continuous type as in Example 7.7.5.

What we would like to do is to change from one set of joint sufficient statistics to another, always reducing the number of statistics involved until we cannot go any further without losing the sufficiency of the resulting statistics. Those statistics that are there at the end of this reduction are called **minimal sufficient statistics**. These are sufficient for the parameters and are functions of every other set of sufficient statistics for those same parameters. Often, if there are k parameters, we can find k joint sufficient statistics that are minimal. In particular, if there is one parameter, we can often find a single sufficient statistic that is minimal. Most of the earlier examples that we have considered illustrate this point, but this is not always the case, as shown by the following example.

Example 7.8.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the uniform distribution over the interval $(\theta-1, \theta+1)$ having pdf

$$f(x;\theta) = (\tfrac{1}{2})I_{(\theta-1,\theta+1)}(x), \quad \text{where} -\infty < \theta < \infty.$$

The joint pdf of $X_1, X_2, \ldots, X_n$ equals the product of $(\frac{1}{2})^n$ and certain indicator functions, namely,

$$\left(\frac{1}{2}\right)^n \prod_{i=1}^{n} I_{(\theta-1,\theta+1)}(x_i) = \left(\frac{1}{2}\right)^n \{I_{(\theta-1,\theta+1)}[\min(x_i)]\}\{I_{(\theta-1,\theta+1)}[\max(x_i)]\},$$

because $\theta - 1 < \min(x_i) \leq x_j \leq \max(x_i) < \theta + 1,\ j = 1, 2, \ldots, n$. Thus the order statistics $Y_1 = \min(X_i)$ and $Y_n = \max(X_i)$ are the sufficient statistics for θ. These two statistics actually are minimal for this one parameter, as we cannot reduce the number of them to less than two and still have sufficiency. ■

There is an observation that helps us see that almost all the sufficient statistics that we have studied thus far are minimal. We have noted that the mle $\hat{\theta}$ of θ is a function of one or more sufficient statistics, when the latter exists. Suppose that this mle $\hat{\theta}$ is also sufficient. Since this sufficient statistic $\hat{\theta}$ is a function of the other sufficient statistics, by Theorem 7.3.2, it must be minimal. For example, we have

1. The mle $\hat{\theta} = \overline{X}$ of θ in $N(\theta, \sigma^2)$, σ^2 known, is a minimal sufficient statistic for θ.

2. The mle $\hat{\theta} = \overline{X}$ of θ in a Poisson distribution with mean θ is a minimal sufficient statistic for θ.

3. The mle $\hat{\theta} = Y_n = \max(X_i)$ of θ in the uniform distribution over $(0, \theta)$ is a minimal sufficient statistic for θ.

4. The maximum likelihood estimators $\hat{\theta}_1 = \overline{X}$ and $\hat{\theta}_2 = [(n-1)/n]S^2$ of θ_1 and θ_2 in $N(\theta_1, \theta_2)$ are joint minimal sufficient statistics for θ_1 and θ_2.

From these examples we see that the minimal sufficient statistics do not need to be unique, for any one-to-one transformation of them also provides minimal sufficient statistics. The linkage between minimal sufficient statistics and the mle, however, does not hold in many interesting instances. We illustrate this in the next two examples.

Example 7.8.2. Consider the model given in Example 7.8.1. There we noted that $Y_1 = \min(X_i)$ and $Y_n = \max(X_i)$ are joint sufficient statistics. Also, we have

$$\theta - 1 < Y_1 < Y_n < \theta + 1$$

or, equivalently,

$$Y_n - 1 < \theta < Y_1 + 1.$$

Hence, to maximize the likelihood function so that it equals $(\frac{1}{2})^n$, θ can be any value between $Y_n - 1$ and $Y_1 + 1$. For example, many statisticians take the mle to be the mean of these two endpoints, namely,

$$\hat{\theta} = \frac{Y_n - 1 + Y_1 + 1}{2} = \frac{Y_1 + Y_n}{2},$$

which is the midrange. We recognize, however, that this mle is not unique. Some might argue that since $\hat{\theta}$ is an mle of θ and since it is a function of the joint sufficient statistics, Y_1 and Y_n, for θ, it is a minimal sufficient statistic. This is not the case at all, for $\hat{\theta}$ is not even sufficient. Note that the mle must itself be a sufficient statistic for the parameter before it can be considered the minimal sufficient statistic. ■

Note that we can model the situation in the last example by

$$X_i = \theta + W_i, \tag{7.8.1}$$

where $W_1, W_2, \ldots, W_n$ are iid with the common uniform$(-1, 1)$ pdf. Hence this is an example of a location model. We discuss these models in general next.

Example 7.8.3. Consider a location model given by

$$X_i = \theta + W_i, \tag{7.8.2}$$

where $W_1, W_2, \ldots, W_n$ are iid with the common pdf $f(w)$ and common continuous cdf $F(w)$. From Example 7.7.5, we know that the order statistics $Y_1 < Y_2 < \cdots < Y_n$ are a set of complete and sufficient statistics for this situation. Can we obtain a smaller set of minimal sufficient statistics? Consider the following four situations:

(a) Suppose $f(w)$ is the $N(0, 1)$ pdf. Then we know that $\overline{X}$ is both the MVUE and mle of θ. Also, $\overline{X} = n^{-1} \sum_{i=1}^{n} Y_i$, i.e., a function of the order statistics. Hence $\overline{X}$ is minimal sufficient.

(b) Suppose $f(w) = \exp\{-w\}$, for $w > 0$, zero elsewhere. Then the statistic Y_1 is a sufficient statistic as well as the mle, and thus is minimal sufficient.

(c) Suppose $f(w)$ is the logistic pdf. As discussed in Example 6.1.2, the mle of θ exists and it is easy to compute. As shown on page 38 of Lehmann and Casella (1998), though, the order statistics are minimal sufficient for this situation. That is, no reduction is possible.

(d) Suppose $f(w)$ is the Laplace pdf. It was shown in Example 6.1.1 that the median, Q_2 is the mle of θ, but it is not a sufficient statistic. Further, similar to the logistic pdf, it can be shown that the order statistics are minimal sufficient for this situation. ■

In general, the situation described in parts (c) and (d), where the mle is obtained rather easily while the set of minimal sufficient statistics is the set of order statistics and no reduction is possible, is the norm for location models.

There is also a relationship between a minimal sufficient statistic and completeness that is explained more fully in Lehmann and Scheffé (1950). Let us say simply and without explanation that for the cases in this book, complete sufficient statistics are minimal sufficient statistics. The converse is not true, however, by noting that in Example 7.8.1, we have

$$E\left[\frac{Y_n - Y_1}{2} - \frac{n-1}{n+1}\right] = 0, \quad \text{for all } \theta.$$

That is, there is a nonzero function of those minimal sufficient statistics, Y_1 and Y_n, whose expectation is zero for all θ.

There are other statistics that almost seem opposites of sufficient statistics. That is, while sufficient statistics contain all the information about the parameters, these other statistics, called **ancillary statistics**, have distributions free of the parameters and seemingly contain no information about those parameters. As an illustration, we know that the variance S^2 of a random sample from $N(\theta, 1)$ has a distribution that does not depend upon θ and hence is an ancillary statistic. Another example is the ratio $Z = X_1/(X_1 + X_2)$, where X_1, X_2 is a random sample from a gamma distribution with known parameter $\alpha > 0$ and unknown parameter $\beta = \theta$, because Z has a beta distribution that is free of θ. There are many examples of ancillary statistics, and we provide some rules that make them rather easy to find with certain models, which we present in the next three examples.

Example 7.8.4 (Location-Invariant Statistics)**.** In Example 7.8.3, we introduced the location model. Recall that a random sample $X_1, X_2, \ldots, X_n$ follows this model if

$$X_i = \theta + W_i, \quad i = 1, \ldots, n, \tag{7.8.3}$$

where $-\infty < \theta < \infty$ is a parameter and $W_1, W_2, \ldots, W_n$ are iid random variables with the pdf $f(w)$, which does not depend on θ. Then the common pdf of X_i is $f(x - \theta)$.

Let $Z = u(X_1, X_2, \ldots, X_n)$ be a statistic such that

$$u(x_1 + d, x_2 + d, \ldots, x_n + d) = u(x_1, x_2, \ldots, x_n),$$

for all real d. Hence

$$Z = u(W_1 + \theta, W_2 + \theta, \ldots, W_n + \theta) = u(W_1, W_2, \ldots, W_n)$$

is a function of $W_1, W_2, \ldots, W_n$ alone (not of θ). Hence Z must have a distribution that does not depend upon θ. We call $Z = u(X_1, X_2, \ldots, X_n)$ a **location-invariant statistic**.

Assuming a location model, the following are some examples of location-invariant statistics: the sample variance $= S^2$, the sample range $= \max\{X_i\} - \min\{X_i\}$, the mean deviation from the sample median $= (1/n)\sum |X_i - \text{median}(X_i)|$, $X_1 + X_2 - X_3 - X_4$, $X_1 + X_3 - 2X_2$, $(1/n)\sum[X_i - \min(X_i)]$, and so on. To see that the range is location-invariant, note that

$$\begin{aligned} \max\{X_i\} - \theta &= \max\{X_i - \theta\} = \max\{W_i\} \\ \min\{X_i\} - \theta &= \min\{X_i - \theta\} = \min\{W_i\}. \end{aligned}$$

So,

$$\text{range} = \max\{X_i\} - \min\{X_i\} = \max\{X_i\} - \theta - (\min\{X_i\} - \theta) = \max\{W_i\} - \min\{W_i\}.$$

Hence the distribution of the range only depends on the distribution of the W_is and, thus, it is location-invariant. For the location invariance of other statistics, see Exercise 7.8.4. ■

Example 7.8.5 (Scale-Invariant Statistics)**.** Consider a random sample $X_1, \ldots, X_n$ that follows a **scale model**, i.e., a model of the form

$$X_i = \theta W_i, \quad i = 1, \ldots, n, \tag{7.8.4}$$

where $\theta > 0$ and $W_1, W_2, \ldots, W_n$ are iid random variables with pdf $f(w)$, which does not depend on θ. Then the common pdf of X_i is $\theta^{-1} f(x/\theta)$. We call θ a scale parameter. Suppose that $Z = u(X_1, X_2, \ldots, X_n)$ is a statistic such that

$$u(cx_1, cx_2, \ldots, cx_n) = u(x_1, x_2, \ldots, x_n)$$

for all $c > 0$. Then

$$Z = u(X_1, X_2, \ldots, X_n) = u(\theta W_1, \theta W_2, \ldots, \theta W_n) = u(W_1, W_2, \ldots, W_n).$$

Since neither the joint pdf of $W_1, W_2, \ldots, W_n$ nor Z contains θ, the distribution of Z must not depend upon θ. We say that Z is a **scale-invariant statistic**.

The following are some examples of scale-invariant statistics: $X_1/(X_1 + X_2)$, $X_1^2/\sum_1^n X_i^2$, $\min(X_i)/\max(X_i)$, and so on. The scale invariance of the first statistic follows from

$$\frac{X_1}{X_1 + X_2} = \frac{(\theta X_1)/\theta}{[(\theta X_1) + (\theta X_2)]/\theta} = \frac{W_1}{W_1 + W_2}.$$

The scale invariance of the other statistics is asked for in Exercise 7.8.5. ■

Example 7.8.6 (Location- and Scale-Invariant Statistics)**.** Finally, consider a random sample $X_1, X_2, \ldots, X_n$ that follows a location and scale model as in Example 7.7.5. That is,

$$X_i = \theta_1 + \theta_2 W_i, \quad i = 1, \ldots, n, \tag{7.8.5}$$

where W_i are iid with the common pdf $f(t)$ which is free of θ_1 and θ_2. In this case, the pdf of X_i is $\theta_2^{-1} f((x - \theta_1)/\theta_2)$. Consider the statistic $Z = u(X_1, X_2, \ldots, X_n)$, where

$$u(cx_1 + d, \ldots, cx_n + d) = u(x_1, \ldots, x_n).$$

Then

$$Z = u(X_1, \ldots, X_n) = u(\theta_1 + \theta_2 W_1, \ldots, \theta_1 + \theta_2 W_n) = u(W_1, \ldots, W_n).$$

Since neither the joint pdf of $W_1, \ldots, W_n$ nor Z contains θ_1 and θ_2, the distribution of Z must not depend upon θ_1 nor θ_2. Statistics such as $Z = u(X_1, X_2, \ldots, X_n)$ are called **location- and scale-invariant statistics**. The following are four examples of such statistics:

(a) $T_1 = [\max(X_i) - \min(X_i)]/S$;

(b) $T_2 = \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2 / S^2$;

(c) $T_3 = (X_i - \overline{X})/S$;

(d) $T_4 = |X_i - X_j|/S,, ; i \neq j$.

Let $\overline{X} - \theta_1 = n^{-1} \sum_{i=1}^{n} (X_i - \theta_1)$. Then the location and scale invariance of the statistic in (d) follows from the two identities

$$\begin{aligned} S^2 &= \theta_2^2 \sum_{i=1}^{n} \left[\frac{X_i - \theta_1}{\theta_2} - \frac{\overline{X} - \theta_1}{\theta_2} \right]^2 = \theta_2^2 \sum_{i=1}^{n} (W_i - \overline{W})^2 \\ X_i - X_j &= \theta_2 \left[\frac{X_i - \theta_1}{\theta_2} - \frac{X_j - \theta_1}{\theta_2} \right] = \theta_2 (W_i - W_j). \end{aligned}$$

See Exercise 7.8.6 for the other statistics. ■

Thus, these location-invariant, scale-invariant, and location- and scale-invariant statistics provide good illustrations, with the appropriate model for the pdf, of ancillary statistics. Since an ancillary statistic and a complete (minimal) sufficient statistic are such opposites, we might believe that there is, in some sense, no relationship between the two. This is true, and in the next section we show that they are independent statistics.

EXERCISES

7.8.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from each of the following distributions involving the parameter θ. In each case find the mle of θ and show that it is a sufficient statistic for θ and hence a minimal sufficient statistic.

(a) $b(1, \theta)$, where $0 \leq \theta \leq 1$.

(b) Poisson with mean $\theta > 0$.

(c) Gamma with $\alpha = 3$ and $\beta = \theta > 0$.

(d) $N(\theta, 1)$, where $-\infty < \theta < \infty$.

(e) $N(0, \theta)$, where $0 < \theta < \infty$.

7.8.2. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from the uniform distribution over the closed interval $[-\theta, \theta]$ having pdf $f(x; \theta) = (1/2\theta) I_{[-\theta,\theta]}(x)$.

(a) Show that Y_1 and Y_n are joint sufficient statistics for θ.

(b) Argue that the mle of θ is $\hat{\theta} = \max(-Y_1, Y_n)$.

(c) Demonstrate that the mle $\hat{\theta}$ is a sufficient statistic for θ and thus is a minimal sufficient statistic for θ.

7.8.3. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from a distribution with pdf

$$f(x; \theta_1, \theta_2) = \left(\frac{1}{\theta_2}\right) e^{-(x-\theta_1)/\theta_2} I_{(\theta_1,\infty)}(x),$$

where $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. Find the joint minimal sufficient statistics for θ_1 and θ_2.

7.8.4. Continuing with Example 7.8.4, show that the following statistics are location-invariant:

(a) The sample variance $= S^2$.

(b) The mean deviation from the sample median $= (1/n) \sum |X_i - \text{median}(X_i)|$.

(c) $(1/n) \sum [X_i - \min(X_i)]$.

7.8.5. In Example 7.8.5, a scale model was presented and scale invariance was defined. Using the notation of this example, show that the following statistics are scale-invariant:

(a) $X_1^2 / \sum_1^n X_i^2$.

(b) $\min\{X_i\} / \max\{X_i\}$.

7.8.6. Obtain the location and scale invariance of the other statistics listed in Example 7.8.6, i.e., the statistics

(a) $T_1 = [\max(X_i) - \min(X_i)]/S$.

(b) $T_2 = \sum_{i=1}^{n-1}(X_{i+1} - X_i)^2/S^2$.

(c) $T_3 = (X_i - \overline{X})/S$.

7.8.7. With random samples from each of the distributions given in Exercises 7.8.1(d), 7.8.2, and 7.8.3, define at least two ancillary statistics that are different from the examples given in the text. These examples illustrate, respectively, location-invariant, scale-invariant, and location- and scale-invariant statistics.

7.9 Sufficiency, Completeness, and Independence

We have noted that if we have a sufficient statistic Y_1 for a parameter θ, $\theta \in \Omega$, then $h(z|y_1)$, the conditional pdf of another statistic Z, given $Y_1 = y_1$, does not depend upon θ. If, moreover, Y_1 and Z are independent, the pdf $g_2(z)$ of Z is such that $g_2(z) = h(z|y_1)$, and hence $g_2(z)$ must not depend upon θ either. So the independence of a statistic Z and the sufficient statistic Y_1 for a parameter θ imply that the distribution of Z does not depend upon $\theta \in \Omega$. That is, Z is an ancillary statistic.

It is interesting to investigate a converse of that property. Suppose that the distribution of an ancillary statistic Z does not depend upon θ; then are Z and the sufficient statistic Y_1 for θ independent? To begin our search for the answer, we know that the joint pdf of Y_1 and Z is $g_1(y_1;\theta)h(z|y_1)$, where $g_1(y_1;\theta)$ and $h(z|y_1)$ represent the marginal pdf of Y_1 and the conditional pdf of Z given $Y_1 = y_1$, respectively. Thus the marginal pdf of Z is

$$\int_{-\infty}^{\infty} g_1(y_1;\theta)h(z|y_1)\,dy_1 = g_2(z),$$

which, by hypothesis, does not depend upon θ. Because

$$\int_{-\infty}^{\infty} g_2(z)g_1(y_1;\theta)\,dy_1 = g_2(z),$$

if follows, by taking the difference of the last two integrals, that

$$\int_{-\infty}^{\infty} [g_2(z) - h(z|y_1)]g_1(y_1;\theta)\,dy_1 = 0 \tag{7.9.1}$$

for all $\theta \in \Omega$. Since Y_1 is sufficient statistic for θ, $h(z|y_1)$ does not depend upon θ. By assumption, $g_2(z)$ and hence $g_2(z) - h(z|y_1)$ do not depend upon θ. Now if the family $\{g_1(y_1;\theta) : \theta \in \Omega\}$ is complete, Equation (7.9.1) would require that

$$g_2(z) - h(z|y_1) = 0 \quad \text{or} \quad g_2(z) = h(z|y_1).$$

That is, the joint pdf of Y_1 and Z must be equal to

$$g_1(y_1;\theta)h(z|y_1) = g_1(y_1;\theta)g_2(z).$$

Accordingly, Y_1 and Z are independent, and we have proved the following theorem, which was considered in special cases by Neyman and Hogg and proved in general by Basu.

Theorem 7.9.1. *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution having a pdf $f(x;\theta)$, $\theta \in \Omega$, where Ω is an interval set. Suppose that the statistic Y_1 is a complete and sufficient statistic for θ. Let $Z = u(X_1, X_2, \ldots, X_n)$ be any other statistic (not a function of Y_1 alone). If the distribution of Z does not depend upon θ, then Z is independent of the sufficient statistic Y_1.*

In the discussion above, it is interesting to observe that if Y_1 is a sufficient statistic for θ, then the independence of Y_1 and Z implies that the distribution of Z does not depend upon θ whether $\{g_1(y_1;\theta) : \theta \in \Omega\}$ is or is not complete. Conversely, to prove the independence from the fact that $g_2(z)$ does not depend upon θ, we definitely need the completeness. Accordingly, if we are dealing with situations in which we know that family $\{g_1(y_1;\theta) : \theta \in \Omega\}$ is complete (such as a regular case of the exponential class), we can say that the statistic Z is independent of the sufficient statistic Y_1 if and only if the distribution of Z does not depend upon θ(i.e., Z is an ancillary statistic).

It should be remarked that the theorem (including the special formulation of it for regular cases of the exponential class) extends immediately to probability density functions that involve m parameters for which there exist m joint sufficient statistics. For example, let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution having the pdf $f(x;\theta_1,\theta_2)$ that represents a regular case of the exponential class so that there are two joint complete sufficient statistics for θ_1 and θ_2. Then any other statistic $Z = u(X_1, X_2, \ldots, X_n)$ is independent of the joint complete sufficient statistics if and only if the distribution of Z does not depend upon θ_1 or θ_2.

We present an example of the theorem that provides an alternative proof of the independence of $\overline{X}$ and S^2, the mean and the variance of a random sample of size n from a distribution that is $N(\mu,\sigma^2)$. This proof is given as if we were unaware that $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$, because that fact and the independence were established in Theorem 3.6.1.

Example 7.9.1. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution that is $N(\mu,\sigma^2)$. We know that the mean $\overline{X}$ of the sample is, for every known σ^2, a complete sufficient statistic for the parameter μ, $-\infty < \mu < \infty$. Consider the statistic

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

which is location-invariant. Thus S^2 must have a distribution that does not depend upon μ; and hence, by the theorem, S^2 and $\overline{X}$, the complete sufficient statistic for μ, are independent. ■

Example 7.9.2. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the distribution having pdf

$$\begin{aligned} f(x;\theta) &= \exp\{-(x-\theta)\}, \quad \theta < x < \infty, \quad -\infty < \theta < \infty, \\ &= 0 \quad \text{elsewhere.} \end{aligned}$$

Here the pdf is of the form $f(x-\theta)$, where $f(w) = e^{-w}$, $0 < w < \infty$, zero elsewhere. Moreover, we know (Exercise 7.4.5) that the first order statistic $Y_1 = \min(X_i)$ is a

complete sufficient statistic for θ. Hence Y_1 must be independent of each location-invariant statistic $u(X_1, X_2, \ldots, X_n)$, enjoying the property that

$$u(x_1 + d, x_2 + d, \ldots, x_n + d) = u(x_1, x_2, \ldots, x_n)$$

for all real d. Illustrations of such statistics are S^2, the sample range, and

$$\frac{1}{n}\sum_{i=1}^{n}[X_i - \min(X_i)]. \quad \blacksquare$$

Example 7.9.3. Let X_1, X_2 denote a random sample of size $n = 2$ from a distribution with pdf

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta}e^{-x/\theta}, \quad 0 < x < \infty, \ \ 0 < \theta < \infty, \\ &= 0 \quad \text{elsewhere.} \end{aligned}$$

The pdf is of the form $(1/\theta)f(x/\theta)$, where $f(w) = e^{-w}$, $0 < w < \infty$, zero elsewhere. We know that $Y_1 = X_1 + X_2$ is a complete sufficient statistic for θ. Hence, Y_1 is independent of every scale-invariant statistic $u(X_1, X_2)$ with the property $u(cx_1, cx_2) = u(x_1, x_2)$. Illustrations of these are X_1/X_2 and $X_1/(X_1 + X_2)$, statistics that have F- and beta distributions, respectively. $\blacksquare$

Example 7.9.4. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $N(\theta_1, \theta_2)$, $-\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$. In Example 7.7.2 it was proved that the mean $\overline{X}$ and the variance S^2 of the sample are joint complete sufficient statistics for θ_1 and θ_2. Consider the statistic

$$Z = \frac{\sum_{1}^{n-1}(X_{i+1} - X_i)^2}{\sum_{1}^{n}(X_i - \overline{X})^2} = u(X_1, X_2, \ldots, X_n),$$

which satisfies the property that $u(cx_1 + d, \ldots, cx_n + d) = u(x_1, \ldots, x_n)$. That is, the ancillary statistic Z is independent of both $\overline{X}$ and S^2. $\blacksquare$

In this section we have given several examples in which the complete sufficient statistics are independent of ancillary statistics. Thus, in those cases, the ancillary statistics provide no information about the parameters. However, if the sufficient statistics are not complete, the ancillary statistics could provide some information as the following example demonstrates.

Example 7.9.5. We refer back to Examples 7.8.1 and 7.8.2. There the first and nth order statistics, Y_1 and Y_n, were minimal sufficient statistics for θ, where the sample arose from an underlying distribution having pdf $(\frac{1}{2})I_{(\theta-1,\theta+1)}(x)$. Often $T_1 = (Y_1 + Y_n)/2$ is used as an estimator of θ, as it is a function of those sufficient statistics that is unbiased. Let us find a relationship between T_1 and the ancillary statistic $T_2 = Y_n - Y_1$.

The joint pdf of Y_1 and Y_n is

$$g(y_1, y_n; \theta) = n(n-1)(y_n - y_1)^{n-2}/2^n, \quad \theta - 1 < y_1 < y_n < \theta + 1,$$

zero elsewhere. Accordingly, the joint pdf of T_1 and T_2 is, since the absolute value of the Jacobian equals 1,

$$h(t_1, t_2; \theta) = n(n-1)t_2^{n-2}/2^n, \quad \theta - 1 + \frac{t_2}{2} < t_1 < \theta + 1 - \frac{t_2}{2}, \quad 0 < t_2 < 2,$$

zero elsewhere. Thus the pdf of T_2 is

$$h_2(t_2; \theta) = n(n-1)t_2^{n-2}(2 - t_2)/2^n, \quad 0 < t_2 < 2,$$

zero elsewhere, which, of course, is free of θ as T_2 is an ancillary statistic. Thus, the conditional pdf of T_1, given $T_2 = t_2$, is

$$h_{1|2}(t_1|t_2; \theta) = \frac{1}{2 - t_2}, \quad \theta - 1 + \frac{t_2}{2} < t_1 < \theta + 1 - \frac{t_2}{2}, \quad 0 < t_2 < 2,$$

zero elsewhere. Note that this is uniform on the interval $(\theta - 1 + t_2/2, \theta + 1 - t_2/2)$; so the conditional mean and variance of T_1 are, respectively,

$$E(T_1|t_2) = \theta \quad \text{and} \quad \text{var}(T_1|t_2) = \frac{(2 - t_2)^2}{12}.$$

Given $T_2 = t_2$, we know something about the conditional variance of T_1. In particular, if that observed value of T_2 is large (close to 2), then that variance is small and we can place more reliance on the estimator T_1. On the other hand, a small value of t_2 means that we have less confidence in T_1 as an estimator of θ. It is extremely interesting to note that this conditional variance does not depend upon the sample size n but only on the given value of $T_2 = t_2$. As the sample size increases, T_2 tends to become larger and, in those cases, T_1 has smaller conditional variance. ■

While Example 7.9.5 is a special one demonstrating mathematically that an ancillary statistic can provide some help in point estimation, this does actually happen in practice, too. For illustration, we know that if the sample size is large enough, then

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

has an approximate standard normal distribution. Of course, if the sample arises from a normal distribution, $\overline{X}$ and S are independent and T has a t-distribution with $n - 1$ degrees of freedom. Even if the sample arises from a symmetric distribution, $\overline{X}$ and S are uncorrelated and T has an approximate t-distribution and certainly an approximate standard normal distribution with sample sizes around 30 or 40. On the other hand, if the sample arises from a highly skewed distribution (say to the right), then $\overline{X}$ and S are highly correlated and the probability $P(-1.96 < T < 1.96)$ is not necessarily close to 0.95 unless the sample size is extremely large (certainly much greater than 30). Intuitively, one can understand why this correlation exists if

the underlying distribution is highly skewed to the right. While S has a distribution free of μ (and hence is an ancillary), a large value of S implies a large value of $\overline{X}$, since the underlying pdf is like the one depicted in Figure 7.9.1. Of course, a small value of $\overline{X}$ (say less than the mode) requires a relatively small value of S. This means that unless n is extremely large, it is risky to say that

$$\overline{x} - \frac{1.96s}{\sqrt{n}}, \quad \overline{x} + \frac{1.96s}{\sqrt{n}}$$

provides an approximate 95% confidence interval with data from a very skewed distribution. As a matter of fact, the authors have seen situations in which this confidence coefficient is closer to 80%, rather than 95%, with sample sizes of 30 to 40.

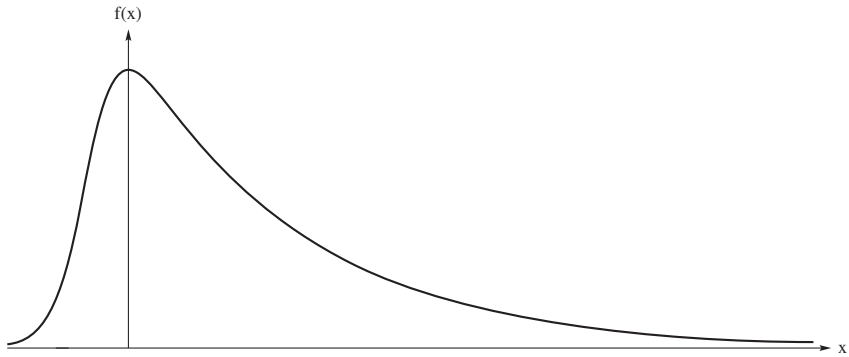

Figure 7.9.1: Graph of a right skewed distribution; see also Exercise 7.9.14.

EXERCISES

7.9.1. Let $Y_1 < Y_2 < Y_3 < Y_4$ denote the order statistics of a random sample of size $n = 4$ from a distribution having pdf $f(x;\theta) = 1/\theta,\ 0 < x < \theta$, zero elsewhere, where $0 < \theta < \infty$. Argue that the complete sufficient statistic Y_4 for θ is independent of each of the statistics Y_1/Y_4 and $(Y_1 + Y_2)/(Y_3 + Y_4)$.
Hint: Show that the pdf is of the form $(1/\theta)f(x/\theta)$, where $f(w) = 1,\ 0 < w < 1$, zero elsewhere.

7.9.2. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample from a $N(\theta, \sigma^2),\ -\infty < \theta < \infty$, distribution. Show that the distribution of $Z = Y_n - \overline{X}$ does not depend upon θ. Thus $\overline{Y} = \sum_1^n Y_i/n$, a complete sufficient statistic for θ is independent of Z.

7.9.3. Let $X_1, X_2, \ldots, X_n$ be iid with the distribution $N(\theta, \sigma^2),\ -\infty < \theta < \infty$. Prove that a necessary and sufficient condition that the statistics $Z = \sum_1^n a_i X_i$ and $Y = \sum_1^n X_i$, a complete sufficient statistic for θ, are independent is that $\sum_1^n a_i = 0$.

7.9.4. Let X and Y be random variables such that $E(X^k)$ and $E(Y^k) \neq 0$ exist for $k = 1, 2, 3, \ldots$. If the ratio X/Y and its denominator Y are independent, prove that $E[(X/Y)^k] = E(X^k)/E(Y^k), k = 1, 2, 3, \ldots$.
Hint: Write $E(X^k) = E[Y^k(X/Y)^k]$.

7.9.5. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from a distribution that has pdf $f(x;\theta) = (1/\theta)e^{-x/\theta},\ 0 < x < \infty,\ 0 < \theta < \infty$, zero elsewhere. Show that the ratio $R = nY_1 / \sum_1^n Y_i$ and its denominator (a complete sufficient statistic for θ) are independent. Use the result of the preceding exercise to determine $E(R^k),\ k = 1, 2, 3, \ldots$.

7.9.6. Let $X_1, X_2, \ldots, X_5$ be iid with pdf $f(x) = e^{-x},\ 0 < x < \infty$, zero elsewhere. Show that $(X_1 + X_2)/(X_1 + X_2 + \cdots + X_5)$ and its denominator are independent.
Hint: The pdf $f(x)$ is a member of $\{f(x;\theta) : 0 < \theta < \infty\}$, where $f(x;\theta) = (1/\theta)e^{-x/\theta},\ 0 < x < \infty$, zero elsewhere.

7.9.7. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample from the normal distribution $N(\theta_1, \theta_2),\ -\infty < \theta_1 < \infty,\ 0 < \theta_2 < \infty$. Show that the joint complete sufficient statistics $\overline{X} = \overline{Y}$ and S^2 for θ_1 and θ_2 are independent of each of $(Y_n - \overline{Y})/S$ and $(Y_n - Y_1)/S$.

7.9.8. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample from a distribution with the pdf

$$f(x;\theta_1,\theta_2) = \frac{1}{\theta_2} \exp\left(-\frac{x-\theta_1}{\theta_2}\right),$$

$\theta_1 < x < \infty$, zero elsewhere, where $-\infty < \theta_1 < \infty,\ 0 < \theta_2 < \infty$. Show that the joint complete sufficient statistics Y_1 and $\overline{X} = \overline{Y}$ for the parameters θ_1 and θ_2 are independent of $(Y_2 - Y_1)/\sum_1^n (Y_i - Y_1)$.

7.9.9. Let $X_1, X_2, \ldots, X_5$ be a random sample of size $n = 5$ from the normal distribution $N(0, \theta)$.

(a) Argue that the ratio $R = (X_1^2 + X_2^2)/(X_1^2 + \cdots + X_5^2)$ and its denominator $(X_1^2 + \cdots + X_5^2)$ are independent.

(b) Does $5R/2$ have an F-distribution with 2 and 5 degrees of freedom? Explain your answer.

(c) Compute $E(R)$ using Exercise 7.9.4.

7.9.10. Referring to Example 7.9.5 of this section, determine c so that

$$P(-c < T_1 - \theta < c | T_2 = t_2) = 0.95.$$

Use this result to find a 95% confidence interval for θ, given $T_2 = t_2$; and note how its length is smaller when the range of t_2 is larger.

7.9.11. Show that $Y = |X|$ is a complete sufficient statistic for $\theta > 0$, where X has the pdf $f_X(x;\theta) = 1/(2\theta)$, for $-\theta < x < \theta$, zero elsewhere. Show that $Y = |X|$ and $Z = \text{sgn}(X)$ are independent.

7.9.12. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample from a $N(\theta, \sigma^2)$ distribution, where σ^2 is fixed but arbitrary. Then $\overline{Y} = \overline{X}$ is a complete sufficient statistic for θ. Consider another estimator T of θ, such as $T = (Y_i + Y_{n+1-i})/2$, for $i = 1, 2, \ldots, [n/2]$, or T could be any weighted average of these latter statistics.

(a) Argue that $T - \overline{X}$ and $\overline{X}$ are independent random variables.

(b) Show that $\text{Var}(T) = \text{Var}(\overline{X}) + \text{Var}(T - \overline{X})$.

(c) Since we know $\text{Var}(\overline{X}) = \sigma^2/n$, it might be more efficient to estimate $\text{Var}(T)$ by estimating the $\text{Var}(T - \overline{X})$ by Monte Carlo methods rather than doing that with $\text{Var}(T)$ directly, because $\text{Var}(T) \geq \text{Var}(T - \overline{X})$. This is often called the *Monte Carlo Swindle*.

7.9.13. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with pdf $f(x; \theta) = (1/2)\theta^3 x^2 e^{-\theta x}$, $0 < x < \infty$, zero elsewhere, where $0 < \theta < \infty$:

(a) Find the mle, $\hat{\theta}$, of θ. Is $\hat{\theta}$ unbiased?
Hint: Find the pdf of $Y = \sum_1^n X_i$ and then compute $E(\hat{\theta})$.

(b) Argue that Y is a complete sufficient statistic for θ.

(c) Find the MVUE of θ.

(d) Show that X_1/Y and Y are independent.

(e) What is the distribution of X_1/Y?

7.9.14. The pdf depicted in Figure 7.9.1 is given by

$$f_{m_2}(x) = e^{-x}(1 + m_2^{-1}e^{-x})^{-(m_2+1)}, \quad -\infty < x < \infty, \tag{7.9.2}$$

where $m_2 > 0$ (the pdf graphed is for $m_2 = 0.1$). This is a member of a large family of pdfs, log F-family, which are useful in survival (lifetime) analysis; see Chapter 3 of Hettmansperger and McKean (2011).

(a) Let W be a random variable with pdf (7.9.2). Show that $W = \log Y$, where Y has an F-distribution with 2 and $2m_2$ degrees of freedom.

(b) Show that the pdf becomes the logistic (6.1.8) if $m_2 = 1$.

(c) Consider the location model where

$$X_i = \theta + W_i \quad i = 1, \ldots, n,$$

where $W_1, \ldots, W_n$ are iid with pdf (7.9.2). Similar to the logistic location model, the order statistics are minimal sufficient for this model. Show, similar to Example 6.1.2, that the mle of θ exists.

This page intentionally left blank

Chapter 8

Optimal Tests of Hypotheses

8.1 Most Powerful Tests

In Section 4.5, we introduced the concept of hypotheses testing and followed it with the introduction of likelihood ratio tests in Chapter 6. In this chapter, we discuss certain best tests.

For convenience to the reader, in the next several paragraphs we quickly review concepts of testing that were presented in Section 4.5. We are interested in a random variable X that has pdf or pmf $f(x;\theta)$, where $\theta \in \Omega$. We assume 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 the hypotheses as

$$H_0 : \theta \in \omega_0 \text{ versus } H_1 : \theta \in \omega_1. \tag{8.1.1}$$

The hypothesis H_0 is referred to as the **null hypothesis**, while H_1 is referred to as the **alternative hypothesis**. The test of H_0 versus H_1 is based on a sample $X_1, \ldots, X_n$ from the distribution of X. In this chapter, we often use the vector $\mathbf{X}' = (X_1, \ldots, X_n)$ to denote the random sample and $\mathbf{x}' = (x_1, \ldots, x_n)$ to denote the values of the sample. Let $\mathcal{S}$ denote the support of the random sample $\mathbf{X}' = (X_1, \ldots, X_n)$.

A **test** of H_0 versus H_1 is based on a subset C of $\mathcal{S}$. This set C is called the **critical region** and its corresponding decision rule is

$$\begin{array}{ll} \text{Reject } H_0 \text{ (Accept } H_1) & \text{if } \mathbf{X} \in C \\ \text{Retain } H_0 \text{ (Reject } H_1) & \text{if } \mathbf{X} \in C^c. \end{array} \tag{8.1.2}$$

Note that a test is defined by its critical region. Conversely, a critical region defines a test.

Recall that the 2×2 decision table, 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 **size** or **significance**

level of the test is the probability of a Type I error; i.e.,

$$\alpha = \max_{\theta \in \omega_0} P_\theta(\mathbf{X} \in C). \tag{8.1.3}$$

Note that $P_\theta(\mathbf{X} \in C)$ should be read as the probability that $\mathbf{X} \in C$ when θ is the true parameter. Subject to tests having size α, we select tests that minimize Type II error or equivalently maximize the probability of rejecting H_0 when $\theta \in \omega_1$. Recall that the **power function** of a test is given by

$$\gamma_C(\theta) = P_\theta(\mathbf{X} \in C); \quad \theta \in \omega_1. \tag{8.1.4}$$

In Chapter 4, we gave examples of tests of hypotheses, while in Sections 6.3 and 6.4, we discussed tests based on maximum likelihood theory. In this chapter, we want to construct best tests for certain situations.

We begin with testing a simple hypothesis H_0 against a simple alternative H_1. Let $f(x;\theta)$ denote the pdf or pmf of a random variable X, where $\theta \in \Omega = \{\theta', \theta''\}$. Let $\omega_0 = \{\theta'\}$ and $\omega_1 = \{\theta''\}$. Let $\mathbf{X}' = (X_1, \ldots, X_n)$ be a random sample from the distribution of X. We now define a best critical region (and hence a best test) for testing the simple hypothesis H_0 against the alternative simple hypothesis H_1.

Definition 8.1.1. *Let C denote a subset of the sample space. Then we say that C is a* **best critical region** *of size α for testing the simple hypothesis $H_0 : \theta = \theta'$ against the alternative simple hypothesis $H_1 : \theta = \theta''$ if*

(a) $P_{\theta'}[\mathbf{X} \in C] = \alpha$.

(b) *And for every subset A of the sample space,*

$$P_{\theta'}[\mathbf{X} \in A] = \alpha \Rightarrow P_{\theta''}[\mathbf{X} \in C] \geq P_{\theta''}[\mathbf{X} \in A].$$

This definition states, in effect, the following: In general, there is a multiplicity of subsets A of the sample space such that $P_{\theta'}[\mathbf{X} \in A] = \alpha$. Suppose that there is one of these subsets, say C, such that when H_1 is true, the power of the test associated with C is at least as great as the power of the test associated with every other A. Then C is defined as a best critical region of size α for testing H_0 against H_1.

As Theorem 8.1.1 shows, there is a best test for this simple versus simple case. But first, we offer a simple example examining this definition in some detail.

Example 8.1.1. Consider the one random variable X that has a binomial distribution with $n = 5$ and $p = \theta$. Let $f(x;\theta)$ denote the pmf of X and let $H_0 : \theta = \frac{1}{2}$ and $H_1 : \theta = \frac{3}{4}$. The following tabulation gives, at points of positive probability density, the values of $f(x;\frac{1}{2})$, $f(x;\frac{3}{4})$, and the ratio $f(x;\frac{1}{2})/f(x;\frac{3}{4})$.

x	0	1	2
$f(x;1/2)$	1/32	5/32	10/32
$f(x;3/4)$	1/1024	15/1024	90/1024
$f(x;1/2)/f(x;3/4)$	32/1	32/3	32/9
x	3	4	5
$f(x;1/2)$	10/32	5/32	1/32
$f(x;3/4)$	270/1024	405/1024	243/1024
$f(x;1/2)/f(x;3/4)$	32/27	32/81	32/243

We shall use one random value of X to test the simple hypothesis $H_0 : \theta = \frac{1}{2}$ against the alternative simple hypothesis $H_1 : \theta = \frac{3}{4}$, and we shall first assign the significance level of the test to be $\alpha = \frac{1}{32}$. We seek a best critical region of size $\alpha = \frac{1}{32}$. If $A_1 = \{x : x = 0\}$ or $A_2 = \{x : x = 5\}$, then $P_{\{\theta=1/2\}}(X \in A_1) = P_{\{\theta=1/2\}}(X \in A_2) = \frac{1}{32}$ and there is no other subset A_3 of the space $\{x : x = 0, 1, 2, 3, 4, 5\}$ such that $P_{\{\theta=1/2\}}(X \in A_3) = \frac{1}{32}$. Then either A_1 or A_2 is the best critical region C of size $\alpha = \frac{1}{32}$ for testing H_0 against H_1. We note that $P_{\{\theta=1/2\}}(X \in A_1) = \frac{1}{32}$ and $P_{\{\theta=3/4\}}(X \in A_1) = \frac{1}{1024}$. Thus, if the set A_1 is used as a critical region of size $\alpha = \frac{1}{32}$, we have the intolerable situation that the probability of rejecting H_0 when H_1 is true (H_0 is false) is much less than the probability of rejecting H_0 when H_0 is true.

On the other hand, if the set A_2 is used as a critical region, then $P_{\{\theta=1/2\}}(X \in A_2) = \frac{1}{32}$ and $P_{\{\theta=3/4\}}(X \in A_2) = \frac{243}{1024}$. That is, the probability of rejecting H_0 when H_1 is true is much greater than the probability of rejecting H_0 when H_0 is true. Certainly, this is a more desirable state of affairs, and actually A_2 is the best critical region of size $\alpha = \frac{1}{32}$. The latter statement follows from the fact that when H_0 is true, there are but two subsets, A_1 and A_2, of the sample space, each of whose probability measure is $\frac{1}{32}$ and the fact that

$$\frac{243}{1024} = P_{\{\theta=3/4\}}(X \in A_2) > P_{\{\theta=3/4\}}(X \in A_1) = \frac{1}{1024}.$$

It should be noted in this problem that the best critical region $C = A_2$ of size $\alpha = \frac{1}{32}$ is found by including in C the point (or points) at which $f(x; \frac{1}{2})$ is *small* in comparison with $f(x; \frac{3}{4})$. This is seen to be true once it is observed that the ratio $f(x; \frac{1}{2})/f(x; \frac{3}{4})$ is a minimum at $x = 5$. Accordingly, the ratio $f(x; \frac{1}{2})/f(x; \frac{3}{4})$, that is given in the last line of the above tabulation, provides us with a precise tool by which to find a best critical region C for certain given values of α. To illustrate this, take $\alpha = \frac{6}{32}$. When H_0 is true, each of the subsets $\{x : x = 0, 1\}$, $\{x : x = 0, 4\}$, $\{x : x = 1, 5\}$, $\{x : x = 4, 5\}$ has probability measure $\frac{6}{32}$. By direct computation it is found that the best critical region of this size is $\{x : x = 4, 5\}$. This reflects the fact that the ratio $f(x; \frac{1}{2})/f(x; \frac{3}{4})$ has its two smallest values for $x = 4$ and $x = 5$. The power of this test, which has $\alpha = \frac{6}{32}$, is

$$P_{\{\theta=3/4\}}(X = 4, 5) = \frac{405}{1024} + \frac{243}{1024} = \frac{648}{1024}. \qquad \blacksquare$$

The preceding example should make the following theorem, due to Neyman and Pearson, easier to understand. It is an important theorem because it provides a systematic method of determining a best critical region.

Theorem 8.1.1. Neyman–Pearson Theorem. *Let $X_1, X_2, \ldots, X_n$, where n is a fixed positive integer, denote a random sample from a distribution that has pdf or pmf $f(x;\theta)$. Then the likelihood of $X_1, X_2, \ldots, X_n$ is*

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta), \quad \text{for } \mathbf{x}' = (x_1, \ldots, x_n).$$

Let θ' and θ'' be distinct fixed values of θ so that $\Omega = \{\theta : \theta = \theta', \theta''\}$, and let k be a positive number. Let C be a subset of the sample space such that

(a) $\dfrac{L(\theta'; \mathbf{x})}{L(\theta''; \mathbf{x})} \leq k$, *for each point* $\mathbf{x} \in C$.

(b) $\dfrac{L(\theta'; \mathbf{x})}{L(\theta''; \mathbf{x})} \geq k$, *for each point* $\mathbf{x} \in C^c$.

(c) $\alpha = P_{H_0}[\mathbf{X} \in C]$.

Then C is a best critical region of size α for testing the simple hypothesis $H_0 : \theta = \theta'$ against the alternative simple hypothesis $H_1 : \theta = \theta''$.

Proof: We shall give the proof when the random variables are of the continuous type. If C is the only critical region of size α, the theorem is proved. If there is another critical region of size α, denote it by A. For convenience, we shall let $\int_R \cdots \int L(\theta; x_1, \ldots, x_n)\, dx_1 \cdots dx_n$ be denoted by $\int_R L(\theta)$. In this notation we wish to show that

$$\int_C L(\theta'') - \int_A L(\theta'') \geq 0.$$

Since C is the union of the disjoint sets $C \cap A$ and $C \cap A^c$ and A is the union of the disjoint sets $A \cap C$ and $A \cap C^c$, we have

$$\begin{aligned} \int_C L(\theta'') - \int_A L(\theta'') &= \int_{C\cap A} L(\theta'') + \int_{C\cap A^c} L(\theta'') - \int_{A\cap C} L(\theta'') - \int_{A\cap C^c} L(\theta'') \\ &= \int_{C\cap A^c} L(\theta'') - \int_{A\cap C^c} L(\theta''). \end{aligned} \tag{8.1.5}$$

However, by the hypothesis of the theorem, $L(\theta'') \geq (1/k)L(\theta')$ at each point of C, and hence at each point of $C \cap A^c$; thus,

$$\int_{C\cap A^c} L(\theta'') \geq \frac{1}{k} \int_{C\cap A^c} L(\theta').$$

But $L(\theta'') \leq (1/k)L(\theta')$ at each point of C^c, and hence at each point of $A \cap C^c$; accordingly,

$$\int_{A\cap C^c} L(\theta'') \leq \frac{1}{k} \int_{A\cap C^c} L(\theta').$$

These inequalities imply that

$$\int_{C\cap A^c} L(\theta'') - \int_{A\cap C^c} L(\theta'') \geq \frac{1}{k} \int_{C\cap A^c} L(\theta') - \frac{1}{k} \int_{A\cap C^c} L(\theta');$$

and, from Equation (8.1.5), we obtain

$$\int_C L(\theta'') - \int_A L(\theta'') \geq \frac{1}{k}\left[\int_{C\cap A^c} L(\theta') - \int_{A\cap C^c} L(\theta')\right]. \tag{8.1.6}$$

However,

$$\begin{aligned}\int_{C\cap A^c} L(\theta') - \int_{A\cap C^c} L(\theta') &= \int_{C\cap A^c} L(\theta') + \int_{C\cap A} L(\theta') \\ &\quad - \int_{A\cap C} L(\theta') - \int_{A\cap C^c} L(\theta') \\ &= \int_C L(\theta') - \int_A L(\theta') = \alpha - \alpha = 0.\end{aligned}$$

If this result is substituted in inequality (8.1.6), we obtain the desired result,

$$\int_C L(\theta'') - \int_A L(\theta'') \geq 0.$$

If the random variables are of the discrete type, the proof is the same with integration replaced by summation. ■

Remark 8.1.1. As stated in the theorem, conditions (a), (b), and (c) are sufficient ones for region C to be a best critical region of size α. However, they are also necessary. We discuss this briefly. Suppose there is a region A of size α that does not satisfy (a) and (b) and that is as powerful at $\theta = \theta''$ as C, which satisfies (a), (b), and (c). Then expression (8.1.5) would be zero, since the power at θ'' using A is equal to that using C. It can be proved that to have expression (8.1.5) equal zero, A must be of the same form as C. As a matter of fact, in the continuous case, A and C would essentially be the same region; that is, they could differ only by a set having probability zero. However, in the discrete case, if $P_{H_0}[L(\theta') = kL(\theta'')]$ is positive, A and C could be different sets, but each would necessarily enjoy conditions (a), (b), and (c) to be a best critical region of size α. ■

It would seem that a test should have the property that its power should never fall below its significance level; otherwise, the probability of falsely rejecting H_0 (level) is higher than the probability of correctly rejecting H_0 (power). We say a test having this property is **unbiased**, which we now formally define:

Definition 8.1.2. *Let X be a random variable which has pdf or pmf $f(x;\theta)$, where $\theta \in \Omega$. Consider the hypotheses given in expression (8.1.1). Let $\mathbf{X}' = (X_1, \ldots, X_n)$ denote a random sample on X. Consider a test with critical region C and level α. We say that this test is* **unbiased** *if*

$$P_\theta(\mathbf{X} \in C) \geq \alpha,$$

for all $\theta \in \omega_1$.

As the next corollary shows, the best test given in Theorem 8.1.1 is an unbiased test.

Corollary 8.1.1. *As in Theorem 8.1.1, let C be the critical region of the best test of $H_0 : \theta = \theta'$ versus $H_1 : \theta = \theta''$. Suppose the significance level of the test is α. Let $\gamma_C(\theta'') = P_{\theta''}[\mathbf{X} \in C]$ denote the power of the test. Then $\alpha \leq \gamma_C(\theta'')$.*

Proof: Consider the "unreasonable" test in which the data are ignored, but a Bernoulli trial is performed which has probability α of success. If the trial ends in success, we reject H_0. The level of this test is α. Because the power of a test is the probability of rejecting H_0 when H_1 is true, the power of this unreasonable test is α also. But C is the best critical region of size α and thus has power greater than or equal to the power of the unreasonable test. That is, $\gamma_C(\theta'') \geq \alpha$, which is the desired result. ■

Another aspect of Theorem 8.1.1 to be emphasized is that if we take C to be the set of all points $\mathbf{x}$ which satisfy

$$\frac{L(\theta'; \mathbf{x})}{L(\theta''; \mathbf{x})} \leq k, \quad k > 0,$$

then, in accordance with the theorem, C is a best critical region. This inequality can frequently be expressed in one of the forms (where c_1 and c_2 are constants)

$$u_1(\mathbf{x}; \theta', \theta'') \leq c_1$$

or

$$u_2(\mathbf{x}; \theta', \theta'') \geq c_2.$$

Suppose that it is the first form, $u_1 \leq c_1$. Since θ' and θ'' are given constants, $u_1(\mathbf{X}; \theta', \theta'')$ is a statistic; and if the pdf or pmf of this statistic can be found when H_0 is true, then the significance level of the test of H_0 against H_1 can be determined from this distribution. That is,

$$\alpha = P_{H_0}[u_1(\mathbf{X}; \theta', \theta'') \leq c_1].$$

Moreover, the test may be based on this statistic; for if the observed vector value of $\mathbf{X}$ is $\mathbf{x}$, we reject H_0 (accept H_1) if $u_1(\mathbf{x}) \leq c_1$.

A positive number k determines a best critical region C whose size is $\alpha = P_{H_0}[\mathbf{X} \in C]$ for that particular k. It may be that this value of α is unsuitable for the purpose at hand; that is, it is too large or too small. However, if there is a statistic $u_1(\mathbf{X})$ as in the preceding paragraph, whose pdf or pmf can be determined when H_0 is true, we need not experiment with various values of k to obtain a desirable significance level. For if the distribution of the statistic is known, or can be found, we may determine c_1 such that $P_{H_0}[u_1(\mathbf{X}) \leq c_1]$ is a desirable significance level.

An illustrative example follows.

Example 8.1.2. Let $\mathbf{X}' = (X_1, \ldots, X_n)$ denote a random sample from the distribution that has the pdf

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right), \quad -\infty < x < \infty.$$

It is desired to test the simple hypothesis $H_0 : \theta = \theta' = 0$ against the alternative simple hypothesis $H_1 : \theta = \theta'' = 1$. Now

$$\begin{aligned} \frac{L(\theta';\mathbf{x})}{L(\theta'';\mathbf{x})} &= \frac{(1/\sqrt{2\pi})^n \exp\left[-\sum_1^n x_i^2/2\right]}{(1/\sqrt{2\pi})^n \exp\left[-\sum_1^n (x_i-1)^2 \Big/ 2\right]} \\ &= \exp\left(-\sum_1^n x_i + \frac{n}{2}\right). \end{aligned}$$

If $k > 0$, the set of all points $(x_1, x_2, \ldots, x_n)$ such that

$$\exp\left(-\sum_1^n x_i + \frac{n}{2}\right) \leq k$$

is a best critical region. This inequality holds if and only if

$$-\sum_1^n x_i + \frac{n}{2} \leq \log k$$

or, equivalently,

$$\sum_1^n x_i \geq \frac{n}{2} - \log k = c.$$

In this case, a best critical region is the set $C = \{(x_1, x_2, \ldots, x_n) : \sum_1^n x_i \geq c\}$, where c is a constant that can be determined so that the size of the critical region is a desired number α. The event $\sum_1^n X_i \geq c$ is equivalent to the event $\overline{X} \geq c/n = c_1$, for example, so the test may be based upon the statistic $\overline{X}$. If H_0 is true, that is, $\theta = \theta' = 0$, then $\overline{X}$ has a distribution that is $N(0, 1/n)$. Given the significance level α, the number c_1 is computed in R as $c_1 = \mathbf{qnorm}(1-\alpha, 0, 1/\sqrt{n})$; hence, $P_{H_0}(\overline{X} \geq c_1) = \alpha$. So, if the experimental values of $X_1, X_2, \ldots, X_n$ were, respectively, $x_1, x_2, \ldots, x_n$, we would compute $\overline{x} = \sum_1^n x_i/n$. If $\overline{x} \geq c_1$, the simple hypothesis $H_0 : \theta = \theta' = 0$ would be rejected at the significance level α; if $\overline{x} < c_1$, the hypothesis H_0 would be accepted. The probability of rejecting H_0 when H_0 is true is α the level of significance. The probability of rejecting H_0, when H_0 is false, is the value of the power of the test at $\theta = \theta'' = 1$, which is,

$$P_{H_1}(\overline{X} \geq c_1) = \int_{c_1}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1/n}} \exp\left[-\frac{(\overline{x}-1)^2}{2(1/n)}\right] d\overline{x}. \tag{8.1.7}$$

For example, if $n = 25$ and α is 0.05, $c_1 =$ `qnorm(0.95,0,1/5) = 0.329`, using R. Hence, the power of the test to detect $\theta = 1$, given in expression (8.1.7), is computed by `1 - pnorm(0.329,1,1/5) = 0.9996`. ■

There is another aspect of this theorem that warrants special mention. It has to do with the number of parameters that appear in the pdf. Our notation suggests that there is but one parameter. However, a careful review of the proof reveals that nowhere was this needed or assumed. The pdf or pmf may depend upon any finite number of parameters. What is essential is that the hypothesis H_0 and the alternative hypothesis H_1 be simple, namely, that they completely specify the distributions. With this in mind, we see that the simple hypotheses H_0 and H_1 do not need to be hypotheses about the parameters of a distribution, nor, as a matter of fact, do the random variables $X_1, X_2, \ldots, X_n$ need to be independent. That is, if H_0 is the simple hypothesis that the joint pdf or pmf is $g(x_1, x_2, \ldots, x_n)$, and if H_1 is the alternative simple hypothesis that the joint pdf or pmf is $h(x_1, x_2, \ldots, x_n)$, then C is a best critical region of size α for testing H_0 against H_1 if, for $k > 0$,

1. $\dfrac{g(x_1, x_2, \ldots, x_n)}{h(x_1, x_2, \ldots, x_n)} \leq k$ for $(x_1, x_2, \ldots, x_n) \in C$.
2. $\dfrac{g(x_1, x_2, \ldots, x_n)}{h(x_1, x_2, \ldots, x_n)} \geq k$ for $(x_1, x_2, \ldots, x_n) \in C^c$.
3. $\alpha = P_{H_0}[(X_1, X_2, \ldots, X_n) \in C]$.

Consider the following example.

Example 8.1.3. Let $X_1, \ldots, X_n$ denote a random sample on X that has pmf $f(x)$ with support $\{0, 1, 2, \ldots\}$. It is desired to test the simple hypothesis

$$H_0 : f(x) = \begin{cases} \frac{e^{-1}}{x!} & x = 0, 1, 2, \ldots \\ 0 & \text{elsewhere,} \end{cases}$$

against the alternative simple hypothesis

$$H_1 : f(x) = \begin{cases} (\frac{1}{2})^{x+1} & x = 0, 1, 2, \ldots \\ 0 & \text{elsewhere.} \end{cases}$$

That is, we want to test whether X has a Poisson distribution with mean $\lambda = 1$ versus X has a geometric distribution with $p = 1/2$. Here

$$\begin{aligned} \frac{g(x_1, \ldots, x_n)}{h(x_1, \ldots, x_n)} &= \frac{e^{-n}/(x_1! x_2! \cdots x_n!)}{(\frac{1}{2})^n (\frac{1}{2})^{x_1 + x_2 + \cdots + x_n}} \\ &= \frac{(2e^{-1})^n 2^{\sum x_i}}{\prod_1^n (x_i!)}. \end{aligned}$$

If $k > 0$, the set of points $(x_1, x_2, \ldots, x_n)$ such that

$$\left(\sum_1^n x_i \right) \log 2 - \log \left[\prod_1^n (x_i!) \right] \leq \log k - n \log(2e^{-1}) = c$$

is a best critical region C. Consider the case of $k = 1$ and $n = 1$. The preceding inequality may be written $2^{x_1}/x_1! \leq e/2$. This inequality is satisfied by all points in the set $C = \{x_1 : x_1 = 0, 3, 4, 5, \ldots\}$. Using R, the level of significance is

$$P_{H_0}(X_1 \in C) = 1 - P_{H_0}(X_1 = 1, 2) = 1 - \mathbf{dpois}(1, 1) - \mathbf{dpois}(2, 1) = 0.4482.$$

The power of the test to detect H_1 is computed as

$$P_{H_1}(X_1 \in C) = 1 - P_{H_1}(X_1 = 1, 2) = 1 - (\tfrac{1}{4} + \tfrac{1}{8}) = 0.625. \quad \blacksquare$$

Note that these results are consistent with Corollary 8.1.1.

Remark 8.1.2. In the notation of this section, say C is a critical region such that

$$\alpha = \int_C L(\theta') \quad \text{and} \quad \beta = \int_{C^c} L(\theta''),$$

where α and β equal the respective probabilities of the Type I and Type II errors associated with C. Let d_1 and d_2 be two given positive constants. Consider a certain linear function of α and β, namely,

$$\begin{aligned} d_1 \int_C L(\theta') + d_2 \int_{C^c} L(\theta'') &= d_1 \int_C L(\theta') + d_2 \left[1 - \int_C L(\theta'')\right] \\ &= d_2 + \int_C [d_1 L(\theta') - d_2 L(\theta'')]. \end{aligned}$$

If we wished to minimize this expression, we would select C to be the set of all $(x_1, x_2, \ldots, x_n)$ such that

$$d_1 L(\theta') - d_2 L(\theta'') < 0$$

or, equivalently,

$$\frac{L(\theta')}{L(\theta'')} < \frac{d_2}{d_1}, \quad \text{for all } (x_1, x_2, \ldots, x_n) \in C,$$

which according to the Neyman–Pearson theorem provides a best critical region with $k = d_2/d_1$. That is, this critical region C is one that minimizes $d_1\alpha + d_2\beta$. There could be others, including points on which $L(\theta')/L(\theta'') = d_2/d_1$, but these would still be best critical regions according to the Neyman–Pearson theorem. $\blacksquare$

EXERCISES

8.1.1. In Example 8.1.2 of this section, let the simple hypotheses read $H_0 : \theta = \theta' = 0$ and $H_1 : \theta = \theta'' = -1$. Show that the best test of H_0 against H_1 may be carried out by use of the statistic $\overline{X}$, and that if $n = 25$ and $\alpha = 0.05$, the power of the test is 0.9996 when H_1 is true.

8.1.2. Let the random variable X have the pdf $f(x;\theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, zero elsewhere. Consider the simple hypothesis $H_0 : \theta = \theta' = 2$ and the alternative hypothesis $H_1 : \theta = \theta'' = 4$. Let X_1, X_2 denote a random sample of size 2 from this distribution. Show that the best test of H_0 against H_1 may be carried out by use of the statistic $X_1 + X_2$.

8.1.3. Repeat Exercise 8.1.2 when $H_1 : \theta = \theta'' = 6$. Generalize this for every $\theta'' > 2$.

8.1.4. Let $X_1, X_2, \ldots, X_{10}$ be a random sample of size 10 from a normal distribution $N(0, \sigma^2)$. Find a best critical region of size $\alpha = 0.05$ for testing $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 2$. Is this a best critical region of size 0.05 for testing $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 4$? Against $H_1 : \sigma^2 = \sigma_1^2 > 1$?

8.1.5. If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution having pdf of the form $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, zero elsewhere, show that a best critical region for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$ is $C = \{(x_1, x_2, \ldots, x_n) : c \leq \prod_{i=1}^{n} x_i\}$.

8.1.6. Let $X_1, X_2, \ldots, X_{10}$ be a random sample from a distribution that is $N(\theta_1, \theta_2)$. Find a best test of the simple hypothesis $H_0 : \theta_1 = \theta_1' = 0$, $\theta_2 = \theta_2' = 1$ against the alternative simple hypothesis $H_1 : \theta_1 = \theta_1'' = 1$, $\theta_2 = \theta_2'' = 4$.

8.1.7. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution $N(\theta, 100)$. Show that $C = \{(x_1, x_2, \ldots, x_n) : c \leq \overline{x} = \sum_1^n x_i/n\}$ is a best critical region for testing $H_0 : \theta = 75$ against $H_1 : \theta = 78$. Find n and c so that

$$P_{H_0}[(X_1, X_2, \ldots, X_n) \in C] = P_{H_0}(\overline{X} \geq c) = 0.05$$

and

$$P_{H_1}[(X_1, X_2, \ldots, X_n) \in C] = P_{H_1}(\overline{X} \geq c) = 0.90,$$

approximately.

8.1.8. If $X_1, X_2, \ldots, X_n$ is a random sample from a beta distribution with parameters $\alpha = \beta = \theta > 0$, find a best critical region for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$.

8.1.9. Let $X_1, X_2, \ldots, X_n$ be iid with pmf $f(x;p) = p^x(1-p)^{1-x}$, $x = 0, 1$, zero elsewhere. Show that $C = \{(x_1, \ldots, x_n) : \sum_1^n x_i \leq c\}$ is a best critical region for testing $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{1}{3}$. Use the Central Limit Theorem to find n and c so that approximately $P_{H_0}(\sum_1^n X_i \leq c) = 0.10$ and $P_{H_1}(\sum_1^n X_i \leq c) = 0.80$.

8.1.10. Let $X_1, X_2, \ldots, X_{10}$ denote a random sample of size 10 from a Poisson distribution with mean θ. Show that the critical region C defined by $\sum_1^{10} x_i \geq 3$ is a best critical region for testing $H_0 : \theta = 0.1$ against $H_1 : \theta = 0.5$. Determine, for this test, the significance level α and the power at $\theta = 0.5$. Use the R function `ppois`.

8.2 Uniformly Most Powerful Tests

This section takes up the problem of a test of a simple hypothesis H_0 against an alternative composite hypothesis H_1. We begin with an example.

Example 8.2.1. Consider the pdf

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

of Exercises 8.1.2 and 8.1.3. It is desired to test the simple hypothesis $H_0 : \theta = 2$ against the alternative composite hypothesis $H_1 : \theta > 2$. Thus $\Omega = \{\theta : \theta \geq 2\}$. A random sample, X_1, X_2, of size $n = 2$ is used, and the critical region is $C = \{(x_1, x_2) : 9.5 \leq x_1 + x_2 < \infty\}$. It was shown in the exercises cited that the significance level of the test is approximately 0.05 and the power of the test when $\theta = 4$ is approximately 0.31. The power function $\gamma(\theta)$ of the test for all $\theta \geq 2$ is

$$\begin{aligned} \gamma(\theta) &= 1 - \int_0^{9.5} \int_0^{9.5-x_2} \frac{1}{\theta^2} \exp\left(-\frac{x_1 + x_2}{\theta}\right) dx_1 dx_2 \\ &= \left(\frac{\theta + 9.5}{\theta}\right) e^{-9.5/\theta}, \quad 2 \leq \theta. \end{aligned}$$

For example, $\gamma(2) = 0.05$, $\gamma(4) = 0.31$, and $\gamma(9.5) = 2/e \approx 0.74$. It is shown (Exercise 8.1.3) that the set $C = \{(x_1, x_2) : 9.5 \leq x_1 + x_2 < \infty\}$ is a best critical region of size 0.05 for testing the simple hypothesis $H_0 : \theta = 2$ against each simple hypothesis in the composite hypothesis $H_1 : \theta > 2$. ■

The preceding example affords an illustration of a test of a simple hypothesis H_0 that is a best test of H_0 against every simple hypothesis in the alternative composite hypothesis H_1. We now define a critical region, when it exists, which is a best critical region for testing a simple hypothesis H_0 against an alternative composite hypothesis H_1. It seems desirable that this critical region should be a best critical region for testing H_0 against each simple hypothesis in H_1. That is, the power function of the test that corresponds to this critical region should be at least as great as the power function of any other test with the same significance level for every simple hypothesis in H_1.

Definition 8.2.1. *The critical region C is a* **uniformly most powerful (UMP) critical region** *of size α for testing the simple hypothesis H_0 against an alternative composite hypothesis H_1 if the set C is a best critical region of size α for testing H_0 against each simple hypothesis in H_1. A test defined by this critical region C is called a* **uniformly most powerful (UMP) test**, *with significance level α, for testing the simple hypothesis H_0 against the alternative composite hypothesis H_1.*

As will be seen presently, uniformly most powerful tests do not always exist. However, when they do exist, the Neyman–Pearson theorem provides a technique for finding them. Some illustrative examples are given here.

Example 8.2.2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $N(0, \theta)$, where the variance θ is an unknown positive number. It will be shown that there exists a uniformly most powerful test with significance level α for testing the simple hypothesis $H_0 : \theta = \theta'$, where θ' is a fixed positive number, against the alternative composite hypothesis $H_1 : \theta > \theta'$. Thus $\Omega = \{\theta : \theta \geq \theta'\}$. The joint pdf of $X_1, X_2, \ldots, X_n$ is

$$L(\theta; x_1, x_2, \ldots, x_n) = \left(\frac{1}{2\pi\theta}\right)^{n/2} \exp\left\{-\frac{1}{2\theta}\sum_{i=1}^{n} x_i^2\right\}.$$

Let θ'' represent a number greater than θ', and let k denote a positive number. Let C be the set of points where

$$\frac{L(\theta'; x_1, x_2, \ldots, x_n)}{L(\theta''; x_1, x_2, \ldots, x_n)} \leq k,$$

that is, the set of points where

$$\left(\frac{\theta''}{\theta'}\right)^{n/2} \exp\left[-\left(\frac{\theta'' - \theta'}{2\theta'\theta''}\right)\sum_{1}^{n} x_i^2\right] \leq k$$

or, equivalently,

$$\sum_{1}^{n} x_i^2 \geq \frac{2\theta'\theta''}{\theta'' - \theta'}\left[\frac{n}{2}\log\left(\frac{\theta''}{\theta'}\right) - \log k\right] = c.$$

The set $C = \{(x_1, x_2, \ldots, x_n) : \sum_1^n x_i^2 \geq c\}$ is then a best critical region for testing the simple hypothesis $H_0 : \theta = \theta'$ against the simple hypothesis $\theta = \theta''$. It remains to determine c, so that this critical region has the desired size α. If H_0 is true, the random variable $\sum_1^n X_i^2/\theta'$ has a chi-square distribution with n degrees of freedom. Since $\alpha = P_{\theta'}(\sum_1^n X_i^2/\theta' \geq c/\theta')$, c/θ' may be computed, for example, by the R code `qchisq`$(1 - \alpha, n)$. Then $C = \{(x_1, x_2, \ldots, x_n) : \sum_1^n x_i^2 \geq c\}$ is a best critical region of size α for testing $H_0 : \theta = \theta'$ against the hypothesis $\theta = \theta''$. Moreover, for each number θ'' greater than θ', the foregoing argument holds. That is, $C = \{(x_1, \ldots, x_n) : \sum_1^n x_i^2 \geq c\}$ is a uniformly most powerful critical region of size α for testing $H_0 : \theta = \theta'$ against $H_1 : \theta > \theta'$. If $x_1, x_2, \ldots, x_n$ denote the experimental values of $X_1, X_2, \ldots, X_n$, then $H_0 : \theta = \theta'$ is rejected at the significance level α, and $H_1 : \theta > \theta'$ is accepted if $\sum_1^n x_i^2 \geq c$; otherwise, $H_0 : \theta = \theta'$ is accepted.

If, in the preceding discussion, we take $n = 15$, $\alpha = 0.05$, and $\theta' = 3$, then the two hypotheses are $H_0 : \theta = 3$ and $H_1 : \theta > 3$. Using R, $c/3$ is computed by `qchisq(0.95,15) = 24.996`. Hence, $c = 74.988$. ■

Example 8.2.3. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $N(\theta, 1)$, where θ is unknown. It will be shown that there is no uniformly most powerful test of the simple hypothesis $H_0 : \theta = \theta'$, where θ' is a fixed number against the alternative composite hypothesis $H_1 : \theta \neq \theta'$. Thus $\Omega = \{\theta : -\infty < \theta < \infty\}$. Let θ'' be a number not equal to θ'. Let k be a positive number and consider

$$\frac{(1/2\pi)^{n/2} \exp\left[-\sum_1^n (x_i - \theta')^2/2\right]}{(1/2\pi)^{n/2} \exp\left[-\sum_1^n (x_i - \theta'')^2/2\right]} \leq k.$$

The preceding inequality may be written as

$$\exp\left\{-(\theta'' - \theta')\sum_1^n x_i + \frac{n}{2}[(\theta'')^2 - (\theta')^2]\right\} \leq k$$

or

$$(\theta'' - \theta')\sum_1^n x_i \geq \frac{n}{2}[(\theta'')^2 - (\theta')^2] - \log k.$$

This last inequality is equivalent to

$$\sum_1^n x_i \geq \frac{n}{2}(\theta'' + \theta') - \frac{\log k}{\theta'' - \theta'},$$

provided that $\theta'' > \theta'$, and it is equivalent to

$$\sum_1^n x_i \leq \frac{n}{2}(\theta'' + \theta') - \frac{\log k}{\theta'' - \theta'}$$

if $\theta'' < \theta'$. The first of these two expressions defines a best critical region for testing $H_0 : \theta = \theta'$ against the hypothesis $\theta = \theta''$ provided that $\theta'' > \theta'$, while the second expression defines a best critical region for testing $H_0 : \theta = \theta'$ against the hypothesis $\theta = \theta''$ provided that $\theta'' < \theta'$. That is, a best critical region for testing the simple hypothesis against an alternative simple hypothesis, say $\theta = \theta' + 1$, does not serve as a best critical region for testing $H_0 : \theta = \theta'$ against the alternative simple hypothesis $\theta = \theta' - 1$. By definition, then, there is no uniformly most powerful test in the case under consideration.

It should be noted that had the alternative composite hypothesis been one-sided, either $H_1 : \theta > \theta'$ or $H_1 : \theta < \theta'$, a uniformly most powerful test would exist in each instance. ■

Example 8.2.4. In Exercise 8.1.10, the reader was asked to show that if a random sample of size $n = 10$ is taken from a Poisson distribution with mean θ, the critical region defined by $\sum_1^n x_i \geq 3$ is a best critical region for testing $H_0 : \theta = 0.1$ against

$H_1 : \theta = 0.5$. This critical region is also a uniformly most powerful one for testing $H_0 : \theta = 0.1$ against $H_1 : \theta > 0.1$ because, with $\theta'' > 0.1$,

$$\frac{(0.1)^{\sum x_i} e^{-10(0.1)}/(x_1!x_2!\cdots x_n!)}{(\theta'')^{\sum x_i} e^{-10(\theta'')}/(x_1!x_2!\cdots x_n!)} \leq k$$

is equivalent to

$$\left(\frac{0.1}{\theta''}\right)^{\sum x_i} e^{-10(0.1-\theta'')} \leq k.$$

The preceding inequality may be written as

$$\left(\sum_1^n x_i\right)(\log 0.1 - \log \theta'') \leq \log k + 10(1 - \theta'')$$

or, since $\theta'' > 0.1$, equivalently as

$$\sum_1^n x_i \geq \frac{\log k + 10 - 10\theta''}{\log 0.1 - \log \theta''}.$$

Of course, $\sum_1^n x_i \geq 3$ is of the latter form. ■

Let us make an important observation, although obvious when pointed out. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has pdf $f(x;\theta)$, $\theta \in \Omega$. Suppose that $Y = u(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for θ. In accordance with the factorization theorem, the joint pdf of $X_1, X_2, \ldots, X_n$ may be written

$$L(\theta; x_1, x_2, \ldots, x_n) = k_1[u(x_1, x_2, \ldots, x_n); \theta]k_2(x_1, x_2, \ldots, x_n),$$

where $k_2(x_1, x_2, \ldots, x_n)$ does not depend upon θ. Consequently, the ratio

$$\frac{L(\theta'; x_1, x_2, \ldots, x_n)}{L(\theta''; x_1, x_2, \ldots, x_n)} = \frac{k_1[u(x_1, x_2, \ldots, x_n); \theta']}{k_1[u(x_1, x_2, \ldots, x_n); \theta'']}$$

depends upon $x_1, x_2, \ldots, x_n$ only through $u(x_1, x_2, \ldots, x_n)$. Accordingly, if there is a sufficient statistic $Y = u(X_1, X_2, \ldots, X_n)$ for θ and if a best test or a uniformly most powerful test is desired, there is no need to consider tests that are based upon any statistic other than the sufficient statistic. This result supports the importance of sufficiency.

In the above examples, we have presented uniformly most powerful tests. For some families of pdfs and hypotheses, we can obtain general forms of such tests. We sketch these results for the general one-sided hypotheses of the form

$$H_0 : \theta \leq \theta' \text{ versus } H_1 : \theta > \theta'. \tag{8.2.1}$$

The other one-sided hypotheses with the null hypothesis $H_0 : \theta \geq \theta'$, is completely analogous. Note that the null hypothesis of (8.2.1) is a composite hypothesis. Recall from Chapter 4 that the level of a test for the hypotheses (8.2.1) is defined by

$\max_{\theta \le \theta'} \gamma(\theta)$, where $\gamma(\theta)$ is the power function of the test. That is, the significance level is the maximum probability of Type I error.

Let $\mathbf{X}' = (X_1, \ldots, X_n)$ be a random sample with common pdf (or pmf) $f(x;\theta)$, $\theta \in \Omega$, and, hence with the likelihood function

$$L(\theta, \mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta), \quad \mathbf{x}' = (x_1, \ldots, x_n) \, .$$

We consider the family of pdfs that has monotone likelihood ratio as defined next.

Definition 8.2.2. *We say that the likelihood $L(\theta, \mathbf{x})$ has* **monotone likelihood ratio** *(***mlr***) in the statistic $y = u(\mathbf{x})$ if, for $\theta_1 < \theta_2$, the ratio*

$$\frac{L(\theta_1, \mathbf{x})}{L(\theta_2, \mathbf{x})} \tag{8.2.2}$$

is a monotone function of $y = u(\mathbf{x})$.

Assume then that our likelihood function $L(\theta, \mathbf{x})$ has a monotone decreasing likelihood ratio in the statistic $y = u(\mathbf{x})$. Then the ratio in (8.2.2) is equal to $g(y)$, where g is a decreasing function. The case where the likelihood function has a monotone increasing likelihood ratio (i.e., g is an increasing function) follows similarly by changing the sense of the inequalities below. Let α denote the significance level. Then we claim that the following test is UMP level α for the hypotheses (8.2.1):

$$\text{Reject } H_0 \text{ if } Y \ge c_Y, \tag{8.2.3}$$

where c_Y is determined by $\alpha = P_{\theta'}[Y \ge c_Y]$. To show this claim, first consider the simple null hypothesis $H_0' : \theta = \theta'$. Let $\theta'' > \theta'$ be arbitrary but fixed. Let C denote the most powerful critical region for θ' versus θ''. By the Neyman–Pearson Theorem, C is defined by

$$\frac{L(\theta', \mathbf{X})}{L(\theta'', \mathbf{X})} \le k \text{ if and only if } \mathbf{X} \in C,$$

where k is determined by $\alpha = P_{\theta'}[\mathbf{X} \in C]$. But by Definition 8.2.2, because $\theta'' > \theta'$,

$$\frac{L(\theta', \mathbf{X})}{L(\theta'', \mathbf{X})} = g(Y) \le k \Leftrightarrow Y \ge g^{-1}(k),$$

where $g^{-1}(k)$ satisfies $\alpha = P_{\theta'}[Y \ge g^{-1}(k)]$; i.e., $c_Y = g^{-1}(k)$. Hence the Neyman–Pearson test is equivalent to the test defined by (8.2.3). Furthermore, the test is UMP for θ' versus $\theta'' > \theta'$ because the test only depends on $\theta'' > \theta'$ and $g^{-1}(k)$ is uniquely determined under θ'.

Let $\gamma_Y(\theta)$ denote the power function of the test (8.2.3). To finish, we need to show that $\max_{\theta \le \theta'} \gamma_Y(\theta) = \alpha$. But this follows immediately if we can show that $\gamma_Y(\theta)$ is a nondecreasing function. To see this, let $\theta_1 < \theta_2$. Note that since $\theta_1 < \theta_2$, the test (8.2.3) is the most powerful test for testing θ_1 versus θ_2 with the level $\gamma_Y(\theta_1)$. By Corollary 8.1.1, the power of the test at θ_2 must not be below the level; i.e., $\gamma_Y(\theta_2) \ge \gamma_Y(\theta_1)$. Hence $\gamma_Y(\theta)$ is a nondecreasing function. Since the power function is nondecreasing, it follows from Definition 8.1.2 that the mlr tests are unbiased tests for the hypotheses (8.2.1); see Exercise 8.2.14.

Example 8.2.5. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter $p = \theta$, where $0 < \theta < 1$. Let $\theta' < \theta''$. Consider the ratio of likelihoods

$$\frac{L(\theta'; x_1, x_2, \ldots, x_n)}{L(\theta''; x_1, x_2, \ldots, x_n)} = \frac{(\theta')^{\sum x_i}(1-\theta')^{n-\sum x_i}}{(\theta'')^{\sum x_i}(1-\theta'')^{n-\sum x_i}} = \left[\frac{\theta'(1-\theta'')}{\theta''(1-\theta')}\right]^{\sum x_i} \left(\frac{1-\theta'}{1-\theta''}\right)^n.$$

Since $\theta'/\theta'' < 1$ and $(1-\theta'')/(1-\theta') < 1$, so that $\theta'(1-\theta'')/\theta''(1-\theta') < 1$, the ratio is a decreasing function of $y = \sum x_i$. Thus we have a monotone likelihood ratio in the statistic $Y = \sum X_i$.

Consider the hypotheses

$$H_0 : \theta \leq \theta' \text{ versus } H_1 : \theta > \theta'. \tag{8.2.4}$$

By our discussion above, the UMP level α decision rule for testing H_0 versus H_1 is given by

$$\text{Reject } H_0 \text{ if } Y = \textstyle\sum_{i=1}^n X_i \geq c,$$

where c is such that $\alpha = P_{\theta'}[Y \geq c]$. ■

In the last example concerning a Bernoulli pmf, we obtained a UMP test by showing that its likelihood possesses mlr. The Bernoulli distribution is a regular case of the exponential family and our argument, under the one assumption below, can be generalized to the entire regular exponential family. To show this, suppose that the random sample $X_1, X_2, \ldots, X_n$ arises from a pdf or pmf representing a regular case of the exponential class, namely,

$$f(x;\theta) = \begin{cases} \exp[p(\theta)K(x) + H(x) + q(\theta)] & x \in \mathcal{S} \\ 0 & \text{elsewhere,} \end{cases}$$

where the support of X, $\mathcal{S}$, is free of θ. Further assume that $p(\theta)$ is an increasing function of θ. Then

$$\begin{aligned} \frac{L(\theta')}{L(\theta'')} &= \frac{\exp\left[p(\theta')\sum_1^n K(x_i) + \sum_1^n H(x_i) + nq(\theta')\right]}{\exp\left[p(\theta'')\sum_1^n K(x_i) + \sum_1^n H(x_i) + nq(\theta'')\right]} \\ &= \exp\left\{[p(\theta') - p(\theta'')]\sum_1^n K(x_i) + n[q(\theta') - q(\theta'')]\right\}. \end{aligned}$$

If $\theta' < \theta''$, $p(\theta)$ being an increasing function, requires this ratio to be a decreasing function of $y = \sum_1^n K(x_i)$. Thus, we have a monotone likelihood ratio in the statistic $Y = \sum_1^n K(X_i)$. Hence consider the hypotheses

$$H_0 : \theta \leq \theta' \text{ versus } H_1 : \theta > \theta'. \tag{8.2.5}$$

By our discussion above concerning mlr, the UMP level α decision rule for testing H_0 versus H_1 is given by

$$\text{Reject } H_0 \text{ if } Y = \sum_{i=1}^{n} K(X_i) \geq c,$$

where c is such that $\alpha = P_{\theta'}[Y \geq c]$. Furthermore, the power function of this test is an increasing function in θ.

For the record, consider the other one-sided alternative hypotheses,

$$H_0 : \theta \geq \theta' \text{ versus } H_1 : \theta < \theta'. \tag{8.2.6}$$

The UMP level α decision rule is, for $p(\theta)$ an increasing function,

$$\text{Reject } H_0 \text{ if } Y = \sum_{i=1}^{n} K(X_i) \leq c,$$

where c is such that $\alpha = P_{\theta'}[Y \leq c]$.

If in the preceding situation with monotone likelihood ratio we test $H_0 : \theta = \theta'$ against $H_1 : \theta > \theta'$, then $\sum K(x_i) \geq c$ would be a uniformly most powerful critical region. From the likelihood ratios displayed in Examples 8.2.2–8.2.5, we see immediately that the respective critical regions

$$\sum_{i=1}^{n} x_i^2 \geq c, \quad \sum_{i=1}^{n} x_i \geq c, \quad \sum_{i=1}^{n} x_i \geq c, \quad \sum_{i=1}^{n} x_i \geq c$$

are uniformly most powerful for testing $H_0 : \theta = \theta'$ against $H_1 : \theta > \theta'$.

There is a final remark that should be made about uniformly most powerful tests. Of course, in Definition 8.2.1, the word *uniformly* is associated with θ; that is, C is a best critical region of size α for testing $H_0 : \theta = \theta_0$ against all θ values given by the composite alternative H_1. However, suppose that the form of such a region is

$$u(x_1, x_2, \ldots, x_n) \leq c.$$

Then this form provides uniformly most powerful critical regions for all attainable α values by, of course, appropriately changing the value of c. That is, there is a certain uniformity property, also associated with α, that is not always noted in statistics texts.

EXERCISES

8.2.1. Let X have the pmf $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$, zero elsewhere. We test the simple hypothesis $H_0 : \theta = \frac{1}{4}$ against the alternative composite hypothesis $H_1 : \theta < \frac{1}{4}$ by taking a random sample of size 10 and rejecting $H_0 : \theta = \frac{1}{4}$ if and only if the observed values $x_1, x_2, \ldots, x_{10}$ of the sample observations are such that $\sum_1^{10} x_i \leq 1$. Find the power function $\gamma(\theta)$, $0 < \theta \leq \frac{1}{4}$, of this test.

8.2.2. Let X have a pdf of the form $f(x;\theta) = 1/\theta,\ 0 < x < \theta$, zero elsewhere. Let $Y_1 < Y_2 < Y_3 < Y_4$ denote the order statistics of a random sample of size 4 from this distribution. Let the observed value of Y_4 be y_4. We reject $H_0 : \theta = 1$ and accept $H_1 : \theta \neq 1$ if either $y_4 \leq \frac{1}{2}$ or $y_4 > 1$. Find the power function $\gamma(\theta),\ 0 < \theta$, of the test.

8.2.3. Consider a normal distribution of the form $N(\theta, 4)$. The simple hypothesis $H_0 : \theta = 0$ is rejected, and the alternative composite hypothesis $H_1 : \theta > 0$ is accepted if and only if the observed mean $\overline{x}$ of a random sample of size 25 is greater than or equal to $\frac{3}{5}$. Find the power function $\gamma(\theta),\ 0 \leq \theta$, of this test.

8.2.4. Consider the distributions $N(\mu_1, 400)$ and $N(\mu_2, 225)$. Let $\theta = \mu_1 - \mu_2$. Let $\overline{x}$ and $\overline{y}$ denote the observed means of two independent random samples, each of size n, from these two distributions. We reject $H_0 : \theta = 0$ and accept $H_1 : \theta > 0$ if and only if $\overline{x} - \overline{y} \geq c$. If $\gamma(\theta)$ is the power function of this test, find n and c so that $\gamma(0) = 0.05$ and $\gamma(10) = 0.90$, approximately.

8.2.5. Consider Example 8.2.2. Show that $L(\theta)$ has a monotone likelihood ratio in the statistic $\sum_{i=1}^{n} X_i^2$. Use this to determine the UMP test for $H_0 : \theta = \theta'$, where θ' is a fixed positive number, versus $H_1 : \theta < \theta'$.

8.2.6. If, in Example 8.2.2 of this section, $H_0 : \theta = \theta'$, where θ' is a fixed positive number, and $H_1 : \theta \neq \theta'$, show that there is no uniformly most powerful test for testing H_0 against H_1.

8.2.7. Let $X_1, X_2, \ldots, X_{25}$ denote a random sample of size 25 from a normal distribution $N(\theta, 100)$. Find a uniformly most powerful critical region of size $\alpha = 0.10$ for testing $H_0 : \theta = 75$ against $H_1 : \theta > 75$.

8.2.8. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution $N(\theta, 16)$. Find the sample size n and a uniformly most powerful test of $H_0 : \theta = 25$ against $H_1 : \theta < 25$ with power function $\gamma(\theta)$ so that approximately $\gamma(25) = 0.10$ and $\gamma(23) = 0.90$.

8.2.9. Consider a distribution having a pmf of the form $f(x;\theta) = \theta^x(1-\theta)^{1-x},\ x = 0, 1$, zero elsewhere. Let $H_0 : \theta = \frac{1}{20}$ and $H_1 : \theta > \frac{1}{20}$. Use the Central Limit Theorem to determine the sample size n of a random sample so that a uniformly most powerful test of H_0 against H_1 has a power function $\gamma(\theta)$, with approximately $\gamma(\frac{1}{20}) = 0.05$ and $\gamma(\frac{1}{10}) = 0.90$.

8.2.10. Illustrative Example 8.2.1 of this section dealt with a random sample of size $n = 2$ from a gamma distribution with $\alpha = 1,\ \beta = \theta$. Thus the mgf of the distribution is $(1 - \theta t)^{-1},\ t < 1/\theta,\ \theta \geq 2$. Let $Z = X_1 + X_2$. Show that Z has a gamma distribution with $\alpha = 2,\ \beta = \theta$. Express the power function $\gamma(\theta)$ of Example 8.2.1 in terms of a single integral. Generalize this for a random sample of size n.

8.2.11. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta) = \theta x^{\theta-1},\ 0 < x < 1$, zero elsewhere, where $\theta > 0$. Show the likelihood has mlr in the statistic $\prod_{i=1}^{n} X_i$. Use this to determine the UMP test for $H_0 : \theta = \theta'$ against $H_1 : \theta < \theta'$, for fixed $\theta' > 0$.

8.2.12. Let X have the pdf $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$, zero elsewhere. We test $H_0 : \theta = \frac{1}{2}$ against $H_1 : \theta < \frac{1}{2}$ by taking a random sample $X_1, X_2, \ldots, X_5$ of size $n = 5$ and rejecting H_0 if $Y = \sum_1^n X_i$ is observed to be less than or equal to a constant c.

(a) Show that this is a uniformly most powerful test.

(b) Find the significance level when $c = 1$.

(c) Find the significance level when $c = 0$.

(d) By using a *randomized test*, as discussed in Example 4.6.4, modify the tests given in parts (b) and (c) to find a test with significance level $\alpha = \frac{2}{32}$.

8.2.13. Let $X_1, \ldots, X_n$ denote a random sample from a gamma-type distribution with $\alpha = 2$ and $\beta = \theta$. Let $H_0 : \theta = 1$ and $H_1 : \theta > 1$.

(a) Show that there exists a uniformly most powerful test for H_0 against H_1, determine the statistic Y upon which the test may be based, and indicate the nature of the best critical region.

(b) Find the pdf of the statistic Y in part (a). If we want a significance level of 0.05, write an equation that can be used to determine the critical region. Let $\gamma(\theta)$, $\theta \geq 1$, be the power function of the test. Express the power function as an integral.

8.2.14. Show that the mlr test defined by expression (8.2.3) is an unbiased test for the hypotheses (8.2.1).

8.3 Likelihood Ratio Tests

In the first section of this chapter, we presented the most powerful tests for simple versus simple hypotheses. In the second section, we extended this theory to uniformly most powerful tests for essentially one-sided alternative hypotheses and families of distributions that have a monotone likelihood ratio. What about the general case? That is, suppose the random variable X has pdf or pmf $f(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters in Ω. Let $\omega \subset \Omega$ and consider the hypotheses

$$H_0 : \boldsymbol{\theta} \in \omega \text{ versus } H_1 : \boldsymbol{\theta} \in \Omega \cap \omega^c. \tag{8.3.1}$$

There are complications in extending the optimal theory to this general situation, which are addressed in more advanced books; see, in particular, Lehmann (1986). We illustrate some of these complications with an example. Suppose X has a $N(\theta_1, \theta_2)$ distribution and that we want to test $\theta_1 = \theta_1'$, where θ_1' is specified. In the notation of (8.3.1), $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $\Omega = \{\boldsymbol{\theta} : -\infty < \theta_1 < \infty, \theta_2 > 0\}$, and $\omega = \{\boldsymbol{\theta} : \theta_1 = \theta_1', \theta_2 > 0\}$. Notice that $H_0 : \boldsymbol{\theta} \in \omega$ is a composite null hypothesis. Let $X_1, \ldots, X_n$ be a random sample on X.

Assume for the moment that θ_2 is known. Then H_0 becomes the simple hypothesis $\theta_1 = \theta_1'$. This is essentially the situation discussed in Example 8.2.3. There

it was shown that no UMP test exists for this situation. If we restrict attention to the class of unbiased tests (Definition 8.1.2), then a theory of best tests can be constructed; see Lehmann (1986). For our illustrative example, as Exercise 8.3.21 shows, the test based on the critical region

$$C_2 = \left\{ |\overline{X} - \theta_1'| > \sqrt{\frac{\theta_2}{n}} z_{\alpha/2} \right\}$$

is unbiased. Then it follows from Lehmann that it is an UMP unbiased level α test.

In practice, though, the variance θ_2 is unknown. In this case, theory for optimal tests can be constructed using the concept of what are called conditional tests. We do not pursue this any further in this text, but refer the interested reader to Lehmann (1986).

Recall from Chapter 6 that the likelihood ratio tests (6.3.3) can be used to test general hypotheses such as (8.3.1). While in general the exact null distribution of the test statistic cannot be determined, under regularity condtions the likelihood ratio test statistic is asymptotically χ^2 under H_0. Hence we can obtain an approximate test in most situations. Although, there is no guarantee that likelihood ratio tests are optimal, similar to tests based on the Neyman–Pearson Theorem, they are based on a ratio of likelihood functions and, in many situations, are asymptotically optimal.

In the example above on testing for the mean of a normal distribution, with known variance, the likelihood ratio test is the same as the UMP unbiased test. When the variance is unknown, the likelihood ratio test results in the one-sample t-test as shown in Example 6.5.1 of Chapter 6. This is the same as the conditional test discussed in Lehmann (1986).

In the remainder of this section, we present likelihood ratio tests for situations when sampling from normal distributions.

8.3.1 Likelihood Ratio Tests for Testing Means of Normal Distributions

In Example 6.5.1 of Chapter 6, we derived the likelihood ratio test for the one-sample t-test to test for the mean of a normal distribution with unknown variance. In the next example, we derive the likelihood ratio test for compairing the means of two independent normal distributions. We then discuss the power functions for both of these tests.

Example 8.3.1. Let the independent random variables X and Y have distributions that are $N(\theta_1, \theta_3)$ and $N(\theta_2, \theta_3)$, where the means θ_1 and θ_2 and common variance θ_3 are unknown. Then $\Omega = \{(\theta_1, \theta_2, \theta_3) : -\infty < \theta_1 < \infty, -\infty < \theta_2 < \infty, 0 < \theta_3 < \infty\}$. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ denote independent random samples from these distributions. The hypothesis $H_0 : \theta_1 = \theta_2$, unspecified, and θ_3 unspecified, is to be tested against all alternatives. Then $\omega = \{(\theta_1, \theta_2, \theta_3) : -\infty < \theta_1 = \theta_2 < \infty, 0 < \theta_3 < \infty\}$. Here $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ are $n + m > 2$ mutually

independent random variables having the likelihood functions

$$L(\omega) = \left(\frac{1}{2\pi\theta_3}\right)^{(n+m)/2} \exp\left\{-\frac{1}{2\theta_3}\left[\sum_1^n (x_i-\theta_1)^2 + \sum_1^m (y_i-\theta_1)^2\right]\right\}$$

and

$$L(\Omega) = \left(\frac{1}{2\pi\theta_3}\right)^{(n+m)/2} \exp\left\{-\frac{1}{2\theta_3}\left[\sum_1^n (x_i-\theta_1)^2 + \sum_1^m (y_i-\theta_2)^2\right]\right\}.$$

If $\partial \log L(\omega)/\partial\theta_1$ and $\partial \log L(\omega)/\partial\theta_3$ are equated to zero, then (Exercise 8.3.2)

$$\begin{aligned} \sum_1^n (x_i-\theta_1) + \sum_1^m (y_i-\theta_1) &= 0 \\ \frac{1}{\theta_3}\left[\sum_1^n (x_i-\theta_1)^2 + \sum_1^m (y_i-\theta_1)^2\right] &= n+m. \end{aligned} \tag{8.3.2}$$

The solutions for θ_1 and θ_3 are, respectively,

$$\begin{aligned} u &= (n+m)^{-1}\left\{\sum_1^n x_i + \sum_1^m y_i\right\} \\ w &= (n+m)^{-1}\left\{\sum_1^n (x_i-u)^2 + \sum_1^m (y_i-u)^2\right\}. \end{aligned}$$

Further, u and w maximize $L(\omega)$. The maximum is

$$L(\hat{\omega}) = \left(\frac{e^{-1}}{2\pi w}\right)^{(n+m)/2}.$$

In a like manner, if

$$\frac{\partial \log L(\Omega)}{\partial\theta_1}, \quad \frac{\partial \log L(\Omega)}{\partial\theta_2}, \quad \frac{\partial \log L(\Omega)}{\partial\theta_3}$$

are equated to zero, then (Exercise 8.3.3)

$$\begin{aligned} \sum_1^n (x_i-\theta_1) &= 0 \\ \sum_1^m (y_i-\theta_2) &= 0 \\ -(n+m) + \frac{1}{\theta_3}\left[\sum_1^n (x_i-\theta_1)^2 + \sum_1^m (y_i-\theta_2)^2\right] &= 0. \end{aligned} \tag{8.3.3}$$

The solutions for $\theta_1,\ \theta_2$, and θ_3 are, respectively,

$$\begin{aligned} u_1 &= n^{-1}\sum_1^n x_i \\ u_2 &= m^{-1}\sum_1^m y_i \\ w' &= (n+m)^{-1}\left[\sum_1^n (x_i-u_1)^2 + \sum_1^m (y_i-u_2)^2\right], \end{aligned}$$

and, further, $u_1,\ u_2$, and w' maximize $L(\Omega)$. The maximum is

$$L(\hat{\Omega}) = \left(\frac{e^{-1}}{2\pi w'}\right)^{(n+m)/2},$$

so that

$$\Lambda(x_1,\ldots,x_n,y_1,\ldots,y_m) = \Lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \left(\frac{w'}{w}\right)^{(n+m)/2}.$$

The random variable defined by $\Lambda^{2/(n+m)}$ is

$$\frac{\sum_1^n (X_i-\overline{X})^2 + \sum_1^m (Y_i-\overline{Y})^2}{\sum_1^n \{X_i - [(n\overline{X}+m\overline{Y})/(n+m)]\}^2 + \sum_1^n \{Y_i - [(n\overline{X}+m\overline{Y})/(n+m)]\}^2}.$$

Now

$$\begin{aligned} \sum_1^n \left(X_i - \frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2 &= \sum_1^n \left[(X_i-\overline{X}) + \left(\overline{X} - \frac{n\overline{X}+m\overline{Y}}{n+m}\right)\right]^2 \\ &= \sum_1^n (X_i-\overline{X})^2 + n\left(\overline{X} - \frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2 \end{aligned}$$

and

$$\begin{aligned} \sum_1^m \left(Y_i - \frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2 &= \sum_1^m \left[(Y_i-\overline{Y}) + \left(\overline{Y} - \frac{n\overline{X}+m\overline{Y}}{n+m}\right)\right]^2 \\ &= \sum_1^m (Y_i-\overline{Y})^2 + m\left(\overline{Y} - \frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2. \end{aligned}$$

But

$$n\left(\overline{X} - \frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2 = \frac{m^2 n}{(n+m)^2}(\overline{X}-\overline{Y})^2$$

and

$$m\left(\overline{Y}-\frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2=\frac{n^2m}{(n+m)^2}(\overline{X}-\overline{Y})^2.$$

Hence the random variable defined by $\Lambda^{2/(n+m)}$ may be written

$$\frac{\sum_1^n(X_i-\overline{X})^2+\sum_1^m(Y_i-\overline{Y})^2}{\sum_1^n(X_i-\overline{X})^2+\sum_1^m(Y_i-\overline{Y})^2+[nm/(n+m)](\overline{X}-\overline{Y})^2}$$

$$=\frac{1}{1+\dfrac{[nm/(n+m)](\overline{X}-\overline{Y})^2}{\sum_1^n(X_i-\overline{X})^2+\sum_1^m(Y_i-\overline{Y})^2}}.$$

If the hypothesis $H_0:\theta_1=\theta_2$ is true, the random variable

$$T=\sqrt{\frac{nm}{n+m}}(\overline{X}-\overline{Y})\left\{(n+m-2)^{-1}\left[\sum_1^n(X_i-\overline{X})^2+\sum_1^m(Y_i-\overline{Y})^2\right]\right\}^{-1/2} \tag{8.3.4}$$

has, in accordance with Section 3.6, a t-distribution with $n+m-2$ degrees of freedom. Thus the random variable defined by $\Lambda^{2/(n+m)}$ is

$$\frac{n+m-2}{(n+m-2)+T^2}.$$

The test of H_0 against all alternatives may then be based on a t-distribution with $n+m-2$ degrees of freedom.

The likelihood ratio principle calls for the rejection of H_0 if and only if $\Lambda\le\lambda_0<1$. Thus the significance level of the test is

$$\alpha=P_{H_0}[\Lambda(X_1,\ldots,X_n,Y_1,\ldots,Y_m)\le\lambda_0].$$

However, $\Lambda(X_1,\ldots,X_n,Y_1,\ldots,Y_m)\le\lambda_0$ is equivalent to $|T|\ge c$, and so

$$\alpha=P(|T|\ge c;H_0).$$

For given values of n and m, the number c is is easily computed. In R, $c=$`qt`$(1-\alpha/2,n+m-2)$. Then H_0 is rejected at a significance level α if and only if $|t|\ge c$, where t is the observed value of T. If, for instance, $n=10$, $m=6$, and $\alpha=0.05$, then $c=$`qt`$(0.975,14)=2.1448$. ■

For this last example as well as the one-sample t-test derived in Example 6.5.1, it was found that the likelihood ratio test could be based on a statistic that, when the hypothesis H_0 is true, has a t-distribution. To help us compute the power functions of these tests at parameter points other than those described by the hypothesis H_0, we turn to the following definition.

Definition 8.3.1. *Let the random variable W be $N(\delta, 1)$; let the random variable V be $\chi^2(r)$, and let W and V be independent. The quotient*

$$T = \frac{W}{\sqrt{V/r}}$$

is said to have a noncentral t-distribution *with r degrees of freedom and noncentrality parameter δ. If $\delta = 0$, we say that T has a central t-distribution.*

In the light of this definition, let us reexamine the t-statistics of Examples 6.5.1 and 8.3.1.

Example 8.3.2 (Power of the One Sample t-Test)**.** For Example 6.5.1, consider a more general situation. Assume that $X_1, \ldots, X_n$ is a random sample on X that has a $N(\mu, \sigma^2)$ distribution. We are interested in testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, where μ_0 is specified. Then from Example 6.5.1, the likelihood ratio test statistic is

$$\begin{aligned} t(X_1, \ldots, X_n) &= \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sqrt{\sum_1^n (X_i - \overline{X})^2/(n-1)}} \\ &= \frac{\sqrt{n}(\overline{X} - \mu_0)/\sigma}{\sqrt{\sum_1^n (X_i - \overline{X})^2/[\sigma^2(n-1)]}}. \end{aligned}$$

The hypothesis H_0 is rejected at level α if $|t| \geq t_{\alpha/2,n-1}$. Suppose $\mu_1 \neq \mu_0$ is an alternative of interest. Because $E_{\mu_1}[\sqrt{n}\overline{X}/\sigma\sqrt{n}\overline{X}/\sigma] = \sqrt{n}(\mu_1 - \mu_0)/\sigma$, the power of the test to detect μ_1 is

$$\gamma(\mu_1) = P(|t| \geq t_{\alpha/2,n-1}) = 1 - P(t \leq t_{\alpha/2,n-1}) + P(t \leq -t_{\alpha/2,n-1}), \qquad (8.3.5)$$

where t has a noncentral t-distribution with noncentrality parameter $\delta = \sqrt{n}(\mu_1 - \mu_0)/\sigma$ and $n-1$ degrees of freedom. This is computed in R by the call

```
1 - pt(tc,n-1,ncp=delta) + pt(-tc,n-1,ncp=delta)
```

where `tc` is $t_{\alpha/2,n-1}$ and `delta` is the noncentrality parameter δ.

The following R code computes a graph of the power curve of this test. Notice that the horizontal range of the plot is the interval $[\mu_0 - 4\sigma/\sqrt{n}, \mu_0 + 4\sigma/\sqrt{n}]$. As indicated the parameters need to be set.

```
## Input mu0, sig, n, alpha.
fse = 4*sig/sqrt(n); maxmu = mu0 + fse; tc = qt(1-(alpha/2),n-1)
minmu = mu0 -fse; mu1 = seq(minmu,maxmu,.1)
delta = (mu1-mu0)/(sig/sqrt(n))
gs = 1 - pt(tc,n-1,ncp=delta) + pt(-tc,n-1,ncp=delta)
plot(gs~mu1,pch=" ",xlab=expression(mu[1]),ylab=expression(gamma))
lines(gs~mu1)
```

This code is the body of the function `tpowerg.R`. Exercise 8.3.5 discusses its use. ■

Example 8.3.3 (Power of the Two Sample t-Test)**.** In Example 8.3.1 we had

$$T = \frac{W_2}{\sqrt{V_2/(n+m-2)}},$$

where

$$W_2 = \sqrt{\frac{nm}{n+m}}(\overline{X} - \overline{Y}) \Big/ \sigma$$

and

$$V_2 = \frac{\sum_1^n (X_i - \overline{X})^2 + \sum_1^m (Y_i - \overline{Y})^2}{\sigma^2}.$$

Here W_2 is $N[\sqrt{nm/(n+m)}(\theta_1 - \theta_2)/\sigma, 1]$, V_2 is $\chi^2(n+m-2)$, and W_2 and V_2 are independent. Accordingly, if $\theta_1 \neq \theta_2$, T has a noncentral t-distribution with $n+m-2$ degrees of freedom and noncentrality parameter $\delta_2 = \sqrt{nm/(n+m)}(\theta_1 - \theta_2)/\sigma$. It is interesting to note that $\delta_1 = \sqrt{n}\theta_1/\sigma$ measures the deviation of θ_1 from $\theta_1 = 0$ in units of the standard deviation $\sigma/\sqrt{n}$ of $\overline{X}$. The noncentrality parameter $\delta_2 = \sqrt{nm/(n+m)}(\theta_1 - \theta_2)/\sigma$ is equal to the deviation of $\theta_1 - \theta_2$ from $\theta_1 - \theta_2 = 0$ in units of the standard deviation $\sigma/\sqrt{(n+m)/mn}$ of $\overline{X} - \overline{Y}$.

As in the last example, it is easy to write R code that evaluates power for this test. For a numerical illustration, assume that the common variance is $\theta_3 = 100$, $n = 20$, and $m = 15$. Suppose $\alpha = 0.05$ and we want to determine the power of the test to detect $\Delta = 5$, where $\Delta = \theta_1 - \theta_2$. In this case the critical value is $t_{0.25,33} = \text{qt}(.975, 33) = 2.0345$ and the noncentrality parameter is $\delta_2 = 1.4639$. The power is computed as

```
1- pt(2.0345,33,ncp=1.4639) + pt(-2.0345,33,ncp=1.4639) = 0.2954
```

Hence, the test has a 29.4% chance of detecting a difference in means of 5. ■

Remark 8.3.1. The one- and two-sample tests for normal means, presented in Examples 6.5.1 and 8.3.1, are the tests for normal means presented in most elementary statistics books. They are based on the assumption of normality. What if the underlying distributions are not normal? In that case, with finite variances, the t-test statistics for these situations are asymptotically correct. For example, consider the one-sample t-test. Suppose $X_1, \ldots, X_n$ are iid with a common nonnormal pdf that has mean θ_1 and finite variance σ^2. The hypotheses remain the same, i.e., $H_0 : \theta_1 = \theta_1'$ versus $H_1 : \theta_1 \neq \theta_1'$. The t-test statistic, T_n, is given by

$$T_n = \frac{\sqrt{n}(\overline{X} - \theta_1')}{S_n}, \tag{8.3.6}$$

where S_n is the sample standard deviation. Our critical region is $C_1 = \{|T_n| \geq t_{\alpha/2,n-1}\}$. Recall that $S_n \to \sigma$ in probability. Hence, by the Central Limit Theorem, under H_0,

$$T_n = \frac{\sigma}{S_n} \frac{\sqrt{n}(\overline{X} - \theta_1')}{\sigma} \xrightarrow{D} Z, \tag{8.3.7}$$

where Z has a standard normal distribution. Hence the asymptotic test would use the critical region $C_2 = \{|T_n| \geq z_{\alpha/2}\}$. By (8.3.7) the critical region C_2 would have approximate size α. In practice, we would use C_1, because t critical values are generally larger than z critical values and, hence, the use of C_1 would be conservative; i.e., the size of C_1 would be slightly smaller than that of C_2. As Exercise 8.3.4 shows, the two-sample t-test is also asymptotically correct, provided the underlying distributions have the *same* variance. ■

For nonnormal situations where the distribution is "close" to the normal distribution, the t-test is essentially valid; i.e., the true level of significance is close to the nominal α. In terms of robustness, we would say that for these situations the t-test possesses **robustness of validity**. But the t-test may not possess **robustness of power**. For nonnormal situations, there are more powerful tests than the t-test; see Chapter 10 for discussion.

For finite sample sizes and for distributions that are decidedly not normal, very skewed for instance, the validity of the t-test may also be questionable, as we illustrate in the following simulation study.

Example 8.3.4 (Skewed Contaminated Normal Family of Distributions)**.** Consider the random variable X given by

$$X = (1 - I_\epsilon)Z + I_\epsilon Y, \tag{8.3.8}$$

where Z has a $N(0,1)$ distribution, Y has a $N(\mu_c, \sigma_c^2)$ distribution, I_ϵ has a $bin(1,\epsilon)$ distribution, and Z, Y, and I_ϵ are mutually independent. Assume that $\epsilon < 0.5$ and $\sigma_c > 1$, so that Y is the contaminating random variable in the mixture. If $\mu_c = 0$, then X has the contaminated normal distribution discussed in Section 3.4.1, which is symmetrically distributed about 0. For $\mu_c \neq 0$, the distribution of X, (8.3.8), is skewed and we call it the **skewed contaminated normal distribution**, $SCN(\epsilon, \sigma_c, \mu_C)$. Note that $E(X) = \epsilon\mu_c$ and in Exercise 8.3.18 the cdf and pdf of X are derived. The R function `rscn` generates random variates from this distribution.

In this example, we show the results of a small simulation study on the validity of the t-test for random samples from the distribution of X. Consider the one-sided hypotheses

$$H_0 : \mu = \mu_X \text{ versus } H_0 : \mu < \mu_X.$$

Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution of X. As a test statistic we consider the t-test discussed in Example 4.5.4, which is also given in expression (8.3.6); that is, the test statistic is $T_n = (\overline{X} - \mu_X)/(S_n/\sqrt{n})$, where $\overline{X}$ and S_n are the sample mean and standard deviation of $X_1, X_2, \ldots, X_n$, respectively. We set the level of significance at $\alpha = 0.05$ and used the decision rule: Reject H_0 if $T_n \leq t_{0.05,n-1}$. For the study, we set $n = 30$, $\epsilon = 0.20$, and $\sigma_c = 25$. For μ_c, we selected the five values of $0, 5, 10, 15$, and 20, as shown in Table 8.3.1. For each of these five situations, we ran 10,000 simulations and recorded $\widehat{\alpha}$, which is the number of rejections of H_0 divided by the number of simulations, i.e., the empirical α level.

For the test to be valid, $\widehat{\alpha}$ should be close to the nominal value of 0.05. As Table 8.3.1 shows, though, for all cases other than $\mu_c = 0$, the t-test is quite liberal; that is, its empirical significance level far exceeds the nominal 0.05 level (as Exercise

Table 8.3.1: Empirical α Levels for the Nominal 0.05 t-Test of Example 8.3.4.

	Empirical α				
μ_c	0	5	10	15	20
$\widehat{\alpha}$	0.0458	0.0961	0.1238	0.1294	0.1301

8.3.19 shows, the sampling error in the table is about 0.004). Note that when $\mu_c = 0$ the distribution of X is symmetric about 0 and in this case the empirical level is close to the nominal value of 0.05. ■

8.3.2 Likelihood Ratio Tests for Testing Variances of Normal Distributions

In this section, we discuss likelihood ratio tests for variances of normal distributions. In the next example, we begin with the two sample problem.

Example 8.3.5. In Example 8.3.1, in testing the equality of the means of two normal distributions, it was assumed that the unknown variances of the distributions were equal. Let us now consider the problem of testing the equality of these two unknown variances. We are given the independent random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from the distributions, which are $N(\theta_1, \theta_3)$ and $N(\theta_2, \theta_4)$, respectively. We have

$$\Omega = \{(\theta_1, \theta_2, \theta_3, \theta_4) : -\infty < \theta_1, \theta_2 < \infty, 0 < \theta_3, \theta_4 < \infty\}.$$

The hypothesis $H_0 : \theta_3 = \theta_4$, unspecified, with θ_1 and θ_2 also unspecified, is to be tested against all alternatives. Then

$$\omega = \{(\theta_1, \theta_2, \theta_3, \theta_4) : -\infty < \theta_1, \theta_2 < \infty, 0 < \theta_3 = \theta_4 < \infty\}.$$

It is easy to show (see Exercise 8.3.11) that the statistic defined by $\Lambda = L(\hat{\omega})/L(\hat{\Omega})$ is a function of the statistic

$$F = \frac{\sum_1^n (X_i - \overline{X})^2/(n-1)}{\sum_1^m (Y_i - \overline{Y})^2/(m-1)}. \tag{8.3.9}$$

If $\theta_3 = \theta_4$, this statistic F has an F-distribution with $n-1$ and $m-1$ degrees of freedom. The hypothesis that $(\theta_1, \theta_2, \theta_3, \theta_4) \in \omega$ is rejected if the computed $F \leq c_1$ or if the computed $F \geq c_2$. The constants c_1 and c_2 are usually selected so that, if $\theta_3 = \theta_4$,

$$P(F \leq c_1) = P(F \geq c_2) = \frac{\alpha_1}{2},$$

where α_1 is the desired significance level of this test. The power function of this test is derived in Exercise 8.3.10. ■

Remark 8.3.2. We caution the reader on this last test for the equality of two variances. In Remark 8.3.1, we discussed that the one- and two-sample t-tests for means are asymptotically correct. The two-sample variance test of the last example is not, however; see, for example, page 143 of Hettmansperger and McKean (2011). If the underlying distributions are not normal, then the F-critical values may be far from valid critical values (unlike the t-critical values for the means tests as discussed in Remark 8.3.1). In a large simulation study, Conover, Johnson, and Johnson (1981) showed that instead of having the nominal size of $\alpha = 0.05$, the F-test for variances using the F-critical values could have significance levels as high as 0.80, in certain nonnormal situations. Thus the two-sample F-test for variances does not possess robustness of validity. It should only be used in situations where the assumption of normality can be justified. See Exercise 8.3.17 for an illustrative data set. ■

The corresponding likelihood ratio test for the variance of a normal distribution based on one sample is discussed in Exercise 8.3.9. The cautions raised in Remark 8.3.1, hold for this test also.

Example 8.3.6. Let the independent random variables X and Y have distributions that are $N(\theta_1, \theta_3)$ and $N(\theta_2, \theta_4)$. In Example 8.3.1, we derived the likelihood ratio test statistic T of the hypothesis $\theta_1 = \theta_2$ when $\theta_3 = \theta_4$, while in Example 8.3.5 we obtained the likelihood ratio test statistic F of the hypothesis $\theta_3 = \theta_4$. The hypothesis that $\theta_1 = \theta_2$ is rejected if the computed $|T| \geq c$, where the constant c is selected so that $\alpha_2 = P(|T| \geq c; \theta_1 = \theta_2, \theta_3 = \theta_4)$ is the assigned significance level of the test. We shall show that, if $\theta_3 = \theta_4$, the likelihood ratio test statistics for equality of variances and equality of means, respectively F and T, are independent. Among other things, this means that if these two tests based on F and T, respectively, are performed sequentially with significance levels α_1 and α_2, the probability of accepting both these hypotheses, when they are true, is $(1 - \alpha_1)(1 - \alpha_2)$. Thus the significance level of this joint test is $\alpha = 1 - (1 - \alpha_1)(1 - \alpha_2)$.

Independence of F and T, when $\theta_3 = \theta_4$, can be established using sufficiency and completeness. The statistics $\overline{X}$, $\overline{Y}$, and $\sum_1^n (X_i - \overline{X})^2 + \sum_1^n (Y_i - \overline{Y})^2$ are joint complete sufficient statistics for the three parameters θ_1, θ_2, and $\theta_3 = \theta_4$. Obviously, the distribution of F does not depend upon θ_1, θ_2, or $\theta_3 = \theta_4$, and hence F is independent of the three joint complete sufficient statistics. However, T is a function of these three joint complete sufficient statistics alone, and, accordingly, T is independent of F. It is important to note that these two statistics are independent whether $\theta_1 = \theta_2$ or $\theta_1 \neq \theta_2$. This permits us to calculate probabilities other than the significance level of the test. For example, if $\theta_3 = \theta_4$ and $\theta_1 \neq \theta_2$, then

$$P(c_1 < F < c_2,\ |T| \geq c) = P(c_1 < F < c_2) P(|T| \geq c).$$

The second factor in the right-hand member is evaluated by using the probabilities of a noncentral t-distribution. Of course, if $\theta_3 = \theta_4$ and the difference $\theta_1 - \theta_2$ is large, we would want the preceding probability to be close to 1 because the event $\{c_1 < F < c_2,\ |T| \geq c\}$ leads to a correct decision, namely, accept $\theta_3 = \theta_4$ and reject $\theta_1 = \theta_2$. ■

EXERCISES

8.3.1. Verzani (2014) discusses a data set on healthy individuals, including their temperatures by gender. The data are in the file `tempbygender.rda` and the variables of interest are `maletemp` and `femaletemp`. Download this file from the site listed in the Preface.

(a) Obtain comparison boxplots. Comment on the plots. Which, if any, gender seems to have lower temperatures? Based on the width of the boxplots, comment on the assumption of equal variances.

(b) As discussed in Example 8.3.3, compute the two-sample, two-sided t-test that there is no difference in the true mean temperatures between genders. Obtain the p-value of the test and conclude in terms of the problem at the nominal α-level of 0.05.

(c) Obtain a 95% confidence interval for the difference in means. What does it mean in terms of the problem?

8.3.2. Verify Equations (8.3.2) of Example 8.3.1 of this section.

8.3.3. Verify Equations (8.3.3) of Example 8.3.1 of this section.

8.3.4. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ follow the location model

$$\begin{aligned} X_i &= \theta_1 + Z_i, \quad i = 1, \ldots, n \\ Y_i &= \theta_2 + Z_{n+i}, \quad i = 1, \ldots, m, \end{aligned}$$

where $Z_1, \ldots, Z_{n+m}$ are iid random variables with common pdf $f(z)$. Assume that $E(Z_i) = 0$ and $\text{Var}(Z_i) = \theta_3 < \infty$.

(a) Show that $E(X_i) = \theta_1$, $E(Y_i) = \theta_2$, and $\text{Var}(X_i) = \text{Var}(Y_i) = \theta_3$.

(b) Consider the hypotheses of Example 8.3.1, i.e.,

$$H_0 : \theta_1 = \theta_2 \text{ versus } H_1 : \theta_1 \neq \theta_2.$$

Show that under H_0, the test statistic T given in expression (8.3.4) has a limiting $N(0, 1)$ distribution.

(c) Using part (b), determine the corresponding large sample test (decision rule) of H_0 versus H_1. (This shows that the test in Example 8.3.1 is asymptotically correct.)

8.3.5. In Example 8.3.2, the power function for the one-sample t-test is discussed.

(a) Plot the power function for the following setup: X has a $N(\mu, \sigma^2)$ distribution; $H_0 : \mu = 50$ versus $H_1 : \mu \neq 50$; $\alpha = 0.05$; $n = 25$; and $\sigma = 10$.

(b) Overlay the power curve in (a) with that for $\alpha = 0.01$. Comment.

(c) Overlay the power curve in (a) with that for $n = 35$. Comment.

(d) Determine the smallest value of n so the power exceeds 0.80 to detect $\mu = 53$. *Hint:* Modify the R function `tpowerg.R` so it returns the power for a specified alternative.

8.3.6. The effect that a certain drug (Drug A) has on increasing blood pressure is a major concern. It is thought that a modification of the drug (Drug B) will lessen the increase in blood pressure. Let μ_A and μ_B be the true mean increases in blood pressure due to Drug A and B, respectively. The hypotheses of interest are $H_0 : \mu_A = \mu_B = 0$ versus $H_1 : \mu_A > \mu_B = 0$. The two-sample t-test statistic discussed in Example 8.3.3 is to be used to conduct the analysis. The nominal level is set at $\alpha = 0.05$ For the experimental design assume that the sample sizes are the same; i.e., $m = n$. Also, based on data from Drug A, $\sigma = 30$ seems to be a reasonable selection for the common standard deviation. Determine the common sample size, so that the difference in means $\mu_A - \mu_B = 12$ has an 80% detection rate. Suppose when the experiment is over, due to patients dropping out, the sample sizes for Drugs A and B are respectively $n = 72$ and $m = 68$. What was the actual power of the experiment to detect the difference of 12?

8.3.7. Show that the likelihood ratio principle leads to the same test when testing a simple hypothesis H_0 against an alternative simple hypothesis H_1, as that given by the Neyman–Pearson theorem. Note that there are only two points in Ω.

8.3.8. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution $N(\theta, 1)$. Show that the likelihood ratio principle for testing $H_0 : \theta = \theta'$, where θ' is specified, against $H_1 : \theta \neq \theta'$ leads to the inequality $|\overline{x} - \theta'| \geq c$.

(a) Is this a uniformly most powerful test of H_0 against H_1?

(b) Is this a uniformly most powerful unbiased test of H_0 against H_1?

8.3.9. Let $X_1, X_2, \ldots, X_n$ be iid $N(\theta_1, \theta_2)$. Show that the likelihood ratio principle for testing $H_0 : \theta_2 = \theta_2'$ specified, and θ_1 unspecified, against $H_1 : \theta_2 \neq \theta_2'$, θ_1 unspecified, leads to a test that rejects when $\sum_1^n (x_i - \overline{x})^2 \leq c_1$ or $\sum_1^n (x_i - \overline{x})^2 \geq c_2$, where $c_1 < c_2$ are selected appropriately.

8.3.10. For the situation discussed in Example 8.3.5, derive the power function for the likelihood ratio test statistic given in expression (8.3.9).

8.3.11. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent random samples from the distributions $N(\theta_1, \theta_3)$ and $N(\theta_2, \theta_4)$, respectively.

(a) Show that the likelihood ratio for testing $H_0 : \theta_1 = \theta_2,\ \theta_3 = \theta_4$ against all alternatives is given by

$$\frac{\left[\sum_1^n (x_i - \overline{x})^2/n\right]^{n/2} \left[\sum_1^m (y_i - \overline{y})^2/m\right]^{m/2}}{\left\{\left[\sum_1^n (x_i - u)^2 + \sum_1^m (y_i - u)^2\right] \Big/ (m+n)\right\}^{(n+m)/2}},$$

where $u = (n\overline{x} + m\overline{y})/(n+m)$.

(b) Show that the likelihood ratio for testing $H_0 : \theta_3 = \theta_4$ with θ_1 and θ_2 unspecified can be based on the test statistic F given in expression (8.3.9).

8.3.12. Let $Y_1 < Y_2 < \cdots < Y_5$ be the order statistics of a random sample of size $n = 5$ from a distribution with pdf $f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}$, $-\infty < x < \infty$, for all real θ. Find the likelihood ratio test Λ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

8.3.13. A random sample $X_1, X_2, \ldots, X_n$ arises from a distribution given by

$$H_0 : f(x;\theta) = \frac{1}{\theta}, \quad 0 < x < \theta, \quad \text{zero elsewhere,}$$

or

$$H_1 : f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}, \quad 0 < x < \infty, \quad \text{zero elsewhere.}$$

Determine the likelihood ratio (Λ) test associated with the test of H_0 against H_1.

8.3.14. Consider a random sample $X_1, X_2, \ldots, X_n$ from a distribution with pdf $f(x;\theta) = \theta(1-x)^{\theta-1}$, $0 < x < 1$, zero elsewhere, where $\theta > 0$.

(a) Find the form of the uniformly most powerful test of $H_0 : \theta = 1$ against $H_1 : \theta > 1$.

(b) What is the likelihood ratio Λ for testing $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$?

8.3.15. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be independent random samples from two normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, where σ^2 is the common but unknown variance.

(a) Find the likelihood ratio Λ for testing $H_0 : \mu_1 = \mu_2 = 0$ against all alternatives.

(b) Rewrite Λ so that it is a function of a statistic Z which has a well-known distribution.

(c) Give the distribution of Z under both null and alternative hypotheses.

8.3.16. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution with $\mu_1, \mu_2, \sigma_1^2 = \sigma_2^2 = \sigma^2, \rho = \frac{1}{2}$, where μ_1, μ_2, and $\sigma^2 > 0$ are unknown real numbers. Find the likelihood ratio Λ for testing $H_0 : \mu_1 = \mu_2 = 0$, σ^2 unknown against all alternatives. The likelihood ratio Λ is a function of what statistic that has a well-known distribution?

8.3.17. Let X be a random variable with pdf $f_X(x) = (2b_X)^{-1}\exp\{-|x|/b_X\}$, for $-\infty < x < \infty$ and $b_X > 0$. First, show that the variance of X is $\sigma_X^2 = 2b_X^2$. Next, let Y, independent of X, have pdf $f_Y(y) = (2b_Y)^{-1}\exp\{-|y|/b_Y\}$, for $-\infty < x < \infty$ and $b_Y > 0$. Consider the hypotheses

$$H_0 : \ \sigma_X^2 = \sigma_Y^2 \text{ versus } H_1 : \ \sigma_X^2 > \sigma_Y^2.$$

To illustrate Remark 8.3.2 for testing these hypotheses, consider the following data set (data are also in the file `exercise8316.rda`). Sample 1 represents the values of a sample drawn on X with $b_X = 1$, while Sample 2 represents the values of a sample drawn on Y with $b_Y = 1$. Hence, in this case H_0 is true.

Sample	−0.389	−2.177	0.813	−0.001
1	−0.110	−0.709	0.456	0.135
Sample	0.763	−0.570	−2.565	−1.733
1	0.403	0.778	−0.115	
Sample	−1.067	−0.577	0.361	−0.680
2	−0.634	−0.996	−0.181	0.239
Sample	−0.775	−1.421	−0.818	0.328
2	0.213	1.425	−0.165	

(a) Obtain *comparison boxplots* of these two samples. Comparison boxplots consist of boxplots of both samples drawn on the same scale. Based on these plots, in particular the interquartile ranges, what do you conclude about H_0?

(b) Obtain the F-test (for a one-sided hypothesis) as discussed in Remark 8.3.2 at level $\alpha = 0.10$. What is your conclusion?

(c) The test in part (b) is not exact. Why?

8.3.18. For the skewed contaminated normal random variable X of Example 8.3.4, derive the cdf, pdf, mean, and variance of X.

8.3.19. For Table 8.3.1 of Example 8.3.4, show that the half-width of the 95% confidence interval for a binomial proportion as given in Chapter 4 is 0.004 at the nominal value of 0.05.

8.3.20. If computational facilities are available, perform a Monte Carlo study of the two-sided t-test for the skewed contaminated normal situation of Example 8.3.4. The R function `rscn.R` generates variates from the distribution of X.

8.3.21. Suppose $X_1, \ldots, X_n$ is a random sample on X which has a $N(\mu, \sigma_0^2)$ distribution, where σ_0^2 is known. Consider the two-sided hypotheses

$$H_0 : \mu = 0 \text{ versus } H_1 : \mu \neq 0.$$

Show that the test based on the critical region $C = \{|\overline{X}| > \sqrt{\sigma_0^2/n} z_{\alpha/2}\}$ is an unbiased level α test.

8.3.22. Assume the same situation as in the last exercise but consider the test with critical region $C^* = \{\overline{X} > \sqrt{\sigma_0^2/n} z_\alpha\}$. Show that the test based on C^* has significance level α but that it is not an unbiased test.

8.4 *The Sequential Probability Ratio Test

Theorem 8.1.1 provides us with a method for determining a best critical region for testing a simple hypothesis against an alternative simple hypothesis. Recall its statement: Let $X_1, X_2, \ldots, X_n$ be a random sample with fixed sample size n from a distribution that has pdf or pmf $f(x; \theta)$, where $\theta = \{\theta : \theta = \theta', \theta''\}$ and θ' and θ''

are known numbers. For this section, we denote the likelihood of $X_1, X_2, \ldots, X_n$ by

$$L(\theta; n) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta),$$

a notation that reveals both the parameter θ and the sample size n. If we reject $H_0 : \theta = \theta'$ and accept $H_1 : \theta = \theta''$ when and only when

$$\frac{L(\theta'; n)}{L(\theta''; n)} \leq k,$$

where $k > 0$, then by Theorem 8.1.1 this is a best test of H_0 against H_1.

Let us now suppose that the sample size n is *not* fixed in advance. In fact, let the sample size be a random variable N with sample space $\{1, 2, , 3, \ldots\}$. An interesting procedure for testing the simple hypothesis $H_0 : \theta = \theta'$ against the simple hypothesis $H_1 : \theta = \theta''$ is the following: Let k_0 and k_1 be two positive constants with $k_0 < k_1$. Observe the independent outcomes $X_1, X_2, X_3, \ldots$ in a sequence, for example, $x_1, x_2, x_3, \ldots$, and compute

$$\frac{L(\theta'; 1)}{L(\theta''; 1)}, \frac{L(\theta'; 2)}{L(\theta''; 2)}, \frac{L(\theta'; 3)}{L(\theta''; 3)}, \ldots.$$

The hypothesis $H_0 : \theta = \theta'$ is rejected (and $H_1 : \theta = \theta''$ is accepted) if and only if there exists a positive integer n so that $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$ belongs to the set

$$C_n = \left\{ \mathbf{x}_n : k_0 < \frac{L(\theta', j)}{L(\theta'', j)} < k_1,\ j = 1, \ldots, n-1, \text{ and } \frac{L(\theta', n)}{L(\theta'', n)} \leq k_0 \right\}. \quad (8.4.1)$$

On the other hand, the hypothesis $H_0 : \theta = \theta'$ is accepted (and $H_1 : \theta = \theta''$ is rejected) if and only if there exists a positive integer n so that $(x_1, x_2, \ldots, x_n)$ belongs to the set

$$B_n = \left\{ \mathbf{x}_n : k_0 < \frac{L(\theta', j)}{L(\theta'', j)} < k_1,\ j = 1, \ldots, n-1, \text{ and } \frac{L(\theta', n)}{L(\theta'', n)} \geq k_1 \right\}. \quad (8.4.2)$$

That is, we continue to observe sample observations as long as

$$k_0 < \frac{L(\theta', n)}{L(\theta'', n)} < k_1. \quad (8.4.3)$$

We stop these observations in one of two ways:

1. With rejection of $H_0 : \theta = \theta'$ as soon as

$$\frac{L(\theta', n)}{L(\theta'', n)} \leq k_0$$

or

2. With acceptance of $H_0 : \theta = \theta'$ as soon as

$$\frac{L(\theta', n)}{L(\theta'', n)} \geq k_1,$$

A test of this kind is called Wald's **sequential probability ratio test**. Frequently, inequality (8.4.3) can be conveniently expressed in an equivalent form:

$$c_0(n) < u(x_1, x_2, \ldots, x_n) < c_1(n), \tag{8.4.4}$$

where $u(X_1, X_2, \ldots, X_n)$ is a statistic and $c_0(n)$ and $c_1(n)$ depend on the constants $k_0, k_1, \theta', \theta''$, and on n. Then the observations are stopped and a decision is reached as soon as

$$u(x_1, x_2, \ldots, x_n) \leq c_0(n) \quad \text{or} \quad u(x_1, x_2, \ldots, x_n) \geq c_1(n).$$

We now give an illustrative example.

Example 8.4.1. Let X have a pmf

$$f(x;\theta) = \begin{cases} \theta^x(1-\theta)^{1-x} & x = 0, 1 \\ 0 & \text{elsewhere.} \end{cases}$$

In the preceding discussion of a sequential probability ratio test, let $H_0 : \theta = \frac{1}{3}$ and $H_1 : \theta = \frac{2}{3}$; then, with $\sum x_i = \sum_{i=1}^n x_i$,

$$\frac{L(\frac{1}{3}, n)}{L(\frac{2}{3}, n)} = \frac{(\frac{1}{3})^{\sum x_i}(\frac{2}{3})^{n-\sum x_i}}{(\frac{2}{3})^{\sum x_i}(\frac{1}{3})^{n-\sum x_i}} = 2^{n-2\sum x_i}.$$

If we take logarithms to the base 2, the inequality

$$k_0 < \frac{L(\frac{1}{3}, n)}{L(\frac{2}{3}, n)} < k_1,$$

with $0 < k_0 < k_1$, becomes

$$\log_2 k_0 < n - 2\sum_1^n x_i < \log_2 k_1,$$

or, equivalently, in the notation of expression (8.4.4),

$$c_0(n) = \frac{n}{2} - \frac{1}{2}\log_2 k_1 < \sum_1^n x_i < \frac{n}{2} - \frac{1}{2}\log_2 k_0 = c_1(n).$$

Note that $L(\frac{1}{3}, n)/L(\frac{2}{3}, n) \leq k_0$ if and only if $c_1(n) \leq \sum_1^n x_i$; and $L(\frac{1}{3}, n)/L(\frac{2}{3}, n) \geq k_1$ if and only if $c_0(n) \geq \sum_1^n x_i$. Thus we continue to observe outcomes as long as $c_0(n) < \sum_1^n x_i < c_1(n)$. The observation of outcomes is discontinued with the first value of n of N for which either $c_1(n) \leq \sum_1^n x_i$ or $c_0(n) \geq \sum_1^n x_i$. The inequality $c_1(n) \leq \sum_1^n x_i$ leads to rejection of $H_0 : \theta = \frac{1}{3}$ (the acceptance of H_1), and the inequality $c_0(n) \geq \sum_1^n x_i$ leads to the acceptance of $H_0 : \theta = \frac{1}{3}$ (the rejection of H_1). ■

Remark 8.4.1. At this point, the reader undoubtedly sees that there are many questions that should be raised in connection with the sequential probability ratio test. Some of these questions are possibly among the following:

1. What is the probability of the procedure continuing indefinitely?
2. What is the value of the power function of this test at each of the points $\theta = \theta'$ and $\theta = \theta''$?
3. If θ'' is one of several values of θ specified by an alternative composite hypothesis, say $H_1 : \theta > \theta'$, what is the power function at each point $\theta \geq \theta'$?
4. Since the sample size N is a random variable, what are some of the properties of the distribution of N? In particular, what is the expected value $E(N)$ of N?
5. How does this test compare with tests that have a fixed sample size n? ■

A course in sequential analysis would investigate these and many other problems. However, in this book our objective is largely that of acquainting the reader with this kind of test procedure. Accordingly, we assert that the answer to question 1 is zero. Moreover, it can be proved that if $\theta = \theta'$ or if $\theta = \theta''$, $E(N)$ is smaller for this sequential procedure than the sample size of a fixed-sample-size test that has the same values of the power function at those points. We now consider question 2 in some detail.

In this section we shall denote the power of the test when H_0 is true by the symbol α and the power of the test when H_1 is true by the symbol $1 - \beta$. Thus α is the probability of committing a Type I error (the rejection of H_0 when H_0 is true), and β is the probability of committing a Type II error (the acceptance of H_0 when H_0 is false). With the sets C_n and B_n as previously defined, and with random variables of the continuous type, we then have

$$\alpha = \sum_{n=1}^{\infty} \int_{C_n} L(\theta', n), \quad 1 - \beta = \sum_{n=1}^{\infty} \int_{C_n} L(\theta'', n).$$

Since the probability is 1 that the procedure terminates, we also have

$$1 - \alpha = \sum_{n=1}^{\infty} \int_{B_n} L(\theta', n), \quad \beta = \sum_{n=1}^{\infty} \int_{B_n} L(\theta'', n).$$

If $(x_1, x_2, \ldots, x_n) \in C_n$, we have $L(\theta', n) \leq k_0 L(\theta'', n)$; hence, it is clear that

$$\alpha = \sum_{n=1}^{\infty} \int_{C_n} L(\theta', n) \leq \sum_{n=1}^{\infty} \int_{C_n} k_0 L(\theta'', n) = k_0(1 - \beta).$$

Because $L(\theta', n) \geq k_1 L(\theta'', n)$ at each point of the set B_n, we have

$$1 - \alpha = \sum_{n=1}^{\infty} \int_{B_n} L(\theta', n) \geq \sum_{n=1}^{\infty} \int_{B_n} k_1 L(\theta'', n) = k_1 \beta.$$

Accordingly, it follows that

$$\frac{\alpha}{1 - \beta} \leq k_0, \quad k_1 \leq \frac{1 - \alpha}{\beta}, \tag{8.4.5}$$

provided that β is not equal to 0 or 1.

Now let α_a and β_a be preassigned proper fractions; some typical values in the applications are 0.01, 0.05, and 0.10. If we take

$$k_0 = \frac{\alpha_a}{1-\beta_a}, \quad k_1 = \frac{1-\alpha_a}{\beta_a},$$

then inequalities (8.4.5) become

$$\frac{\alpha}{1-\beta} \leq \frac{\alpha_a}{1-\beta_a}, \quad \frac{1-\alpha_a}{\beta_a} \leq \frac{1-\alpha}{\beta}; \tag{8.4.6}$$

or, equivalently,

$$\alpha(1-\beta_a) \leq (1-\beta)\alpha_a, \quad \beta(1-\alpha_a) \leq (1-\alpha)\beta_a.$$

If we add corresponding members of the immediately preceding inequalities, we find that

$$\alpha + \beta - \alpha\beta_a - \beta\alpha_a \leq \alpha_a + \beta_a - \beta\alpha_a - \alpha\beta_a$$

and hence

$$\alpha + \beta \leq \alpha_a + \beta_a;$$

that is, the sum $\alpha + \beta$ of the probabilities of the two kinds of errors is bounded above by the sum $\alpha_a + \beta_a$ of the preassigned numbers. Moreover, since α and β are positive proper fractions, inequalities (8.4.6) imply that

$$\alpha \leq \frac{\alpha_a}{1-\beta_a}, \qquad \beta \leq \frac{\beta_a}{1-\alpha_a};$$

consequently, we have an upper bound on each of α and β. Various investigations of the sequential probability ratio test seem to indicate that in most practical cases, the values of α and β are quite close to α_a and β_a. This prompts us to approximate the power function at the points $\theta = \theta'$ and $\theta = \theta''$ by α_a and $1 - \beta_a$, respectively.

Example 8.4.2. Let X be $N(\theta, 100)$. To find the sequential probability ratio test for testing $H_0 : \theta = 75$ against $H_1 : \theta = 78$ such that each of α and β is approximately equal to 0.10, take

$$k_0 = \frac{0.10}{1-0.10} = \frac{1}{9}, \qquad k_1 = \frac{1-0.10}{0.10} = 9.$$

Since

$$\frac{L(75,n)}{L(78,n)} = \frac{\exp\left[-\sum(x_i-75)^2/2(100)\right]}{\exp\left[-\sum(x_i-78)^2/2(100)\right]} = \exp\left(-\frac{6\sum x_i - 459n}{200}\right),$$

the inequality

$$k_0 = \frac{1}{9} < \frac{L(75,n)}{L(78,n)} < 9 = k_1$$

can be rewritten, by taking logarithms, as

$$-\log 9 < \frac{6\sum x_i - 459n}{200} < \log 9.$$

This inequality is equivalent to the inequality

$$c_0(n) = \frac{153}{2}n - \frac{100}{3}\log 9 < \sum_1^n x_i < \frac{153}{2}n + \frac{100}{3}\log 9 = c_1(n).$$

Moreover, $L(75, n)/L(78, n) \leq k_0$ and $L(75, n)/L(78, n) \geq k_1$ are equivalent to the inequalities $\sum_1^n x_i \geq c_1(n)$ and $\sum_1^n x_i \leq c_0(n)$, respectively. Thus the observation of outcomes is discontinued with the first value of n of N for which either $\sum_1^n x_i \geq c_1(n)$ or $\sum_1^n x_i \leq c_0(n)$. The inequality $\sum_1^n x_i \geq c_1(n)$ leads to the rejection of $H_0 : \theta = 75$, and the inequality $\sum_1^n x_i \leq c_0(n)$ leads to the acceptance of $H_0 : \theta = 75$. The power of the test is approximately 0.10 when H_0 is true, and approximately 0.90 when H_1 is true. ■

Remark 8.4.2. It is interesting to note that a sequential probability ratio test can be thought of as a *random-walk procedure*. To illustrate, the final inequalities of Examples 8.4.1 and 8.4.2 can be written as

$$-\log_2 k_1 < \sum_1^n 2(x_i - 0.5) < -\log_2 k_0$$

and

$$-\frac{100}{3}\log 9 < \sum_1^n (x_i - 76.5) < \frac{100}{3}\log 9,$$

respectively. In each instance, think of starting at the point zero and taking random steps until one of the boundaries is reached. In the first situation the random steps are $2(X_1 - 0.5), 2(X_2 - 0.5), 2(X_3 - 0.5), \ldots$, which have the same length, 1, but with random directions. In the second instance, both the length and the direction of the steps are random variables, $X_1 - 76.5, X_2 - 76.5, X_3 - 76.5, \ldots$. ■

In recent years, there has been much attention devoted to improving quality of products using statistical methods. One such simple method was developed by Walter Shewhart in which a sample of size n of the items being produced is taken and they are measured, resulting in n values. The mean $\overline{X}$ of these n measurements has an approximate normal distribution with mean μ and variance σ^2/n. In practice, μ and σ^2 must be estimated, but in this discussion, we assume that they are known. From theory we know that the probability is 0.997 that $\overline{x}$ is between

$$\text{LCL} = \mu - \frac{3\sigma}{\sqrt{n}} \quad \text{and} \quad \text{UCL} = \mu + \frac{3\sigma}{\sqrt{n}}.$$

These two values are called the lower (LCL) and upper (UCL) control limits, respectively. Samples like these are taken periodically, resulting in a sequence of means,

say $\overline{x}_1, \overline{x}_2, \overline{x}_3, \ldots$. These are usually plotted; and if they are between the LCL and UCL, we say that the process is **in control**. If one falls outside the limits, this would suggest that the mean μ has shifted, and the process would be investigated.

It was recognized by some that there could be a shift in the mean, say from μ to $\mu + (\sigma/\sqrt{n})$; and it would still be difficult to detect that shift with a single sample mean, for now the probability of a single $\overline{x}$ exceeding UCL is only about 0.023. This means that we would need about $1/0.023 \approx 43$ samples, each of size n, on the average before detecting such a shift. This seems too long; so statisticians recognized that they should be cumulating experience as the sequence $\overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots$ is observed in order to help them detect the shift sooner. It is the practice to compute the standardized variable $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$; thus, we state the problem in these terms and provide the solution given by a sequential probability ratio test.

Here Z is $N(\theta, 1)$, and we wish to test $H_0 : \theta = 0$ against $H_1 : \theta = 1$ using the sequence of iid random variables $Z_1, Z_2, \ldots, Z_m, \ldots$. We use m rather than n, as the latter is the size of the samples taken periodically. We have

$$\frac{L(0,m)}{L(1,m)} = \frac{\exp\left[-\sum z_i^2/2\right]}{\exp\left[-\sum (z_i - 1)^2/2\right]} = \exp\left[-\sum_{i=1}^{m}(z_i - 0.5)\right].$$

Thus

$$k_0 < \exp\left[-\sum_{i=1}^{m}(z_i - 0.5)\right] < k_1$$

can be written as

$$h = -\log k_0 > \sum_{i=1}^{m}(z_i - 0.5) > -\log k_1 = -h.$$

It is true that $-\log k_0 = \log k_1$ when $\alpha_a = \beta_a$. Often, $h = -\log k_0$ is taken to be about 4 or 5, suggesting that $\alpha_a = \beta_a$ is small, like 0.01. As $\sum(z_i - 0.5)$ is cumulating the sum of $z_i - 0.5$, $i = 1, 2, 3, \ldots$, these procedures are often called CUSUMS. If the CUSUM $= \sum(z_i - 0.5)$ exceeds h, we would investigate the process, as it seems that the mean has shifted upward. If this shift is to $\theta = 1$, the theory associated with these procedures shows that we need only eight or nine samples on the average, rather than 43, to detect this shift. For more information about these methods, the reader is referred to one of the many books on quality improvement through statistical methods. What we would like to emphasize here is that through sequential methods (not only the sequential probability ratio test), we should take advantage of all past experience that we can gather in making inferences.

EXERCISES

8.4.1. Let X be $N(0, \theta)$ and, in the notation of this section, let $\theta' = 4$, $\theta'' = 9$, $\alpha_a = 0.05$, and $\beta_a = 0.10$. Show that the sequential probability ratio test can be based upon the statistic $\sum_1^n X_i^2$. Determine $c_0(n)$ and $c_1(n)$.

8.4.2. Let X have a Poisson distribution with mean θ. Find the sequential probability ratio test for testing $H_0 : \theta = 0.02$ against $H_1 : \theta = 0.07$. Show that this test can be based upon the statistic $\sum_1^n X_i$. If $\alpha_a = 0.20$ and $\beta_a = 0.10$, find $c_0(n)$ and $c_1(n)$.

8.4.3. Let the independent random variables Y and Z be $N(\mu_1, 1)$ and $N(\mu_2, 1)$, respectively. Let $\theta = \mu_1 - \mu_2$. Let us observe independent observations from each distribution, say $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$. To test sequentially the hypothesis $H_0 : \theta = 0$ against $H_1 : \theta = \frac{1}{2}$, use the sequence $X_i = Y_i - Z_i,\ i = 1, 2, \ldots$. If $\alpha_a = \beta_a = 0.05$, show that the test can be based upon $\overline{X} = \overline{Y} - \overline{Z}$. Find $c_0(n)$ and $c_1(n)$.

8.4.4. Suppose that a manufacturing process makes about 3% defective items, which is considered satisfactory for this particular product. The managers would like to decrease this to about 1% and clearly want to guard against a substantial increase, say to 5%. To monitor the process, periodically $n = 100$ items are taken and the number X of defectives counted. Assume that X is $b(n = 100, p = \theta)$. Based on a sequence $X_1, X_2, \ldots, X_m, \ldots$, determine a sequential probability ratio test that tests $H_0 : \theta = 0.01$ against $H_1 : \theta = 0.05$. (Note that $\theta = 0.03$, the present level, is in between these two values.) Write this test in the form

$$h_0 > \sum_{i=1}^{m} (x_i - nd) > h_1$$

and determine d, h_0, and h_1 if $\alpha_a = \beta_a = 0.02$.

8.4.5. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x; \theta) = \theta x^{\theta - 1},\ 0 < x < 1$, zero elsewhere.

(a) Find a complete sufficient statistic for θ.

(b) If $\alpha_a = \beta_a = \frac{1}{10}$, find the sequential probability ratio test of $H_0 : \theta = 2$ against $H_1 : \theta = 3$.

8.5 *Minimax and Classification Procedures

We have considered several procedures that may be used in problems of point estimation. Among these were decision function procedures (in particular, minimax decisions). In this section, we apply minimax procedures to the problem of testing a simple hypothesis H_0 against an alternative simple hypothesis H_1. It is important to observe that these procedures yield, in accordance with the Neyman–Pearson theorem, a best test of H_0 against H_1. We end this section with a discussion on an application of these procedures to a classification problem.

8.5.1 Minimax Procedures

We first investigate the decision function approach to the problem of testing a simple null hypothesis against a simple alternative hypothesis. Let the joint pdf of the n

random variables $X_1, X_2, \ldots, X_n$ depend upon the parameter θ. Here n is a fixed positive integer. This pdf is denoted by $L(\theta; x_1, x_2, \ldots, x_n)$ or, for brevity, by $L(\theta)$. Let θ' and θ'' be distinct and fixed values of θ. We wish to test the simple hypothesis $H_0 : \theta = \theta'$ against the simple hypothesis $H_1 : \theta = \theta''$. Thus the parameter space is $\Omega = \{\theta : \theta = \theta', \theta''\}$. In accordance with the decision function procedure, we need a function δ of the observed values of $X_1, \ldots, X_n$ (or, of the observed value of a statistic Y) that decides which of the two values of θ, θ' or θ'', to accept. That is, the function δ selects either $H_0 : \theta = \theta'$ or $H_1 : \theta = \theta''$. We denote these decisions by $\delta = \theta'$ and $\delta = \theta''$, respectively. Let $\mathcal{L}(\theta, \delta)$ represent the loss function associated with this decision problem. Because the pairs $(\theta = \theta', \delta = \theta')$ and $(\theta = \theta'', \delta = \theta'')$ represent correct decisions, we shall always take $\mathcal{L}(\theta', \theta') = \mathcal{L}(\theta'', \theta'') = 0$. On the other hand, if either $\delta = \theta''$ when $\theta = \theta'$ or $\delta = \theta'$ when $\theta = \theta''$, then a positive value should be assigned to the loss function; that is, $\mathcal{L}(\theta', \theta'') > 0$ and $\mathcal{L}(\theta'', \theta') > 0$.

It has previously been emphasized that a test of $H_0 : \theta = \theta'$ against $H_1 : \theta = \theta''$ can be described in terms of a critical region in the sample space. We can do the same kind of thing with the decision function. That is, we can choose a subset of C of the sample space and if $(x_1, x_2, \ldots, x_n) \in C$, we can make the decision $\delta = \theta''$; whereas if $(x_1, x_2, \ldots, x_n) \in C^c$, the complement of C, we make the decision $\delta = \theta'$. Thus a given critical region C determines the decision function. In this sense, we may denote the risk function by $R(\theta, C)$ instead of $R(\theta, \delta)$. That is, in a notation used in Section 7.1,

$$R(\theta, C) = R(\theta, \delta) = \int_{C \cup C^c} \mathcal{L}(\theta, \delta) L(\theta).$$

Since $\delta = \theta''$ if $(x_1, \ldots, x_n) \in C$ and $\delta = \theta'$ if $(x_1, \ldots, x_n) \in C^c$, we have

$$R(\theta, C) = \int_C \mathcal{L}(\theta, \theta'') L(\theta) + \int_{C^c} \mathcal{L}(\theta, \theta') L(\theta). \tag{8.5.1}$$

If, in Equation (8.5.1), we take $\theta = \theta'$, then $\mathcal{L}(\theta', \theta') = 0$ and hence

$$R(\theta', C) = \int_C \mathcal{L}(\theta', \theta'') L(\theta') = \mathcal{L}(\theta', \theta'') \int_C L(\theta').$$

On the other hand, if in Equation (8.5.1) we let $\theta = \theta''$, then $\mathcal{L}(\theta'', \theta'') = 0$ and, accordingly,

$$R(\theta'', C) = \int_{C^c} \mathcal{L}(\theta'', \theta') L(\theta'') = \mathcal{L}(\theta'', \theta') \int_{C^c} L(\theta'').$$

It is enlightening to note that if $\gamma(\theta)$ is the power function of the test associated with the critical region C, then

$$R(\theta', C) = \mathcal{L}(\theta', \theta'') \gamma(\theta') = \mathcal{L}(\theta', \theta'') \alpha,$$

where $\alpha = \gamma(\theta')$ is the significance level; and

$$R(\theta'', C) = \mathcal{L}(\theta'', \theta')[1 - \gamma(\theta'')] = \mathcal{L}(\theta'', \theta') \beta,$$

where $\beta = 1 - \gamma(\theta'')$ is the probability of the type II error.

Let us now see if we can find a minimax solution to our problem. That is, we want to find a critical region C so that

$$\max[R(\theta', C), R(\theta'', C)]$$

is minimized. We shall show that the solution is the region

$$C = \left\{(x_1, \ldots, x_n) : \frac{L(\theta'; x_1, \ldots, x_n)}{L(\theta''; x_1, \ldots, x_n)} \leq k\right\},$$

provided the positive constant k is selected so that $R(\theta', C) = R(\theta'', C)$. That is, if k is chosen so that

$$\mathcal{L}(\theta', \theta'') \int_C L(\theta') = \mathcal{L}(\theta'', \theta') \int_{C^c} L(\theta''),$$

then the critical region C provides a minimax solution. In the case of random variables of the continuous type, k can always be selected so that $R(\theta', C) = R(\theta'', C)$. However, with random variables of the discrete type, we may need to consider an auxiliary random experiment when $L(\theta')/L(\theta'') = k$ in order to achieve the exact equality $R(\theta', C) = R(\theta'', C)$.

To see that C is the minimax solution, consider every other region A for which $R(\theta', C) \geq R(\theta', A)$. A region A for which $R(\theta', C) < R(\theta', A)$ is not a candidate for a minimax solution, for then $R(\theta', C) = R(\theta'', C) < \max[R(\theta', A), R(\theta'', A)]$. Since $R(\theta', C) \geq R(\theta', A)$ means that

$$\mathcal{L}(\theta', \theta'') \int_C L(\theta') \geq \mathcal{L}(\theta', \theta'') \int_A L(\theta'),$$

we have

$$\alpha = \int_C L(\theta') \geq \int_A L(\theta');$$

that is, the significance level of the test associated with the critical region A is less than or equal to α. But C, in accordance with the Neyman–Pearson theorem, is a best critical region of size α. Thus

$$\int_C L(\theta'') \geq \int_A L(\theta'')$$

and

$$\int_{C^c} L(\theta'') \leq \int_{A^c} L(\theta'').$$

Accordingly,

$$\mathcal{L}(\theta'', \theta') \int_{C^c} L(\theta'') \leq \mathcal{L}(\theta'', \theta') \int_{A^c} L(\theta''),$$

or, equivalently,

$$R(\theta'', C) \leq R(\theta'', A).$$

That is,

$$R(\theta', C) = R(\theta'', C) \leq R(\theta'', A).$$

This means that

$$\max[R(\theta', C), R(\theta'', C)] \leq R(\theta'', A).$$

Then certainly,

$$\max[R(\theta', C), R(\theta'', C)] \leq \max[R(\theta', A), R(\theta'', A)],$$

and the critical region C provides a minimax solution, as we wanted to show.

Example 8.5.1. Let $X_1, X_2, \ldots, X_{100}$ denote a random sample of size 100 from a distribution that is $N(\theta, 100)$. We again consider the problem of testing $H_0 : \theta = 75$ against $H_1 : \theta = 78$. We seek a minimax solution with $\mathcal{L}(75, 78) = 3$ and $\mathcal{L}(78, 75) = 1$. Since $L(75)/L(78) \leq k$ is equivalent to $\overline{x} \geq c$, we want to determine c, and thus k, so that

$$3P(\overline{X} \geq c; \theta = 75) = P(\overline{X} < c; \theta = 78). \tag{8.5.2}$$

Because $\overline{X}$ is $N(\theta, 1)$, the preceding equation can be rewritten as

$$3[1 - \Phi(c - 75)] = \Phi(c - 78).$$

As requested in Exercise 8.5.4, the reader can show by using Newton's algorithm that the solution to one place is $c = 76.8$. The significance level of the test is $1 - \Phi(1.8) = 0.036$, approximately, and the power of the test when H_1 is true is $1 - \Phi(-1.2) = 0.885$, approximately. ■

8.5.2 Classification

The summary above has an interesting application to the problem of **classification**, which can be described as follows. An investigator makes a number of measurements on an item and wants to place it into one of several categories (or classify it). For convenience in our discussion, we assume that only two measurements, say X and Y, are made on the item to be classified. Moreover, let X and Y have a joint pdf $f(x, y; \theta)$, where the parameter θ represents one or more parameters. In our simplification, suppose that there are only two possible joint distributions (categories) for X and Y, which are indexed by the parameter values θ' and θ'', respectively. In this case, the problem then reduces to one of observing $X = x$ and $Y = y$ and then testing the hypothesis $\theta = \theta'$ against the hypothesis $\theta = \theta''$, with the classification of X and Y being in accord with which hypothesis is accepted. From the Neyman–Pearson theorem, we know that a best decision of this sort is of the following form: If

$$\frac{f(x, y; \theta')}{f(x, y; \theta'')} \leq k,$$

choose the distribution indexed by θ''; that is, we classify (x, y) as coming from the distribution indexed by θ''. Otherwise, choose the distribution indexed by θ'; that is, we classify (x, y) as coming from the distribution indexed by θ'. Some discussion on the choice of k follows in the next remark.

Remark 8.5.1 (On the Choice of k)**.** Consider the following probabilities:

$$\begin{aligned}\pi' &= P[(X,Y) \text{ is drawn from the distribution with pdf } f(x,y;\theta')]\\ \pi'' &= P[(X,Y) \text{ is drawn from the distribution with pdf } f(x,y;\theta'')].\end{aligned}$$

Note that $\pi' + \pi'' = 1$. Then it can be shown that the optimal classification rule is determined by taking $k = \pi''/\pi'$; see, for instance, Seber (1984). Hence, if we have prior information on how likely the item is drawn from the distribution with parameter θ', then we can obtain the classification rule. In practice, it is common for each distribution to be equilikely, in which case, $\pi' = \pi'' = 1/2$ and, hence, $k = 1$. ■

Example 8.5.2. Let (x, y) be an observation of the random pair (X, Y), which has a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ. In Section 3.5 that joint pdf is given by

$$f(x,y;\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-q(x,y;\mu_1,\mu_2)/2},$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$, where $\sigma_1 > 0,\ \sigma_2 > 0,\ -1 < \rho < 1$, and

$$q(x,y;\mu_1,\mu_2) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

Assume that σ_1^2, σ_2^2, and ρ are known but that we do not know whether the respective means of (X, Y) are (μ_1', μ_2') or (μ_1'', μ_2''). The inequality

$$\frac{f(x,y;\mu_1',\mu_2',\sigma_1^2,\sigma_2^2,\rho)}{f(x,y;\mu_1'',\mu_2'',\sigma_1^2,\sigma_2^2,\rho)} \le k$$

is equivalent to

$$\tfrac{1}{2}[q(x,y;\mu_1'',\mu_2'') - q(x,y;\mu_1',\mu_2')] \le \log k.$$

Moreover, it is clear that the difference in the left-hand member of this inequality does not contain terms involving x^2, xy, and y^2. In particular, this inequality is the same as

$$\begin{aligned}\frac{1}{1-\rho^2}\left\{\left[\frac{\mu_1'-\mu_1''}{\sigma_1^2} - \frac{\rho(\mu_2'-\mu_2'')}{\sigma_1\sigma_2}\right]x + \left[\frac{\mu_2'-\mu_2''}{\sigma_2^2} - \frac{\rho(\mu_1'-\mu_1'')}{\sigma_1\sigma_2}\right]y\right\}\\ \le \log k + \tfrac{1}{2}[q(0,0;\mu_1',\mu_2') - q(0,0;\mu_1'',\mu_2'')],\end{aligned}$$

or, for brevity,

$$ax + by \le c. \tag{8.5.3}$$

That is, if this linear function of x and y in the left-hand member of inequality (8.5.3) is less than or equal to a constant, we classify (x, y) as coming from the bivariate normal distribution with means μ_1'' and μ_2''. Otherwise, we classify (x, y) as arising from the bivariate normal distribution with means μ_1' and μ_2'. Of course, if the prior probabilities can be assigned as discussed in Remark 8.5.1 then k and thus c can be found easily; see Exercise 8.5.3. ■

Once the rule for classification is established, the statistician might be interested in the two probabilities of misclassifications using that rule. The first of these two is associated with the classification of (x, y) as arising from the distribution indexed by θ'' if, in fact, it comes from that index by θ'. The second misclassification is similar, but with the interchange of θ' and θ''. In the preceding example, the probabilities of these respective misclassifications are

$$P(aX + bY \le c; \mu_1', \mu_2') \quad \text{and} \quad P(aX + bY > c; \mu_1'', \mu_2'').$$

The distribution of $Z = aX + bY$ is obtained from Theorem 3.5.2. It follows that the distribution of $Z = aX + bY$ is given by

$$N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + 2ab\rho\sigma_1\sigma_2 + b^2\sigma_2^2).$$

With this information, it is easy to compute the probabilities of misclassifications; see Exercise 8.5.3.

One final remark must be made with respect to the use of the important classification rule established in Example 8.5.2. In most instances the parameter values μ_1', μ_2' and μ_1'', μ_2'' as well as σ_1^2, σ_2^2, and ρ are unknown. In such cases the statistician has usually observed a random sample (frequently called a *training sample*) from each of the two distributions. Let us say the samples have sizes n' and n'', respectively, with sample characteristics

$$\overline{x}', \overline{y}', (s_x')^2, (s_y')^2, r' \quad \text{and} \quad \overline{x}'', \overline{y}'', (s_x'')^2, (s_y'')^2, r''.$$

The statistics r' and r'' are the sample correlation coefficients, as defined in expression (9.7.1) of Section 9.7. The sample correlation coefficient is the mle for the correlation parameter ρ of a bivariate normal distribution; see Section 9.7. If in inequality (8.5.3) the parameters $\mu_1', \mu_2', \mu_1'', \mu_2'', \sigma_1^2, \sigma_2^2$, and $\rho\sigma_1\sigma_2$ are replaced by the unbiased estimates

$$\overline{x}', \overline{y}', \overline{x}'', \overline{y}'', \frac{(n'-1)(s_x')^2 + (n''-1)(s_x'')^2}{n' + n'' - 2}, \frac{(n'-1)(s_y')^2 + (n''-1)(s_y'')^2}{n' + n'' - 2},$$
$$\frac{(n'-1)r's_x's_y' + (n''-1)r''s_x''s_y''}{n' + n'' - 2},$$

the resulting expression in the left-hand member is frequently called Fisher's **linear discriminant function**. Since those parameters have been estimated, the distribution theory associated with $aX + bY$ does provide an approximation.

Although we have considered only bivariate distributions in this section, the results can easily be extended to multivariate normal distributions using the results of Section 3.5; see also Chapter 6 of Seber (1984).

EXERCISES

8.5.1. Let $X_1, X_2, \ldots, X_{20}$ be a random sample of size 20 from a distribution that is $N(\theta, 5)$. Let $L(\theta)$ represent the joint pdf of $X_1, X_2, \ldots, X_{20}$. The problem is to test $H_0 : \theta = 1$ against $H_1 : \theta = 0$. Thus $\Omega = \{\theta : \theta = 0, 1\}$.

(a) Show that $L(1)/L(0) \leq k$ is equivalent to $\overline{x} \leq c$.

(b) Find c so that the significance level is $\alpha = 0.05$. Compute the power of this test if H_1 is true.

(c) If the loss function is such that $\mathcal{L}(1,1) = \mathcal{L}(0,0) = 0$ and $\mathcal{L}(1,0) = \mathcal{L}(0,1) > 0$, find the minimax test. Evaluate the power function of this test at the points $\theta = 1$ and $\theta = 0$.

8.5.2. Let $X_1, X_2, \ldots, X_{10}$ be a random sample of size 10 from a Poisson distribution with parameter θ. Let $L(\theta)$ be the joint pdf of $X_1, X_2, \ldots, X_{10}$. The problem is to test $H_0 : \theta = \frac{1}{2}$ against $H_1 : \theta = 1$.

(a) Show that $L(\frac{1}{2})/L(1) \leq k$ is equivalent to $y = \sum_1^n x_i \geq c$.

(b) In order to make $\alpha = 0.05$, show that H_0 is rejected if $y > 9$ and, if $y = 9$, reject H_0 with probability $\frac{1}{2}$ (using some auxiliary random experiment).

(c) If the loss function is such that $\mathcal{L}(\frac{1}{2}, \frac{1}{2}) = \mathcal{L}(1,1) = 0$ and $\mathcal{L}(\frac{1}{2}, 1) = 1$ and $\mathcal{L}(1, \frac{1}{2}) = 2$, show that the minimax procedure is to reject H_0 if $y > 6$ and, if $y = 6$, reject H_0 with probability 0.08 (using some auxiliary random experiment).

8.5.3. In Example 8.5.2 let $\mu_1' = \mu_2' = 0$, $\mu_1'' = \mu_2'' = 1$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, and $\rho = \frac{1}{2}$.

(a) Find the distribution of the linear function $aX + bY$.

(b) With $k = 1$, compute $P(aX + bY \leq c; \mu_1' = \mu_2' = 0)$ and $P(aX + bY > c; \mu_1'' = \mu_2'' = 1)$.

8.5.4. Determine Newton's algorithm to find the solution of Equation (8.5.2). If software is available, write a program that performs your algorithm and then show that the solution is $c = 76.8$. If software is not available, solve (8.5.2) by "trial and error."

8.5.5. Let X and Y have the joint pdf

$$f(x, y; \theta_1, \theta_2) = \frac{1}{\theta_1 \theta_2} \exp\left(-\frac{x}{\theta_1} - \frac{y}{\theta_2}\right), \quad 0 < x < \infty, \ \ 0 < y < \infty,$$

zero elsewhere, where $0 < \theta_1, \ 0 < \theta_2$. An observation (x, y) arises from the joint distribution with parameters equal to either $(\theta_1' = 1, \theta_2' = 5)$ or $(\theta_1'' = 3, \theta_2'' = 2)$. Determine the form of the classification rule.

8.5.6. Let X and Y have a joint bivariate normal distribution. An observation (x, y) arises from the joint distribution with parameters equal to either

$$\mu_1' = \mu_2' = 0, \;\; (\sigma_1^2)' = (\sigma_2^2)' = 1, \;\; \rho' = \tfrac{1}{2}$$

or

$$\mu_1'' = \mu_2'' = 1, \;\; (\sigma_1^2)'' = 4, \;\; (\sigma_2^2)'' = 9, \;\; \rho'' = \tfrac{1}{2}.$$

Show that the classification rule involves a second-degree polynomial in x and y.

8.5.7. Let $\boldsymbol{W}' = (W_1, W_2)$ be an observation from one of two bivariate normal distributions, I and II, each with $\mu_1 = \mu_2 = 0$ but with the respective variance-covariance matrices

$$\boldsymbol{V}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \quad \text{and} \quad \boldsymbol{V}_2 = \begin{pmatrix} 3 & 0 \\ 0 & 12 \end{pmatrix}.$$

How would you classify $\boldsymbol{W}$ into I or II?

Chapter 9

Inferences About Normal Linear Models

9.1 Introduction

In this chapter, we consider analyses of some of the most widely used linear models. These models include one- and two-way analysis of variance (ANOVA) models and regression and correlation models. We generally assume normally distributed random errors for these models. The inference procedures that we discuss are, for the most part, based on maximum likelihood procedures. The theory requires some discussion of quadratic forms which we briefly introduce next.

Consider polynomials of degree 2 in n variables, $X_1, \ldots, X_n$, of the form

$$q(X_1, \ldots, X_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} X_i a_{ij} X_j,$$

for n^2 constants a_{ij}. We call this form a **quadratic form** in the variables $X_1, \ldots, X_n$. If both the variables and the coefficients are real, it is called a **real quadratic** form. Only real quadratic forms are considered in this book. To illustrate, the form $X_1^2 + X_1X_2 + X_2^2$ is a quadratic form in the two variables X_1 and X_2; the form $X_1^2 + X_2^2 + X_3^2 - 2X_1X_2$ is a quadratic form in the three variables X_1, X_2, and X_3; but the form $(X_1 - 1)^2 + (X_2 - 2)^2 = X_1^2 + X_2^2 - 2X_1 - 4X_2 + 5$ is not a quadratic form in X_1 and X_2, although it is a quadratic form in the variables $X_1 - 1$ and $X_2 - 2$.

Let $\overline{X}$ and S^2 denote, respectively, the mean and variance of a random sample

$X_1, X_2, \ldots, X_n$ from an arbitrary distribution. Thus

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - \frac{n}{n^2}\left(\sum_{i=1}^{n} X_i\right)^2 \\ &= \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j\right) \\ &= \frac{n-1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{i<j} X_i X_j.\end{aligned}$$

So the sample variance is a quadratic form in the variables $X_1, \ldots, X_n$.

9.2 One-Way ANOVA

Consider b independent random variables that have normal distributions with unknown means $\mu_1, \mu_2, \ldots, \mu_b$, respectively, and unknown but common variance σ^2. For each $j = 1, 2, \ldots, b$, let $X_{1j}, X_{2j}, \ldots, X_{n_j j}$ represent a random sample of size n_j from the normal distribution with mean μ_j and variance σ^2. The appropriate model for the observations is

$$X_{ij} = \mu_j + e_{ij}\ ; \quad i = 1, \ldots, n_j\ , j = 1, \ldots, b, \tag{9.2.1}$$

where e_{ij} are iid $N(0, \sigma^2)$. Let $n = \sum_{j=1}^{b} n_j$ denote the total sample size. Suppose that it is desired to test the composite hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_b \text{ versus } H_1 : \mu_j \neq \mu_{j'}, \text{ for some } j \neq j'. \tag{9.2.2}$$

We derive the likelihood ratio test for these hypotheses.

Such problems often arise in practice. For example, suppose for a certain type of disease there are b drugs that can be used to treat it and we are interested in determining which drug is best in terms of a certain response. Let X_j denote this response when drug j is applied and let $\mu_j = E(X_j)$. If we assume that X_j is $N(\mu_j, \sigma^2)$, then the above null hypothesis says that all the drugs are equally effective; see Exercise 9.2.6 for a numerical illustration of this situation involving drugs that are intended to lower cholesterol. In general, we often summarize this problem by saying that we have one factor at b levels. In this case the factor is the treatment of the disease and each level corresponds to one of the treatment drugs.

Model (9.2.1) is called a **one-way** model. As shown, the likelihood ratio test can be thought of in terms of estimates of variance. Hence, this is an example of an

analysis of variance (**ANOVA**). In short, we say that this example is a one-way ANOVA problem.

Here the **full model** parameter space is

$$\Omega = \{(\mu_1, \mu_2, \ldots, \mu_b, \sigma^2) : -\infty < \mu_j < \infty,\ 0 < \sigma^2 < \infty\},$$

while the **reduced model** (full model under H_0) parameter space is

$$\omega = \{(\mu_1, \mu_2, \ldots, \mu_b, \sigma^2) : -\infty < \mu_1 = \mu_2 = \cdots = \mu_b = \mu < \infty,\ 0 < \sigma^2 < \infty\}.$$

The likelihood functions, denoted by $L(\Omega)$ and $L(\omega)$ are, respectively,

$$L(\Omega) = \left(\frac{1}{2\pi\sigma^2}\right)^{ab/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \mu_j)^2\right].$$

and

$$L(\omega) = \left(\frac{1}{2\pi\sigma^2}\right)^{ab/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \mu)^2\right]$$

We first consider the reduced model. Notice that it is just a one sample model with sample size n from a $N(\mu, \sigma^2)$ distribution. We have derived the mles in Example 4.1.3 of Chapter 4, which, in this notation, are given by

$$\hat{\mu}_\omega = \tfrac{1}{n}\textstyle\sum_{j=1}^{b}\sum_{i=1}^{n_j} x_{ij} = \overline{x}_{\cdot\cdot} \text{ and } \hat{\sigma}^2_\omega = \tfrac{1}{n}\textstyle\sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_{\cdot\cdot})^2. \tag{9.2.3}$$

The notation $\overline{x}_{\cdot\cdot}$ denotes that the mean is taken over both subscripts. This is often called the **grand mean**. Evaluating $L(\omega)$ at the mles, we obtain after simplification:

$$L(\hat{\omega}) = \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{1}{\hat{\sigma}^2_\omega}\right)^{n/2} e^{-n/2}. \tag{9.2.4}$$

Next, we consider the full model. The log of its likelihood is

$$\log\ L(\Omega) = -(n/2)\log(2\pi) - (n/2)\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \mu_j)^2. \tag{9.2.5}$$

For $j = 1, \ldots, b$, the partial of the log of $L(\Omega)$ with respect to μ_j results in

$$\frac{\partial \log L(\Omega}{\partial \mu_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n_j}(x_{ij} - \mu_j).$$

Setting this partial to 0 and solving for μ_j, we obtain the mle of μ_j which we denote by

$$\hat{\mu}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij} = \overline{x}_{\cdot j}, \quad j = 1, \ldots, b. \tag{9.2.6}$$

Since this derivation did not depend on σ, to find the mle of σ, we substitute $\overline{x}_{\cdot j}$ for μ_j in the $\log L(\Omega)$. Taking the partial derivative with respect to σ we then get

$$\frac{\partial \log L(\Omega}{\partial \sigma} = -(n/2)\frac{2\sigma}{\sigma^2} + \frac{1}{\sigma^3}\sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_{\cdot j})^2.$$

Solving this for σ^2, we obtain[1] the mle

$$\hat{\sigma}^2_{\Omega} = \frac{1}{n}\sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_{\cdot j})^2. \tag{9.2.7}$$

Substituting these mles for their respective parameters in $L(\Omega)$, after some simplification, leads to

$$L(\hat{\Omega}) = \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{1}{\hat{\sigma}^2_{\Omega}}\right)^{n/2} e^{-n/2}. \tag{9.2.8}$$

Hence, the likelihood ratio test rejects H_0 in favor of H_1 for small values of the statistic $\hat{\Lambda} = L(\hat{\omega})/L(\hat{\Omega})$ or equivalently, for large values of $\hat{\Lambda}^{-2/n}$. We can express this test statistic as a ratio of two quadratic forms Q_3 and Q as

$$\begin{aligned}\hat{\Lambda}^{n/2} &= \frac{\hat{\sigma}^2_{\Omega}}{\hat{\sigma}^2_{\omega}} = \frac{\sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_{\cdot j})^2}{\sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_{\cdot\cdot})^2} \\ &=_{\text{dfn}} \frac{Q_3}{Q}.\end{aligned} \tag{9.2.9}$$

In order to rewrite the test statistic in terms of an F-statistic, consider the identity involving Q, Q_3, and another quadratic form Q_4 given by:

$$\begin{aligned}Q &= \sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_{\cdot\cdot})^2 = \sum_{j=1}^{b}\sum_{i=1}^{n_j}[(x_{ij} - \overline{x}_{\cdot j}) + (\overline{x}_{\cdot j} - \overline{x}_{\cdot\cdot})]^2 \\ &= \sum_{j=1}^{b}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_{\cdot j})^2 + \sum_{j=1}^{b} n_j(\overline{x}_{\cdot j} - \overline{x}_{\cdot\cdot})^2 \\ &=_{\text{dfn}} Q_3 + Q_4.\end{aligned} \tag{9.2.10}$$

This derivation follows because the cross product term in the second line is 0. Using this identity, the test statistic $\hat{\Lambda}^{-2/n}$ can be expressed as

$$\hat{\Lambda}^{-2/n} = \frac{Q_3 + Q_4}{Q_3} = 1 + \frac{Q_4}{Q_3}.$$

As the final version, note that the test rejects H_0 if F is too large where

$$F = \frac{Q_4/(b-1)}{Q_3/(n-b)}. \tag{9.2.11}$$

[1] We are using the fact that the mle of σ^2 is the square of the mle of σ.