# Chapter 4

# TESTING AND CONFIDENCE REGIONS: BASIC THEORY

## 4.1 INTRODUCTION

In Sections 1.3, 3.2, and 3.3 we defined the testing problem abstractly, treating it as a decision theory problem in which we are to decide whether $P \in \mathcal{P}_0$ or $\mathcal{P}_1$ or, parametrically, whether $\theta \in \Theta_0$ or $\Theta_1$ if $\mathcal{P}_j = \{P_\theta : \theta \in \Theta_j\}$, where $\mathcal{P}_0, \mathcal{P}_1$ or $\Theta_0, \Theta_1$ are a partition of the model $\mathcal{P}$ or, respectively, the parameter space $\Theta$.

This framework is natural if, as is often the case, we are trying to get a yes or no answer to important questions in science, medicine, public policy, and indeed most human activities, and we have data providing some evidence one way or the other.

As we have seen, in examples such as 1.1.3 the questions are sometimes simple and the type of data to be gathered under our control. Does a new drug improve recovery rates? Does a new car seat design improve safety? Does a new marketing policy increase market share? We can design a clinical trial, perform a survey, or more generally construct an experiment that yields data $X$ in $\mathcal{X} \subset R^q$, modeled by us as having distribution $P_\theta$, $\theta \in \Theta$, where $\Theta$ is partitioned into $\{\Theta_0, \Theta_1\}$ with $\Theta_0$ and $\Theta_1$ corresponding, respectively, to answering "no" or "yes" to the preceding questions.

Usually, the situation is less simple. The design of the experiment may not be under our control, what is an appropriate stochastic model for the data may be questionable, and what $\Theta_0$ and $\Theta_1$ correspond to in terms of the stochastic model may be unclear. Here are two examples that illustrate these issues.

**Example 4.1.1.** *Sex Bias in Graduate Admissions at Berkeley.* The Graduate Division of the University of California at Berkeley attempted to study the possibility that sex bias operated in graduate admissions in 1973 by examining admissions data. They initially tabulated $N_{m1}, N_{f1}$, the numbers of admitted male and female applicants, and the corresponding numbers $N_{m0}, N_{f0}$ of denied applicants. If $n$ is the total number of applicants, it might be tempting to model $(N_{m1}, N_{m0}, N_{f1}, N_{f0})$ by a multinomial, $\mathcal{M}(n, p_{m1}, p_{m0}, p_{f1}, p_{f0})$, distribution. But this model is suspect because in fact we are looking at the population of all applicants here, not a sample. Accepting this model provisionally, what does the hypothesis of no sex bias correspond to? Again it is natural to translate this into

$$P[\text{Admit} \mid \text{Male}] = \frac{p_{m1}}{p_{m1}+p_{m0}} = P[\text{Admit} \mid \text{Female}] = \frac{p_{f1}}{p_{f1}+p_{f0}}.$$

But is this a correct translation of what absence of bias means? Only if admission is determined centrally by the toss of a coin with probability

$$\frac{p_{m1}}{p_{m1}+p_{m0}} = \frac{p_{f1}}{p_{f1}+p_{f0}}.$$

In fact, as is discussed in a paper by Bickel, Hammel, and O'Connell (1975), admissions are performed at the departmental level and rates of admission differ significantly from department to department. If departments "use different coins," then the data are naturally decomposed into $\mathbf{N} = (N_{m1d}, N_{m0d}, N_{f1d}, N_{f0d},\ d = 1, \ldots, D)$, where $N_{m1d}$ is the number of male admits to department $d$, and so on. Our multinomial assumption now becomes $\mathbf{N}_d = (N_{m1d}, N_{m0d}, N_{f1d}, N_{f0d})$ are independent with corresponding distributions $\mathcal{M}(n_d, p_{m1d}, p_{m0d}, p_{f1d}, p_{f0d})$, $d = 1, \ldots, D$. In these terms the hypothesis of "no bias" can now be translated into:

$$H : \frac{p_{m1d}}{p_{m1d}+p_{m0d}} = \frac{p_{f1d}}{p_{f1d}+p_{f0d}}$$

for $d = 1, \ldots, D$. This is *not* the same as our previous hypothesis unless all departments have the same number of applicants or all have the same admission rate,

$$\frac{p_{m1}+p_{f1}}{p_{m1}+p_{f1}+p_{m0}+p_{f0}}.$$

In fact, the same data can lead to opposite conclusions regarding these hypotheses—a phenomenon called *Simpson's paradox.* The example illustrates both the difficulty of specifying a stochastic model and translating the question one wants to answer into a statistical hypothesis. □

**Example 4.1.2.** *Mendel's Peas.* In one of his famous experiments laying the foundation of the quantitative theory of genetics, Mendel crossed peas heterozygous for a trait with two alleles, one of which was dominant. The progeny exhibited approximately the expected ratio of one homozygous dominant to two heterozygous dominants (to one recessive). In a modern formulation, if there were $n$ dominant offspring (seeds), the natural model is to assume, if the inheritance ratio can be arbitrary, that $N_{AA}$, the number of homozygous dominants, has a binomial $(n, p)$ distribution. The hypothesis of dominant inheritance corresponds to $H : p = \frac{1}{3}$ with the alternative $K : p \neq \frac{1}{3}$. It was noted by Fisher as reported in Jeffreys (1961) that in this experiment the observed fraction $\frac{m}{n}$ was much closer to $\frac{1}{3}$ than might be expected under the hypothesis that $N_{AA}$ has a binomial, $\mathcal{B}\left(n, \frac{1}{3}\right)$, distribution,

$$P\left[\left|\frac{N_{AA}}{n} - \frac{1}{3}\right| \leq \left|\frac{m}{n} - \frac{1}{3}\right|\right] = 7 \times 10^{-5}.$$

Fisher conjectured that rather than believing that such a very extraordinary event occurred it is more likely that the numbers were made to "agree with theory" by an overzealous assistant. That is, either $N_{AA}$ cannot really be thought of as stochastic or any stochastic model needs to permit distributions other than $\mathcal{B}(n, p)$, for instance, $(1-\epsilon)\delta_{\frac{n}{3}} + \epsilon\mathcal{B}(n, p)$, where $1-\epsilon$ is the probability that the assistant fudged the data and $\delta_{\frac{n}{3}}$ is point mass at $\frac{n}{3}$. □

What the second of these examples suggests is often the case. The set of distributions corresponding to one answer, say $\Theta_0$, is better defined than the alternative answer $\Theta_1$. That a treatment has no effect is easier to specify than what its effect is; see, for instance, our discussion of constant treatment effect in Example 1.1.3. In science generally a theory typically closely specifies the type of distribution $P$ of the data $X$ as, say, $P = P_\theta$, $\theta \in \Theta_0$. If the theory is false, it's not clear what $P$ should be as in the preceding Mendel example. These considerations lead to the asymmetric formulation that saying $P \in \mathcal{P}_0$ $(\theta \in \Theta_0)$ corresponds to *acceptance* of the *hypothesis* $H : P \in \mathcal{P}_0$ and $P \in \mathcal{P}_1$ corresponds to *rejection* sometimes written as $K : P \in \mathcal{P}_1$.[(1)]

As we have stated earlier, acceptance and rejection can be thought of as actions $a = 0$ or 1, and we are then led to the natural $0-1$ loss $l(\theta, a) = 0$ if $\theta \in \Theta_a$ and 1 otherwise. Moreover, recall that a decision procedure in the case of a *test* is described by a test function $\delta : x \to \{0, 1\}$ or *critical region* $C \equiv \{x : \delta(x) = 1\}$, the set of points for which we reject.

It is convenient to distinguish between two structural possibilities for $\Theta_0$ and $\Theta_1$: If $\Theta_0$ consists of only one point, we call $\Theta_0$ and $H$ *simple*. When $\Theta_0$ contains more than one point, $\Theta_0$ and $H$ are called *composite*. The same conventions apply to $\Theta_1$ and $K$.

We illustrate these ideas in the following example.

**Example 4.1.3.** Suppose we have discovered a new drug that we believe will increase the rate of recovery from some disease over the recovery rate when an old established drug is applied. Our hypothesis is then the *null hypothesis* that the new drug does not improve on the old drug. Suppose that we know from past experience that a fixed proportion $\theta_0 = 0.3$ recover from the disease with the old drug. What our hypothesis means is that the chance that an individual randomly selected from the ill population will recover is the same with the new and old drug. To investigate this question we would have to perform a random experiment. Most simply we would sample $n$ patients, administer the new drug, and then base our decision on the observed sample $\mathbf{X} = (X_1, \dots, X_n)$, where $X_i$ is 1 if the $i$th patient recovers and 0 otherwise. Thus, suppose we observe $S = \Sigma X_i$, the number of recoveries among the $n$ randomly selected patients who have been administered the new drug.[(2)] If we let $\theta$ be the probability that a patient to whom the new drug is administered recovers and the population of (present and future) patients is thought of as infinite, then $S$ has a $\mathcal{B}(n, \theta)$ distribution. If we suppose the new drug is at least as effective as the old, then $\Theta = [\theta_0, 1]$, where $\theta_0$ is the probability of recovery using the old drug. Now $\Theta_0 = \{\theta_0\}$ and $H$ is simple; $\Theta_1$ is the interval $(\theta_0, 1]$ and $K$ is composite. In situations such as this one we shall simplify notation and write $H : \theta = \theta_0$, $K : \theta > \theta_0$. If we allow for the possibility that the new drug is less effective than the old, then $\Theta_0 = [0, \theta_0]$ and $\Theta_0$ is composite. It will turn out that in most cases the solution to testing problems with $\Theta_0$ simple also solves the composite $\Theta_0$ problem. See Remark 4.1.

In this example with $\Theta_0 = \{\theta_0\}$ it is reasonable to reject $H$ if $S$ is "much" larger than what would be expected by chance if $H$ is true and the value of $\theta$ is $\theta_0$. Thus, we reject $H$ if $S$ exceeds or equals some integer, say $k$, and accept $H$ otherwise. That is, in the terminology of Section 1.3, our critical region $C$ is $\{\mathbf{X} : S \geq k\}$ and the test function or rule is $\delta_k(\mathbf{X}) = 1\{S \geq k\}$ with

$$P_I = \text{probability of type I error} = P_{\theta_0}(S \geq k)$$
$$P_{II} = \text{probability of type II error} = P_\theta(S < k),\ \theta > \theta_0.$$

The constant $k$ that determines the critical region is called the *critical value*. □

In most problems it turns out that the tests that arise naturally have the kind of structure we have just described. There is a statistic $T$ that "tends" to be small, if $H$ is true, and large, if $H$ is false. We call $T$ a *test statistic*. (Other authors consider test statistics $T$ that tend to be small, when $H$ is false. $-T$ would then be a test statistic in our sense.) We select a number $c$ and our test is to calculate $T(x)$ and then reject $H$ if $T(x) \geq c$ and accept $H$ otherwise. The value $c$ that completes our specification is referred to as the *critical value* of the test. Note that a test statistic generates a family of possible tests as $c$ varies. We will discuss the fundamental issue of how to choose $T$ in Sections 4.2, 4.3, and later chapters.

We now turn to the prevalent point of view on how to choose $c$.

### The Neyman Pearson Framework

As we discussed in Section 1.3, the Neyman Pearson approach rests on the idea that, of the two errors, one can be thought of as more important. By convention this is chosen to be the type I error and that in turn determines what we call $H$ and what we call $K$. Given this position, how reasonable is this point of view?

In the medical setting of Example 4.1.3 this asymmetry appears reasonable. It has also been argued that, generally in science, announcing that a new phenomenon has been observed when in fact nothing has happened (the so-called null hypothesis) is more serious than missing something new that has in fact occurred. We do not find this persuasive, but if this view is accepted, it again reasonably leads to a Neyman Pearson formulation.

As we noted in Examples 4.1.1 and 4.1.2, asymmetry is often also imposed because one of $\Theta_0, \Theta_1$, is much better defined than its complement and/or the distribution of statistics $T$ under $\Theta_0$ is easy to compute. In that case rejecting the hypothesis at level $\alpha$ is interpreted as a measure of the weight of evidence we attach to the falsity of $H$. For instance, testing techniques are used in searching for regions of the genome that resemble other regions that are known to have significant biological activity. One way of doing this is to align the known and unknown regions and compute statistics based on the number of matches. To determine significant values of these statistics a (more complicated) version of the following is done. Thresholds (critical values) are set so that if the matches occur at random (i.e., matches at one position are independent of matches at other positions) and the probability of a match is $\frac{1}{2}$, then the probability of exceeding the threshold (type I) error is smaller than $\alpha$. No one really believes that $H$ is true and possible types of alternatives are vaguely known at best, but computation under $H$ is easy.

The Neyman Pearson framework is still valuable in these situations by at least making us think of possible alternatives and then, as we shall see in Sections 4.2 and 4.3, suggesting what test statistics it is best to use.

There is an important class of situations in which the Neyman Pearson framework is inappropriate, such as the quality control Example 1.1.1. Indeed, it is too limited in any situation in which, even though there are just two actions, we can attach, even nominally, numbers to the two losses that are not equal and/or depend on $\theta$. See Problem 3.2.9. Finally, in the Bayesian framework with a prior distribution on the parameter, the approach of Example 3.2.2(b) is the one to take in all cases with $\Theta_0$ and $\Theta_1$ simple.

Here are the elements of the Neyman Pearson story. Begin by specifying a small number $\alpha > 0$ such that probabilities of type I error greater than $\alpha$ are undesirable. Then restrict attention to tests that in fact have the probability of rejection less than or equal to $\alpha$ for all $\theta \in \Theta_0$. As we have noted in Section 1.3, such tests are said to have *level (of significance)* $\alpha$, and we speak of rejecting $H$ at level $\alpha$. The values $\alpha = 0.01$ and $0.05$ are commonly used in practice. Because a test of level $\alpha$ is also of level $\alpha' > \alpha$, it is convenient to give a name to the smallest level of significance of a test. This quantity is called the *size* of the test and is the maximum probability of type I error. That is, if we have a test statistic $T$ and use critical value $c$, our test has size $\alpha(c)$ given by

$$\alpha(c) = \sup\{P_\theta[T(X) \geq c] : \theta \in \Theta_0\}. \tag{4.1.1}$$

Now $\alpha(c)$ is nonincreasing in $c$ and typically $\alpha(c) \uparrow 1$ as $c \downarrow -\infty$ and $\alpha(c) \downarrow 0$ as $c \uparrow \infty$. In that case, if $0 < \alpha < 1$, there exists a unique smallest $c$ for which $\alpha(c) \leq \alpha$. This is the critical value we shall use, if our test statistic is $T$ and we want level $\alpha$. It is referred to as the *level $\alpha$ critical value*. In Example 4.1.3 with $\delta(\mathbf{X}) = 1\{S \geq k\}$, $\theta_0 = 0.3$ and $n = 10$, we find from binomial tables the level 0.05 critical value 6 and the test has size $\alpha(6) = P_{\theta_0}(S \geq 6) = 0.0473$.

Once the level or critical value is fixed, the probabilities of type II error as $\theta$ ranges over $\Theta_1$ are determined. By convention $1 - P$ [type II error] is usually considered. Specifically,

**Definition 4.1.1.** The *power* of a test against the alternative $\theta$ is the probability of rejecting $H$ when $\theta$ is true.

Thus, the power is 1 minus the probability of type II error. It can be thought of as the probability that the test will "detect" that the alternative $\theta$ holds. The power is a function of $\theta$ on $\Theta_1$. If $\Theta_0$ is composite as well, then the probability of type I error is also a function of $\theta$. Both the power and the probability of type I error are contained in the *power function*, which is defined *for all* $\theta \in \Theta$ by

$$\beta(\theta) = \beta(\theta, \delta) = P_\theta[\text{Rejection}] = P_\theta[\delta(X) = 1] = P_\theta[T(X) \geq c].$$

If $\theta \in \Theta_0$, $\beta(\theta, \delta)$ is just the probability of type I error, whereas if $\theta \in \Theta_1$, $\beta(\theta, \delta)$ is the power against $\theta$.

**Example 4.1.3 (continued).** Here

$$\beta(\theta, \delta_k) = P(S \geq k) = \sum_{j=k}^{n} \binom{n}{j} \theta^j (1-\theta)^{n-j}.$$

A plot of this function for $n = 10$, $\theta_0 = 0.3$, $k = 6$ is given in Figure 4.1.1.

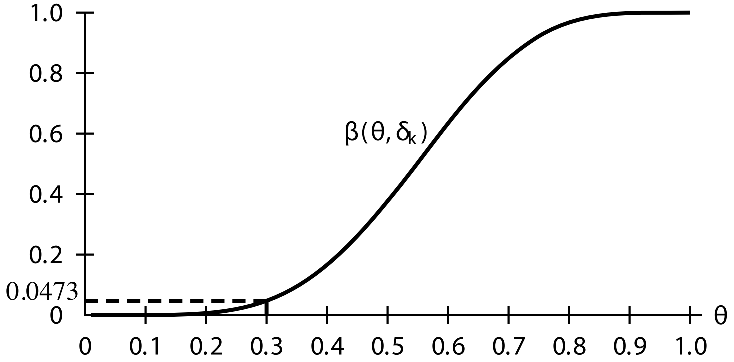

**Figure 4.1.1.** Power function of the level $0.05$ one-sided test $\delta_k$ of $H : \theta = 0.3$ versus $K : \theta > 0.3$ for the $\mathcal{B}(10, \theta)$ family of distributions. The power is plotted as a function of $\theta, k = 6$ and the size is $0.0473$.

Note that in this example the power at $\theta = \theta_1 > 0.3$ is the probability that the level $0.05$ test will detect an improvement of the recovery rate from $0.3$ to $\theta_1 > 0.3$. When $\theta_1$ is $0.5$, a $67\%$ improvement, this probability is only $.3770$. What is needed to improve on this situation is a larger sample size $n$. One of the most important uses of power is in the selection of sample sizes to achieve reasonable chances of detecting interesting alternatives. We return to this question in Section 4.3. □

**Remark 4.1.** From Figure 4.1.1 it appears that the power function is increasing (a proof will be given in Section 4.3). It follows that the size of the test is unchanged if instead of $\Theta_0 = \{\theta_0\}$ we used $\Theta_0 = [0, \theta_0]$. That is,

$$\alpha(k) = \sup\{P_\theta[T(X) \geq k] : \theta \in \Theta_0\} = P_{\theta_0}[T(X) \geq k].$$

**Example 4.1.4.** *One-Sided Tests for the Mean of a Normal Distribution with Known Variance.* Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ is a sample from $\mathcal{N}(\mu, \sigma^2)$ population with $\sigma^2$ is known. (The $\sigma^2$ unknown case is treated in Section 4.5.) We want to test $H : \mu \leq 0$ versus $K : \mu > 0$. This problem arises when we want to compare two treatments or a treatment and control (nothing) and both treatments are administered to the same subject. For instance, suppose we want to see if a drug induces sleep. We might, for each of a group of $n$ randomly selected patients, record sleeping time without the drug (or after the administration of a placebo) and then after some time administer the drug and record sleeping time again. Let $X_i$ be the difference between the time slept after administration of the drug and time slept without administration of the drug by the $i$th patient. If we assume $X_1, \ldots, X_n$ are normally distributed with mean $\mu$ and variance $\sigma^2$, then the drug effect is measured by $\mu$ and $H$ is the hypothesis that the drug has no effect or is detrimental, whereas $K$ is the alternative that it has some positive effect.

Because $\bar{X}$ tends to be larger under $K$ than under $H$, it is natural to reject $H$ for large values of $\bar{X}$. It is convenient to replace $\bar{X}$ by the test statistic $T(\mathbf{X}) = \sqrt{n}\bar{X}/\sigma$, which generates the same family of critical regions. The power function of the test with critical value $c$ is

$$\begin{aligned} \beta(\mu) = P_\mu[T(\mathbf{X}) \geq c] &= P_\mu\left[\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \geq c - \frac{\sqrt{n}\mu}{\sigma}\right] \\ &= 1 - \Phi\left(c - \frac{\sqrt{n}\mu}{\sigma}\right) = \Phi\left(-c + \frac{\sqrt{n}\mu}{\sigma}\right) \end{aligned} \tag{4.1.2}$$

because $\Phi(z) = 1 - \Phi(-z)$. Because $\beta(\mu)$ is increasing,

$$\alpha(c) = \sup\{\beta(\mu) : \mu \leq 0\} = \beta(0) = \Phi(-c).$$

The smallest $c$ for which $\Phi(-c) \leq \alpha$ is obtained by setting $\Phi(-c) = \alpha$ or

$$c = -z(\alpha)$$

where $-z(\alpha) = z(1-\alpha)$ is the $(1-\alpha)$ quantile of the $\mathcal{N}(0,1)$ distribution. □

### The Heuristics of Test Construction

When hypotheses are expressed in terms of an estimable parameter $H : \theta \in \Theta_0 \subset \mathcal{R}^p$, and we have available a good estimate $\widehat{\theta}$ of $\theta$, it is clear that a reasonable test statistic is $d(\widehat{\theta}, \Theta_0)$, where $d$ is the Euclidean (or some other) distance and $d(x, S) \equiv \inf\{d(x,y) : y \in S\}$. This *minimum distance principle* is essentially what underlies Examples 4.1.2 and 4.1.3. In Example 4.1.2, $p = P[AA]$, $\frac{N_{AA}}{n}$ is the MLE of $p$ and $d\left(\frac{N_{AA}}{n}, \Theta_0\right) = \left|\frac{N_{AA}}{n} - \frac{1}{3}\right|$. In Example 4.1.3, $\frac{X}{n}$ estimates $\theta$ and $d\left(\frac{X}{n}, [0, \theta_0]\right) = \left(\frac{X}{n} - \theta_0\right)_+$ where $y_+ = y1(y \geq 0)$. Rejecting for large values of this statistic is equivalent to rejecting for large values of $X$.

Given a test statistic $T(X)$ we need to determine critical values and eventually the power of the resulting tests. The task of finding a critical value is greatly simplified if $\mathcal{L}_\theta(T(\mathbf{X}))$ doesn't depend on $\theta$ for $\theta \in \Theta_0$. This occurs if $\Theta_0$ is simple as in Example 4.1.3. But it occurs also in more interesting situations such as testing $\mu = \mu_0$ versus $\mu \neq \mu_0$ if we have $\mathcal{N}(\mu, \sigma^2)$ observations with both parameters unknown (the $t$ tests of Example 4.5.1 and Example 4.1.5). In all of these cases, $\mathcal{L}_0$, the common distribution of $T(\mathbf{X})$ under $\theta \in \Theta_0$, has a closed form and is tabled. However, in any case, critical values yielding correct type I probabilities are easily obtained by *Monte Carlo methods*. That is, if we generate i.i.d. $T(\mathbf{X}^{(1)}), \dots, T(\mathbf{X}^{(B)})$ from $\mathcal{L}_0$, then the test that rejects iff $T(\mathbf{X}) > T_{((B+1)(1-\alpha))}$, where $T_{(1)} \leq \cdots \leq T_{(B+1)}$ are the ordered $T(\mathbf{X}), T(\mathbf{X}^{(1)}), \dots, T(\mathbf{X}^{(B)})$, has level $\alpha$ if $\mathcal{L}_0$ is continuous and $(B+1)(1-\alpha)$ is an integer (Problem 4.1.9).

The key feature of situations in which $\mathcal{L}_\theta(T_n) \equiv \mathcal{L}_0$ for $\theta \in \Theta_0$ is usually invariance under the action of a group of transformations. See Lehmann (1997) and Volume II for discussions of this property.

Here are two examples of testing hypotheses in a nonparametric context in which the minimum distance principle is applied and calculation of a critical value is straightforward.

**Example 4.1.5.** *Goodness of Fit Tests.* Let $X_1, \dots, X_n$ be i.i.d. as $X \sim F$, where $F$ is continuous. Consider the problem of testing $H : F = F_0$ versus $K : F \neq F_0$. Let $\widehat{F}$ denote the empirical distribution and consider the sup distance between the hypothesis $F_0$ and the plug-in estimate of $F$, the empirical distribution function $\widehat{F}$, as a test statistic

$$D_n = \sup_x |\widehat{F}(x) - F_0(x)|.$$

It can be shown (Problem 4.1.7) that $D_n$, which is called the *Kolmogorov statistic*, can be written as

$$D_n = \max_{i=1,\dots,n} \max \left\{ \frac{i}{n} - F_0(x_{(i)}),\ F_0(x_{(i)}) - \frac{(i-1)}{n} \right\} \tag{4.1.3}$$

where $x_{(1)} < \cdots < x_{(n)}$ is the ordered observed sample, that is, *the order statistics*. This statistic has the following *distribution-free* property:

**Proposition 4.1.1.** *The distribution of $D_n$ under $H$ is the same for all continuous $F_0$. In particular, $P_{F_0}(D_n \leq d) = P_U(D_n \leq d)$, where $U$ denotes the $\mathcal{U}(0,1)$ distribution.*

***Proof.*** Set $U_i = F_0(X_i)$, then by Problem B.3.4, $U_i \sim \mathcal{U}(0,1)$. Also

$$\begin{aligned} \widehat{F}(x) &= n^{-1}\Sigma 1\{X_i \leq x\} = n^{-1}\Sigma 1\{F_0(X_i) \leq F_0(x)\} \\ &= n^{-1}\Sigma 1\{U_i \leq F_0(x)\} = \widehat{U}(F_0(x)) \end{aligned}$$

where $\widehat{U}$ denotes the empirical distribution function of $U_1, \dots, U_n$. As $x$ ranges over $R$, $u = F_0(x)$ ranges over $(0,1)$, thus,

$$D_n = \sup_{0<u<1} |\widehat{U}(u) - u|$$

and the result follows. □

Note that the hypothesis here is simple so that for any one of these hypotheses $F = F_0$, the distribution can be simulated (or exhibited in closed form). What is remarkable is that it is independent of which $F_0$ we consider. This is again a consequence of invariance properties (Lehmann, 1997).

The distribution of $D_n$ has been thoroughly studied for finite and large $n$. In particular, for $n > 80$, and

$$h_n(t) = t/(\sqrt{n} + 0.12 + 0.11/\sqrt{n})$$

close approximations to the size $\alpha$ critical values $k_\alpha$ are $h_n(1.628)$, $h_n(1.358)$, and $h_n(1.224)$ for $\alpha = .01$, .05, and .10 respectively. □

**Example 4.1.6.** *Goodness of Fit to the Gaussian Family.* Suppose $X_1, \dots, X_n$ are i.i.d. $F$ and the hypothesis is $H : F = \Phi\left(\frac{\cdot - \mu}{\sigma}\right)$ for some $\mu, \sigma$, which is evidently composite. We can proceed as in Example 4.1.5 rewriting $H : F(\mu + \sigma x) = \Phi(x)$ for all $x$ where $\mu = E_F(X_1)$, $\sigma^2 = \text{Var}_F(X_1)$. The natural estimate of the parameter $F(\mu + \sigma x)$ is

$\widehat{F}(\bar{X} + \widehat{\sigma}x)$ where $\bar{X}$ and $\widehat{\sigma}^2$ are the MLEs of $\mu$ and $\sigma^2$. Applying the sup distance again, we obtain the statistic

$$\begin{aligned} T_n &= \sup_x |\widehat{F}(\bar{X} + \widehat{\sigma}x) - \Phi(x)| \\ &= \sup_x |\widehat{G}(x) - \Phi(x)| \end{aligned}$$

where $\widehat{G}$ is the empirical distribution of $(\Delta_1, \dots, \Delta_n)$ with $\Delta_i \equiv (X_i - \bar{X})/\widehat{\sigma}$. But, under $H$, the joint distribution of $(\Delta_1, \dots, \Delta_n)$ doesn't depend on $\mu, \sigma^2$ and is that of $(Z_i - \bar{Z}) \Big/ \left(\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})^2\right)^{\frac{1}{2}}$, $1 \le i \le n$, where $Z_1, \dots, Z_n$ are i.i.d. $\mathcal{N}(0,1)$. (See Section B.3.2.) Thus, $T_n$ has the same distribution $\mathcal{L}_0$ under $H$, whatever be $\mu$ and $\sigma^2$, and the critical value may be obtained by simulating i.i.d. observations $Z_i$, $1 \le i \le n$, from $\mathcal{N}(0,1)$, then computing the $T_n$ corresponding to those $Z_i$. We do this $B$ times independently, thereby obtaining $T_{n1}, \dots, T_{nB}$. Now the Monte Carlo critical value is the $[(B+1)(1-\alpha)+1]$th order statistic among $T_n, T_{n1}, \dots, T_{nB}$.(3) □

### The $p$-Value: The Test Statistic as Evidence

Different individuals faced with the same testing problem may have different criteria of size. Experimenter I may be satisfied to reject the hypothesis $H$ using a test with size $\alpha = 0.05$, whereas experimenter II insists on using $\alpha = 0.01$. It is then possible that experimenter I rejects the hypothesis $H$, whereas experimenter II accepts $H$ on the basis of the same outcome $\mathbf{x}$ of an experiment. If the two experimenters can agree on a common test statistic $T$, this difficulty may be overcome by reporting the outcome of the experiment in terms of the *observed size* or *p-value* or *significance probability* of the test. This quantity is a statistic that is defined as the smallest level of significance $\alpha$ at which an experimenter using $T$ would reject on the *basis of the observed outcome* $x$. That is, if the experimenter's critical value corresponds to a test of size less than the $p$-value, $H$ is not rejected; otherwise, $H$ is rejected.

Consider, for instance, Example 4.1.4. If we observe $\mathbf{X} = \mathbf{x} = (x_1, \dots, x_n)$, we would reject $H$ if, and only if, $\alpha$ satisfies

$$T(\mathbf{x}) = \frac{\sqrt{n}\bar{x}}{\sigma} \ge -z(\alpha)$$

or upon applying $\Phi$ to both sides if, and only if,

$$\alpha \ge \Phi(-T(\mathbf{x})).$$

Therefore, if $\mathbf{X} = \mathbf{x}$, the $p$-value is

$$\Phi(-T(\mathbf{x})) = \Phi\left(\frac{-\sqrt{n}\bar{x}}{\sigma}\right). \tag{4.1.4}$$

Considered as a statistic the $p$-value is $\Phi(-\sqrt{n}\bar{X}/\sigma)$.

In general, let $X$ be a $q$ dimensional random vector. We will show that we can express the $p$-value simply in terms of the function $\alpha(\cdot)$ defined in (4.1.1). Suppose that we observe $X = x$. Then if we use critical value $c$, we would reject $H$ if, and only if,

$$T(x) \geq c.$$

Thus, the largest critical value $c$ for which we would reject is $c = T(x)$. But the size of a test with critical value $c$ is just $\alpha(c)$ and $\alpha(c)$ is decreasing in $c$. Thus, the smallest $\alpha$ for which we would reject corresponds to the largest $c$ for which we would reject and is just $\alpha(T(x))$. We have proved the following.

**Proposition 4.1.2.** *The $p$-value is $\alpha(T(X))$.*

This is in agreement with (4.1.4). Similarly in Example 4.1.3,

$$\alpha(k) = \sum_{j=k}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j}$$

and the $p$-value is $\alpha(s)$ where $s$ is the observed value of $X$. The normal approximation is used for the $p$-value also. Thus, for $\min\{n\theta_0, n(1-\theta_0)\} \geq 5$,

$$\alpha(s) = P_{\theta_0}(S \geq s) \simeq 1 - \Phi\left(\frac{s - \frac{1}{2} - n\theta_0}{[n\theta_0(1-\theta_0)]^{\frac{1}{2}}}\right). \qquad (4.1.5)$$

The $p$-value is used extensively in situations of the type we described earlier, when $H$ is well defined, but $K$ is not, so that type II error considerations are unclear. In this context, to quote Fisher (1958), "The actual value of $p$ obtainable from the table by interpolation indicates the strength of the evidence against the null hypothesis" (p. 80).

The $p$-value can be thought of as a standardized version of our original statistic; that is, $\alpha(T)$ is on the unit interval and when $H$ is simple and $T$ has a continuous distribution, $\alpha(T)$ has a uniform, $\mathcal{U}(0,1)$, distribution (Problem 4.1.5).

It is possible to use $p$-values to combine the evidence relating to a given hypothesis $H$ provided by several different independent experiments producing different kinds of data. For example, if $r$ experimenters use continuous test statistics $T_1, \dots, T_r$ to produce $p$-values $\alpha(T_1), \dots, \alpha(T_r)$, then if $H$ is simple Fisher (1958) proposed using

$$\widehat{T} = -2\sum_{j=1}^{r} \log \alpha(T_j) \qquad (4.1.6)$$

to test $H$. The statistic $\widehat{T}$ has a chi-square distribution with $2r$ degrees of freedom (Problem 4.1.6). Thus, $H$ is rejected if $\widehat{T} \geq x_{1-\alpha}$ where $x_{1-\alpha}$ is the $1-\alpha$th quantile of the $\chi^2_{2r}$ distribution. Various methods of combining the data from different experiments in this way are discussed by van Zwet and Osterhoff (1967). More generally, these kinds of issues are currently being discussed under the rubric of *data-fusion* and *meta-analysis* (e.g., see Hedges and Olkin, 1985).

The preceding paragraph gives an example in which the hypothesis specifies a distribution completely; that is, under $H$, $\alpha(T_i)$ has a $\mathcal{U}(0,1)$ distribution. This is an instance of testing *goodness of fit*; that is, we test whether the distribution of $X$ is different from a specified $F_0$.

**Summary.** We introduce the basic concepts and terminology of testing statistical hypotheses and give the Neyman–Pearson framework. In particular, we consider experiments in which important questions about phenomena can be turned into questions about whether a parameter $\theta$ belongs to $\Theta_0$ or $\Theta_1$, where $\Theta_0$ and $\Theta_1$ are disjoint subsets of the parameter space $\Theta$. We introduce the basic concepts of simple and composite hypotheses, (null) hypothesis $H$ and alternative (hypothesis) $K$, test functions, critical regions, test statistics, type I error, type II error, significance level, size, power, power function, and $p$-value. In the Neyman–Pearson framework, we specify a small number $\alpha$ and construct tests that have at most probability (significance level) $\alpha$ of rejecting $H$ (deciding $K$) when $H$ is true; then, subject to this restriction, we try to maximize the probability (power) of rejecting $H$ when $K$ is true.

## 4.2 CHOOSING A TEST STATISTIC: THE NEYMAN–PEARSON LEMMA

We have seen how a hypothesis-testing problem is defined and how performance of a given test $\delta$, or equivalently, a given test statistic $T$, is measured in the Neyman–Pearson theory. Typically a test statistic is not given but must be chosen on the basis of its performance. In Sections 3.2 and 3.3 we derived test statistics that are best in terms of minimizing Bayes risk and maximum risk. In this section we will consider the problem of finding the level $\alpha$ test that has the highest possible power. Such a test and the corresponding test statistic are called *most powerful* (MP).

We start with the problem of testing a simple hypothesis $H : \theta = \theta_0$ versus a simple alternative $K : \theta = \theta_1$. In this case the Bayes principle led to procedures based on the *simple likelihood ratio statistic* defined by

$$L(x, \theta_0, \theta_1) = \frac{p(x, \theta_1)}{p(x, \theta_0)}$$

where $p(x, \theta)$ is the density or frequency function of the random vector $X$. The statistic $L$ takes on the value $\infty$ when $p(x, \theta_1) > 0$, $p(x, \theta_0) = 0$; and, by convention, equals 0 when both numerator and denominator vanish.

The statistic $L$ is reasonable for testing $H$ versus $K$ with large values of $L$ favoring $K$ over $H$. For instance, in the binomial example (4.1.3),

$$\begin{aligned} L(\mathbf{x}, \theta_0, \theta_1) &= (\theta_1/\theta_0)^S[(1-\theta_1)/(1-\theta_0)]^{n-S} \\ &= [\theta_1(1-\theta_0)/\theta_0(1-\theta_1)]^S[(1-\theta_1)/(1-\theta_0)]^n, \end{aligned} \tag{4.2.1}$$

which is large when $S = \Sigma X_i$ is large, and $S$ tends to be large when $K : \theta = \theta_1 > \theta_0$ is true.

We call $\varphi_k$ a *likelihood ratio* or *Neyman–Pearson (NP) test (function)* if for some $0 \le k \le \infty$ we can write the test function $\varphi_k$ as

$$\varphi_k(x) = \begin{array}{ll} 1 & \text{if } L(x, \theta_0, \theta_1) > k \\ 0 & \text{if } L(x, \theta_0, \theta_1) < k \end{array}$$

with $\varphi_k(x)$ any value in $(0, 1)$ if equality occurs. Note (Section 3.2) that $\varphi_k$ is a Bayes rule with $k = \pi/(1 - \pi)$, where $\pi$ denotes the prior probability of $\{\theta_0\}$. We show that in addition to being Bayes optimal, $\varphi_k$ is MP for level $E_{\theta_0}\varphi_k(X)$.

Because we want results valid for all possible test sizes $\alpha$ in $[0, 1]$, we consider *randomized tests* $\varphi$, which are tests that may take values in $(0, 1)$. If $0 < \varphi(x) < 1$ for the observation vector $x$, the interpretation is that we toss a coin with probability of heads $\varphi(x)$ and reject $H$ iff the coin shows heads. (See also Section 1.3.) For instance, if want size $\alpha = .05$ in Example 4.1.3 with $n = 10$ and $\theta_0 = 0.3$, we choose $\varphi(x) = 0$ if $S < 5$, $\varphi(x) = 1$ if $S > 5$, and

$$\varphi(x) = [0.05 - P(S > 5)]/P(S = 5) = .0262$$

if $S = 5$. Such randomized tests are not used in practice. They are only used to show that with randomization, likelihood ratio tests are unbeatable no matter what the size $\alpha$ is.

**Theorem 4.2.1. (Neyman–Pearson Lemma).**

(a) *If $\alpha > 0$ and $\varphi_k$ is a size $\alpha$ likelihood ratio test, then $\varphi_k$ is MP in the class of level $\alpha$ tests.*

(b) *For each $0 \le \alpha \le 1$ there exists an MP size $\alpha$ likelihood ratio test provided that randomization is permitted, $0 < \varphi(x) < 1$, for some $x$.*

(c) *If $\varphi$ is an MP level $\alpha$ test, then it must be a level $\alpha$ likelihood ratio test; that is, there exists $k$ such that*

$$P_\theta[\varphi(X) \ne \varphi_k(X), L(X, \theta_0, \theta_1) \ne k] = 0 \tag{4.2.2}$$

*for $\theta = \theta_0$ and $\theta = \theta_1$.*

***Proof.*** (a) Let $E_i$ denote $E_{\theta_i}$, $i = 0, 1$, and suppose $\varphi$ is a level $\alpha$ test, then

$$E_0\varphi_k(X) = \alpha, \ E_0\varphi(X) \le \alpha. \tag{4.2.3}$$

We want to show $E_1[\varphi_k(X) - \varphi(X)] \ge 0$. To this end consider

$$\begin{aligned} &E_1[\varphi_k(X) - \varphi(X)] - kE_0[\varphi_k(X) - \varphi(X)] \\ &= E_0[\varphi_k(X) - \varphi(X)]\left[\tfrac{p(X,\theta_1)}{p(X,\theta_0)} - k\right] + E_1[\varphi_k(X) - \varphi(X)]1\{p(X, \theta_0) = 0\} \\ &= I + II \text{ (say), where} \\ &I = E_0\{\varphi_k(X)[L(X, \theta_0, \theta_1) - k] - \varphi(X)[L(X, \theta_0, \theta_1) - k]\}. \end{aligned}$$

Because $L(x, \theta_0, \theta_1) - k$ is $< 0$ or $\ge 0$ according as $\varphi_k(x)$ is 0 or in $(0, 1]$, and because $0 \le \varphi(x) \le 1$, then $I \ge 0$. Note that $\alpha > 0$ implies $k < \infty$ and, thus, $\varphi_k(x) = 1$ if $p(x, \theta_0) = 0$. It follows that $II \ge 0$. Finally, using (4.2.3), we have shown that

$$E_1[\varphi_k(X) - \varphi(X)] \ge kE_0[\varphi_k(X) - \varphi(X)] \ge 0. \tag{4.2.4}$$

(b) If $\alpha = 0$, $k = \infty$ makes $\varphi_k$ MP size $\alpha$. If $\alpha = 1$, $k = 0$ makes $E_1\varphi_k(X) = 1$ and $\varphi_k$ is MP size $\alpha$. Next consider $0 < \alpha < 1$. Let $P_i$ denote $P_{\theta_i}$, $i = 0, 1$. Because $P_0[L(X, \theta_0, \theta_1) = \infty] = 0$, then there exists $k < \infty$ such that

$$P_0[L(X, \theta_0, \theta_1) > k] \le \alpha \text{ and } P_0[L(X, \theta_0, \theta_1) \ge k] \ge \alpha.$$

If $P_0[L(X, \theta_0, \theta_1) = k] = 0$, then $\varphi_k$ is MP size $\alpha$. If not, define

$$\varphi_k(x) = \frac{\alpha - P_0[L(X, \theta_0, \theta_1) > k]}{P_0[L(X, \theta_0, \theta_1) = k]}$$

on the set $\{x : L(x, \theta_0, \theta_1) = k\}$. Now $\varphi_k$ is MP size $\alpha$.

(c) Let $x \in \{x : p(x, \theta_1) > 0\}$, then to have equality in (4.2.4) we need to have $\varphi(x) = \varphi_k(x) = 1$ when $L(x, \theta_0, \theta_1) > k$ and have $\varphi(x) = \varphi_k(x) = 0$ when $L(x, \theta_0, \theta_1) < k$. It follows that (4.2.2) holds for $\theta = \theta_1$. The same argument works for $x \in \{x : p(x, \theta_0) > 0\}$ and $\theta = \theta_0$. □

It follows from the Neyman–Pearson lemma that an MP test has power at least as large as its level; that is,

**Corollary 4.2.1.** *If $\varphi$ is an MP level $\alpha$ test, then $E_{\theta_1}\varphi(X) \ge \alpha$ with equality iff $p(\cdot, \theta_0) = p(\cdot, \theta_1)$.*

***Proof.*** See Problem 4.2.7.

**Remark 4.2.1.** Let $\pi$ denote the prior probability of $\theta_0$ so that $(1-\pi)$ is the prior probability of $\theta_1$. Then the posterior probability of $\theta_1$ is

$$\pi(\theta_1 \mid x) = \frac{(1-\pi)p(x, \theta_1)}{(1-\pi)p(x, \theta_1) + \pi p(x, \theta_0)} = \frac{(1-\pi)L(x, \theta_0, \theta_1)}{(1-\pi)L(x, \theta_0, \theta_1) + \pi}. \tag{4.2.5}$$

If $\delta_\pi$ denotes the Bayes procedure of Example 3.2.2(b), then, when $\pi = k/(k+1)$, $\delta_\pi = \varphi_k$. Moreover, we conclude from (4.2.5) that this $\delta_\pi$ decides $\theta_1$ or $\theta_0$ according as $\pi(\theta_1 \mid x)$ is larger than or smaller than $1/2$.

Part (a) of the lemma can, for $0 < \alpha < 1$, also be easily argued from this Bayes property of $\varphi_k$ (Problem 4.2.10).

Here is an example illustrating calculation of the most powerful level $\alpha$ test $\varphi_k$.

**Example 4.2.1.** Consider Example 3.3.2 where $\mathbf{X} = (X_1, \dots, X_n)$ is a sample of $n$ $\mathcal{N}(\mu, \sigma^2)$ random variables with $\sigma^2$ known and we test $H : \mu = 0$ versus $K : \mu = v$, where $v$ is a known signal. We found

$$L(\mathbf{X}, 0, v) = \exp\left\{\frac{v}{\sigma^2}\sum_{i=1}^n X_i - \frac{nv^2}{2\sigma^2}\right\}.$$

Note that any strictly increasing function of an optimal statistic is optimal because the two statistics generate the same family of critical regions. Therefore,

$$T(\mathbf{X}) = \sqrt{n}\frac{\bar{X}}{\sigma} = \frac{\sigma}{v\sqrt{n}}\left[\log L(\mathbf{X}, 0, v) + \frac{nv^2}{2\sigma^2}\right]$$

is also optimal for this problem. But $T$ is the test statistic we proposed in Example 4.1.4. From our discussion there we know that for any specified $\alpha$, the test that rejects if, and only if,

$$T \geq z(1-\alpha) \tag{4.2.6}$$

has probability of type I error $\alpha$.

The power of this test is, by (4.1.2), $\Phi(z(\alpha) + (v\sqrt{n}/\sigma))$. By the Neyman–Pearson lemma this is the largest power available with a level $\alpha$ test. Thus, if we want the probability of detecting a signal $v$ to be at least a preassigned value $\beta$ (say, .90 or .95), then we solve $\Phi(z(\alpha)+(v\sqrt{n}/\sigma)) = \beta$ for $n$ and find that we need to take $n = (\sigma/v)^2[z(1-\alpha)+z(\beta)]^2$. This is the smallest possible $n$ for any size $\alpha$ test. □

An interesting feature of the preceding example is that the test defined by (4.2.6) that is MP for a specified signal $v$ does not depend on $v$: The same test maximizes the power for all possible signals $v > 0$. Such a test is called uniformly most powerful (UMP).

We will discuss the phenomenon further in the next section. The following important example illustrates, among other things, that the UMP test phenomenon is largely a feature of one-dimensional parameter problems.

**Example 4.2.2.** *Simple Hypothesis Against Simple Alternative for the Multivariate Normal: Fisher's Discriminant Function.* Suppose $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 0, 1$. The likelihood ratio test for $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $K : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ is based on

$$L(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \frac{\det^{\frac{1}{2}}(\Sigma_0)\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}}{\det^{\frac{1}{2}}(\Sigma_1)\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right\}}.$$

Rejecting $H$ for $L$ large is equivalent to rejecting for

$$Q \equiv (\mathbf{X}-\boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\mathbf{X}-\boldsymbol{\mu}_0) - (\mathbf{X}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{X}-\boldsymbol{\mu}_1)$$

large. Particularly important is the case $\Sigma_0 = \Sigma_1$ when "$Q$ large" is equivalent to "$F \equiv (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\Sigma_0^{-1}\mathbf{X}$ large." The function $F$ is known as the Fisher discriminant function. It is used in a classification context in which $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ correspond to two known populations and we desire to classify a new observation $\mathbf{X}$ as belonging to one or the other. We return to this in Volume II. Note that in general the test statistic $L$ depends intrinsically on $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$. However if, say, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \lambda\boldsymbol{\Delta}_0$, $\lambda > 0$ and $\Sigma_1 = \Sigma_0$, then, if $\boldsymbol{\mu}_0, \boldsymbol{\Delta}_0, \Sigma_0$ are known, a UMP (for all $\lambda$) test exists and is given by: Reject if

$$\boldsymbol{\Delta}_0^T\Sigma_0^{-1}(\mathbf{X}-\boldsymbol{\mu}_0) \geq c \tag{4.2.7}$$

where $c = z(1-\alpha)[\boldsymbol{\Delta}_0^T\Sigma_0^{-1}\boldsymbol{\Delta}_0]^{\frac{1}{2}}$ (Problem 4.2.8). If $\boldsymbol{\Delta}_0 = (1, 0, \ldots, 0)^T$ and $\Sigma_0 = I$, then this test rule is to reject $H$ if $X_1$ is large; however, if $\Sigma_0 \neq I$, this is no longer the case (Problem 4.2.9). In this example we have assumed that $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ for the two populations are known. If this is not the case, they are estimated with their empirical versions with sample means estimating population means and sample covariances estimating population covariances.

**Summary.** We introduce the simple likelihood ratio statistic and simple likelihood ratio (SLR) test for testing the simple hypothesis $H : \theta = \theta_0$ versus the simple alternative $K : \theta = \theta_1$. The Neyman–Pearson lemma, which states that the size $\alpha$ SLR test is uniquely most powerful (MP) in the class of level $\alpha$ tests, is established.

We note the connection of the MP test to the Bayes procedure of Section 3.2 for deciding between $\theta_0$ and $\theta_1$. Two examples in which the MP test does not depend on $\theta_1$ are given. Such tests are said to be UMP (uniformly most powerful).

## 4.3 UNIFORMLY MOST POWERFUL TESTS AND MONOTONE LIKELIHOOD RATIO MODELS

We saw in the two Gaussian examples of Section 4.2 that UMP tests for one-dimensional parameter problems exist. This phenomenon is not restricted to the Gaussian case as the next example illustrates. Before we give the example, here is the general definition of UMP:

**Definition 4.3.1.** A level $\alpha$ test $\varphi^*$ is *uniformly most powerful* (UMP) for $H : \theta \in \Theta_0$ versus $K : \theta \in \Theta_1$ if

$$\beta(\theta, \varphi^*) \geq \beta(\theta, \varphi) \text{ for all } \theta \in \Theta_1, \tag{4.3.1}$$

for any other level $\alpha$ test $\varphi$.

**Example 4.3.1.** *Testing for a Multinomial Vector.* Suppose that $(N_1, \dots, N_k)$ has a multinomial $\mathcal{M}(n, \theta_1, \dots, \theta_k)$ distribution with frequency function,

$$p(n_1, \dots, n_k, \theta) = \frac{n!}{n_1! \dots n_k!} \theta_1^{n_1} \dots \theta_k^{n_k}$$

where $n_1, \dots, n_k$ are integers summing to $n$. With such data we often want to test a simple hypothesis $H : \theta_1 = \theta_{10}, \dots, \theta_k = \theta_{k0}$. For instance, if a match in a genetic breeding experiment can result in $k$ types, $n$ offspring are observed, and $N_i$ is the number of offspring of type $i$, then $(N_1, \dots, N_k) \sim \mathcal{M}(n, \theta_1, \dots, \theta_k)$. The simple hypothesis would correspond to the theory that the expected proportion of offspring of types $1, \dots, k$ are given by $\theta_{10}, \dots, \theta_{k0}$. Usually the alternative to $H$ is composite. However, there is sometimes a simple alternative theory $K : \theta_1 = \theta_{11}, \dots, \theta_k = \theta_{k1}$. In this case, the likelihood ratio $L$ is

$$L = \prod_{i=1}^{k} \left( \frac{\theta_{i1}}{\theta_{i0}} \right)^{N_i}.$$

Here is an interesting special case: Suppose $\theta_{j0} > 0$ for all $j$, $0 < \epsilon < 1$ and for some fixed integer $l$ with $1 \leq l \leq k$

$$\theta_{l1} = \epsilon \theta_{l0};\ \theta_{j1} = \rho \theta_{j0},\ j \neq l, \tag{4.3.2}$$

where

$$\rho = (1 - \theta_{l0})^{-1}(1 - \epsilon \theta_{l0}).$$

That is, under the alternative, type $l$ is less frequent than under $H$ and the conditional probabilities of the other types given that type $l$ has not occurred are the same under $K$ as they are under $H$. Then

$$L = \rho^{n-N_l}\epsilon^{N_l} = \rho^n(\epsilon/\rho)^{N_l}.$$

Because $\epsilon < 1$ implies that $\rho > \epsilon$, we conclude that the MP test rejects $H$, if and only if, $N_l \leq c$. Critical values for level $\alpha$ are easily determined because $N_l \sim \mathcal{B}(n, \theta_{l0})$ under $H$. Moreover, for $\alpha = P(N_l \leq c)$, this test is UMP for testing $H$ versus $K : \theta \in \Theta_1 = \{\boldsymbol{\theta} : \boldsymbol{\theta}$ is of the form (4.3.2) with $0 < \epsilon < 1\}$. Note that because $l$ can be any of the integers $1, \ldots, k$, we get radically different best tests depending on which $\theta_i$ we assume to be $\theta_{l0}$ under $H$. □

Typically the MP test of $H : \theta = \theta_0$ versus $K : \theta = \theta_1$ depends on $\theta_1$ and the test is not UMP. However, we have seen three models where, in the case of a real parameter, there is a statistic $T$ such that the test with critical region $\{x : T(x) \geq c\}$ is UMP. This is part of a general phenomena we now describe.

**Definition 4.3.2.** The family of models $\{P_\theta : \theta \in \Theta\}$ with $\Theta \subset R$ is said to be a monotone likelihood ratio (MLR) family in $T$ if for $\theta_1 < \theta_2$ the distributions $P_{\theta_1}$ and $P_{\theta_2}$ are distinct and there exists a statistic $T(x)$ such that the ratio $p(x, \theta_2)/p(x, \theta_1)$ is an increasing function of $T(x)$. □

**Example 4.3.2 (Example 4.1.3 continued).** In this i.i.d. Bernoulli case, set $s = \sum_{i=1}^n x_i$, then

$$p(\mathbf{x}, \theta) = \theta^s(1-\theta)^{n-s} = (1-\theta)^n[\theta/(1-\theta)]^s$$

and the model is by (4.2.1) MLR in $s$. □

**Example 4.3.3.** Consider the one-parameter exponential family model

$$p(x, \theta) = h(x)\exp\{\eta(\theta)T(x) - B(\theta)\}.$$

If $\eta(\theta)$ is strictly increasing in $\theta \in \Theta$, then this family is MLR. Example 4.2.1 is of this form with $T(\mathbf{x}) = \sqrt{n}\bar{x}/\sigma$ and $\eta(\mu) = \sqrt{n}\mu/\sigma$, where $\sigma$ is known. □

Define the Neyman–Pearson (NP) test function

$$\delta_t(x) = \begin{matrix} 1 & \text{if } T(x) > t \\ 0 & \text{if } T(x) < t \end{matrix} \tag{4.3.3}$$

with $\delta_t(x)$ any value in $(0, 1)$ if $T(x) = t$. Consider the problem of testing $H : \theta = \theta_0$ versus $K : \theta = \theta_1$ with $\theta_0 < \theta_1$. If $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset R$, is an MLR family in $T(x)$, then $L(x, \theta_0, \theta_1) = h(T(x))$ for some increasing function $h$. Thus, $\delta_t$ equals the likelihood ratio test $\varphi_{h(t)}$ and is MP. Because $\delta_t$ does not depend on $\theta_1$, it is UMP at level $\alpha = E_{\theta_0}\delta_t(x)$ for testing $H : \theta = \theta_0$ versus $K : \theta > \theta_0$, in fact.

**Theorem 4.3.1.** *Suppose $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset R$, is an MLR family in $T(x)$.*

(1) *For each $t \in (0, \infty)$, the power function $\beta(\theta) = E_\theta\delta_t(X)$ is increasing in $\theta$.*

(2) *If $E_{\theta_0}\delta_t(X) = \alpha > 0$, then $\delta_t$ is UMP level $\alpha$ for testing $H : \theta \le \theta_0$ versus $K : \theta > \theta_0$.*

***Proof.*** (1) follows from $\delta_t = \varphi_{h(t)}$ and Corollary 4.2.1 by noting that for any $\theta_1 < \theta_2$, $\delta_t$ is MP at level $E_{\theta_1}\delta_t(X)$ for testing $H : \theta = \theta_1$ versus $K : \theta = \theta_2$. To show (2), recall that we have seen that $\delta_t$ maximizes the power for testing $H : \theta = \theta_0$ versus $K : \theta > \theta_0$ among the class of tests with level $\alpha = E_{\theta_0}\delta_t(X)$. If $\theta < \theta_0$, then by (1), $E_\theta\delta_t(X) \le \alpha$ and $\delta_t$ is of level $\alpha$ for $H : \theta \le \theta_0$. Because the class of tests with level $\alpha$ for $H : \theta \le \theta_0$ is contained in the class of tests with level $\alpha$ for $H : \theta = \theta_0$, and because $\delta_t$ maximizes the power over this larger class, $\delta_t$ is UMP for $H : \theta \le \theta_0$ versus $K : \theta > \theta_0$. □

The following useful result follows immediately.

**Corollary 4.3.1.** *Suppose $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset R$, is an MLR family in $T(x)$. If the distribution function $F_0$ of $T(X)$ under $X \sim P_{\theta_0}$ is continuous and if $t(1-\alpha)$ is a solution of $F_0(t) = 1 - \alpha$, then the test that rejects H if and only if $T(x) \ge t(1-\alpha)$ is UMP level $\alpha$ for testing $H : \theta \le \theta_0$ versus $K : \theta > \theta_0$.*

**Example 4.3.4.** *Testing Precision.* Suppose $X_1, \dots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma^2)$ population, where $\mu$ is a known standard, and we are interested in the precision $\sigma^{-1}$ of the measurements $X_1, \dots, X_n$. For instance, we could be interested in the precision of a new measuring instrument and test it by applying it to a known standard. Because the more serious error is to judge the precision adequate when it is not, we test $H : \sigma \ge \sigma_0$ versus $K : \sigma < \sigma_0$, where $\sigma_0^{-1}$ represents the minimum tolerable precision. Let $S = \sum_{i=1}^n (X_i - \mu)^2$, then

$$p(\mathbf{x}, \theta) = \exp\left\{-\frac{1}{2\sigma^2}S - \frac{n}{2}\log(2\pi\sigma^2)\right\}.$$

This is a one-parameter exponential family and is MLR in $T = -S$. The UMP level $\alpha$ test rejects $H$ if and only if $S \le s(\alpha)$ where $s(\alpha)$ is such that $P_{\sigma_0}(S \le s(\alpha)) = \alpha$. If we write

$$\frac{S}{\sigma_0^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma_0}\right)^2$$

we see that $S/\sigma_0^2$ has a $\chi_n^2$ distribution. Thus, the critical constant $s(\alpha)$ is $\sigma_0^2 x_n(\alpha)$, where $x_n(\alpha)$ is the $\alpha$th quantile of the $\chi_n^2$ distribution. □

**Example 4.3.5.** *Quality Control.* Suppose that, as in Example 1.1.1, $X$ is the observed number of defectives in a sample of $n$ chosen at random without replacement from a lot of $N$ items containing $b$ defectives, where $b = N\theta$. If the inspector making the test considers lots with $b_0 = N\theta_0$ defectives or more unsatisfactory, she formulates the hypothesis $H$ as $\theta \ge \theta_0$, the alternative $K$ as $\theta < \theta_0$, and specifies an $\alpha$ such that the probability of rejecting $H$ (keeping a bad lot) is at most $\alpha$. If $\alpha$ is a value taken on by the distribution of $X$, we now show that the test $\delta^*$ which rejects $H$ if, and only if, $X \le h(\alpha)$, where $h(\alpha)$ is the $\alpha$th quantile of the hypergeometric, $\mathcal{H}(N\theta_0, N, n)$, distribution, is UMP level $\alpha$. For simplicity suppose that $b_0 \geq n$, $N - b_0 \geq n$. Then, if $N\theta_1 = b_1 < b_0$ and $0 \leq x \leq b_1$, (1.1.1) yields

$$L(x, \theta_0, \theta_1) = \frac{b_1(b_1 - 1)\dots(b_1 - x + 1)(N - b_1)\dots(N - b_1 - n + x + 1)}{b_0(b_0 - 1)\dots(b_0 - x + 1)(N - b_0)\dots(N - b_0 - n + x + 1)}.$$

Note that $L(x, \theta_0, \theta_1) = 0$ for $b_1 < x \leq n$. Thus, for $0 \leq x \leq b_1 - 1$,

$$\frac{L(x+1, \theta_0, \theta_1)}{L(x, \theta_0, \theta_1)} = \left(\frac{b_1 - x}{b_0 - x}\right) \frac{(N - n + 1) - (b_0 - x)}{(N - n + 1) - (b_1 - x)} < 1.$$

Therefore, $L$ is decreasing in $x$ and the hypergeometric model is an MLR family in $T(x) = -x$. It follows that $\delta^*$ is UMP level $\alpha$. The critical values for the hypergeometric distribution are available on statistical calculators and software. □

## Power and Sample Size

In the Neyman–Pearson framework we choose the test whose size is small. That is, we choose the critical constant so that the maximum probability of falsely rejecting the null hypothesis $H$ is small. On the other hand, we would also like large power $\beta(\theta)$ when $\theta \in \Theta_1$; that is, we want the probability of correctly detecting an alternative $K$ to be large. However, as seen in Figure 4.1.1 and formula (4.1.2), this is, in general, not possible for all parameters in the alternative $\Theta_1$. In both these cases, $H$ and $K$ are of the form $H : \theta \leq \theta_0$ and $K : \theta > \theta_0$, and the powers are continuous increasing functions with $\lim_{\theta \downarrow \theta_0} \beta(\theta) = \alpha$. By Corollary 4.3.1, this is a general phenomenon in MLR family models with $p(\mathbf{x}, \theta)$ continuous in $\theta$.

This continuity of the power shows that not too much significance can be attached to acceptance of $H$, if all points in the alternative are of equal significance: We can find $\theta > \theta_0$ sufficiently close to $\theta_0$ so that $\beta(\theta)$ is arbitrarily close to $\beta(\theta_0) = \alpha$. For such $\theta$ the probability of falsely accepting $H$ is almost $1 - \alpha$.

This is not serious in practice if we have an *indifference region*. This is a subset of the alternative on which we are willing to tolerate low power. In our normal example 4.1.4 we might be uninterested in values of $\mu$ in $(0, \Delta)$ for some small $\Delta > 0$ because such improvements are negligible. Thus, $(0, \Delta)$ would be our indifference region. Off the indifference region, we want guaranteed power as well as an upper bound on the probability of type I error. In our example this means that in addition to the indifference region and level $\alpha$, we specify $\beta$ close to 1 and would like to have $\beta(\mu) \geq \beta$ for all $\mu \geq \Delta$. This is possible for arbitrary $\beta < 1$ only by making the sample size $n$ large enough. In Example 4.1.4 because $\beta(\mu)$ is increasing, the appropriate $n$ is obtained by solving

$$\beta(\Delta) = \Phi(z(\alpha) + \sqrt{n}\Delta/\sigma) = \beta$$

for sample size $n$. This equation is equivalent to

$$z(\alpha) + \sqrt{n}\Delta/\sigma = z(\beta)$$

whose solution is

$$n = (\Delta/\sigma)^{-2}[z(1 - \alpha) + z(\beta)]^2.$$

Note that a small signal-to-noise ratio $\Delta/\sigma$ will require a large sample size $n$.

Dual to the problem of not having enough power is that of having too much. It is natural to associate statistical significance with practical significance so that a very low $p$-value is interpreted as evidence that the alternative that holds is physically significant, that is, far from the hypothesis. Formula $(4.1.2)$ shows that, if $n$ is very large and/or $\sigma$ is small, we can have very great power for alternatives very close to $0$. This problem arises particularly in goodness-of-fit tests (see Example $4.1.5$), when we test the hypothesis that a very large sample comes from a particular distribution. Such hypotheses are often rejected even though for practical purposes "the fit is good enough." The reason is that $n$ is so large that unimportant small discrepancies are picked up. There are various ways of dealing with this problem. They often reduce to adjusting the critical value so that the probability of rejection for parameter value at the boundary of some indifference region is $\alpha$. In Example $4.1.4$ this would mean rejecting $H$ if, and only if,

$$\sqrt{n}\frac{\bar{X}}{\sigma} \geq z(1-\alpha) + \sqrt{n}\frac{\Delta}{\sigma}.$$

As a further example and precursor to Section $5.4.4$, we next show how to find the sample size that will "approximately" achieve desired power $\beta$ for the size $\alpha$ test in the binomial example.

**Example 4.3.6 (Example 4.1.3 continued).** Our discussion uses the classical normal approximation to the binomial distribution. First, to achieve approximate size $\alpha$, we solve $\beta(\theta_0) = P_{\theta_0}(S \geq s)$ for $s$ using $(4.1.4)$ and find the approximate critical value

$$s_0 = n\theta_0 + \frac{1}{2} + z(1-\alpha)[n\theta_0(1-\theta_0)]^{1/2}.$$

Again using the normal approximation, we find

$$\beta(\theta) = P_\theta(S \geq s_0) = \Phi\left(\frac{n\theta + \frac{1}{2} - s_0}{[n\theta(1-\theta)]^{1/2}}\right).$$

Now consider the indifference region $(\theta_0, \theta_1)$, where $\theta_1 = \theta_0 + \Delta$, $\Delta > 0$. We solve $\beta(\theta_1) = \beta$ for $n$ and find the approximate solution

$$n = (\theta_1 - \theta_0)^{-2}\{z(1-\alpha)[\theta_0(1-\theta_0)]^{1/2} + z(\beta)[\theta_1(1-\theta_1)]^{1/2}\}^2.$$

For instance, if $\alpha = .05$, $\beta = .90$, $\theta_0 = 0.3$, and $\theta_1 = 0.35$, we need

$$n = (0.05)^{-2}\{1.645 \times 0.3(0.7) + 1.282 \times 0.35(0.65)\}^2 = 162.4.$$

Thus, the size .05 binomial test of $H : \theta = 0.3$ requires approximately 163 observations to have probability .90 of detecting the 17% increase in $\theta$ from $0.3$ to $0.35$. The power achievable (exactly, using the SPLUS package) for the level .05 test for $\theta = .35$ and $n = 163$ is 0.86. □

Our discussion can be generalized. Suppose $\theta$ is a vector. Often there is a function $q(\theta)$ such that $H$ and $K$ can be formulated as $H : q(\theta) \leq q_0$ and $K : q(\theta) > q_0$. Now let

$q_1 > q_0$ be a value such that we want to have power $\beta(\theta)$ at least $\beta$ when $q(\theta) \geq q_1$. The set $\{\theta : q_0 < q(\theta) < q_1\}$ is our indifference region. For each $n$ suppose we have a level $\alpha$ test for $H$ versus $K$ based on a suitable test statistic $T$. Suppose that $\beta(\theta)$ depends on $\theta$ only through $q(\theta)$ and is a continuous increasing function of $q(\theta)$, and also increases to 1 for fixed $\theta \in \Theta_1$ as $n \to \infty$. To achieve level $\alpha$ and power at least $\beta$, first let $c_0$ be the smallest number $c$ such that

$$P_{\theta_0}[T \geq c] \leq \alpha.$$

Then let $n$ be the smallest integer such that

$$P_{\theta_1}[T \geq c_0] \geq \beta$$

where $\theta_0$ is such that $q(\theta_0) = q_0$ and $\theta_1$ is such that $q(\theta_1) = q_1$. This procedure can be applied, for instance, to the $F$ test of the linear model in Section 6.1 by taking $q(\theta)$ equal to the noncentrality parameter governing the distribution of the statistic under the alternative.

Implicit in this calculation is the assumption that $P_{\theta_1}[T \geq c_0]$ is an increasing function of $n$.

We have seen in Example 4.1.5 that a particular test statistic can have a fixed distribution $\mathcal{L}_0$ under the hypothesis. It may also happen that the distribution of $T_n$ as $\theta$ ranges over $\Theta_1$ is determined by a one-dimensional parameter $\lambda(\theta)$ so that $\Theta_0 = \{\theta : \lambda(\theta) = 0\}$ and $\Theta_1 = \{\theta : \lambda(\theta) > 0\}$ and $\mathcal{L}_0(T_n) = \mathcal{L}_{\lambda(\theta)}(T_n)$ for all $\theta$. The theory we have developed demonstrates that if $\mathcal{L}_\lambda(T_n)$ is an MLR family, then rejecting for large values of $T_n$ is UMP among all tests based on $T_n$. Reducing the problem to choosing among such tests comes from invariance considerations that we do not enter into until Volume II. However, we illustrate what can happen with a simple example.

**Example 4.3.7.** *Testing Precision Continued.* Suppose that in the Gaussian model of Example 4.3.4, $\mu$ is unknown. Then the MLE of $\sigma^2$ is $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ as in Example 2.2.9. Although $H : \sigma = \sigma_0$ is now composite, the distribution of $T_n \equiv n\widehat{\sigma}^2/\sigma_0^2$ is $\chi^2_{n-1}$, independent of $\mu$. Thus, the critical value for testing $H : \sigma = \sigma_0$ versus $K : \sigma < \sigma_0$ and rejecting $H$ if $T_n$ is small, is the $\alpha$ percentile of $\chi^2_{n-1}$. It is evident from the argument of Example 4.3.3 that this test is UMP for $H : \sigma \geq \sigma_0$ versus $K : \sigma < \sigma_0$ *among* all tests depending on $\widehat{\sigma}^2$ only. □

## Complete Families of Tests

The Neyman–Pearson framework is based on using the 0-1 loss function. We may ask whether decision procedures other than likelihood ratio tests arise if we consider loss functions $l(\theta, a)$, $a \in \mathcal{A} = \{0, 1\}$, $\theta \in \Theta$, that are not 0-1. For instance, for $\Theta_1 = (\theta_0, \infty)$, we may consider $l(\theta, 0) = (\theta - \theta_0)$, $\theta \in \Theta_1$. In general, when testing $H : \theta \leq \theta_0$ versus $K : \theta > \theta_0$, a reasonable class of loss functions are those that satisfy

$$\begin{aligned} l(\theta,1) - l(\theta,0) > 0 \quad & \text{for } \theta < \theta_0 \\ l(\theta,1) - l(\theta,0) < 0 \quad & \text{for } \theta > \theta_0. \end{aligned} \tag{4.3.4}$$

The class $\mathcal{D}$ of decision procedures is said to be *complete*$^{(1),(2)}$ if for any decision rule $\varphi$ there exists $\delta \in \mathcal{D}$ such that

$$R(\theta, \delta) \leq R(\theta, \varphi) \text{ for all } \theta \in \Theta. \tag{4.3.5}$$

That is, if the model is correct and loss function is appropriate, then any procedure not in the complete class can be matched or improved *at all* $\theta$ by one in the complete class. Thus, it isn't worthwhile to look outside of complete classes. In the following the decision procedures are test functions.

**Theorem 4.3.2.** *Suppose* $\{P_\theta : \theta \in \Theta\}$*,* $\Theta \subset R$*, is an MLR family in* $T(x)$ *and suppose the loss function* $l(\theta, a)$ *satisfies* (4.3.4)*, then the class of tests of the form* (4.3.3) *with* $E\delta_t(X) = \alpha$*,* $0 \leq \alpha \leq 1$*, is complete.*

***Proof.*** The risk function of any test rule $\varphi$ is

$$\begin{aligned} R(\theta, \varphi) &= E_\theta\{\varphi(X)l(\theta, 1) + [1 - \varphi(X)]l(\theta, 0)\} \\ &= E_\theta\{l(\theta, 0) + [l(\theta, 1) - l(\theta, 0)]\varphi(X)\}. \end{aligned}$$

Let $\delta_t(X)$ be such that, for some $\theta_0$, $E_{\theta_0}\delta_t(X) = E_{\theta_0}\varphi(X) > 0$. If $E_\theta\varphi(X) \equiv 0$ for all $\theta$ then $\delta_\infty(X)$ clearly satisfies (4.3.5). Now $\delta_t$ is UMP for $H : \theta \leq \theta_0$ versus $K : \theta > \theta_0$ by Theorem 4.3.1 and, hence,

$$R(\theta, \delta_t) - R(\theta, \varphi) = (l(\theta, 1) - l(\theta, 0))(E_\theta(\delta_t(x)) - E_\theta(\varphi(X))) \leq 0 \text{ for } \theta > \theta_0. \tag{4.3.6}$$

But $1 - \delta_t$ is similarly UMP for $H : \theta \geq \theta_0$ versus $K : \theta < \theta_0$ (Problem 4.3.12) and, hence, $E_\theta(1 - \delta_t(X)) = 1 - E_\theta\delta_t(X) \geq 1 - E_\theta\varphi(X)$ for $\theta < \theta_0$. Thus, (4.3.5) holds for all $\theta$. □

**Summary.** We consider models $\{P_\theta : \theta \in \Theta\}$ for which there exist tests that are most powerful for every $\theta$ in a composite alternative $\Theta_1$ (UMP tests). For $\theta$ real, a model is said to be monotone likelihood ratio (MLR) if the simple likelihood ratio statistic for testing $\theta_0$ versus $\theta_1$ is an increasing function of a statistic $T(x)$ for every $\theta_0 < \theta_1$. For MLR models, the test that rejects $H : \theta \leq \theta_0$ for large values of $T(x)$ is UMP for $K : \theta > \theta_0$. In such situations we show how sample size can be chosen to guarantee minimum power for alternatives a given distance from $H$. Finally, we show that for MLR models, the class of most powerful Neyman Pearson (NP) tests is complete in the sense that for loss functions other than the 0-1 loss function, the risk of any procedure can be matched or improved by an NP test.

## 4.4 CONFIDENCE BOUNDS, INTERVALS, AND REGIONS

We have in Chapter 2 considered the problem of obtaining precise estimates of parameters and we have in this chapter treated the problem of deciding whether the parameter $\theta$ is a member of a specified set $\Theta_0$. Now we consider the problem of giving confidence bounds, intervals, or sets that constrain the parameter with prescribed probability $1 - \alpha$. As an illustration consider Example 4.1.4 where $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with $\sigma^2$ known. Suppose that $\mu$ represents the mean increase in sleep among patients administered a drug. Then we can use the experimental outcome $\mathbf{X} = (X_1, \dots, X_n)$ to establish a lower bound $\underline{\mu}(\mathbf{X})$ for $\mu$ with a prescribed probability $(1 - \alpha)$ of being correct. In the non-Bayesian framework, $\mu$ is a constant, and we look for a statistic $\underline{\mu}(\mathbf{X})$ that satisfies $P(\underline{\mu}(\mathbf{X}) \le \mu) = 1 - \alpha$ with $1 - \alpha$ equal to .95 or some other desired level of confidence. In our example this is achieved by writing

$$P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \le z(1 - \alpha)\right) = 1 - \alpha.$$

By solving the inequality inside the probability for $\mu$, we find

$$P(\bar{X} - \sigma z(1 - \alpha)/\sqrt{n} \le \mu) = 1 - \alpha$$

and

$$\underline{\mu}(\mathbf{X}) = \bar{X} - \sigma z(1 - \alpha)/\sqrt{n}$$

is a lower bound with $P(\underline{\mu}(\mathbf{X}) \le \mu) = 1 - \alpha$. We say that $\underline{\mu}(\mathbf{X})$ is a *lower confidence bound with confidence level* $1 - \alpha$.

Similarly, as in (1.3.8), we may be interested in an upper bound on a parameter. In the $\mathcal{N}(\mu, \sigma^2)$ example this means finding a statistic $\bar{\mu}(\mathbf{X})$ such that $P(\bar{\mu}(\mathbf{X}) \ge \mu) = 1 - \alpha$; and a solution is

$$\bar{\mu}(\mathbf{X}) = \bar{X} + \sigma z(1 - \alpha)/\sqrt{n}.$$

Here $\bar{\mu}(\mathbf{X})$ is called an *upper level* $(1 - \alpha)$ *confidence bound* for $\mu$.

Finally, in many situations where we want an indication of the accuracy of an estimator, we want both lower and upper bounds. That is, we want to find $a$ such that the probability that the interval $[\bar{X} - a, \bar{X} + a]$ contains $\mu$ is $1 - \alpha$. We find such an interval by noting

$$P\left(\frac{\sqrt{n}|\bar{X} - \mu|}{\sigma} \le z\left(1 - \tfrac{1}{2}\alpha\right)\right) = 1 - \alpha$$

and solving the inequality inside the probability for $\mu$. This gives

$$P(\mu^-(\mathbf{X}) \le \mu \le \mu^+(\mathbf{X})) = 1 - \alpha$$

where

$$\mu^{\pm}(\mathbf{X}) = \bar{X} \pm \sigma z\left(1 - \tfrac{1}{2}\alpha\right)/\sqrt{n}.$$

We say that $[\mu^-(\mathbf{X}), \mu^+(\mathbf{X})]$ is a *level* $(1 - \alpha)$ *confidence interval* for $\mu$.

In general, if $\nu = \nu(P)$, $P \in \mathcal{P}$, is a parameter, and $X \sim P$, $X \in R^q$, it may not be possible for a bound or interval to achieve exactly probability $(1 - \alpha)$ for a prescribed $(1 - \alpha)$ such as .95. In this case, we settle for a probability at least $(1 - \alpha)$. That is,

**Definition 4.4.1.** A statistic $\underline{\nu}(X)$ is called a *level* $(1-\alpha)$ *lower confidence bound* for $\nu$ if for every $P \in \mathcal{P}$,

$$P[\underline{\nu}(X) \leq \nu] \geq 1 - \alpha.$$

Similarly, $\bar{\nu}(X)$ is called a *level* $(1-\alpha)$ *upper confidence bound* for $\nu$ if for every $P \in \mathcal{P}$,

$$P[\bar{\nu}(X) \geq \nu] \geq 1 - \alpha.$$

Moreover, the random interval $[\underline{\nu}(X), \bar{\nu}(X)]$ formed by a pair of statistics $\underline{\nu}(X)$, $\bar{\nu}(X)$ is a *level* $(1-\alpha)$ or a $100(1-\alpha)\%$ *confidence interval* for $\nu$ if, for all $P \in \mathcal{P}$,

$$P[\underline{\nu}(X) \leq \nu \leq \bar{\nu}(X)] \geq 1 - \alpha.$$

The quantities on the left are called the *probabilities of coverage* and $(1-\alpha)$ is called a *confidence level.*

For a given bound or interval, the confidence level is clearly not unique because any number $(1-\alpha') \leq (1-\alpha)$ will be a confidence level if $(1-\alpha)$ is. In order to avoid this ambiguity it is convenient to define the *confidence coefficient* to be the largest possible confidence level. Note that in the case of intervals this is just

$$\inf\{P[\underline{\nu}(X) \leq \nu \leq \bar{\nu}(X), P \in \mathcal{P}]\}$$

(i.e., the minimum probability of coverage). For the normal measurement problem we have just discussed the probability of coverage is independent of $P$ and equals the confidence coefficient.

**Example 4.4.1.** *The (Student) t Interval and Bounds.* Let $X_1, \ldots, X_n$ be a sample from a $\mathcal{N}(\mu, \sigma^2)$ population, and assume initially that $\sigma^2$ is known. In the preceding discussion we used the fact that $Z(\mu) = \sqrt{n}(\bar{X} - \mu)/\sigma$ has a $\mathcal{N}(0,1)$ distribution to obtain a confidence interval for $\mu$ by solving $-z\left(1 - \frac{1}{2}\alpha\right) \leq Z(\mu) \leq z\left(1 - \frac{1}{2}\alpha\right)$ for $\mu$. In this process $Z(\mu)$ is called a *pivot*. In general, finding confidence intervals (or bounds) often involves finding appropriate pivots. Now we turn to the $\sigma^2$ unknown case and propose the pivot $T(\mu)$ obtained by replacing $\sigma$ in $Z(\mu)$ by its estimate $s$, where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

That is, we will need the distribution of

$$T(\mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{s}.$$

Now $Z(\mu) = \sqrt{n}(\bar{X} - \mu)/\sigma$ has a $\mathcal{N}(0,1)$ distribution and is, by Theorem B.3.3, independent of $V = (n-1)s^2/\sigma^2$, which has a $\chi^2_{n-1}$ distribution. We conclude from the definition of the (Student) $t$ distribution in Section B.3.1 that $Z(\mu)/\sqrt{V/(n-1)} = T(\mu)$

has the $t$ distribution $\mathcal{T}_{n-1}$ whatever be $\mu$ and $\sigma^2$. Let $t_k(p)$ denote the $p$th quantile of the $\mathcal{T}_k$ distribution. Then

$$P\left(-t_{n-1}\left(1-\tfrac{1}{2}\alpha\right) \le T(\mu) \le t_{n-1}\left(1-\tfrac{1}{2}\alpha\right)\right) = 1-\alpha.$$

Solving the inequality inside the probability for $\mu$, we find

$$P\left[\bar{X} - st_{n-1}\left(1-\tfrac{1}{2}\alpha\right)/\sqrt{n} \le \mu \le \bar{X} + st_{n-1}\left(1-\tfrac{1}{2}\alpha\right)/\sqrt{n}\right] = 1-\alpha.$$

The shortest level $(1-\alpha)$ confidence interval of the type $\bar{X} \pm sc/\sqrt{n}$ is, thus,

$$\left[\bar{X} - st_{n-1}\left(1-\tfrac{1}{2}\alpha\right)/\sqrt{n},\ \bar{X} + st_{n-1}\left(1-\tfrac{1}{2}\alpha\right)/\sqrt{n}\right]. \tag{4.4.1}$$

Similarly, $\bar{X} - st_{n-1}(1-\alpha)/\sqrt{n}$ and $\bar{X} + st_{n-1}(1-\alpha)/\sqrt{n}$ are natural lower and upper confidence bounds with confidence coefficients $(1-\alpha)$.

To calculate the coefficients $t_{n-1}\left(1-\frac{1}{2}\alpha\right)$ and $t_{n-1}(1-\alpha)$, we use a calculator, computer software, or Tables I and II. For instance, if $n = 9$ and $\alpha = 0.01$, we enter Table II to find that the probability that a $\mathcal{T}_{n-1}$ variable exceeds 3.355 is .005. Hence, $t_{n-1}\left(1-\frac{1}{2}\alpha\right) = 3.355$ and

$$[\bar{X} - 3.355s/3,\ \bar{X} + 3.355s/3]$$

is the desired level 0.99 confidence interval.

From the results of Section B.7 (see Problem B.7.12), we see that as $n \to \infty$ the $\mathcal{T}_{n-1}$ distribution converges in law to the standard normal distribution. For the usual values of $\alpha$, we can reasonably replace $t_{n-1}(p)$ by the standard normal quantile $z(p)$ for $n > 120$.

Up to this point, we have assumed that $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. It turns out that the distribution of the pivot $T(\mu)$ is fairly close to the $\mathcal{T}_{n-1}$ distribution if the $X$'s have a distribution that is nearly symmetric and whose tails are not much heavier than the normal. In this case the interval (4.4.1) has confidence coefficient close to $1-\alpha$. On the other hand, for very skew distributions such as the $\chi^2$ with few degrees of freedom, or very heavy-tailed distributions such as the Cauchy, the confidence coefficient of (4.4.1) can be much smaller than $1-\alpha$. The properties of confidence intervals such as (4.4.1) in non-Gaussian situations can be investigated using the asymptotic and Monte Carlo methods introduced in Chapter 5. See Figure 5.3.1. If we assume $\sigma^2 < \infty$, the interval will have probability $(1-\alpha)$ in the limit as $n \to \infty$. □

**Example 4.4.2.** *Confidence Intervals and Bounds for the Variance of a Normal Distribution.* Suppose that $X_1, \dots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma^2)$ population. By Theorem B.3.1, $V(\sigma^2) = (n-1)s^2/\sigma^2$ has a $\chi^2_{n-1}$ distribution and can be used as a pivot. Thus, if we let $x_{n-1}(p)$ denote the $p$th quantile of the $\chi^2_{n-1}$ distribution, and if $\alpha_1 + \alpha_2 = \alpha$, then

$$P(x(\alpha_1) \le V(\sigma^2) \le x(1-\alpha_2)) = 1-\alpha.$$

By solving the inequality inside the probability for $\sigma^2$ we find that

$$[(n-1)s^2/x(1-\alpha_2),\ (n-1)s^2/x(\alpha_1)] \tag{4.4.2}$$

is a confidence interval with confidence coefficient $(1-\alpha)$.

The length of this interval is random. There is a unique choice of $\alpha_1$ and $\alpha_2$, which uniformly minimizes expected length among all intervals of this type. It may be shown that for $n$ large, taking $\alpha_1 = \alpha_2 = \frac{1}{2}\alpha$ is not far from optimal (Tate and Klett, 1959).

The pivot $V(\sigma^2)$ similarly yields the respective lower and upper confidence bounds $(n-1)s^2/x(1-\alpha)$ and $(n-1)s^2/x(\alpha)$.

In contrast to Example 4.4.1, if we drop the normality assumption, the confidence interval and bounds for $\sigma^2$ do not have confidence coefficient $1-\alpha$ even in the limit as $n \to \infty$. Asymptotic methods and Monte Carlo experiments as described in Chapter 5 have shown that the confidence coefficient may be arbitrarily small depending on the underlying true distribution, which typically is unknown. In Problem 4.4.16 we give an interval with correct limiting coverage probability. □

The method of pivots works primarily in problems related to sampling from normal populations. If we consider "approximate" pivots, the scope of the method becomes much broader. We illustrate by an example.

**Example 4.4.3.** *Approximate Confidence Bounds and Intervals for the Probability of Success in $n$ Bernoulli Trials.* If $X_1, \dots, X_n$ are the indicators of $n$ Bernoulli trials with probability of success $\theta$, then $\bar{X}$ is the MLE of $\theta$. There is no natural "exact" pivot based on $\bar{X}$ and $\theta$. However, by the De Moivre–Laplace theorem, $\sqrt{n}(\bar{X}-\theta)/\sqrt{\theta(1-\theta)}$ has approximately a $\mathcal{N}(0,1)$ distribution. If we use this function as an "approximate" pivot and let $\approx$ denote "approximate equality," we can write

$$P\left[\left|\frac{\sqrt{n}(\bar{X}-\theta)}{\sqrt{\theta(1-\theta)}}\right| \le z\left(1-\tfrac{1}{2}\alpha\right)\right] \approx 1-\alpha.$$

Let $k_\alpha = z\left(1-\frac{1}{2}\alpha\right)$ and observe that this is equivalent to

$$P\left[(\bar{X}-\theta)^2 \le \frac{k_\alpha^2}{n}\theta(1-\theta)\right] = P[g(\theta,\bar{X}) \le 0] \approx 1-\alpha$$

where

$$g(\theta,\bar{X}) = \left(1+\frac{k_\alpha^2}{n}\right)\theta^2 - \left(2\bar{X}+\frac{k_\alpha^2}{n}\right)\theta + \bar{X}^2.$$

For fixed $0 \le \bar{X} \le 1$, $g(\theta,\bar{X})$ is a quadratic polynomial with two real roots. In terms of $S = n\bar{X}$, they are$^{(1)}$

$$\begin{aligned} \underline{\theta}(\mathbf{X}) &= \left\{S+\frac{k_\alpha^2}{2} - k_\alpha\sqrt{[S(n-S)/n]+k_\alpha^2/4}\right\}\Big/ (n+k_\alpha^2) \\ \bar{\theta}(\mathbf{X}) &= \left\{S+\frac{k_\alpha^2}{2} + k_\alpha\sqrt{[S(n-S)/n]+k_\alpha^2/4}\right\}\Big/ (n+k_\alpha^2). \end{aligned} \tag{4.4.3}$$

Because the coefficient of $\theta^2$ in $g(\theta,\bar{X})$ is greater than zero,

$$[\theta : g(\theta,\bar{X}) \le 0] = [\underline{\theta}(\mathbf{X}) \le \theta \le \bar{\theta}(\mathbf{X})], \tag{4.4.4}$$

so that $[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$ is an approximate level $(1-\alpha)$ confidence interval for $\theta$. We can similarly show that the endpoints of the level $(1-2\alpha)$ interval are approximate upper and lower level $(1-\alpha)$ confidence bounds. These intervals and bounds are satisfactory in practice for the usual levels, if the smaller of $n\theta, n(1-\theta)$ is at least $6$. For small $n$, it is better to use the exact level $(1-\alpha)$ procedure developed in Section $4.5$. A discussion is given in Brown, Cai, and Das Gupta (2001).

Note that in this example we can determine the sample size needed for desired accuracy. For instance, consider the market researcher whose interest is the proportion $\theta$ of a population that will buy a product. He draws a sample of $n$ potential customers, calls willingness to buy success, and uses the preceding model. He can then determine how many customers should be sampled so that $(4.4.4)$ has length $0.02$ and is a confidence interval with confidence coefficient approximately $0.95$. To see this, note that the length, say $l$, of the interval is

$$l = 2k_\alpha\{\sqrt{[S(n-S)/n] + k_\alpha^2/4}\}(n+k_\alpha^2)^{-1}.$$

Now use the fact that

$$S(n-S)/n = \tfrac{1}{4}n - n^{-1}\left(S - \tfrac{1}{2}n\right)^2 \le \tfrac{1}{4}n \tag{4.4.5}$$

to conclude that

$$l \le k_\alpha/\sqrt{n+k_\alpha^2}. \tag{4.4.6}$$

Thus, to bound $l$ above by $l_0 = 0.02$, we choose $n$ so that $k_\alpha(n+k_\alpha^2)^{-\frac{1}{2}} = l_0$. That is, we choose

$$n = \left(\frac{k_\alpha}{l_0}\right)^2 - k_\alpha^2.$$

In this case, $1-\frac{1}{2}\alpha = 0.975$, $k_\alpha = z\left(1-\frac{1}{2}\alpha\right) = 1.96$, and we can achieve the desired length $0.02$ by choosing $n$ so that

$$n = \left(\frac{1.96}{0.02}\right)^2 - (1.96)^2 = 9,600.16, \text{ or } n = 9,601.$$

This formula for the sample size is very crude because $(4.4.5)$ is used and it is only good when $\theta$ is near $1/2$. Better results can be obtained if one has upper or lower bounds on $\theta$ such as $\theta \le \theta_0 < \frac{1}{2}, \theta \ge \theta_1 > \frac{1}{2}$. See Problem $4.4.4$.

Another approximate pivot for this example is $\sqrt{n}(\bar{X}-\theta)/\sqrt{\bar{X}(1-\bar{X})}$. This leads to the simple interval

$$\bar{X} \pm k_\alpha\sqrt{\bar{X}(1-\bar{X})}/\sqrt{n}. \tag{4.4.7}$$

See Brown, Cai, and Das Gupta (2001) for a discussion. □

### Confidence Regions for Functions of Parameters

We define level $(1-\alpha)$ confidence regions for a function $q(\theta)$ as random subsets of the range of $q$ that cover the true value of $q(\theta)$ with probability at least $(1-\alpha)$. Note that if $C(\mathbf{X})$ is a level $(1-\alpha)$ confidence region for $\theta$, then $q(C(\mathbf{X})) = \{q(\theta) : \theta \in C(\mathbf{X})\}$ is a level $(1-\alpha)$ confidence region for $q(\theta)$.

If the distribution of $T_{\mathbf{X}}(\theta)$ does not invlove $\theta$, then it is called a *pivot*. In this case, if $P_\theta(T_{\mathbf{X}}(\theta) \in A) \geq 1-\alpha$, then $T_{\mathbf{X}}^{-1}(A)$ is a level $(1-\alpha)$ confidence region for $\theta$.

**Example 4.4.4.** Let $X_1, \ldots, X_n$ denote the number of hours a sample of Internet subscribers spend per week on the Internet. Suppose $X_1, \ldots, X_n$ is modeled as a sample from an exponential, $\mathcal{E}(\theta^{-1})$, distribution, and suppose we want a confidence interval for the population proportion $P(X \geq x)$ of subscribers that spend at least $x$ hours per week on the Internet. Here $q(\theta) = 1 - F(x) = \exp\{-x/\theta\}$. By Problem B.3.4, $2n\bar{X}/\theta$ has a chi-square, $\chi^2_{2n}$, distribution. By using $2n\bar{X}/\theta$ as a pivot we find the $(1-\alpha)$ confidence interval

$$2n\bar{X}/x\left(1-\tfrac{1}{2}\alpha\right) \leq \theta \leq 2n\bar{X}/x\left(\tfrac{1}{2}\alpha\right)$$

where $x(\beta)$ denotes the $\beta$th quantile of the $\chi^2_{2n}$ distribution. Let $\underline{\theta}$ and $\bar{\theta}$ denote the lower and upper boundaries of this interval, then

$$\exp\{-x/\underline{\theta}\} \leq q(\theta) \leq \exp\{-x/\bar{\theta}\}$$

is a confidence interval for $q(\theta)$ with confidence coefficient $(1-\alpha)$.

If $q$ is not $1-1$, this technique is typically wasteful. That is, we can find confidence regions for $q(\theta)$ entirely contained in $q(C(\mathbf{X}))$ with confidence level $(1-\alpha)$. For instance, we will later give confidence regions $C(\mathbf{X})$ for pairs $\theta = (\theta_1, \theta_2)^T$. In this case, if $q(\theta) = \theta_1$, $q(C(\mathbf{X}))$ is larger than the confidence set obtained by focusing on $\theta_1$ alone.

### Confidence Regions of Higher Dimension

We can extend the notion of a confidence interval for one-dimensional functions $q(\theta)$ to $r$-dimensional vectors $\mathbf{q}(\theta) = (q_1(\theta), \ldots, q_r(\theta))$. Suppose $\underline{q}_j(X)$ and $\bar{q}_j(X)$ are real-valued. Then the $r$-dimensional random rectangle

$$I(X) = \{\mathbf{q}(\theta) : \underline{q}_j(X) \leq q_j(\theta) \leq \bar{q}_j(X),\ j = 1, \ldots, r\}$$

is said to be a level $(1-\alpha)$ confidence region, if the probability that it covers the unknown but fixed true $(q_1(\theta), \ldots, q_r(\theta))$ is at least $(1-\alpha)$. We write this as

$$P[\mathbf{q}(\theta) \in I(X)] \geq 1-\alpha.$$

Note that if $I_j(X) = [\underline{T}_j, \bar{T}_j]$ is a level $(1-\alpha_j)$ confidence interval for $q_j(\theta)$ and if the pairs $(\underline{T}_1, \bar{T}_1), \ldots, (\underline{T}_r, \bar{T}_r)$ are independent, then the rectangle $I(X) = I_1(X) \times \cdots \times I_r(X)$ has level

$$\prod_{j=1}^{r}(1-\alpha_j). \tag{4.4.8}$$

Thus, an $r$-dimensional confidence rectangle is in this case automatically obtained from the one-dimensional intervals. Moreover, if we choose $\alpha_j = 1 - (1-\alpha)^{\frac{1}{r}}$, then $I(X)$ has confidence level $1-\alpha$.

An approach that works even if the $I_j$ are not independent is to use Bonferroni's inequality (A.2.7). According to this inequality,

$$P[\mathbf{q}(\theta) \in I(X)] \geq 1 - \sum_{j=1}^{r} P[q_j(\theta) \notin I_j(X)] \geq 1 - \sum_{j=1}^{r} \alpha_j.$$

Thus, if we choose $\alpha_j = \alpha/r$, $j = 1, \ldots, r$, then $I(X)$ has confidence level $(1-\alpha)$.

**Example 4.4.5.** *Confidence Rectangle for the Parameters of a Normal Distribution.* Suppose $X_1, \ldots, X_n$ is a $\mathcal{N}(\mu, \sigma^2)$ sample and we want a confidence rectangle for $(\mu, \sigma^2)$. From Example 4.4.1

$$I_1(X) = \bar{X} \pm s t_{n-1}\left(1 - \tfrac{1}{4}\alpha\right)/\sqrt{n}$$

is a confidence interval for $\mu$ with confidence coefficient $\left(1 - \frac{1}{2}\alpha\right)$. From Example 4.4.2,

$$I_2(X) = \left[\frac{(n-1)s^2}{x_{n-1}\left(1-\frac{1}{4}\alpha\right)}, \frac{(n-1)s^2}{x_{n-1}\left(\frac{1}{4}\alpha\right)}\right]$$

is a reasonable confidence interval for $\sigma^2$ with confidence coefficient $\left(1 - \frac{1}{2}\alpha\right)$. Thus, $I_1(X) \times I_2(X)$ is a level $(1-\alpha)$ confidence rectangle for $(\mu, \sigma^2)$. The exact confidence coefficient is given in Problem 4.4.15. □

The method of pivots can also be applied to $\infty$-dimensional parameters such as $F$.

**Example 4.4.6.** Suppose $X_1, \ldots, X_n$ are i.i.d. as $X \sim P$, and we are interested in the distribution function $F(t) = P(X \leq t)$; that is, $\nu(P) = F(\cdot)$. We assume that $F$ is continuous, in which case (Proposition 4.1.1) the distribution of

$$D_n(F) = \sup_{t \in R} |\widehat{F}(t) - F(t)|$$

does not depend on $F$ and is known (Example 4.1.5). That is, $D_n(F)$ is a pivot. Let $d_\alpha$ be chosen such that $P_F(D_n(F) \leq d_\alpha) = 1 - \alpha$. Then by solving $D_n(F) \leq d_\alpha$ for $F$, we find that a simultaneous in $t$ size $1-\alpha$ confidence region $C(\mathbf{x})(\cdot)$ is the confidence band which, for each $t \in R$, consists of the interval

$$C(\mathbf{x})(t) = (\max\{0, \widehat{F}(t) - d_\alpha\}, \ \min\{1, \widehat{F}(t) + d_\alpha\}).$$

We have shown

$$P(C(\mathbf{X})(t) \supset F(t) \text{ for all } t \in R) = 1 - \alpha$$

for all $P \in \mathcal{P} =$ set of $P$ with $P(-\infty, t]$ continuous in $t$. □

We can apply the notions studied in Examples 4.4.4 and 4.4.5 to give confidence regions for scalar or vector parameters in nonparametric models.

**Example 4.4.7.** *A Lower Confidence Bound for the Mean of a Nonnegative Random Variable.* Suppose $X_1, \dots, X_n$ are i.i.d. as $X$ and that $X$ has a density $f(t) = F'(t)$, which is zero for $t < 0$ and nonzero for $t > 0$. By integration by parts, if $\mu = \mu(F) = \int_0^\infty tf(t)dt$ exists, then

$$\mu = \int_0^\infty [1 - F(t)]dt.$$

Let $\widehat{F}^-(t)$ and $\widehat{F}^+(t)$ be the lower and upper simultaneous confidence boundaries of Example 4.4.6. Then a $(1-\alpha)$ lower confidence bound for $\mu$ is $\underline{\mu}$ given by

$$\underline{\mu} = \int_0^\infty [1 - \widehat{F}^+(t)]dt = \sum_{i \leq n(1-d_\alpha)} \left[1 - \left(\frac{i}{n} + d_\alpha\right)\right][x_{(i+1)} - x_{(i)}] \tag{4.4.9}$$

because for $C(\mathbf{X})$ as in Example 4.4.6, $\underline{\mu} = \inf\{\mu(F) : F \in C(\mathbf{X})\} = \mu(\widehat{F}^+)$ and $\sup\{\mu(F) : F \in C(\mathbf{X})\} = \mu(\widehat{F}^-) = \infty$—see Problem 4.4.19.

Intervals for the case of $F$ supported on an interval (see Problem 4.4.18) arise in accounting practice (see Bickel, 1992, where such bounds are discussed and shown to be asymptotically strictly conservative). □

**Summary.** We define lower and upper confidence bounds (LCBs and UCBs), confidence intervals, and more generally confidence regions. In a parametric model $\{P_\theta : \theta \in \Theta\}$, a level $1-\alpha$ confidence region for a parameter $q(\theta)$ is a set $C(x)$ depending only on the data $x$ such that the probability under $P_\theta$ that $C(X)$ covers $q(\theta)$ is at least $1-\alpha$ for all $\theta \in \Theta$. For a nonparametric class $\mathcal{P} = \{P\}$ and parameter $\nu = \nu(P)$, we similarly require $P(C(X) \supset \nu) \geq 1-\alpha$ for all $P \in \mathcal{P}$. We derive the (Student) $t$ interval for $\mu$ in the $\mathcal{N}(\mu, \sigma^2)$ model with $\sigma^2$ unknown, and we derive an exact confidence interval for the binomial parameter. In a nonparametric setting we derive a simultaneous confidence interval for the distribution function $F(t)$ and the mean of a positive variable $X$.

## 4.5 THE DUALITY BETWEEN CONFIDENCE REGIONS AND TESTS

Confidence regions are random subsets of the parameter space that contain the true parameter with probability at least $1-\alpha$. Acceptance regions of statistical tests are, for a given hypothesis $H$, subsets of the sample space with probability of accepting $H$ at least $1-\alpha$ when $H$ is true. We shall establish a duality between confidence regions and acceptance regions for families of hypotheses.

We begin by illustrating the duality in the following example.

**Example 4.5.1.** *Two-Sided Tests for the Mean of a Normal Distribution.* Suppose that an established theory postulates the value $\mu_0$ for a certain physical constant. A scientist has reasons to believe that the theory is incorrect and measures the constant $n$ times obtaining measurements $X_1, \dots, X_n$. Knowledge of his instruments leads him to assume that the $X_i$ are independent and identically distributed normal random variables with mean $\mu$ and variance $\sigma^2$. If any value of $\mu$ other than $\mu_0$ is a possible alternative, then it is reasonable to formulate the problem as that of testing $H : \mu = \mu_0$ versus $K : \mu \neq \mu_0$.

We can base a size $\alpha$ test on the level $(1-\alpha)$ confidence interval $(4.4.1)$ we constructed for $\mu$ as follows. We accept $H$, if and only if, the postulated value $\mu_0$ is a member of the level $(1-\alpha)$ confidence interval

$$[\bar{X} - st_{n-1}\left(1 - \tfrac{1}{2}\alpha\right)/\sqrt{n},\ \bar{X} + st_{n-1}\left(1 - \tfrac{1}{2}\alpha\right)/\sqrt{n}]. \tag{4.5.1}$$

If we let $T = \sqrt{n}(\bar{X} - \mu_0)/s$, then our test accepts $H$, if and only if, $-t_{n-1}\left(1 - \frac{1}{2}\alpha\right) \leq T \leq t_{n-1}\left(1 - \frac{1}{2}\alpha\right)$. Because $P_\mu[|T| = t_{n-1}\left(1 - \frac{1}{2}\alpha\right)] = 0$ the test is equivalently characterized by rejecting $H$ when $|T| \geq t_{n-1}\left(1 - \frac{1}{2}\alpha\right)$. This test is called *two-sided* because it rejects for both large and small values of the statistic $T$. In contrast to the tests of Example 4.1.4, it has power against parameter values on either side of $\mu_0$.

Because the same interval $(4.5.1)$ is used for every $\mu_0$ we see that we have, in fact, generated a family of level $\alpha$ tests $\{\delta(\mathbf{X}, \mu)\}$ where

$$\begin{aligned} \delta(\mathbf{X}, \mu) &= 1 \text{ if } \sqrt{n}\tfrac{|\bar{X}-\mu|}{s} \geq t_{n-1}\left(1 - \tfrac{1}{2}\alpha\right) \\ &= 0 \text{ otherwise.} \end{aligned} \tag{4.5.2}$$

These tests correspond to different hypotheses, $\delta(\mathbf{X}, \mu_0)$ being of size $\alpha$ only for the hypothesis $H : \mu = \mu_0$.

Conversely, by starting with the test $(4.5.2)$ we obtain the confidence interval $(4.5.1)$ by finding the set of $\mu$ where $\delta(\mathbf{X}, \mu) = 0$.

We achieve a similar effect, generating a family of level $\alpha$ tests, if we start out with (say) the level $(1-\alpha)$ LCB $\bar{X} - t_{n-1}(1-\alpha)s/\sqrt{n}$ and define $\delta^*(\mathbf{X}, \mu)$ to equal 1 if, and only if, $\bar{X} - t_{n-1}(1-\alpha)s/\sqrt{n} \geq \mu$. Evidently,

$$P_{\mu_0}[\delta^*(\mathbf{X}, \mu_0) = 1] = P_{\mu_0}\left[\sqrt{n}\frac{(\bar{X} - \mu_0)}{s} \geq t_{n-1}(1-\alpha)\right] = \alpha.$$

□

These are examples of a general phenomenon. Consider the general framework where the random vector $X$ takes values in the sample space $\mathcal{X} \subset R^q$ and $X$ has distribution $P \in \mathcal{P}$. Let $\nu = \nu(P)$ be a parameter that takes values in the set $\mathcal{N}$. For instance, in Example 4.4.1, $\mu = \mu(P)$ takes values in $\mathcal{N} = (-\infty, \infty)$, in Example 4.4.2, $\sigma^2 = \sigma^2(P)$ takes values in $\mathcal{N} = (0, \infty)$, and in Example 4.4.5, $(\mu, \sigma^2)$ takes values in $\mathcal{N} = (-\infty, \infty) \times (0, \infty)$. For a function space example, consider $\nu(P) = F$, as in Example 4.4.6, where $F$ is the distribution function of $X_i$. Here an example of $\mathcal{N}$ is the class of all continuous distribution functions. Let $S = S(X)$ be a map from $\mathcal{X}$ to subsets of $\mathcal{N}$, then $S$ is a $(1-\alpha)$ *confidence region* for $\nu$ if the probability that $S(X)$ contains $\nu$ is at least $(1-\alpha)$, that is

$$P[\nu \in S(X)] \geq 1 - \alpha, \text{ all } P \in \mathcal{P}.$$

Next consider the testing framework where we test the hypothesis $H = H_{\nu_0} : \nu = \nu_0$ for some specified value $\nu_0$. Suppose we have a test $\delta(X, \nu_0)$ with level $\alpha$. Then the acceptance region

$$A(\nu_0) = \{x : \delta(x, \nu_0) = 0\}$$

is a subset of $\mathcal{X}$ with probability at least $1 - \alpha$. For some specified $\nu_0$, $H$ may be accepted, for other specified $\nu_0$, $H$ may be rejected. Consider the set of $\nu_0$ for which $H_{\nu_0}$ is accepted; this is a random set contained in $\mathcal{N}$ with probability at least $1 - \alpha$ of containing the true value of $\nu(P)$ whatever be $P$. Conversely, if $S(X)$ is a level $1 - \alpha$ confidence region for $\nu$, then the test that accepts $H_{\nu_0}$ if and only if $\nu_0$ is in $S(\mathbf{X})$, is a level $\alpha$ test for $H_{\nu_0}$.

Formally, let $\mathcal{P}_{\nu_0} = \{P : \nu(P) = \nu_0 : \nu_0 \in \mathcal{N}\}$. We have the following.

**Duality Theorem.** *Let $S(X) = \{\nu_0 \in \mathcal{N} : X \in A(\nu_0)\}$, then*

$$P[X \in A(\nu_0)] \geq 1 - \alpha \text{ for all } P \in \mathcal{P}_{\nu_0}$$

*if and only if $S(X)$ is a $1 - \alpha$ confidence region for $\nu$.*

We next apply the duality theorem to MLR families:

**Theorem 4.5.1.** *Suppose $X \sim P_\theta$ where $\{P_\theta : \theta \in \Theta\}$ is MLR in $T = T(X)$ and suppose that the distribution function $F_\theta(t)$ of $T$ under $P_\theta$ is continuous in each of the variables $t$ and $\theta$ when the other is fixed. If the equation $F_\theta(t) = 1 - \alpha$ has a solution $\underline{\theta}_\alpha(t)$ in $\Theta$, then $\underline{\theta}_\alpha(T)$ is a lower confidence bound for $\theta$ with confidence coefficient $1 - \alpha$. Similarly, any solution $\bar{\theta}_\alpha(T)$ of $F_\theta(T) = \alpha$ with $\bar{\theta}_\alpha \in \Theta$ is an upper confidence bound for $\theta$ with coefficient $(1 - \alpha)$. Moreover, if $\alpha_1 + \alpha_2 < 1$, then $[\underline{\theta}_{\alpha_1}, \bar{\theta}_{\alpha_2}]$ is confidence interval for $\theta$ with confidence coefficient $1 - (\alpha_1 + \alpha_2)$.*

***Proof.*** By Corollary 4.3.1, the acceptance region of the UMP size $\alpha$ test of $H : \theta = \theta_0$ versus $K : \theta > \theta_0$ can be written

$$A(\theta_0) = \{x : T(x) \leq t_{\theta_0}(1 - \alpha)\}$$

where $t_{\theta_0}(1 - \alpha)$ is the $1 - \alpha$ quantile of $F_{\theta_0}$. By the duality theorem, if

$$S(t) = \{\theta \in \Theta : t \leq t_\theta(1 - \alpha)\},$$

then $S(T)$ is a $1 - \alpha$ confidence region for $\theta$. By applying $F_\theta$ to both sides of $t \leq t_\theta(1 - \alpha)$, we find

$$S(t) = \{\theta \in \Theta : F_\theta(t) \leq 1 - \alpha\}.$$

By Theorem 4.3.1, the power function $P_\theta(T \geq t) = 1 - F_\theta(t)$ for a test with critical constant $t$ is increasing in $\theta$. That is, $F_\theta(t)$ is decreasing in $\theta$. It follows that $F_\theta(t) \leq 1 - \alpha$ iff $\theta \geq \underline{\theta}_\alpha(t)$ and $S(t) = [\underline{\theta}_\alpha, \infty)$. The proofs for the upper confidence bound and interval follow by the same type of argument. □

We next give connections between confidence bounds, acceptance regions, and $p$-values for MLR families: Let $t$ denote the observed value $t = T(x)$ of $T(X)$ for the datum $x$, let

$\alpha(t, \theta_0)$ denote the $p$-value for the UMP size $\alpha$ test of $H : \theta = \theta_0$ versus $K : \theta > \theta_0$, and let

$$A^*(\theta) = T(A(\theta)) = \{T(x) : x \in A(\theta)\}.$$

**Corollary 4.5.1.** *Under the conditions of Theorem* 4.4.1*,*

$$\begin{aligned} A^*(\theta) &= \{t : \alpha(t, \theta) \geq \alpha\} = (-\infty, t_\theta(1-\alpha)] \\ S(t) &= \{\theta : \alpha(t, \theta) \geq \alpha\} = [\underline{\theta}_\alpha(t), \infty). \end{aligned}$$

***Proof.*** The $p$-value is

$$\alpha(t, \theta) = P_\theta(T \geq t) = 1 - F_\theta(t).$$

We have seen in the proof of Theorem 4.3.1 that $1 - F_\theta(t)$ is increasing in $\theta$. Because $F_\theta(t)$ is a distribution function, $1 - F_\theta(t)$ is decreasing in $t$. The result follows. □

In general, let $\alpha(t, \nu_0)$ denote the $p$-value of a level $\alpha$ test $\delta(T, \nu_0) = 1[T \geq c]$ of $H : \nu = \nu_0$ based on a statistic $T = T(X)$ with observed value $t = T(x)$. Then the set

$$C = \{(t, \nu) : \alpha(t, \nu) \geq \alpha\} = \{(t, \nu) : \delta(t, \nu) = 0\}$$

gives the pairs $(t, \nu)$ where, for the given $t$, $\nu$ will be accepted; and for the given $\nu$, $t$ is in the acceptance region. We call $C$ the set of *compatible* $(t, \theta)$ points. In the $(t, \theta)$ plane, vertical sections of $C$ are the confidence regions $S(t)$ whereas horizontal sections are the acceptance regions $A^*(\nu) = \{t : \delta(t, \nu) = 0\}$. We illustrate these ideas using the example of testing $H : \mu = \mu_0$ when $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with $\sigma^2$ known. Let $T = \bar{X}$, then

$$C = \left\{(t, \mu) : |t - \mu| \leq \sigma z\left(1 - \tfrac{1}{2}\alpha\right) / \sqrt{n}\right\}.$$

Figure 4.5.1 shows the set $C$, a confidence region $S(t_0)$, and an acceptance set $A^*(\mu_0)$ for this example.

**Example 4.5.2.** *Exact Confidence Bounds and Intervals for the Probability of Success in $n$ Bernoulli Trials.* Let $X_1, \ldots, X_n$ be the indicators of $n$ Bernoulli trials with probability of success $\theta$. For $\alpha \in (0, 1)$, we seek reasonable *exact* level $(1 - \alpha)$ upper and lower confidence bounds and confidence intervals for $\theta$. To find a lower confidence bound for $\theta$ our preceding discussion leads us to consider level $\alpha$ tests for $H : \theta \leq \theta_0$, $\theta_0 \in (0, 1)$. We shall use some of the results derived in Example 4.1.3. Let $k(\theta_0, \alpha)$ denote the critical constant of a level $\alpha$ test of $H$. The corresponding level $(1 - \alpha)$ confidence region is given by

$$C(X_1, \ldots, X_n) = \{\theta : S \leq k(\theta, \alpha) - 1\},$$

where $S = \Sigma_{i=1}^n X_i$.

To analyze the structure of the region we need to examine $k(\theta, \alpha)$. We claim that

(i) $k(\theta, \alpha)$ is nondecreasing in $\theta$.

(ii) $k(\theta, \alpha) \to k(\theta_0, \alpha)$ if $\theta \uparrow \theta_0$.

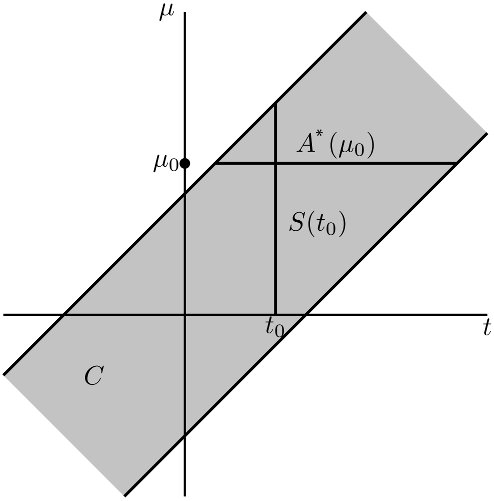

**Figure 4.5.1.** The shaded region is the compatibility set $C$ for the two-sided test of $H_{\mu_0} : \mu = \mu_0$ in the normal model. $S(t_0)$ is a confidence interval for $\mu$ for a given value $t_0$ of $T$, whereas $A^*(\mu_0)$ is the acceptance region for $H_{\mu_0}$.

(iii) $k(\theta, \alpha)$ increases by exactly 1 at its points of discontinuity.

(iv) $k(0, \alpha) = 1$ and $k(1, \alpha) = n + 1$.

To prove (i) note that it was shown in Theorem 4.3.1(1) that $P_\theta[S \geq j]$ is nondecreasing in $\theta$ for fixed $j$. Clearly, it is also nonincreasing in $j$ for fixed $\theta$. Therefore, $\theta_1 < \theta_2$ and $k(\theta_1, \alpha) > k(\theta_2, \alpha)$ would imply that

$$\alpha \geq P_{\theta_2}[S \geq k(\theta_2, \alpha)] \geq P_{\theta_2}[S \geq k(\theta_2, \alpha) - 1] \geq P_{\theta_1}[S \geq k(\theta_1, \alpha) - 1] > \alpha,$$

a contradiction.

The assertion (ii) is a consequence of the following remarks. If $\theta_0$ is a discontinuity point of $k(\theta, \alpha)$, let $j$ be the limit of $k(\theta, \alpha)$ as $\theta \uparrow \theta_0$. Then $P_\theta[S \geq j] \leq \alpha$ for all $\theta < \theta_0$ and, hence, $P_{\theta_0}[S \geq j] \leq \alpha$. On the other hand, if $\theta > \theta_0$, $P_\theta[S \geq j] > \alpha$. Therefore, $P_{\theta_0}[S \geq j] = \alpha$ and $j = k(\theta_0, \alpha)$. The claims (iii) and (iv) are left as exercises.

From (i), (ii), (iii), and (iv) we see that, if we define

$$\underline{\theta}(S) = \inf\{\theta : k(\theta, \alpha) = S + 1\},$$

then

$$C(\mathbf{X}) = \begin{cases} (\underline{\theta}(S), 1] & \text{if } S > 0 \\ [0, 1] & \text{if } S = 0 \end{cases}$$

and $\underline{\theta}(S)$ is the desired level $(1-\alpha)$ LCB for $\theta$.[(1)] Figure 4.5.2 portrays the situation. From our discussion, when $S > 0$, then $k(\underline{\theta}(S), \alpha) = S$ and, therefore, we find $\underline{\theta}(S)$ as the unique solution of the equation,

$$\sum_{r=S}^{n} \binom{n}{r} \theta^r (1-\theta)^{n-r} = \alpha.$$

When $S = 0$, $\underline{\theta}(S) = 0$.

Similarly, we define

$$\bar{\theta}(S) = \sup\{\theta : j(\theta, \alpha) = S - 1\}$$

where $j(\theta, \alpha)$ is given by,

$$\sum_{r=0}^{j(\theta,\alpha)} \binom{n}{r} \theta^r (1-\theta)^{n-r} \le \alpha < \sum_{r=0}^{j(\theta,\alpha)+1} \binom{n}{r} \theta^r (1-\theta)^{n-r}.$$

Then $\bar{\theta}(S)$ is a level $(1-\alpha)$ UCB for $\theta$ and when $S < n$, $\bar{\theta}(S)$ is the unique solution of

$$\sum_{r=0}^{S} \binom{n}{r} \theta^r (1-\theta)^{n-r} = \alpha.$$

When $S = n$, $\bar{\theta}(S) = 1$. Putting the bounds $\underline{\theta}(S)$, $\bar{\theta}(S)$ together we get the confidence interval $[\underline{\theta}(S), \bar{\theta}(S)]$ of level $(1-2\alpha)$. These intervals can be obtained from computer packages that use algorithms based on the preceding considerations. As might be expected, if $n$ is large, these bounds and intervals differ little from those obtained by the first approximate method in Example 4.4.3.

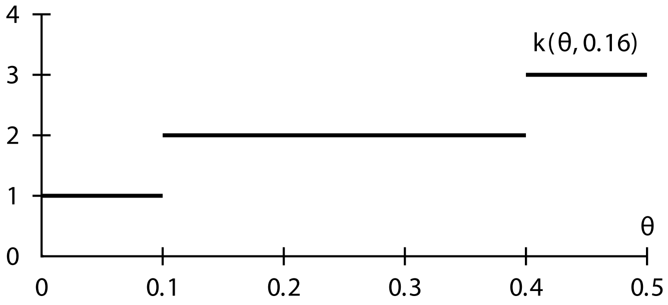

**Figure 4.5.2.** Plot of $k(\theta, 0.16)$ for $n = 2$.

## Applications of Confidence Intervals to Comparisons and Selections

We have seen that confidence intervals lead naturally to two-sided tests. However, two-sided tests seem incomplete in the sense that if $H : \theta = \theta_0$ is rejected in favor of $H : \theta \neq \theta_0$, we usually want to know whether $H : \theta > \theta_0$ or $H : \theta < \theta_0$.

For instance, suppose $\theta$ is the expected difference in blood pressure when two treatments, A and B, are given to high blood pressure patients. Because we do not know whether A or B is to be preferred, we test $H : \theta = 0$ versus $K : \theta \neq 0$. If $H$ is rejected, it is natural to carry the comparison of A and B further by asking whether $\theta < 0$ or $\theta > 0$. If we decide $\theta < 0$, then we select A as the better treatment, and vice versa.

The problem of deciding whether $\theta = \theta_0, \theta < \theta_0$, or $\theta > \theta_0$ is an example of a three-decision problem and is a special case of the decision problems in Section 1.4, and 3.1–3.3. Here we consider the simple solution suggested by the level $(1-\alpha)$ confidence interval $I$:

1. Make no judgment as to whether $\theta < \theta_0$ or $\theta > \theta_0$ if $I$ contains $\theta_0$;
2. Decide $\theta < \theta_0$ if $I$ is entirely to the left of $\theta_0$; and (4.5.3)
3. Decide $\theta > \theta_0$ if $I$ is entirely to the right of $\theta_0$.

**Example 4.5.2.** Suppose $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with $\sigma^2$ known. In Section 4.4 we considered the level $(1-\alpha)$ confidence interval $\bar{X} \pm \sigma z(1-\frac{1}{2}\alpha)/\sqrt{n}$ for $\mu$. Using this interval and (4.5.3) we obtain the following three decision rule based on $T = \sqrt{n}(\bar{X} - \mu_0)/\sigma$:

Do not reject $H : \mu = \mu_0$ if $|T| \leq z(1-\frac{1}{2}\alpha)$.
Decide $\mu < \mu_0$ if $T < -z(1-\frac{1}{2}\alpha)$.
Decide $\mu > \mu_0$ if $T > z(1-\frac{1}{2}\alpha)$.

Thus, the two-sided test can be regarded as the first step in the decision procedure where if $H$ is not rejected, we make no claims of significance, but if $H$ is rejected, we decide whether this is because $\mu$ is smaller or larger than $\mu_0$. For this three-decision rule, the probability of falsely claiming significance of either $\mu < \mu_0$ or $\mu > \mu_0$ is bounded above by $\frac{1}{2}\alpha$. To see this consider first the case $\mu \geq \mu_0$. Then the wrong decision "$\mu < \mu_0$" is made when $T < -z(1-\frac{1}{2}\alpha)$. This event has probability

$$P[T < -z(1-\tfrac{1}{2}\alpha)] = \Phi\left(-z(1-\tfrac{1}{2}\alpha) - \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\right) \leq \Phi(-z(1-\tfrac{1}{2}\alpha)) = \tfrac{1}{2}\alpha.$$

Similarly, when $\mu \leq \mu_0$, the probability of the wrong decision is at most $\frac{1}{2}\alpha$. Therefore, by using this kind of procedure in a comparison or selection problem, we can control the probabilities of a wrong selection by setting the $\alpha$ of the parent test or confidence interval. We can use the two-sided tests and confidence intervals introduced in later chapters in similar fashions.

**Summary.** We explore the connection between tests of statistical hypotheses and confidence regions. If $\delta(x, \nu_0)$ is a level $\alpha$ test of $H : \nu = \nu_0$, then the set $S(x)$ of $\nu_0$ where

$\delta(x, \nu_0) = 0$ is a level $(1-\alpha)$ confidence region for $\nu_0$. If $S(x)$ is a level $(1-\alpha)$ confidence region for $\nu$, then the test that accepts $H : \nu = \nu_0$ when $\nu_0 \in S(x)$ is a level $\alpha$ test. We give explicitly the construction of exact upper and lower confidence bounds and intervals for the parameter in the binomial distribution. We also give a connection between confidence intervals, two-sided tests, and the three-decision problem of deciding whether a parameter $\theta$ is $\theta_0$, less than $\theta_0$, or larger than $\theta_0$, where $\theta_0$ is a specified value.

## 4.6 UNIFORMLY MOST ACCURATE CONFIDENCE BOUNDS

In our discussion of confidence bounds and intervals so far we have not taken their accuracy into account. We next show that for a certain notion of accuracy of confidence bounds, which is connected to the power of the associated one-sided tests, optimality of the tests translates into accuracy of the bounds.

If $\underline{\theta}$ and $\underline{\theta}^*$ are two competing level $(1-\alpha)$ lower confidence bounds for $\theta$, they are both very likely to fall below the true $\theta$. But we also want the bounds to be close to $\theta$. Thus, we say that the bound with the smaller probability of being far below $\theta$ is more accurate. Formally, for $X \in \mathcal{X} \subset R^q$, the following is true.

**Definition 4.6.1.** A level $(1-\alpha)$ LCB $\underline{\theta}^*$ of $\theta$ is said to be more *accurate* than a competing level $(1-\alpha)$ LCB $\underline{\theta}$ if, and only if, for any fixed $\theta$ and all $\theta' < \theta$,

$$P_\theta[\underline{\theta}^*(X) \le \theta'] \le P_\theta[\underline{\theta}(X) \le \theta']. \tag{4.6.1}$$

Similarly, a level $(1-\alpha)$ UCB $\overline{\theta}^*$ is more accurate than a competitor $\overline{\theta}$ if, and only if, for any fixed $\theta$ and all $\theta' > \theta$,

$$P_\theta[\overline{\theta}^*(X) \ge \theta'] \le P_\theta[\overline{\theta}(X) \ge \theta']. \tag{4.6.2}$$

Lower confidence bounds $\underline{\theta}^*$ satisfying $(4.6.1)$ for *all* competitors are called *uniformly most accurate* as are upper confidence bounds satisfying $(4.6.2)$ for all competitors. Note that $\underline{\theta}^*$ is a uniformly most accurate level $(1-\alpha)$ LCB for $\theta$, if and only if $-\underline{\theta}^*$ is a uniformly most accurate level $(1-\alpha)$ UCB for $-\theta$.

**Example 4.6.1 (Examples 3.3.2 and 4.2.1 continued).** Suppose $\mathbf{X} = (X_1, \dots, X_n)$ is a sample of a $\mathcal{N}(\mu, \sigma^2)$ random variables with $\sigma^2$ known. A level $\alpha$ test of $H : \mu = \mu_0$ vs $K : \mu > \mu_0$ rejects $H$ when $\sqrt{n}(\bar{X} - \mu_0)/\sigma \ge z(1-\alpha)$. The dual lower confidence bound is $\underline{\mu}_1(\boldsymbol{X}) = \bar{X} - z(1-\alpha)\sigma/\sqrt{n}$. Using Problem 4.5.6, we find that a competing lower confidence bound is $\underline{\mu}_2(\boldsymbol{X}) = X_{(k)}$, where $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$ denotes the ordered $X_1, \dots, X_n$ and $k$ is defined to be the largest integer such that $P(S \ge k) \ge 1 - \alpha$ for a binomial, $B(n, \frac{1}{2})$, random variable $S$. Which lower bound is more accurate? It does turn out that $\underline{\mu}_1(\mathbf{X})$ is more accurate than $\underline{\mu}_2(\mathbf{X})$ and is, in fact, uniformly most accurate in the $\mathcal{N}(\mu, \sigma^2)$ model. This is a consequence of the following theorem, which reveals that $(4.6.1)$ is nothing more than a comparison of power functions. □

**Theorem 4.6.1.** *Let $\underline{\theta}^*$ be a level $(1-\alpha)$ LCB for $\theta$, a real parameter, such that for each $\theta_0$ the associated test whose critical function $\delta^*(x, \theta_0)$ is given by*

$$\begin{aligned} \delta^*(x,\theta_0) &= 1 \textit{ if } \underline{\theta}^*(x) > \theta_0 \\ &= 0 \textit{ otherwise} \end{aligned}$$

*is UMP level $\alpha$ for $H : \theta = \theta_0$ versus $K : \theta > \theta_0$. Then $\underline{\theta}^*$ is uniformly most accurate at level $(1-\alpha)$.*

***Proof.*** Let $\underline{\theta}$ be a competing level $(1-\alpha)$ LCB $\theta_0$. Defined $\delta(x, \theta_0)$ by

$$\delta(x,\theta_0) = 0 \text{ if, and only if, } \underline{\theta}(x) \le \theta_0.$$

Then $\delta(X, \theta_0)$ is a level $\alpha$ test for $H : \theta = \theta_0$ versus $K : \theta > \theta_0$. Because $\delta^*(X, \theta_0)$ is UMP level $\alpha$ for $H : \theta = \theta_0$ versus $K : \theta > \theta_0$, for $\theta_1 > \theta_0$ we must have

$$E_{\theta_1}(\delta(X,\theta_0)) \le E_{\theta_1}(\delta^*(X,\theta_0))$$

or

$$P_{\theta_1}[\underline{\theta}(X) > \theta_0] \le P_{\theta_1}[\underline{\theta}^*(X) > \theta_0].$$

Identify $\theta_0$ with $\theta'$ and $\theta_1$ with $\theta$ in the statement of Definition 4.4.2 and the result follows. □

If we apply the result and Example 4.2.1 to Example 4.6.1, we find that $\bar{x} - z(1-\alpha)\sigma/\sqrt{n}$ is uniformly most accurate. However, $X_{(k)}$ does have the advantage that we don't have to know $\sigma$ or even the shape of the density $f$ of $X_i$ to apply it. Also, the robustness considerations of Section 3.5 favor $X_{(k)}$ (see Example 3.5.2).

Uniformly most accurate (UMA) bounds turn out to have related nice properties. For instance (see Problem 4.6.7 for the proof), they have the smallest expected "distance" to $\theta$:

**Corollary 4.6.1.** *Suppose $\underline{\theta}^*(\mathbf{X})$ is a UMA level $(1-\alpha)$ lower confidence bound for $\theta$. Let $\underline{\theta}(\mathbf{X})$ be any other $(1-\alpha)$ lower confidence bound, then*

$$E_\theta\{(\theta - \underline{\theta}^*(\mathbf{X}))^+\} \le E_\theta\{(\theta - \underline{\theta}(\mathbf{X}))^+\}$$

*for all $\theta$ where $a^+ = a$, if $a \ge 0$, and $0$ otherwise.*

We can extend the notion of accuracy to confidence bounds for real-valued functions of an arbitrary parameter. We define $\underline{q}^*$ to be a uniformly most accurate level $(1-\alpha)$ LCB for $q(\theta)$ if, and only if, for any other level $(1-\alpha)$ LCB $\underline{q}$,

$$P_\theta[\underline{q}^* \le q(\theta')] \le P_\theta[\underline{q} \le q(\theta')]$$

whenever $q(\theta') < q(\theta)$. Most accurate upper confidence bounds are defined similarly.

**Example 4.6.2.** *Bounds for the Probability of Early Failure of Equipment.* Let $X_1, \ldots, X_n$ be the times to failure of $n$ pieces of equipment where we assume that the $X_i$ are independent $\mathcal{E}(\lambda)$ variables. We want a uniformly most accurate level $(1-\alpha)$ upper confidence bound $\bar{q}^*$ for $q(\lambda) = 1 - e^{-\lambda t_0}$, the probability of early failure of a piece of equipment.

We begin by finding a uniformly most accurate level $(1-\alpha)$ UCB $\bar{\lambda}^*$ for $\lambda$. To find $\bar{\lambda}^*$ we invert the family of UMP level $\alpha$ tests of $H : \lambda \geq \lambda_0$ versus $K : \lambda < \lambda_0$. By Problem 4.6.8, the UMP test accepts $H$ if

$$\sum_{i=1}^{n} X_i < \chi_{2n}(1-\alpha)/2\lambda_0 \qquad (4.6.3)$$

or equivalently if

$$\lambda_0 < \frac{\chi_{2n}(1-\alpha)}{2\sum_{i=1}^{n} X_i}$$

where $\chi_{2n}(1-\alpha)$ is the $(1-\alpha)$ quantile of the $\chi^2_{2n}$ distribution. Therefore, the confidence region corresponding to this test is $(0, \bar{\lambda}^*)$ where $\bar{\lambda}^*$ is by Theorem 4.6.1, a uniformly most accurate level $(1-\alpha)$ UCB for $\lambda$ and, because $q$ is strictly increasing in $\lambda$, it follows that $q(\bar{\lambda}^*)$ is a uniformly most accurate level $(1-\alpha)$ UCB for the probability of early failure. □

## Discussion

We have only considered confidence bounds. The situation with confidence intervals is more complicated. Considerations of accuracy lead us to ask that, subject to the requirement that the confidence level is $(1-\alpha)$, the confidence interval be as short as possible. Of course, the length $\bar{T} - \underline{T}$ is random and it can be shown that in most situations there is no confidence interval of level $(1-\alpha)$ that has uniformly minimum length among all such intervals. There are, however, some large sample results in this direction (see Wilks, 1962, pp. 374–376). If we turn to the expected length $E_\theta(\bar{T} - \underline{T})$ as a measure of precision, the situation is still unsatisfactory because, in general, there does not exist a member of the class of level $(1-\alpha)$ intervals that has minimum expected length for all $\theta$. However, as in the estimation problem, we can restrict attention to certain reasonable subclasses of level $(1-\alpha)$ intervals for which members with uniformly smallest expected length exist. Thus, Neyman defines *unbiased* confidence intervals of level $(1-\alpha)$ by the property that

$$P_\theta[\underline{T} \leq q(\theta) \leq \bar{T}] \geq P_\theta[\underline{T} \leq q(\theta') \leq \bar{T}]$$

for every $\theta, \theta'$. That is, the interval must be at least as likely to cover the true value of $q(\theta)$ as any other value. Pratt (1961) showed that in many of the classical problems of estimation there exist level $(1-\alpha)$ confidence intervals that have uniformly minimum expected length among all level $(1-\alpha)$ unbiased confidence intervals. In particular, the intervals developed in Example 4.5.1 have this property.

Confidence intervals obtained from two-sided tests that are uniformly most powerful within a restricted class of procedures can be shown to have optimality properties within restricted classes. These topics are discussed in Lehmann (1997).

**Summary.** By using the duality between one-sided tests and confidence bounds we show that confidence bounds based on UMP level $\alpha$ tests are uniformly most accurate (UMA) level $(1-\alpha)$ in the sense that, in the case of lower bounds, they are less likely than other level $(1-\alpha)$ lower confidence bounds to fall below any value $\theta'$ below the true $\theta$.

## 4.7 FREQUENTIST AND BAYESIAN FORMULATIONS

We have so far focused on the frequentist formulation of confidence bounds and intervals where the data $X \in \mathcal{X} \subset R^q$ are random while the parameters are fixed but unknown. A consequence of this approach is that once a numerical interval has been computed from experimental data, no probability statement can be attached to this interval. Instead, the interpretation of a $100(1-\alpha)\%$ confidence interval is that if we repeated an experiment indefinitely each time computing a $100(1-\alpha)\%$ confidence interval, then $100(1-\alpha)\%$ of the intervals would contain the true unknown parameter value.

In the Bayesian formulation of Sections 1.2 and 1.6.3, what are called level $(1-\alpha)$ credible bounds and intervals are subsets of the parameter space which are given probability at least $(1-\alpha)$ by the posterior distribution of the parameter given the data. Suppose that, given $\theta$, $X$ has distribution $P_\theta, \theta \in \Theta \subset R$, and that $\theta$ has the prior probability distribution $\Pi$.

**Definition 4.7.1.** Let $\Pi(\cdot|x)$ denote the posterior probability distribution of $\theta$ given $X = x$, then $\underline{\theta}$ and $\bar{\theta}$ are level $(1-\alpha)$ *lower* and *upper credible* bounds for $\theta$ if they respectively satisfy

$$\Pi(\underline{\theta} \leq \theta|x) \geq 1-\alpha, \quad \Pi(\theta \leq \bar{\theta}|x) \geq 1-\alpha.$$

Turning to Bayesian credible intervals and regions, it is natural to consider the collection of $\theta$ that is "most likely" under the distribution $\Pi(\theta|x)$. Thus,

**Definition 4.7.2.** Let $\pi(\cdot|x)$ denote the density of $\theta$ given $X = x$, then

$$C_k = \{\theta : \pi(\theta|x) \geq k\}$$

is called a *level* $(1-\alpha)$ *credible region* for $\theta$ if $\Pi(C_k|x) \geq 1-\alpha$ .

If $\pi(\theta|x)$ is unimodal, then $C_k$ will be an interval of the form $[\underline{\theta}, \bar{\theta}]$. We next give such an example.

**Example 4.7.1.** Suppose that given $\mu$, $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$ with $\sigma_0^2$ known, and that $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$, with $\mu_0$ and $\tau_0^2$ known. Then, from Example 1.1.12, the posterior distribution of $\mu$ given $X_1, \dots, X_n$ is $\mathcal{N}(\widehat{\mu}_B, \widehat{\sigma}_B^2)$, with

$$\widehat{\mu}_B = \frac{\frac{n}{\sigma_0^2}\bar{X} + \frac{1}{\tau_0^2}\mu_0}{\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}}, \widehat{\sigma}_B^2 = \frac{1}{\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}}$$

It follows that the level $1-\alpha$ lower and upper credible bounds for $\mu$ are

$$\underline{\mu} = \widehat{\mu}_B - z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}\left(1 + \frac{\sigma_0^2}{n\tau_0^2}\right)^{\frac{1}{2}}}$$

$$\overline{\mu} = \widehat{\mu}_B + z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}\left(1 + \frac{\sigma_0^2}{n\tau_0^2}\right)^{\frac{1}{2}}}$$

while the level $(1-\alpha)$ credible interval is $[\mu^-, \mu^+]$ with

$$\mu^{\pm} = \widehat{\mu}_B \pm z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}\left(1+\frac{\sigma_0^2}{n\tau_0^2}\right)^{\frac{1}{2}}}.$$

Compared to the frequentist interval $\bar{X} \pm z_{1-\frac{\alpha}{2}}\sigma_0/\sqrt{n}$, the center $\widehat{\mu}_B$ of the Bayesian interval is pulled in the direction of $\mu_0$, where $\mu_0$ is a prior guess of the value of $\mu$. See Example 1.3.4 for sources of such prior guesses. Note that as $\tau_0 \to \infty$, the Bayesian interval tends to the frequentist interval; however, the interpretations of the intervals are different. □

**Example 4.7.2.** Suppose that given $\sigma^2$, $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu_0, \sigma^2)$ where $\mu_0$ is known. Let $\lambda = \sigma^{-2}$ and suppose $\lambda$ has the gamma $\Gamma(\frac{1}{2}a, \frac{1}{2}b)$ density

$$\pi(\lambda) \propto \lambda^{\frac{1}{2}(a-2)} \exp\left\{-\tfrac{1}{2}b\lambda\right\},\ \lambda > 0$$

where $a > 0, b > 0$ are known parameters. Then, by Problem 1.2.12, given $x_1, \dots, x_n$, $(t+b)\lambda$ has a $\chi^2_{a+n}$ distribution, where $t = \sum(x_i - \mu_0)^2$. Let $x_{a+n}(\alpha)$ denote the $\alpha$th quantile of the $\chi^2_{a+n}$ distribution, then $\underline{\lambda} = x_{a+n}(\alpha)/(t+b)$ is a level $(1-\alpha)$ lower credible bound for $\lambda$ and

$$\bar{\sigma}^2_B = (t+b)/x_{a+n}(\alpha)$$

is a level $(1-\alpha)$ upper credible bound for $\sigma^2$. Compared to the frequentist bound $t/x_n(\alpha)$, $\bar{\sigma}^2_B$ is shifted in the direction of the reciprocal $b/a$ of the mean of $\pi(\lambda)$.

We shall analyze Bayesian credible regions further in Chapter 5.

**Summary.** In the Bayesian framework we define bounds and intervals, called level $(1-\alpha)$ credible bounds and intervals, that determine subsets of the parameter space that are assigned probability at least $(1-\alpha)$ by the posterior distribution of the parameter $\boldsymbol{\theta}$ given the data $x$. In the case of a normal prior $\pi(\theta)$ and normal model $p(x \mid \theta)$, the level $(1-\alpha)$ credible interval is similar to the frequentist interval except it is pulled in the direction $\mu_0$ of the prior mean and it is a little narrower. However, the interpretations are different: In the frequentist confidence interval, the probability of coverage is computed with the data $X$ random and $\theta$ fixed, whereas in the Bayesian credible interval, the probability of coverage is computed with $X = x$ fixed and $\boldsymbol{\theta}$ random with probability distribution $\Pi(\theta \mid X = x)$.

## 4.8 PREDICTION INTERVALS

In Section 1.4 we discussed situations in which we want to predict the value of a random variable $Y$. In addition to point prediction of $Y$, it is desirable to give an interval $[\underline{Y}, \bar{Y}]$ that contains the unknown value $Y$ with prescribed probability $(1-\alpha)$. For instance, a doctor administering a treatment with delayed effect will give patients a time interval $[\underline{T}, \bar{T}]$ in which the treatment is likely to take effect. Similarly, we may want an interval for the future GPA of a student or a future value of a portfolio. We define a *level* $(1-\alpha)$ *prediction interval* as an interval $[\underline{Y}, \bar{Y}]$ based on data $X$ such that $P(\underline{Y} \le Y \le \bar{Y}) \ge 1-\alpha$. The problem of finding prediction intervals is similar to finding confidence intervals using a pivot:

**Example 4.8.1.** *The (Student) t Prediction Interval.* As in Example 4.4.1, let $X_1, \dots, X_n$ be i.i.d. as $X \sim \mathcal{N}(\mu, \sigma^2)$. We want a prediction interval for $Y = X_{n+1}$, which is assumed to be also $\mathcal{N}(\mu, \sigma^2)$ and independent of $X_1, \dots, X_n$. Let $\widehat{Y} = \widehat{Y}(\mathbf{X})$ denote a predictor based on $\mathbf{X} = (X_1, \dots, X_n)$. Then $\widehat{Y}$ and $Y$ are independent and the mean squared prediction error (MSPE) of $\widehat{Y}$ is

$$MSPE(\widehat{Y}) = E(\widehat{Y} - Y)^2 = E([\widehat{Y} - \mu] - [Y - \mu])^2 = E[\widehat{Y} - \mu]^2 + \sigma^2.$$

Note that $\widehat{Y}$ can be regarded as both a predictor of $Y$ and as an estimate of $\mu$, and when we do so, $MSPE(\widehat{Y}) = MSE(\widehat{Y}) + \sigma^2$, where MSE denotes the estimation theory mean squared error. It follows that, in this case, the optimal estimator when it exists is also the optimal predictor. In Example 3.4.8, we found that in the class of unbiased estimators, $\bar{X}$ is the optimal estimator. We define a predictor $Y^*$ to be *prediction unbiased* for $Y$ if $E(Y^* - Y) = 0$, and can conclude that in the class of prediction unbiased predictors, the optimal MSPE predictor is $\widehat{Y} = \bar{X}$.

We next use the prediction error $\widehat{Y} - Y$ to construct a pivot that can be used to give a prediction interval. Note that

$$\widehat{Y} - Y = \bar{X} - X_{n+1} \sim \mathcal{N}(0, [n^{-1} + 1]\sigma^2).$$

Moreover, $s^2 = (n-1)^{-1} \sum_1^n (X_i - \bar{X})$ is independent of $\bar{X}$ by Theorem B.3.3 and independent of $X_{n+1}$ by assumption. It follows that

$$Z_p(Y) = \frac{\widehat{Y} - Y}{\sqrt{n^{-1} + 1}\sigma}$$

has a $\mathcal{N}(0,1)$ distribution and is independent of $V = (n-1)s^2/\sigma^2$, which has a $\chi^2_{n-1}$ distribution. Thus, by the definition of the (Student) $t$ distribution in Section B.3,

$$T_p(Y) \equiv \frac{Z_p(Y)}{\sqrt{\frac{V}{n-1}}} = \frac{\widehat{Y} - Y}{\sqrt{n^{-1} + 1}s}$$

has the $t$ distribution, $\mathcal{T}_{n-1}$. By solving $-t_{n-1}\left(1 - \frac{1}{2}\alpha\right) \le T_p(Y) \le t_{n-1}\left(1 - \frac{1}{2}\alpha\right)$ for $Y$, we find the $(1-\alpha)$ prediction interval

$$Y = \bar{X} \pm \sqrt{n^{-1} + 1}st_{n-1}\left(1 - \tfrac{1}{2}\alpha\right). \tag{4.8.1}$$

Note that $T_p(Y)$ acts as a prediction interval pivot in the same way that $T(\mu)$ acts as a confidence interval pivot in Example 4.4.1. Also note that the prediction interval is much wider than the confidence interval (4.4.1). In fact, it can be shown using the methods of

Chapter 5 that the width of the confidence interval $(4.4.1)$ tends to zero in probability at the rate $n^{-\frac{1}{2}}$, whereas the width of the prediction interval tends to $2\sigma z\left(1-\frac{1}{2}\alpha\right)$. Moreover, the confidence level of $(4.4.1)$ is approximately correct for large $n$ even if the sample comes from a nonnormal distribution, whereas the level of the prediction interval $(4.8.1)$ is not $(1-\alpha)$ in the limit as $n \to \infty$ for samples from non-Gaussian distributions. □

We next give a prediction interval that is valid for samples from any population with a continuous distribution.

**Example 4.8.2.** Suppose $X_1, \dots, X_n$ are i.i.d. as $X \sim F$, where $F$ is a continuous distribution function with positive density $f$ on $(a, b)$, $-\infty \le a < b \le \infty$. Let $X_{(1)} < \cdots < X_{(n)}$ denote the order statistics of $X_1, \dots, X_n$. We want a prediction interval for $Y = X_{n+1} \sim F$, where $X_{n+1}$ is independent of the data $X_1, \dots, X_n$. Set $U_i = F(X_i)$, $i = 1, \dots, n+1$, then, by Problem B.2.12, $U_1, \dots, U_{n+1}$ are i.i.d. uniform, $\mathcal{U}(0,1)$. Let $U_{(1)} < \cdots < U_{(n)}$ be $U_1, \dots, U_n$ ordered, then

$$\begin{aligned} P(X_{(j)} \le X_{n+1} \le X_{(k)}) &= P(U_{(j)} \le U_{n+1} \le U_{(k)}) \\ &= \int P(u \le U_{n+1} \le v \mid U_{(j)} = u,\ U_{(k)} = v) dH(u, v) \\ &= \int (v-u) dH(u,v) = E(U_{(k)}) - E(U_{(j)}) \end{aligned}$$

where $H$ is the joint distribution of $U_{(j)}$ and $U_{(k)}$. By Problem B.2.9, $E(U_{(i)}) = i/(n+1)$; thus,

$$P(X_{(j)} \le X_{n+1} \le X_{(k)}) = \frac{k-j}{n+1}. \tag{4.8.2}$$

It follows that $[X_{(j)}, X_{(k)}]$ with $k = n+1-j$ is a level $\alpha = (n+1-2j)/(n+1)$ prediction interval for $X_{n+1}$. This interval is a *distribution-free* prediction interval. See Problem $4.8.5$ for a simpler proof of $(4.8.2)$.

## Bayesian Predictive Distributions

Suppose that $\boldsymbol{\theta}$ is random with $\boldsymbol{\theta} \sim \pi$ and that given $\boldsymbol{\theta} = \theta$, $X_1, \dots, X_{n+1}$ are i.i.d. $p(X \mid \theta)$. Here $X_1, \dots, X_n$ are observable and $X_{n+1}$ is to be predicted. The *posterior predictive distribution* $Q(\cdot \mid \mathbf{x})$ of $X_{n+1}$ is defined as the conditional distribution of $X_{n+1}$ given $\mathbf{x} = (x_1, \dots, x_n)$; that is, $Q(\cdot \mid \mathbf{x})$ has in the continuous case density

$$q(x_{n+1} \mid \mathbf{x}) = \int_\Theta \prod_{i=1}^{n+1} p(x_i \mid \theta)\pi(\theta)d\theta \Big/ \int_\Theta \prod_{i=1}^{n} p(x_i \mid \theta)\pi(\theta)d\theta$$

with a sum replacing the integral in the discrete case. Now $[\underline{Y}_B, \bar{Y}_B]$ is said to be a level $(1-\alpha)$ *Bayesian prediction interval* for $Y = X_{n+1}$ if

$$Q(\underline{Y}_B \le Y \le \bar{Y}_B \mid \mathbf{x}) \ge 1 - \alpha.$$

**Example 4.8.3.** Consider Example 3.2.1 where $(X_i \mid \theta) \sim \mathcal{N}(\theta, \sigma_0^2)$, $\sigma_0^2$ known, and $\pi(\theta)$ is $\mathcal{N}(\eta_0, \tau^2)$, $\tau^2$ known. A sufficient statistic based on the observables $X_1, \dots, X_n$ is $T = \bar{X} = n^{-1}\sum_{i=1}^n X_i$, and it is enough to derive the marginal distribution of $Y = X_{n+1}$ from the joint distribution of $\bar{X}$, $X_{n+1}$ and $\boldsymbol{\theta}$, where $\bar{X}$ and $X_{n+1}$ are independent. Note that

$$E[(X_{n+1} - \boldsymbol{\theta})\boldsymbol{\theta}] = E\{E(X_{n+1} - \theta)\theta \mid \boldsymbol{\theta} = \theta\} = 0.$$

Thus, $X_{n+1} - \boldsymbol{\theta}$ and $\boldsymbol{\theta}$ are uncorrelated and, by Theorem B.4.1, independent.

To obtain the predictive distribution, note that given $\bar{X} = t$, $X_{n+1} - \boldsymbol{\theta}$ and $\boldsymbol{\theta}$ are still uncorrelated and independent. Thus, if we let $\mathcal{L}$ denote "distribution of," then

$$\mathcal{L}\{X_{n+1} \mid \bar{X} = t\} = \mathcal{L}\{(X_{n+1} - \boldsymbol{\theta}) + \boldsymbol{\theta} \mid \bar{X} = t\} = \mathcal{N}(\widehat{\mu}_B, \sigma_0^2 + \widehat{\sigma}_B^2)$$

where, from Example 4.7.1,

$$\widehat{\sigma}_B^2 = \frac{1}{\frac{n}{\sigma_0^2} + \frac{1}{\tau^2}}, \ \widehat{\mu}_B = (\widehat{\sigma}_B^2/\tau^2)\eta_0 + (n\widehat{\sigma}_B^2/\sigma_0^2)\bar{x}.$$

It follows that a level $(1-\alpha)$ Bayesian prediction interval for $Y$ is $[Y_B^-, Y_B^+]$ with

$$Y_B^{\pm} = \widehat{\mu}_B \pm z\left(1 - \tfrac{1}{2}\alpha\right)\sqrt{\sigma_0^2 + \widehat{\sigma}_B^2}. \tag{4.8.3}$$

To consider the frequentist properties of the Bayesian prediction interval (4.8.3) we compute its probability limit under the assumption that $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\theta, \sigma_0^2)$. Because $\widehat{\sigma}_B^2 \to 0$, $(n\widehat{\sigma}_B^2/\sigma_0^2) \to 1$, and $\bar{X} \xrightarrow{P} \theta$ as $n \to \infty$, we find that the interval (4.8.3) converges in probability to $\theta \pm z\left(1 - \frac{1}{2}\alpha\right)\sigma_0$ as $n \to \infty$. This is the same as the probability limit of the frequentist interval (4.8.1). □

The posterior predictive distribution is also used to check whether the model and the prior give a reasonable description of the uncertainty in a study (see Box, 1983).

**Summary.** We consider intervals based on observable random variables that contain an unobservable random variable with probability at least $(1-\alpha)$. In the case of a normal sample of size $n+1$ with only $n$ variables observable, we construct the Student $t$ prediction interval for the unobservable variable. For a sample of size $n+1$ from a continuous distribution we show how the order statistics can be used to give a distribution-free prediction interval. The Bayesian formulation is based on the posterior predictive distribution which is the conditional distribution of the unobservable variable given the observable variables. The Bayesian prediction interval is derived for the normal model with a normal prior.

## 4.9 LIKELIHOOD RATIO PROCEDURES

### 4.9.1 Introduction

Up to this point, the results and examples in this chapter deal mostly with one-parameter problems in which it sometimes is possible to find optimal procedures. However, even in the case in which $\theta$ is one-dimensional, optimal procedures may not exist. For instance, if $X_1, \dots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma^2)$ population with $\sigma^2$ known, there is no UMP test for testing $H : \mu = \mu_0$ vs $K : \mu \neq \mu_0$. To see this, note that it follows from Example 4.2.1 that if $\mu_1 > \mu_0$, the MP level $\alpha$ test $\delta_\alpha(\mathbf{X})$ rejects $H$ for $T > z(1 - \frac{1}{2}\alpha)$, where $T = \sqrt{n}(\bar{X} - \mu_0)/\sigma$. On the other hand, if $\mu_1 < \mu_0$, the MP level $\alpha$ test $\varphi_\alpha(\mathbf{X})$ rejects $H$ if $T \leq z(\alpha)$. Because $\delta_\alpha(\mathbf{x}) \neq \varphi_\alpha(\mathbf{x})$, by the uniqueness of the NP test (Theorem 4.2.1(c)), there can be no UMP test of $H : \mu = \mu_0$ vs $H : \mu \neq \mu_0$.

In this section we introduce intuitive and efficient procedures that can be used when no optimal methods are available and that are natural for multidimensional parameters. The efficiency is in an approximate sense that will be made clear in Chapters 5 and 6. We start with a generalization of the Neyman-Pearson statistic $p(x, \theta_1)/p(x, \theta_0)$. Suppose that $\mathbf{X} = (X_1, \dots, X_n)$ has density or frequency function $p(\mathbf{x}, \theta)$ and we wish to test $H : \theta \in \Theta_0$ vs $K : \theta \in \Theta_1$. The test statistic we want to consider is the *likelihood ratio* given by

$$L(\mathbf{x}) = \frac{\sup\{p(\mathbf{x}, \theta) : \theta \in \Theta_1\}}{\sup\{p(\mathbf{x}, \theta) : \theta \in \Theta_0\}}.$$

Tests that reject $H$ for large values of $L(\mathbf{x})$ are called *likelihood ratio tests.*

To see that this is a plausible statistic, recall from Section 2.2.2 that we think of the likelihood function $L(\theta, \mathbf{x}) = p(\mathbf{x}, \theta)$ as a measure of how well $\theta$ "explains" the given sample $\mathbf{x} = (x_1, \dots, x_n)$. So, if $\sup\{p(\mathbf{x}, \theta) : \theta \in \Theta_1\}$ is large compared to $\sup\{p(\mathbf{x}, \theta) : \theta \in \Theta_0\}$, then the observed sample is best explained by some $\theta \in \Theta_1$, and conversely. Also note that $L(\mathbf{x})$ coincides with the optimal test statistic $p(x, \theta_1)/p(x, \theta_0)$ when $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$. In particular cases, and for large samples, likelihood ratio tests have weak optimality properties to be discussed in Chapters 5 and 6.

In the cases we shall consider, $p(\mathbf{x}, \theta)$ is a continuous function of $\theta$ and $\Theta_0$ is of smaller dimension than $\Theta = \Theta_0 \cup \Theta_1$ so that the likelihood ratio equals the test statistic

$$\lambda(\mathbf{x}) = \frac{\sup\{p(\mathbf{x}, \theta) : \theta \in \Theta\}}{\sup\{p(\mathbf{x}, \theta) : \theta \in \Theta_0\}} \tag{4.9.1}$$

whose computation is often simple. Note that in general

$$\lambda(\mathbf{x}) = \max(L(\mathbf{x}), 1).$$

We are going to derive likelihood ratio tests in several important testing problems. Although the calculations differ from case to case, the basic steps are always the same.

1. Calculate the MLE $\widehat{\theta}$ of $\theta$.

2. Calculate the MLE $\widehat{\theta}_0$ of $\theta$ where $\theta$ may vary only over $\Theta_0$.

3. Form $\lambda(\mathbf{x}) = p(\mathbf{x}, \widehat{\theta})/p(\mathbf{x}, \widehat{\theta}_0)$.

4. Find a function $h$ that is strictly increasing on the range of $\lambda$ such that $h(\lambda(\mathbf{X}))$ has a simple form and a tabled distribution under $H$. Because $h(\lambda(\mathbf{X}))$ is equivalent to $\lambda(\mathbf{X})$, we specify the size $\alpha$ likelihood ratio test through the test statistic $h(\lambda(\mathbf{X}))$ and its $(1 - \alpha)$th quantile obtained from the table.

We can also invert families of likelihood ratio tests to obtain what we shall call *likelihood confidence regions, bounds,* and so on. For instance, we can invert the family of size $\alpha$ likelihood ratio tests of the point hypothesis $H : \theta = \theta_0$ and obtain the level $(1 - \alpha)$ confidence region

$$C(\mathbf{x}) = \{\theta : p(\mathbf{x}, \theta) \geq [c(\theta)]^{-1} \sup_{\theta} p(\mathbf{x}, \theta)\} \tag{4.9.2}$$

where $\sup_\theta$ denotes sup over $\theta \in \Theta$ and the critical constant $c(\theta)$ satisfies

$$P_{\theta_0}\left[\frac{\sup_\theta p(\mathbf{X}, \theta)}{p(\mathbf{X}, \theta_0)} \geq c(\theta_0)\right] = \alpha.$$

It is often approximately true (see Chapter 6) that $c(\theta)$ is independent of $\theta$. In that case, $C(\mathbf{x})$ is just the set of all $\theta$ whose likelihood is on or above some fixed value dependent on the data. An example is discussed in Section 4.9.2.

This section includes situations in which $\theta = (\theta_1, \theta_2)$ where $\theta_1$ is the parameter of interest and $\theta_2$ is a nuisance parameter. We shall obtain likelihood ratio tests for hypotheses of the form $H : \theta_1 = \theta_{10}$, which are composite because $\theta_2$ can vary freely. The family of such level $\alpha$ likelihood ratio tests obtained by varying $\theta_{10}$ can also be inverted and yield confidence regions for $\theta_1$. To see how the process works we refer to the specific examples in Sections 4.9.2–4.9.5.

### 4.9.2 Tests for the Mean of a Normal Distribution-Matched Pair Experiments

Suppose $X_1, \dots, X_n$ form a sample from a $\mathcal{N}(\mu, \sigma^2)$ population in which both $\mu$ and $\sigma^2$ are unknown. An important class of situations for which this model may be appropriate occurs in *matched pair experiments.* Here are some examples. Suppose we want to study the effect of a treatment on a population of patients whose responses are quite variable because the patients differ with respect to age, diet, and other factors. We are interested in expected differences in responses due to the treatment effect. In order to reduce differences due to the extraneous factors, we consider pairs of patients matched so that within each pair the patients are as alike as possible with respect to the extraneous factors. We can regard twins as being matched pairs. After the matching, the experiment proceeds as follows. In the $i$th pair one patient is picked at random (i.e., with probability $\frac{1}{2}$) and given the treatment, while the second patient serves as control and receives a placebo. Response measurements are taken on the treated and control members of each pair.

Studies in which subjects serve as their own control can also be thought of as matched pair experiments. That is, we measure the response of a subject when under treatment and when not under treatment. Examples of such measurements are hours of sleep when receiving a drug and when receiving a placebo, sales performance before and after a course in salesmanship, mileage of cars with and without a certain ingredient or adjustment, and so on.

Let $X_i$ denote the difference between the treated and control responses for the $i$th pair. If the treatment and placebo have the same effect, the difference $X_i$ has a distribution that is symmetric about zero. To this we have added the normality assumption. The test we derive will still have desirable properties in an approximate sense to be discussed in Chapter 5 if the normality assumption is not satisfied. Let $\mu = E(X_1)$ denote the mean difference between the response of the treated and control subjects. We think of $\mu$ as representing the treatment effect. Our null hypothesis of no treatment effect is then $H : \mu = 0$. However, for the purpose of referring to the duality between testing and confidence procedures, we test $H : \mu = \mu_0$, where we think of $\mu_0$ as an established standard for an old treatment.

### Two-Sided Tests

We begin by considering $K : \mu \neq \mu_0$. This corresponds to the alternative "The treatment has some effect, good or bad." However, as discussed in Section 4.5, the test can be modified into a three-decision rule that decides whether there is a significant positive or negative effect.

### Form of the Two-Sided Tests

Let $\theta = (\mu, \sigma^2)$, $\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0\}$.

Under our assumptions,

$$p(\mathbf{x}, \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$

The problem of finding the supremum of $p(\mathbf{x}, \theta)$ was solved in Example 2.2.9. We found that

$$\sup\{p(\mathbf{x}, \theta) : \theta \in \Theta\} = p(\mathbf{x}, \widehat{\theta}),$$

where

$$\widehat{\theta} = (\bar{x}, \widehat{\sigma}^2) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,\right)$$

is the maximum likelihood estimate of $\theta$.

Finding $\sup\{p(\mathbf{x}, \theta) : \theta \in \Theta_0\}$ boils down to finding the maximum likelihood estimate $\widehat{\sigma}_0^2$ of $\sigma^2$ when $\mu = \mu_0$ is known and then evaluating $p(\mathbf{x}, \theta)$ at $(\mu_0, \widehat{\sigma}_0^2)$. The likelihood equation is

$$\frac{\partial}{\partial\sigma^2}\log p(\mathbf{x}, \theta) = \frac{1}{2}\left[\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu_0)^2 - \frac{n}{\sigma^2}\right] = 0,$$

which has the immediate solution

$$\widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2.$$

By Theorem 2.3.1, $\widehat{\sigma}_0^2$ gives the maximum of $p(\mathbf{x}, \theta)$ for $\theta \in \Theta_0$. The test statistic $\lambda(\mathbf{x})$ is equivalent to $\log \lambda(\mathbf{x})$, which thus equals

$$\begin{aligned}\log \lambda(\mathbf{x}) &= \log p(\mathbf{x}, \widehat{\theta}) - \log p(\mathbf{x}, (\mu_0, \widehat{\sigma}_0^2)) \\ &= \left\{-\frac{n}{2}[(\log 2\pi) + (\log \widehat{\sigma}^2)] - \frac{n}{2}\right\} - \left\{-\frac{n}{2}[(\log 2\pi) + (\log \widehat{\sigma}_0^2)] - \frac{n}{2}\right\} \\ &= \frac{n}{2}\log(\widehat{\sigma}_0^2/\widehat{\sigma}^2).\end{aligned}$$

Our test rule, therefore, rejects $H$ for large values of $(\widehat{\sigma}_0^2/\widehat{\sigma}^2)$. To simplify the rule further we use the following equation, which can be established by expanding both sides.

$$\widehat{\sigma}_0^2 = \widehat{\sigma}^2 + (\bar{x} - \mu_0)^2$$

Therefore,

$$(\widehat{\sigma}_0^2/\widehat{\sigma}^2) = 1 + (\bar{x} - \mu_0)^2/\widehat{\sigma}^2.$$

Because $s^2 = (n-1)^{-1}\sum(x_i - \bar{x})^2 = n\widehat{\sigma}^2/(n-1)$, $\widehat{\sigma}_0^2/\widehat{\sigma}^2$ is a monotone increasing function of $|T_n|$ where

$$T_n = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}.$$

Therefore, the likelihood ratio tests reject for large values of $|T_n|$. Because $T_n$ has a $\mathcal{T}$ distribution under $H$ (see Example 4.4.1), the size $\alpha$ critical value is $t_{n-1}(1 - \frac{1}{2}\alpha)$ and we can use calculators or software that gives quantiles of the $t$ distribution, or Table III, to find the critical value. For instance, suppose $n = 25$ and we want $\alpha = 0.05$. Then we would reject $H$ if, and only if, $|T_n| \geq 2.064$.

## One-Sided Tests

The two-sided formulation is natural if two treatments, A and B, are considered to be equal before the experiment is performed. However, if we are comparing a treatment and control, the relevant question is whether the treatment creates an improvement. Thus, the testing problem $H : \mu \leq \mu_0$ versus $K : \mu > \mu_0$ (with $\mu_0 = 0$) is suggested. The statistic $T_n$ is equivalent to the likelihood ratio statistic $\lambda$ for this problem. A proof is sketched in Problem 4.9.2. In Problem 4.9.11 we argue that $P_\delta[T_n \geq t]$ is increasing in $\delta$, where $\delta = (\mu - \mu_0)/\sigma$. Therefore, the test that rejects $H$ for

$$T_n \geq t_{n-1}(1 - \alpha),$$

is of size $\alpha$ for $H : \mu \leq \mu_0$. Similarly, the size $\alpha$ likelihood ratio test for $H : \mu \geq \mu_0$ versus $K : \mu < \mu_0$ rejects $H$ if, and only if,

$$T_n \leq t_{n-1}(\alpha).$$

## Power Functions

To discuss the power of these tests, we need to introduce the *noncentral $t$ distribution* with $k$ degrees of freedom and noncentrality parameter $\delta$. This distribution, denoted by $\mathcal{T}_{k,\delta}$, is by definition the distribution of $Z/\sqrt{V/k}$ where $Z$ and $V$ are independent and have $\mathcal{N}(\delta, 1)$ and $\chi^2_k$ distributions, respectively. The density of $Z/\sqrt{V/k}$ is given in Problem 4.9.12. To derive the distribution of $T_n$, note that from Section B.3, we know that $\sqrt{n}(\bar{X} - \mu)/\sigma$ and $(n-1)s^2/\sigma^2$ are independent and that $(n-1)s^2/\sigma^2$ has a $\chi^2_{n-1}$ distribution. Because $E[\sqrt{n}(\bar{X} - \mu_0)/\sigma] = \sqrt{n}(\mu - \mu_0)/\sigma$ and

$$\text{Var}(\sqrt{n}(\bar{X} - \mu_0)/\sigma) = 1,$$

$\sqrt{n}(\bar{X} - \mu_0)/\sigma$ has $\mathcal{N}(\delta, 1)$ distribution, with $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$. Thus, the ratio

$$\frac{\sqrt{n}(\bar{X} - \mu_0)/\sigma}{\sqrt{\frac{(n-1)s^2/\sigma^2}{n-1}}} = T_n$$

has a $\mathcal{T}_{n-1,\delta}$ distribution, and the power can be obtained from computer software or tables of the noncentral $t$ distribution. Note that the distribution of $T_n$ depends on $\theta = (\mu, \sigma^2)$ only through $\delta$.

The power functions of the one-sided tests are monotone in $\delta$ (Problem 4.9.11) just as the power functions of the corresponding tests of Example 4.2.1 are monotone in $\sqrt{n}\mu/\sigma$.

We can control both probabilities of error by selecting the sample size $n$ large provided we consider alternatives of the form $|\delta| \geq \delta_1 > 0$ in the two-sided case and $\delta \geq \delta_1$ or $\delta \leq \delta_1$ in the one-sided cases. Computer software will compute $n$.

If we consider alternatives of the form $(\mu - \mu_0) \geq \Delta$, say, we can no longer control both probabilities of error by choosing the sample size. The reason is that, whatever be $n$, by making $\sigma$ sufficiently large we can force the noncentrality parameter $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$ as close to 0 as we please and, thus, bring the power arbitrarily close to $\alpha$. We have met similar difficulties (Problem 4.4.7) when discussing confidence intervals. A Stein solution, in which we estimate $\sigma$ for a first sample and use this estimate to decide how many more observations we need to obtain guaranteed power against all alternatives with $|\mu - \mu_0| \geq \Delta$, is possible (Lehmann, 1997, p. 260, Problem 17). With this solution, however, we may be required to take more observations than we can afford on the second stage.

## Likelihood Confidence Regions

If we invert the two-sided tests, we obtain the confidence region

$$C(\mathbf{X}) = \{\mu : |\sqrt{n}(\bar{X} - \mu)/s| \leq t_{n-1}(1 - \tfrac{1}{2}\alpha)\}.$$

We recognize $C(\mathbf{X})$ as the confidence interval of Example 4.4.1. Similarly the one-sided tests lead to the lower and upper confidence bounds of Example 4.4.1.

### Data Example

As an illustration of these procedures, consider the following data due to Cushny and Peebles (see Fisher, 1958, p. 121) giving the difference $B - A$ in sleep gained using drugs $A$ and $B$ on 10 patients. This is a matched pair experiment with each subject serving as its own control.

| Patient $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $A$ | 0.7 | -1.6 | -0.2 | -1.2 | -0.1 | 3.4 | 3.7 | 0.8 | 0.0 | 2.0 |
| $B$ | 1.9 | 0.8 | 1.1 | 0.1 | -0.1 | 4.4 | 5.5 | 1.6 | 4.6 | 3.4 |
| $B - A$ | 1.2 | 2.4 | 1.3 | 1.3 | 0.0 | 1.0 | 1.8 | 0.8 | 4.6 | 1.4 |

If we denote the difference as $x$'s, then $\bar{x} = 1.58$, $s^2 = 1.513$, and $|T_n| = 4.06$. Because $t_9(0.995) = 3.25$, we conclude at the 1% level of significance that the two drugs are significantly different. The 0.99 confidence interval for the mean difference $\mu$ between treatments is [0.32,2.84]. It suggests that not only are the drugs different but in fact $B$ is better than $A$ because no hypothesis $\mu = \mu' < 0$ is accepted at this level. (See also (4.5.3).)

## 4.9.3 Tests and Confidence Intervals for the Difference in Means of Two Normal Populations

We often want to compare two populations with distribution $F$ and $G$ on the basis of two independent samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, one from each population. For instance, suppose we wanted to test the effect of a certain drug on some biological variable (e.g., blood pressure). Then $X_1, \ldots, X_{n_1}$ could be blood pressure measurements on a sample of patients given a placebo, while $Y_1, \ldots, Y_{n_2}$ are the measurements on a sample given the drug. For quantitative measurements such as blood pressure, height, weight, length, volume, temperature, and so forth, it is usually assumed that $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are independent samples from $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ populations, respectively.

The preceding assumptions were discussed in Example 1.1.3. A discussion of the consequences of the violation of these assumptions will be postponed to Chapters 5 and 6.

*Tests*

We first consider the problem of testing $H : \mu_1 = \mu_2$ versus $K : \mu_1 \neq \mu_2$. In the control versus treatment example, this is the problem of determining whether the treatment has any effect.

Let $\theta = (\mu_1, \mu_2, \sigma^2)$. Then $\Theta_0 = \{\theta : \mu_1 = \mu_2\}$ and $\Theta_1 = \{\theta : \mu_1 \neq \mu_2\}$. The log of the likelihood of $(\mathbf{X}, \mathbf{Y}) = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ is

$$\log p(\mathbf{x}, \mathbf{y}, \theta) = -(n/2)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i - \mu_1)^2 + \sum_{j=1}^{n_2}(y_j - \mu_2)^2\right)$$

where $n = n_1 + n_2$. As in Section 4.9.2 the likelihood function and its log are maximized over $\Theta$ by the maximum likelihood estimate $\widehat{\theta}$. In Problem 4.9.6, it is shown that $\widehat{\theta} =$

$(\bar{X}, \bar{Y}, \widetilde{\sigma}^2)$, where

$$\widetilde{\sigma}^2 = \frac{1}{n}\left[\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2\right]$$

When $\mu_1 = \mu_2 = \mu$, our model reduces to the one-sample model of Section 4.9.2. Thus, the maximum of $p$ over $\Theta_0$ is obtained for $\theta = (\widehat{\mu}, \widehat{\mu}, \widetilde{\sigma}_0^2)$, where

$$\widehat{\mu} = \frac{1}{n}\left[\sum_{i=1}^{n_1} X_i + \sum_{j=1}^{n_2} Y_j\right]$$

and

$$\widetilde{\sigma}_0^2 = \frac{1}{n}\left[\sum_{i=1}^{n_1}(X_i - \widehat{\mu})^2 + \sum_{j=1}^{n_2}(Y_j - \widehat{\mu})^2\right].$$

If we use the identities

$$\frac{1}{n}\sum_{i=1}^{n_1}(X_i - \widehat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \frac{n_1}{n}(\bar{X} - \widehat{\mu})^2$$

$$\frac{1}{n}\sum_{i=1}^{n_2}(Y_i - \widehat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n_1}(Y_i - \bar{Y})^2 + \frac{n_2}{n}(\bar{Y} - \widehat{\mu})^2$$

obtained by writing $[X_i - \widehat{\mu}]^2 = [(X_i - \bar{X}) + (\bar{X} - \widehat{\mu})]^2$ and expanding, we find that the log likelihood ratio statistic

$$\log \lambda(\mathbf{x}, \mathbf{y}) = \frac{n}{2}\log(\widetilde{\sigma}_0^2/\widetilde{\sigma}^2)$$

is equivalent to the test statistic $|T|$ where

$$T = \sqrt{\frac{n_1 n_2}{n}}\left(\frac{\bar{Y} - \bar{X}}{s}\right)$$

and

$$s^2 = n\widetilde{\sigma}^2/(n-2) = \frac{1}{n-2}\left[\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2\right].$$

To complete specification of the size $\alpha$ likelihood ratio test we show that $T$ has $\mathcal{T}_{n-2}$ distribution when $\mu_1 = \mu_2$. By Theorem B.3.3

$$\frac{1}{\sigma}\bar{X}, \frac{1}{\sigma}\bar{Y}, \frac{1}{\sigma^2}\sum_{i=1}^{n_1}(X_i - \bar{X})^2 \text{ and } \frac{1}{\sigma^2}\sum_{j=1}^{n_2}(Y_j - \bar{Y})^2$$

are independent and distributed as $\mathcal{N}(\mu_1/\sigma, 1/n_1), \mathcal{N}(\mu_2/\sigma, 1/n_2), \chi^2_{n_1-1}, \chi^2_{n_2-1}$, respectively. We conclude from this remark and the additive property of the $\chi^2$ distribution that $(n-2)s^2/\sigma^2$ has a $\chi^2_{n-2}$ distribution, and that $\sqrt{n_1n_2/n}(\bar{Y}-\bar{X})/\sigma$ has a $\mathcal{N}(0,1)$ distribution and is independent of $(n-2)s^2/\sigma^2$. Therefore, by definition, $T \sim \mathcal{T}_{n-2}$ under $H$ and the resulting *two-sided, two-sample* $t$ test rejects if, and only if,

$$|T| \geq t_{n-2}(1-\frac{1}{2}\alpha)$$

As usual, corresponding to the two-sided test, there are two one-sided tests with critical regions,

$$T \geq t_{n-2}(1-\alpha)$$

for $H : \mu_2 \leq \mu_1$ and

$$T \leq t_{n-2}(\alpha)$$

for $H : \mu_1 \leq \mu_2$.

We can show that these tests are likelihood ratio tests for these hypotheses. It is also true that these procedures are of size $\alpha$ for their respective hypotheses. As in the one-sample case, this follows from the fact that, if $\mu_1 \neq \mu_2$, $T$ has a noncentral $t$ distribution with noncentrality parameter,

$$\delta_2 = E\left(\sqrt{n_1n_2/n}\frac{(\bar{Y}-\bar{X})}{\sigma}\right) = \sqrt{n_1n_2/n}\frac{(\mu_2-\mu_1)}{\sigma}.$$

## Confidence Intervals

To obtain confidence intervals for $\mu_2 - \mu_1$ we naturally look at likelihood ratio tests for the family of testing problems $H : \mu_2 - \mu_1 = \Delta$ versus $K : \mu_2 - \mu_1 \neq \Delta$. As for the special case $\Delta = 0$, we find a simple equivalent statistic $|T(\Delta)|$ where

$$T(\Delta) = \sqrt{n_1n_2/n}(\bar{Y}-\bar{X}-\Delta)/s.$$

If $\mu_2 - \mu_1 = \Delta$, $T(\Delta)$ has a $\mathcal{T}_{n-2}$ distribution and inversion of the tests leads to the interval

$$\bar{Y}-\bar{X} \pm t_{n-2}(1-\frac{1}{2}\alpha)s\sqrt{n/n_1n_2}, \tag{4.9.3}$$

for $\mu_2 - \mu_1$. Similarly, one-sided tests lead to the upper and lower endpoints of the interval as $1-\frac{1}{2}\alpha$ likelihood confidence bounds.

### Data Example

As an illustration, consider the following experiment designed to study the permeability (tendency to leak water) of sheets of building material produced by two different machines. From past experience, it is known that the log of permeability is approximately normally distributed and that the variability from machine to machine is the same. The results in terms of logarithms were (from Hald, 1952, p. 472)

| | | | |
|---|---|---|---|
| $x$ (machine 1) | 1.845 | 1.790 | 2.042 |
| $y$ (machine 2) | 1.583 | 1.627 | 1.282 |

.

We test the hypothesis $H$ of no difference in expected log permeability. $H$ is rejected if $|T| \geq t_{n-2}(1-\frac{1}{2}\alpha)$. Here $\bar{y}-\bar{x}=-0.395, s^2=0.0264$, and $T=-2.977$. Because $t_4(0.975)=2.776$, we conclude at the 5% level of significance that there is a significant difference between the expected log permeability for the two machines. The level 0.95 confidence interval for the difference in mean log permeability is

$$-0.395 \pm 0.368.$$

On the basis of the results of this experiment, we would select machine 2 as producing the smaller permeability and, thus, the more waterproof material. Again we can show that the selection procedure based on the level $(1-\alpha)$ confidence interval has probability at most $\frac{1}{2}\alpha$ of making the wrong selection.

## 4.9.4 The Two-Sample Problem with Unequal Variances

In two-sample problems of the kind mentioned in the introduction to Section 4.9.3, it may happen that the $X$'s and $Y$'s have different variances. For instance, a treatment that increases mean response may increase the variance of the responses. If normality holds, we are led to a model where $X_1,\dots,X_{n_1}; Y_1,\dots,Y_{n_2}$ are two independent $\mathcal{N}(\mu_1,\sigma_1^2)$ and $\mathcal{N}(\mu_2,\sigma_2^2)$ samples, respectively. As a first step we may still want to compare mean responses for the $X$ and $Y$ populations. This is the *Behrens-Fisher problem.*

Suppose first that $\sigma_1^2$ and $\sigma_2^2$ are known. The log likelihood, except for an additive constant, is

$$-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}(x_i-\mu_1)^2-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n_2}(y_j-\mu_2)^2.$$

The MLEs of $\mu_1$ and $\mu_2$ are, thus, $\widehat{\mu}_1=\bar{x}$ and $\widehat{\mu}_2=\bar{y}$ for $(\mu_1,\mu_2)\in R\times R$. When $\mu_1=\mu_2=\mu$, setting the derivative of the log likelihood equal to zero yields the MLE

$$\widehat{\mu}=\frac{\sum_{i=1}^{n_1}x_i+\gamma\sum_{j=1}^{n_2}y_j}{n_1+\gamma n_2}$$

where $\gamma = \sigma_1^2/\sigma_2^2$. It follows that

$$\begin{aligned}\lambda(\mathbf{x},\mathbf{y}) &= \frac{p(\mathbf{x},\mathbf{y},\bar{x},\bar{y})}{p(\mathbf{x},\mathbf{y},\widehat{\mu},\widehat{\mu})}\\ &= \exp\left\{\frac{1}{2\sigma_1^2}\left[\sum_{i=1}^{n_1}(x_i-\widehat{\mu})^2-\sum_{i=1}^{n_1}(x_i-\bar{x})^2\right]\right.\\ &+ \left.\frac{1}{2\sigma_2^2}\left[\sum_{j=1}^{n_2}(y_j-\widehat{\mu})^2-\sum_{j=1}^{n_2}(y_j-\bar{y})^2\right]\right\}.\end{aligned}$$

By writing

$$\begin{aligned}\sum_{i=1}^{n_1}(x_i-\widehat{\mu})^2 &= \sum_{i=1}^{n_1}(x_i-\bar{x})^2+n_1(\widehat{\mu}-\bar{x})^2\\ \sum_{j=1}^{n_2}(y_i-\widehat{\mu})^2 &= \sum_{j=1}^{n_2}(y_i-\bar{y})^2+n_2(\widehat{\mu}-\bar{y})^2\end{aligned}$$

we obtain

$$\lambda(\mathbf{x},\mathbf{y}) = \exp\left\{\frac{n_1}{2\sigma_1^2}(\widehat{\mu}-\bar{x})^2+\frac{n_2}{2\sigma_2^2}(\widehat{\mu}-\bar{y})^2\right\}.$$

Next we compute

$$\begin{aligned}\widehat{\mu}-\bar{x} &= \frac{n_1\sum_{i=1}^{n_1}x_i+n_1\gamma\sum_{j=1}^{n_2}y_j-n_1\sum_{i=1}^{n_1}x_i-\gamma n_2\sum_{j=1}^{n_2}y_j}{n_1(n_1+\gamma n_2)}\\ &= \gamma n_2(\bar{y}-\bar{x})/(n_1+\gamma n_2).\end{aligned}$$

Similarly, $\widehat{\mu}-\bar{y} = n_2(\bar{x}-\bar{y})/(n_1+\gamma n_2)$. It follows that the likelihood ratio test is equivalent to the statistic $|D|/\sigma_D$, where $D = \bar{Y}-\bar{X}$ and $\sigma_D^2$ is the variance of $D$, that is

$$\sigma_D^2 = \text{Var}(\bar{X})+\text{Var}(\bar{Y}) = \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}.$$

Thus, $D/\sigma_D$ has a $\mathcal{N}(\Delta,1)$ distribution, where $\Delta = \mu_2-\mu_1$. Because $\sigma_D^2$ is unknown, it must be estimated. An unbiased estimate is

$$s_D^2 = \frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}.$$

It is natural to try and use $D/s_D$ as a test statistic for the one-sided hypothesis $H:\mu_2\le\mu_1$, $|D|/s_D$ as a test statistic for $H:\mu_1=\mu_2$, and more generally $(D-\Delta)/s_D$ to generate confidence procedures. Unfortunately the distribution of $(D-\Delta)/s_D$ depends on $\sigma_1^2/\sigma_2^2$ for fixed $n_1,n_2$. For large $n_1,n_2$, by Slutsky's theorem and the central limit theorem, $(D-\Delta)/s_D$ has approximately a standard normal distribution (Problem 5.3.28). For small and moderate $n_1,n_2$ an approximation to the distribution of $(D-\Delta)/s_D$ due to Welch (1949) works well.

Let $c = s_1^2/n_1 s_D^2$. Then Welch's approximation is $\mathcal{T}_k$ where

$$k = \left[\frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}\right]^{-1}$$

When $k$ is not an integer, the critical value is obtained by linear interpolation in the $t$ tables or using computer software. The tests and confidence intervals resulting from this approximation are called Welch's solutions to the Behrens-Fisher problem. Wang (1971) has shown the approximation to be very good for $\alpha = 0.05$ and $\alpha = 0.01$, the maximum error in size being bounded by 0.003.

Note that Welch's solution works whether the variances are equal or not. The LR procedure derived in Section 4.9.3, which works well if the variances are equal or $n_1 = n_2$, can unfortunately be very misleading if $\sigma_1^2 \neq \sigma_2^2$ and $n_1 \neq n_2$. See Figure 5.3.3 and Problem 5.3.28.

### 4.9.5 Likelihood Ratio Procedures for Bivariate Normal Distributions

If $n$ subjects (persons, machines, fields, mice, etc.) are sampled from a population and two numerical characteristics are measured on each case, then we end up with a bivariate random sample $(X_1, Y_1), \dots, (X_n, Y_n)$. Empirical data sometimes suggest that a reasonable model is one in which the two characteristics $(X, Y)$ have a joint bivariate normal distribution, $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, with $\sigma_1^2 > 0, \sigma_2^2 > 0$.

#### Testing Independence, Confidence Intervals for $\rho$

The question "Are two random variables $X$ and $Y$ independent?" arises in many statistical studies. Some familiar examples are: $X$ = weight, $Y$ = blood pressure; $X$ = test score on mathematics exam, $Y$ = test score on English exam; $X$ = percentage of fat in diet, $Y$ = cholesterol level in blood; $X$ = average cigarette consumption per day in grams, $Y$ = age at death. If we have a sample as before and assume the bivariate normal model for $(X, Y)$, our problem becomes that of testing $H : \rho = 0$.

#### Two-Sided Tests

If we are interested in all departures from $H$ equally, it is natural to consider the two-sided alternative $K : \rho \neq 0$. We derive the likelihood ratio test for this problem. Let $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Then $\Theta_0 = \{\theta : \rho = 0\}$ and $\Theta_1 = \{\theta : \rho \neq 0\}$. From (B.4.9), the log of the likelihood function of $\mathbf{X} = ((X_1, Y_1), \dots, (X_n, Y_n))$ is

$$\log p(\mathbf{x}, \theta) = -n[\log(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})]$$
$$-\frac{1}{2(1-\rho^2)}\left[\frac{\sum_{i=1}^n (x_i - \mu_1)^2}{\sigma_1^2} - \frac{2\rho\sum_{i=1}^n (x_i - \mu_1)(y_i - \mu_2)}{\sigma_1\sigma_2} + \frac{\sum_{i=1}^n (y_i - \mu_2)^2}{\sigma_2^2}\right].$$

The unrestricted maximum likelihood estimate $\widehat{\theta}$ was given in Problem 2.3.13 as $(\bar{x}, \bar{y}, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\rho})$, where

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \widehat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

$$\widehat{\rho} = \left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right] / n\widehat{\sigma}_1\widehat{\sigma}_2.$$

$\widehat{\rho}$ is called the *sample correlation coefficient* and satisfies $-1 \leq \widehat{\rho} \leq 1$ (Problem 2.1.8). When $\rho = 0$ we have two independent samples, and $\widehat{\theta}_0$ can be obtained by separately maximizing the likelihood of $X_1, \ldots, X_n$ and that of $Y_1, \ldots, Y_n$. We have $\widehat{\theta}_0 = (\bar{x}, \bar{y}, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, 0)$ and the log of the likelihood ratio statistic becomes

$$\begin{aligned}\log \lambda(\mathbf{x}) &= \{\log p(\mathbf{x}, \widehat{\theta})\} - \{\log p(\mathbf{x}, \widehat{\theta}_0)\} \\ &= \left\{-n[\log(2\pi\widehat{\sigma}_1\widehat{\sigma}_2)] - \frac{n}{2}[\log(1 - \widehat{\rho}^2)] - n\right\} \\ &\quad \{-n[\log(2\pi\widehat{\sigma}_1\widehat{\sigma}_2)] - n\} \\ &= -\frac{n}{2}\log(1 - \widehat{\rho}^2).\end{aligned} \tag{4.9.4}$$

Thus, $\log \lambda(\mathbf{x})$ is an increasing function of $\widehat{\rho}^2$, and the likelihood ratio tests reject $H$ for large values of $|\widehat{\rho}|$.

To obtain critical values we need the distribution of $\widehat{\rho}$ or an equivalent statistic under $H$. Now,

$$\widehat{\rho} = \sum_{i=1}^{n}(U_i - \bar{U})(V_i - \bar{V}) / \left[\sum_{i=1}^{n}(U_i - \bar{U})^2\right]^{1/2} \left[\sum_{i=1}^{n}(V_i - \bar{V})^2\right]^{1/2}$$

where $U_i = (X_i - \mu_1)/\sigma_1, V_i = (Y_i - \mu_2)/\sigma_2$. Because $(U_1, V_1), \ldots, (U_n, V_n)$ is a sample from the $\mathcal{N}(0, 0, 1, 1, \rho)$ distribution, the distribution of $\widehat{\rho}$ depends on $\rho$ only. If $\rho = 0$, then by Problem B.4.8,

$$T_n = \frac{\sqrt{n-2}\widehat{\rho}}{\sqrt{1 - \widehat{\rho}^2}} \tag{4.9.5}$$

has a $\mathcal{T}_{n-2}$ distribution. Because $|T_n|$ is an increasing function of $|\widehat{\rho}|$, the two-sided likelihood ratio tests can be based on $|T_n|$ and the critical values obtained from Table II.

When $\rho = 0$, the distribution of $\widehat{\rho}$ is available on computer packages. There is no simple form for the distribution of $\widehat{\rho}$ (or $T_n$) when $\rho \neq 0$. A normal approximation is available. See Example 5.3.6.

Qualitatively, for any $\alpha$, the power function of the LR test is symmetric about $\rho = 0$ and increases continuously from $\alpha$ to 1 as $\rho$ goes from 0 to 1. Therefore, if we specify indifference regions, we can control probabilities of type II error by increasing the sample size.

## One-Sided Tests

In many cases, only one-sided alternatives are of interest. For instance, if we want to decide whether increasing fat in a diet significantly increases cholesterol level in the blood, we would test $H : \rho = 0$ versus $K : \rho > 0$ or $H : \rho \leq 0$ versus $K : \rho > 0$. It can be shown that $\widehat{\rho}$ is equivalent to the likelihood ratio statistic for testing $H : \rho \leq 0$ versus $K : \rho > 0$ and similarly that $-\widehat{\rho}$ corresponds to the likelihood ratio statistic for $H : \rho \geq 0$ versus $K : \rho < 0$. We can show that $P_\theta[\widehat{\rho} \geq c]$ is an increasing function of $\rho$ for fixed $c$ (Problem 4.9.15). Therefore, we obtain size $\alpha$ tests for each of these hypotheses by setting the critical value so that the probability of type I error is $\alpha$ when $\rho = 0$. The power functions of these tests are monotone.

## Confidence Bounds and Intervals

Usually testing independence is not enough and we want bounds and intervals for $\rho$ giving us an indication of what departure from independence is present. To obtain lower confidence bounds, we can start by constructing size $\alpha$ likelihood ratio tests of $H : \rho = \rho_0$ versus $K : \rho > \rho_0$. These tests can be shown to be of the form "Accept if, and only if, $\widehat{\rho} \leq c(\rho_0)$" where $P_{\rho_0}[\widehat{\rho} \leq c(\rho_0)] = 1 - \alpha$. We obtain $c(\rho)$ either from computer software or by the approximation of Chapter 5. Because $c$ can be shown to be monotone increasing in $\rho$, inversion of this family of tests leads to level $1 - \alpha$ lower confidence bounds. We can similarly obtain $1 - \alpha$ upper confidence bounds and, by putting two level $1 - \frac{1}{2}\alpha$ bounds together, we obtain a commonly used confidence interval for $\rho$. These intervals do not correspond to the inversion of the size $\alpha$ LR tests of $H : \rho = \rho_0$ versus $K : \rho \neq \rho_0$ but rather of the "equal-tailed" test that rejects if, and only if $\widehat{\rho} \geq d(\rho_0)$ or $\widehat{\rho} \leq c(\rho_0)$ where $P_{\rho_0}[\widehat{\rho} \geq d(\rho_0)] = P_{\rho_0}[\widehat{\rho} \leq c(\rho_0)] = 1 - \frac{1}{2}\alpha$. However, for large $n$ the equal tails and LR confidence intervals approximately coincide with each other.

## Data Example

As an illustration, consider the following bivariate sample of weights $x_i$ of young rats at a certain age and the weight increase $y_i$ during the following week. We want to know whether there is a correlation between the initial weights and the weight increase and formulate the hypothesis $H : \rho = 0$.

| $x_i$ | 383.2 | 356.4 | 362.5 | 397.4 | 356.0 | 387.6 | 385.1 | 346.6 | 370.7 |
|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 27.3 | 41.0 | 38.4 | 24.4 | 25.9 | 21.9 | 13.1 | 28.5 | 18.1 |

Here $\widehat{\rho} = -0.53$ and $T_n = -1.67$. Thus, by using the two-sided test and referring to the $\mathcal{T}_7$ tables, we find that there is no evidence of correlation: the $p$-value is bigger than 0.05.

**Summary.** The likelihood ratio test statistic $\lambda$ is the ratio of the maximum value of the likelihood under the general model to the maximum value of the likelihood under the model specified by the hypothesis. We find the likelihood ratio tests and associated confidence procedures for four classical normal models:

(1) Matched pair experiments in which differences are modeled as $\mathcal{N}(\mu, \sigma^2)$ and we test the hypothesis that the mean difference $\mu$ is zero. The likelihood ratio test is equivalent to the one-sample (Student) $t$ test.

(2) Two-sample experiments in which two independent samples are modeled as coming from $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ populations, respectively. We test the hypothesis that the means are equal and find that the likelihood ratio test is equivalent to the two-sample (Student) $t$ test.

(3) Two-sample experiments in which two independent samples are modeled as coming from $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ populations, respectively. When $\sigma_1^2$ and $\sigma_2^2$ are known, the likelihood ratio test is equivalent to the test based on $|\bar{Y} - \bar{X}|$. When $\sigma_1^2$ and $\sigma_2^2$ are unknown, we use $(\bar{Y} - \bar{X})/s_D$, where $s_D$ is an estimate of the standard deviation of $D = \bar{Y} - \bar{X}$. Approximate critical values are obtained using Welch's $t$ distribution approximation.

(4) Bivariate sampling experiments in which we have two measurements $X$ and $Y$ on each case in a sample of $n$ cases. We test the hypothesis that $X$ and $Y$ are independent and find that the likelihood ratio test is equivalent to the test based on $|\widehat{\rho}|$, where $\widehat{\rho}$ is the sample correlation coefficient. We also find that the likelihood ratio statistic is equivalent to a $t$ statistic with $n-2$ degrees of freedom.

## 4.10 PROBLEMS AND COMPLEMENTS

### Problems for Section 4.1

**1.** Suppose that $X_1, \dots, X_n$ are independently and identically distributed according to the uniform distribution $U(0, \theta)$. Let $M_n = \max(X_1, \dots, X_n)$ and let

$$\begin{aligned} \delta_c(X) &= 1 \text{ if } M_n \geq c \\ &= 0 \text{ otherwise.} \end{aligned}$$

**(a)** Compute the power function of $\delta_c$ and show that it is a monotone increasing function of $\theta$.

**(b)** In testing $H : \theta \leq \frac{1}{2}$ versus $K : \theta > \frac{1}{2}$, what choice of $c$ would make $\delta_c$ have size exactly 0.05?

**(c)** Draw a rough graph of the power function of $\delta_c$ specified in (b) when $n = 20$.

**(d)** How large should $n$ be so that the $\delta_c$ specified in (b) has power $0.98$ for $\theta = \frac{3}{4}$?

**(e)** If in a sample of size $n = 20$, $M_n = 0.48$, what is the $p$-value?

**2.** Let $X_1, \dots, X_n$ denote the times in days to failure of $n$ similar pieces of equipment. Assume the model where $\mathbf{X} = (X_1, \dots, X_n)$ is an $\mathcal{E}(\lambda)$ sample. Consider the hypothesis $H$ that the mean life $1/\lambda = \mu \leq \mu_0$.

**(a)** Use the result of Problem B.3.4 to show that the test with critical region

$$[\bar{X} \geq \mu_0 x(1-\alpha)/2n],$$

where $x(1-\alpha)$ is the $(1-\alpha)$th quantile of the $\chi^2_{2n}$ distribution, is a size $\alpha$ test.

**(b)** Give an expression of the power in terms of the $\chi^2_{2n}$ distribution.

**(c)** Use the central limit theorem to show that $\Phi[(\mu_0 z(\alpha)/\mu) + \sqrt{n}(\mu - \mu_0)/\mu]$ is an approximation to the power of the test in part (a). Draw a graph of the approximate power function.

*Hint:* Approximate the critical region by $[\bar{X} \geq \mu_0(1 + z(1-\alpha)/\sqrt{n})]$

**(d)** The following are days until failure of air monitors at a nuclear plant. If $\mu_0 = 25$, give a normal approximation to the significance probability. Days until failure:

3 150 40 34 32 37 34 2 31 6 5 14 150 27 4 6 27 10 30 37

Is $H$ rejected at level $\alpha = 0.05$?

**3.** Let $X_1, \ldots, X_n$ be a $\mathcal{P}(\theta)$ sample.

**(a)** Use the MLE $\bar{X}$ of $\theta$ to construct a level $\alpha$ test for $H : \theta \leq \theta_0$ versus $K : \theta > \theta_0$.

**(b)** Show that the power function of your test is increasing in $\theta$.

**(c)** Give an approximate expression for the critical value if $n$ is large and $\theta$ not too close to 0 or $\infty$. (Use the central limit theorem.)

**4.** Let $X_1, \ldots, X_n$ be a sample from a population with the Rayleigh density

$$f(x, \theta) = (x/\theta^2)\exp\{-x^2/2\theta^2\},\ x > 0, \theta > 0.$$

**(a)** Construct a test of $H : \theta = 1$ versus $K : \theta > 1$ with approximate size $\alpha$ using a sufficient statistic for this model.

*Hint:* Use the central limit theorem for the critical value.

**(b)** Check that your test statistic has greater expected value under $K$ than under $H$.

**5.** Show that if $H$ is simple and the test statistic $T$ has a continuous distribution, then the $p$-value $\alpha(T)$ has a uniform, $\mathcal{U}(0,1)$, distribution.

*Hint:* See Problem B.2.12.

**6.** Suppose that $T_1, \ldots, T_r$ are independent test statistics for the same simple $H$ and that each $T_j$ has a continuous distribution, $j = 1, \ldots, r$. Let $\alpha(T_j)$ denote the $p$-value for $T_j$, $j = 1, \ldots, r$.

Show that, under $H$, $\widehat{T} = -2\sum_{j=1}^{r} \log \alpha(T_j)$ has a $\chi^2_{2r}$ distribution.

*Hint:* See Problem B.3.4.

**7.** Establish (4.1.3). Assume that $F_0$ and $F$ are continuous.

**8. (a)** Show that the power $P_F[D_n \geq k_\alpha]$ of the Kolmogorov test is bounded below by

$$\sup_x P_F[|\widehat{F}(x) - F_0(x)| \geq k_\alpha].$$

*Hint:* $D_n \geq |\widehat{F}(x) - F_0(x)|$ for each $x$.

**(b)** Suppose $F_0$ is $\mathcal{N}(0,1)$ and $F(x) = (1+\exp(-x/\tau))^{-1}$ where $\tau = \sqrt{3}/\pi$ is chosen so that $\int_{-\infty}^{\infty} x^2 dF(x) = 1$. (This is the logistic distribution with mean zero and variance 1.) Evaluate the bound $P_F(|\widehat{F}(x) - F_0(x)| \geq k_\alpha)$ for $\alpha = 0.10$, $n = 80$ and $x = 0.5$, 1, and 1.5 using the normal approximation to the binomial distribution of $n\widehat{F}(x)$ and the approximate critical value in Example 4.1.5.

**(c)** Show that if $F$ and $F_0$ are continuous and $F \neq F_0$, then the power of the Kolmogorov test tends to 1 as $n \to \infty$.

**9.** Let $X_1, \ldots, X_n$ be i.i.d. with distribution function $F$ and consider $H : F = F_0$. Suppose that the distribution $\mathcal{L}_0$ of the statistic $T = T(\mathbf{X})$ is continuous under $H$ and that $H$ is rejected for large values of $T$. Let $T^{(1)}, \ldots, T^{(B)}$ be $B$ independent Monte Carlo simulated values of $T$. (In practice these can be obtained by drawing $B$ independent samples $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(B)}$ from $F_0$ on the computer and computing $T^{(j)} = T(\mathbf{X}^{(j)})$, $j = 1, \ldots, B$. Here, to get $X$ with distribution $F_0$, generate a $\mathcal{U}(0,1)$ variable on the computer and set $X = F_0^{-1}(U)$ as in Problem B.2.12(b).) Next let $T_{(1)}, \ldots, T_{(B+1)}$ denote $T, T^{(1)}, \ldots, T^{(B)}$ ordered. Show that the test that rejects $H$ iff $T \geq T_{(B+2-m)}$ has level $\alpha = m/(B+1)$.

*Hint:* If $H$ is true $T(\mathbf{X}), T(\mathbf{X}^{(1)}), \ldots, T(\mathbf{X}^{(B)})$ is a sample of size $B+1$ from $\mathcal{L}_0$. Use the fact that $T(\mathbf{X})$ is equally likely to be any particular order statistic.

**10. (a)** Show that the statistic $T_n$ of Example 4.1.6 is invariant under location and scale. That is, if $X_i' = (X_i - a)/b$, $b > 0$, then $T_n(\mathbf{X}') = T_n(\mathbf{X})$.

**(b)** Use part (a) to conclude that $\mathcal{L}_{\mathcal{N}(\mu,\sigma^2)}(T_n) = \mathcal{L}_{\mathcal{N}(0,1)}(T_n)$.

**11.** In Example 4.1.5, let $\psi(u)$ be a function from $(0,1)$ to $(0,\infty)$, and let $\alpha > 0$. Define the statistics

$$\begin{aligned} S_{\psi,\alpha} &= \sup_x \psi(F_0(x))|\widehat{F}(x) - F_0(x)|^\alpha \\ T_{\psi,\alpha} &= \sup_x \psi(\widehat{F}(x))|\widehat{F}(x) - F_0(x)|^\alpha \end{aligned}$$

$$\begin{aligned} U_{\psi,\alpha} &= \int \psi(F_0(x))|\widehat{F}(x) - F_0(x)|^\alpha dF_0(x) \\ V_{\psi,\alpha} &= \int \psi(\widehat{F}(x))|\widehat{F}(x) - F_0(x)|^\alpha d\widehat{F}(x). \end{aligned}$$

**(a)** For each of these statistics show that the distribution under $H$ does not depend on $F_0$.

**(b)** When $\psi(u) = 1$ and $\alpha = 2$, $V_{\psi,\alpha}$ is called the Cramér–von Mises statistic. Express the Cramer–von Mises statistic as a sum.

**(c)** Are any of the four statistics in (a) invariant under location and scale. (See Problem 4.1.10.)

**12.** *Expected p-values.* Consider a test with critical region of the form $\{T \geq c\}$ for testing $H : \theta = \theta_0$ versus $K : \theta > \theta_0$. Without loss of generality, take $\theta_0 = 0$. Suppose that $T$ has a continuous distribution $F_\theta$, then the $p$-value is

$$U = 1 - F_0(T).$$

**(a)** Show that if the test has level $\alpha$, the power is

$$\beta(\theta) = P(U \leq \alpha) = 1 - F_\theta(F_0^{-1}(1 - \alpha))$$

where $F_0^{-1}(u) = \inf\{t : F_0(t) \geq u\}$.

**(b)** Define the expected $p$-value as $EPV(\theta) = E_\theta U$. Let $T_0$ denote a random variable with distribution $F_0$, which is independent of $T$. Show that $EPV(\theta) = P(T_0 \geq T)$.

*Hint:* $P(T_0 \geq T) = \int P(T_0 \geq t \mid T = t) f_\theta(t) dt$ where $f_\theta(t)$ is the density of $F_\theta(t)$.

**(c)** Suppose that for each $\alpha \in (0, 1)$, the UMP test is of the form $1\{T \geq c\}$. Show that the $EPV(\theta)$ for $1\{T \geq c\}$ is uniformly minimal in $\theta > 0$ when compared to the $EPV(\theta)$ for any other test.

*Hint:* $P(T \leq t_0 \mid T_0 = t_0)$ is 1 minus the power of a test with critical value $t_0$.

**(d)** Consider the problem of testing $H : \mu = \mu_0$ versus $K : \mu > \mu_0$ on the basis of the $\mathcal{N}(\mu, \sigma^2)$ sample $X_1, \ldots, X_n$, where $\sigma$ is known. Let $T = \bar{X} - \mu_0$ and $\theta = \mu - \mu_0$. Show that $EPV(\theta) = \Phi(-\sqrt{n}\theta/\sqrt{2}\sigma)$, where $\Phi$ denotes the standard normal distribution. (For a recent review of expected $p$ values see Sackrowitz and Samuel–Cahn, 1999.)

## Problems for Section 4.2

**1.** Consider Examples 3.3.2 and 4.2.1. You want to buy one of two systems. One has signal-to-noise ratio $v/\sigma_0 = 2$, the other has $v/\sigma_0 = 1$. The first system costs $\$10^6$, the other $\$10^5$. One second of transmission on either system costs $\$10^3$ each. Whichever system you buy during the year, you intend to test the satellite 100 times. If each time you test, you want the number of seconds of response sufficient to ensure that both probabilities of error are $\leq 0.05$, which system is cheaper on the basis of a year's operation?

**2.** Consider a population with three kinds of individuals labeled 1, 2, and 3 occuring in the Hardy–Weinberg proportions $f(1, \theta) = \theta^2$, $f(2, \theta) = 2\theta(1 - \theta)$, $f(3, \theta) = (1 - \theta)^2$. For a sample $X_1, \ldots, X_n$ from this population, let $N_1$, $N_2$, and $N_3$ denote the number of $X_j$ equal to 1, 2, and 3, respectively. Let $0 < \theta_0 < \theta_1 < 1$.

**(a)** Show that $L(\mathbf{x}, \theta_0, \theta_1)$ is an increasing function of $2N_1 + N_2$.

**(b)** Show that if $c > 0$ and $\alpha \in (0, 1)$ satisfy $P_{\theta_0}[2N_1 + N_2 \geq c] = \alpha$, then the test that rejects $H$ if, and only if, $2N_1 + N_2 \geq c$ is MP for testing $H : \theta = \theta_0$ versus $K : \theta = \theta_1$.

**3.** A gambler observing a game in which a single die is tossed repeatedly gets the impression that 6 comes up about 18% of the time, 5 about 14% of the time, whereas the other four numbers are equally likely to occur (i.e., with probability .17). Upon being asked to play, the gambler asks that he first be allowed to test his hypothesis by tossing the die $n$ times.

**(a)** What test statistic should he use if the only alternative he considers is that the die is fair?

**(b)** Show that if $n = 2$ the most powerful level .0196 test rejects if, and only if, two 5's are obtained.

**(c)** Using the fact that if $(N_1, \ldots, N_k) \sim \mathcal{M}(n, \theta_1, \ldots, \theta_k)$, then $a_1N_1 + \cdots + a_kN_k$ has approximately a $\mathcal{N}(n\mu, n\sigma^2)$ distribution, where $\mu = \sum_{i=1}^{k} a_i\theta_i$ and $\sigma^2 = \sum_{i=1}^{k} \theta_i(a_i - \mu)^2$, find an approximation to the critical value of the MP level $\alpha$ test for this problem.

**4.** A formulation of goodness of fit tests specifies that a test is best if the maximum probability of error (of either type) is as small as possible.

**(a)** Show that if in testing $H : \theta = \theta_0$ versus $K : \theta = \theta_1$ there exists a critical value $c$ such that

$$P_{\theta_0}[L(\mathbf{X}, \theta_0, \theta_1) \geq c] = 1 - P_{\theta_1}[L(\mathbf{X}, \theta_0, \theta_1) \geq c]$$

then the likelihood ratio test with critical value $c$ is best in this sense.

**(b)** Find the test that is best in this sense for Example 4.2.1.

**5.** A newly discovered skull has cranial measurements $(X, Y)$ known to be distributed either (as in population 0) according to $\mathcal{N}(0, 0, 1, 1, 0.6)$ or (as in population 1) according to $\mathcal{N}(1, 1, 1, 1, 0.6)$ where all parameters are known. Find a statistic $T(X, Y)$ and a critical value $c$ such that if we use the classification rule, $(X, Y)$ belongs to population 1 if $T \geq c$, and to population 0 if $T < c$, then the maximum of the two *probabilities of misclassification* $P_0[T \geq c]$, $P_1[T < c]$ is as small as possible.

*Hint:* Use Problem 4.2.4 and recall (Proposition B.4.2) that linear combinations of bivariate normal random variables are normally distributed.

**6.** Show that if randomization is permitted, MP-sized $\alpha$ likelihood ratio tests based on $X_1, \ldots, X_n$ with $0 < \alpha < 1$ have power nondecreasing in $n$.

**7.** Prove Corollary 4.2.1.

*Hint:* The MP test has power at least that of the test with test function $\delta(x) = \alpha$.

**8.** In Examle 4.2.2, derive the UMP test defined by (4.2.7).

**9.** In Example 4.2.2, if $\mathbf{\Delta}_0 = (1, 0, \ldots, 0)^T$ and $\Sigma_0 \neq I$, find the MP test for testing $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $K : \boldsymbol{\theta} = \boldsymbol{\theta}_1$.

**10.** For $0 < \alpha < 1$, prove Theorem 4.2.1(a) using the connection between likelihood ratio tests and Bayes tests given in Remark 4.2.1.

## Problems for Section 4.3

**1.** Let $X_i$ be the number of arrivals at a service counter on the $i$th of a sequence of $n$ days. A possible model for these data is to assume that customers arrive according to a homogeneous Poisson process and, hence, that the $X_i$ are a sample from a Poisson distribution with parameter $\theta$, the expected number of arrivals per day. Suppose that if $\theta \le \theta_0$ it is not worth keeping the counter open.

**(a)** Exhibit the optimal (UMP) test statistic for $H : \theta \le \theta_0$ versus $K : \theta > \theta_0$.

**(b)** For what levels can you exhibit a UMP test?

**(c)** What distribution tables would you need to calculate the power function of the UMP test?

**2.** Consider the foregoing situation of Problem 4.3.1. You want to ensure that if the arrival rate is $\le 10$, the probability of your deciding to stay open is $\le 0.01$, but if the arrival rate is $\ge 15$, the probability of your deciding to close is also $\le 0.01$. How many days must you observe to ensure that the UMP test of Problem 4.3.1 achieves this? (Use the normal approximation.)

**3.** In Example 4.3.4, show that the power of the UMP test can be written as

$$\beta(\sigma) = G_n(\sigma_0^2 x_n(\alpha)/\sigma^2)$$

where $G_n$ denotes the $\chi_n^2$ distribution function.

**4.** Let $X_1, \ldots, X_n$ be the times in months until failure of $n$ similar pieces of equipment. If the equipment is subject to wear, a model often used (see Barlow and Proschan, 1965) is the one where $X_1, \ldots, X_n$ is a sample from a Weibull distribution with density $f(x, \lambda) = \lambda c x^{c-1} e^{-\lambda x^c}$, $x > 0$. Here $c$ is a known positive constant and $\lambda > 0$ is the parameter of interest.

**(a)** Show that $\sum_{i=1}^n X_i^c$ is an optimal test statistic for testing $H : 1/\lambda \le 1/\lambda_0$ versus $K : 1/\lambda > 1/\lambda_0$.

**(b)** Show that the critical value for the size $\alpha$ test with critical region $[\sum_{i=1}^n X_i^c \ge k]$ is $k = x_{2n}(1-\alpha)/2\lambda_0$ where $x_{2n}(1-\alpha)$ is the $(1-\alpha)$th quantile of the $\chi_{2n}^2$ distribution and that the power function of the UMP level $\alpha$ test is given by

$$1 - G_{2n}(\lambda x_{2n}(1-\alpha)/\lambda_0)$$

where $G_{2n}$ denotes the $\chi_{2n}^2$ distribution function.
*Hint:* Show that $X_i^c \sim \mathcal{E}(\lambda)$.

**(c)** Suppose $1/\lambda_0 = 12$. Find the sample size needed for a level $0.01$ test to have power at least $0.95$ at the alternative value $1/\lambda_1 = 15$. Use the normal approximation to the critical value and the probability of rejection.

**5.** Show that if $X_1, \ldots, X_n$ is a sample from a *truncated binomial* distribution with

$$p(x, \theta) = \binom{k}{x} \theta^x (1-\theta)^{k-x} / [1 - (1-\theta)^k],\ x = 1, \ldots, k,$$

then $\sum_{i=1}^{n} X_i$ is an optimal test statistic for testing $H : \theta = \theta_0$ versus $K : \theta > \theta_0$.

**6.** Let $X_1, \ldots, X_n$ denote the incomes of $n$ persons chosen at random from a certain population. Suppose that each $X_i$ has the Pareto density

$$f(x, \theta) = c^\theta \theta x^{-(1+\theta)}, \; x > c$$

where $\theta > 1$ and $c > 0$.

**(a)** Express mean income $\mu$ in terms of $\theta$.

**(b)** Find the optimal test statistic for testing $H : \mu = \mu_0$ versus $K : \mu > \mu_0$.

**(c)** Use the central limit theorem to find a normal approximation to the critical value of test in part (b).

*Hint:* Use the results of Theorem 1.6.2 to find the mean and variance of the optimal test statistic.

**7.** In the goodness-of-fit Example 4.1.5, suppose that $F_0(x)$ has a nonzero density on some interval $(a, b)$, $-\infty \leq a < b \leq \infty$, and consider the alternative with distribution function $F(x, \theta) = F_0^\theta(x)$, $0 < \theta < 1$. Show that the UMP test for testing $H : \theta \geq 1$ versus $K : \theta < 1$ rejects $H$ if $-2\Sigma \log F_0(X_i) \geq x_{1-\alpha}$, where $x_{1-\alpha}$ is the $(1-\alpha)$th quantile of the $\chi^2_{2n}$ distribution. (See Problem 4.1.6.) It follows that Fisher's method for combining $p$-values (see 4.1.6) is UMP for testing that the $p$-values are uniformly distributed against $F(u) = u^\theta, 0 < \theta < 1$.

**8.** Let the distribution of survival times of patients receiving a standard treatment be the known distribution $F_0$, and let $Y_1, \ldots, Y_n$ be the i.i.d. survival times of a sample of patients receiving an experimental treatment.

**(a)** *Lehmann Alternative.* In Problem 1.1.12, we derived the model

$$G(y, \Delta) = 1 - [1 - F_0(y)]^{\Delta^{-1}}, \; y > 0, \; \Delta > 0.$$

To test whether the new treatment is beneficial we test $H : \Delta \leq 1$ versus $K : \Delta > 1$. Assume that $F_0$ has a density $f_0$. Find the UMP test. Show how to find critical values.

**(b)** *Nabeya–Miura Alternative.* For the purpose of modeling, imagine a sequence $X_1, X_2, \ldots$ of i.i.d. survival times with distribution $F_0$. Let $N$ be a zero-truncated Poisson, $\mathcal{P}(\lambda)$, random variable, which is independent of $X_1, X_2, \ldots$.

(i) Show that if we model the distribution of $Y$ as $\mathcal{L}(\max\{X_1, \ldots, X_N\})$, then

$$P(Y \leq y) = \frac{e^{\lambda F_0(y)} - 1}{e^\lambda - 1}, \; y > 0, \; \lambda \geq 0.$$

(ii) Show that if we model the distribution of $Y$ as $\mathcal{L}(\min\{X_1, \ldots, X_N\})$, then

$$P(Y \leq y) = \frac{e^{-\lambda F_0(y)} - 1}{e^{-\lambda} - 1}, \; y > 0, \; \lambda \geq 0.$$

(iii) Consider the model

$$\begin{aligned} G(y,\theta) &= \frac{e^{\theta F_0(y)}-1}{e^{\theta}-1},\ \theta \neq 0 \\ &= F_0(y),\ \theta = 0. \end{aligned}$$

To see whether the new treatment is beneficial, we test $H : \theta \leq 0$ versus $K : \theta > 0$. Assume that $F_0$ has a density $f_0(y)$. Show that the UMP test is based on the statistic $\sum_{i=1}^{n} F_0(Y_i)$.

**9.** Let $X_1, \dots, X_n$ be i.i.d. with distribution function $F(x)$. We want to test whether $F$ is exponential, $F(x) = 1 - \exp(-x)$, $x > 0$, or Weibull, $F(x) = 1 - \exp(-x^{\theta})$, $x > 0$, $\theta > 0$. Find the MP test for testing $H : \theta = 1$ versus $K : \theta = \theta_1 > 1$. Show that the test is not UMP.

**10.** Show that under the assumptions of Theorem 4.3.2 the class of all Bayes tests is complete.

*Hint:* Consider the class of all Bayes tests of $H : \theta = \theta_0$ versus $K : \theta = \theta_1$ where $\pi\{\theta_0\} = 1 - \pi\{\theta_1\}$ varies between 0 and 1.

**11.** Show that under the assumptions of Theorem 4.3.1 and 0-1 loss, every Bayes test for $H : \theta \leq \theta_0$ versus $K : \theta > \theta_1$ is of the form $\delta_t$ for some $t$.

*Hint:* A Bayes test rejects (accepts) $H$ if

$$\int_{\theta_1}^{\infty} p(x,\theta)d\pi(\theta) / \int_{-\infty}^{\theta_0} p(x,\theta)d\pi(\theta) \overset{>}{(<)} 1.$$

The left-hand side equals

$$\frac{\int_{\theta_1}^{\infty} L(x,\theta,\theta_0)d\pi(\theta)}{\int_{-\infty}^{\theta_0} L(x,\theta,\theta_0)d\pi(\theta)}.$$

The numerator is an increasing function of $T(x)$, the denominator decreasing.

**12.** Show that under the assumptions of Theorem 4.3.1, $1 - \delta_t$ is UMP for testing $H : \theta \geq \theta_0$ versus $K : \theta < \theta_0$.

## Problems for Section 4.4

**1.** Let $X_1, \dots, X_n$ be a sample from a normal population with unknown mean $\mu$ and unknown variance $\sigma^2$. Using a pivot based on $\Sigma_{i=1}^{n}(X_i - \bar{X})^2$,

**(a)** Show how to construct level $(1-\alpha)$ confidence intervals of fixed finite length for $\log \sigma^2$.

**(b)** Suppose that $\Sigma_{i=1}^{n}(X_i - \bar{X})^2 = 16.52$, $n = 2$, $\alpha = 0.01$. What would you announce as your level $(1-\alpha)$ UCB for $\sigma^2$?

**2.** Let $X_i = (\theta/2)t_i^2 + \epsilon_i$, $i = 1, \dots, n$, where the $\epsilon_i$ are independent normal random variables with mean 0 and known variance $\sigma^2$ (cf. Problem 2.2.1).

**(a)** Using a pivot based on the MLE $(2\Sigma_{i=1}^n t_i^2 X_i)/\Sigma_{i=1}^n t_i^4$ of $\theta$, find a fixed length level $(1-\alpha)$ confidence interval for $\theta$.

**(b)** If $0 \le t_i \le 1$, $i = 1, \dots, n$, but we may otherwise choose the $t_i$ freely, what values should we use for the $t_i$ so as to make our interval as short as possible for given $\alpha$?

**3.** Let $X_1, \dots, X_n$ be as in Problem 4.4.1. Suppose that an experimenter thinking he knows the value of $\sigma^2$ uses a lower confidence bound for $\mu$ of the form $\underline{\mu}(\mathbf{X}) = \bar{X} - c$, where $c$ is chosen so that the confidence level under the assumed value of $\sigma^2$ is $1-\alpha$. What is the actual confidence coefficient of $\underline{\mu}$, if $\sigma^2$ can take on all positive values?

**4.** Suppose that in Example 4.4.3 we know that $\theta \le 0.1$.

**(a)** Justify the interval $[\underline{\theta}, \min(\bar{\theta}, 0.1)]$ if $\underline{\theta} < 0.1$, $[0.1, 0.1]$ if $\underline{\theta} \ge 0.1$, where $\underline{\theta}, \bar{\theta}$ are given by $(4.4.3)$.

**(b)** Calculate the smallest $n$ needed to bound the length of the 95% interval of part (a) by $0.02$. Compare your result to the $n$ needed for $(4.4.3)$.

**5.** Show that if $\underline{q}(\mathbf{X})$ is a level $(1-\alpha_1)$ LCB and $\bar{q}(\mathbf{X})$ is a level $(1-\alpha_2)$ UCB for $q(\theta)$, then $[\underline{q}(\mathbf{X}), \bar{q}(\mathbf{X})]$ is a level $(1-(\alpha_1+\alpha_2))$ confidence interval for $q(\theta)$. (Define the interval arbitrarily if $\underline{q} > \bar{q}$.)

*Hint:* Use (A.2.7).

**6.** Show that if $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ and $\alpha_1 + \alpha_2 \le \alpha$, then the shortest level $(1-\alpha)$ interval of the form

$$\left[\bar{X} - z(1-\alpha_1)\frac{\sigma}{\sqrt{n}},\ \bar{X} + z(1-\alpha_2)\frac{\sigma}{\sqrt{n}}\right]$$

is obtained by taking $\alpha_1 = \alpha_2 = \alpha/2$ (assume $\sigma^2$ known).

*Hint:* Reduce to $\alpha_1+\alpha_2 = \alpha$ by showing that if $\alpha_1+\alpha_2 < \alpha$, there is a shorter interval with $\alpha_1 + \alpha_2 = \alpha$. Use calculus.

**7.** Suppose we want to select a sample size $N$ such that the interval $(4.4.1)$ based on $n = N$ observations has length at most $l$ for some preassigned length $l = 2d$. Stein's (1945) two-stage procedure is the following. Begin by taking a fixed number $n_0 \ge 2$ of observations and calculate $\bar{X}_0 = (1/n_0)\Sigma_{i=1}^{n_0} X_i$ and

$$s_0^2 = (n_0-1)^{-1}\Sigma_{i=1}^{n_0}(X_i - \bar{X}_0)^2.$$

Then take $N - n_0$ further observations, with $N$ being the smallest integer greater than $n_0$ and greater than or equal to

$$\left[s_0 t_{n_0-1}\left(1-\tfrac{1}{2}\alpha\right)/d\right]^2.$$

Show that, although $N$ is random, $\sqrt{N}(\bar{X}-\mu)/s_0$, with $\bar{X} = \Sigma_{i=1}^N X_i/N$, has a $\mathcal{T}_{n_0-1}$ distribution. It follows that

$$\left[\bar{X} - s_0 t_{n_0-1}\left(1-\tfrac{1}{2}\alpha\right)/\sqrt{N},\ \bar{X} + s_0 t_{n_0-1}\left(1-\tfrac{1}{2}\alpha\right)/\sqrt{N}\right]$$

is a confidence interval with confidence coefficient $(1-\alpha)$ for $\mu$ of length at most $2d$. (The sticky point of this approach is that we have no control over $N$, and, if $\sigma$ is large, we may very likely be forced to take a prohibitively large number of observations. The reader interested in pursuing the study of sequential procedures such as this one is referred to the book of Wetherill and Glazebrook, 1986, and the fundamental monograph of Wald, 1947.)

*Hint:* Note that $\bar{X} = (n_0/N)\bar{X}_0 + (1/N)\Sigma_{i=n_0+1}^{N} X_i$. By Theorem B.3.3, $s_0$ is independent of $\bar{X}_0$. Because $N$ depends only on $s_0$, given $N = k$, $\bar{X}$ has a $\mathcal{N}(\mu, \sigma^2/k)$ distribution. Hence, $\sqrt{N}(\bar{X}-\mu)$ has a $\mathcal{N}(0,\sigma^2)$ distribution and is independent of $s_0$.

**8.** **(a)** Show that in Problem 4.4.6, in order to have a level $(1-\alpha)$ confidence interval of length at most $2d$ when $\sigma^2$ is known, it is necessary to take at least $z^2\left(1-\frac{1}{2}\alpha\right)\sigma^2/d^2$ observations.

*Hint:* Set up an inequality for the length and solve for $n$.

**(b)** What would be the minimum sample size in part (a) if $\alpha = 0.001$, $\sigma^2 = 5$, $d = 0.05$?

**(c)** Suppose that $\sigma^2$ is not known exactly, but we are sure that $\sigma^2 \le \sigma_1^2$. Show that $n \ge z^2\left(1-\frac{1}{2}\alpha\right)\sigma_1^2/d^2$ observations are necessary to achieve the aim of part (a).

**9.** Let $S \sim \mathcal{B}(n,\theta)$ and $\bar{X} = S/n$.

**(a)** Use (A.14.18) to show that $\sin^{-1}(\sqrt{\bar{X}}) \pm z\left(1-\frac{1}{2}\alpha\right)/\,2\sqrt{n}$ is an approximate level $(1-\alpha)$ confidence interval for $\sin^{-1}(\sqrt{\theta})$.

**(b)** If $n = 100$ and $\bar{X} = 0.1$, use the result in part (a) to compute an approximate level 0.95 confidence interval for $\theta$.

**10.** Let $X_1,\dots,X_{n_1}$ and $Y_1,\dots,Y_{n_2}$ be two independent samples from $\mathcal{N}(\mu,\sigma^2)$ and $\mathcal{N}(\eta,\tau^2)$ populations, respectively.

**(a)** If all parameters are unknown, find ML estimates of $\mu$, $\nu$, $\sigma^2$, $\tau^2$. Show that this quadruple is sufficient.

**(b)** Exhibit a level $(1-\alpha)$ confidence interval for $\tau^2/\sigma^2$ using a pivot based on the statistics of part (a). Indicate what tables you would need to calculate the interval.

**(c)** If $\sigma^2, \tau^2$ are known, exhibit a fixed length level $(1-\alpha)$ confidence interval for $(\eta-\mu)$.

Such two sample problems arise in comparing the precision of two instruments and in determining the effect of a treatment.

**11.** Show that the endpoints of the approximate level $(1-\alpha)$ interval defined by (4.4.3) are indeed approximate level $\left(1-\frac{1}{2}\alpha\right)$ upper and lower bounds.

*Hint:* $[\underline{\theta}(\mathbf{X}) \le \theta] = \left[\sqrt{n}(\bar{X}-\theta)/[\theta(1-\theta)]^{\frac{1}{2}} \le z\left(1-\frac{1}{2}\alpha\right)\right]$.

**12.** Let $S \sim \mathcal{B}(n,\theta)$. Suppose that it is known that $\theta \le \frac{1}{4}$.

**(a)** Show that $\bar{X} \pm \sqrt{3}z\left(1-\frac{1}{2}\alpha\right)/\,4\sqrt{n}$ is an approximate level $(1-\alpha)$ confidence interval for $\theta$.

**(b)** What sample size is needed to guarantee that this interval has length at most $0.02$?

**13.** Suppose that a new drug is tried out on a sample of 64 patients and that $S = 25$ cures are observed. If $S \sim \mathcal{B}(64, \theta)$, give a 95% confidence interval for the true proportion of cures $\theta$ using (a) $(4.4.3)$, and (b) $(4.4.7)$.

**14.** Suppose that 25 measurements on the breaking strength of a certain alloy yield $\bar{x} = 11.1$ and $s = 3.4$. Assuming that the sample is from a $\mathcal{N}(\mu, \sigma^2)$ population, find

**(a)** A level $0.9$ confidence interval for $\mu$.

**(b)** A level $0.9$ confidence interval for $\sigma$.

**(c)** A level $0.9$ confidence region for $(\mu, \sigma)$.

**(d)** A level $0.9$ confidence interval for $\mu + \sigma$.

**15.** Show that the confidence coefficient of the rectangle of Example $4.4.5$ is

$$\int_a^b 2\left[\Phi\left(\sqrt{\frac{1}{n-1}}\sqrt{\tau}\cdot c\right) - 1\right] g(\tau)d\tau$$

where $a$ is the $\frac{\alpha}{4}$th quantile of the $\chi^2_{n-1}$ distribution; $b$ is the $(1-\frac{\alpha}{4})$th quantile of the $\chi^2_{n-1}$ distribution; $c$ is the $(1-\frac{\alpha}{4})$th quantile of $t$-distribution; and $g(\cdot)$ is the density function of the $\chi^2_{n-1}$ distribution.

*Hint:* Use Theorem B.3.3 and (B.1.30). Condition on $[(n-1)s^2/\sigma^2] = \tau$.

**16.** In Example 4.4.2,

**(a)** Show that $x(\alpha_1)$ and $x(1-\alpha_2)$ can be approximated by $x(\alpha_1) \cong (n-1) + \sqrt{2}(n-1)^{\frac{1}{2}}z(\alpha_1)$ and $x(1-\alpha_2) \cong (n-1) + \sqrt{2}(n-1)^{\frac{1}{2}}z(1-\alpha_2)$.

*Hint:* By B.3.1, $V(\sigma^2)$ can be written as a sum of squares of $n-1$ independent $\mathcal{N}(0,1)$ random variables, $\sum_{i=1}^{n-1} Z_i^2$. Now use the central limit theorem.

**(b)** Suppose that $X_i$ does not necessarily have a normal distribution, but assume that $\mu_4 = E(X_i - \mu)^4 < \infty$ and that $\kappa = \text{Var}[(X_i - \mu)/\sigma]^2 = (\mu_4/\sigma^4) - 1$ is known. Find the limit of the distribution of $n^{-\frac{1}{2}}\{[(n-1)s^2/\sigma^2] - n\}$ and use this distribution to find an approximate $1-\alpha$ confidence interval for $\sigma^2$. (In practice, $\kappa$ is replaced by its MOM estimate. See Problem $5.3.30$.)

*Hint:* $(n-1)s^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2$. Now use the law of large numbers, Slutsky's theorem, and the central limit theorem as given in Appendix A.

**(c)** Suppose $X_i$ has a $\chi^2_k$ distribution. Compute the ($\kappa$ known) confidence intervals of part (b) when $k = 1, 10, 100$, and $10\,000$. Compare them to the approximate interval given in part (a). $\kappa - 2$ is known as the *kurtosis coefficient*. In the case where $X_i$ is normal, it equals $0$. See A.11.11.

*Hint:* Use Problem B.2.4 and the fact that $\chi^2_k = \Gamma\left(k, \frac{1}{2}\right)$.

**17.** Consider Example 4.4.6 with $x$ fixed. That is, we want a level $(1-\alpha)$ confidence interval for $F(x)$. In this case $n\widehat{F}(x) = \#[X_i \le x]$ has a binomial distribution and

$$\frac{\sqrt{n}|\widehat{F}(x)-F(x)|}{\sqrt{F(x)[1-F(x)]}}$$

is the approximate pivot given in Example 4.4.3 for deriving a confidence interval for $\theta = F(x)$.

**(a)** For $0 < a < b < 1$, define

$$A_n(F) = \sup\left\{\frac{\sqrt{n}|\widehat{F}(x)-F(x)|}{\sqrt{F(x)[1-F(x)]}},\ F^{-1}(a) \le x \le F^{-1}(b)\right\}.$$

Typical choices of $a$ and $b$ are .05 and .95. Show that for $F$ continuous

$$P_F(A_n(F) \le t) = P_U(A_n(U) \le t)$$

where $U$ denotes the uniform, $\mathcal{U}(0,1)$, distribution function. It follows that the binomial confidence intervals for $\theta$ in Example 4.4.3 can be turned into simultaneous confidence intervals for $F(x)$ by replacing $z\left(1-\frac{1}{2}\alpha\right)$ by the value $u_\alpha$ determined by $P_U(A_n(U) \le u) = 1-\alpha$.

**(b)** For $0 < a < b < 1$, define

$$B_n(F) = \sup\left\{\frac{\sqrt{n}|\widehat{F}(x)-F(x)|}{\sqrt{\widehat{F}(x)[1-\widehat{F}(x)]}},\ \widehat{F}^{-1}(a) \le x \le \widehat{F}^{-1}(b)\right\}.$$

Show that for $F$ continuous,

$$P_F(B_n(F) \le t) = P_U(B_n(U) \le t).$$

**(c)** For testing $H_0 : F = F_0$ with $F_0$ continuous, indicate how critical values $u_\alpha$ and $t_\alpha$ for $A_n(F_0)$ and $B_n(F_0)$ can be obtained using the Monte Carlo method of Section 4.1.

**18.** Suppose $X_1, \ldots, X_n$ are i.i.d. as $X$ and that $X$ has density $f(t) = F'(t)$. Assume that $f(t) > 0$ iff $t \in (a,b)$ for some $-\infty < a \le 0 < b < \infty$.

**(a)** Show that $\mu = -\int_a^0 F(x)dx + \int_0^b [1-F(x)]dx$.

**(b)** Using Example 4.4.6, find a level $(1-\alpha)$ confidence interval for $\mu$.

**19.** In Example 4.4.7, verify the lower bounary $\underline{\mu}$ given by (4.4.9) and the upper boundary $\bar{\mu} = \infty$.

## Problems for Section 4.5

**1.** Let $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ be independent exponential $\mathcal{E}(\theta)$ and $\mathcal{E}(\lambda)$ samples, respectively, and let $\Delta = \theta/\lambda$.

**(a)** If $f(\alpha)$ denotes the $\alpha$th quantile of the $\mathcal{F}_{2n_1,2n_2}$ distribution, show that $[\bar{Y}f(\frac{1}{2}\alpha)/\bar{X}, \bar{Y}f(1-\frac{1}{2}\alpha)/\bar{X}]$ is a confidence interval for $\Delta$ with confidence coefficient $1-\alpha$.

*Hint:* Use the results of Problems B.3.4 and B.3.5.

**(b)** Show that the test with acceptance region $[f(\frac{1}{2}\alpha) \leq \bar{X}/\bar{Y} \leq f(1-\frac{1}{2}\alpha)]$ has size $\alpha$ for testing $H : \Delta = 1$ versus $K : \Delta \neq 1$.

**(c)** The following are times until breakdown in days of air monitors operated under two different maintenance policies at a nuclear power plant. Experience has shown that the exponential assumption is warranted. Give a $90\%$ confidence interval for the ratio $\Delta$ of mean life times.

| | |
|---|---|
| $x$ | 3 150 40 34 32 37 34 2 31 6 5 14 150 27 4 6 27 10 30 37 |
| $y$ | 8 26 10 8 29 20 10 |

Is $H : \Delta = 1$ rejected at level $\alpha = 0.10$?

**2.** Show that if $\bar{\theta}(\mathbf{X})$ is a level $(1-\alpha)$ UCB for $\theta$, then the test that accepts, if and only if $\bar{\theta}(\mathbf{X}) \geq \theta_0$, is of level $\alpha$ for testing $H : \theta \geq \theta_0$.

*Hint:* If $\theta > \theta_0$, $[\bar{\theta}(\mathbf{X}) < \theta] \supset [\bar{\theta}(\mathbf{X}) < \theta_0]$.

**3. (a)** Deduce from Problem 4.5.2 that the tests of $H : \sigma^2 = \sigma_0^2$ based on the level $(1-\alpha)$ UCBs of Example 4.4.2 are level $\alpha$ for $H : \sigma^2 \geq \sigma_0^2$.

**(b)** Give explicitly the power function of the test of part (a) in terms of the $\chi^2_{n-1}$ distribution function.

**(c)** Suppose that $n = 16, \alpha = 0.05, \sigma_0^2 = 1$. How small must an alternative $\sigma^2$ be before the size $\alpha$ test given in part (a) has power 0.90?

**4. (a)** Find $c$ such that $\delta_c$ of Problem 4.1.1 has size $\alpha$ for $H : \theta \leq \theta_0$.

**(b)** Derive the level $(1-\alpha)$ LCB corresponding to $\delta_c$ of part (a).

**(c)** Similarly derive the level $(1-\alpha)$ UCB for this problem and exhibit the confidence intervals obtained by putting two such bounds of level $(1-\alpha_1)$ and $(1-\alpha_2)$ together.

**(d)** Show that $[M_n, M_n/\alpha^{1/n}]$ is the shortest such confidence interval.

**5.** Let $X_1, X_2$ be independent $\mathcal{N}(\theta_1, \sigma^2), \mathcal{N}(\theta_2, \sigma^2)$, respectively, and consider the problem of testing $H : \theta_1 = \theta_2 = 0$ versus $K : \theta_1^2 + \theta_2^2 > 0$ when $\sigma^2$ is known.

**(a)** Let $\delta_c(X_1, X_2) = 1$ if and only if $X_1^2 + X_2^2 \geq c$. What value of $c$ gives size $\alpha$?

**(b)** Using Problems B.3.12 and B.3.13 show that the power $\beta(\theta_1, \theta_2)$ is an increasing function of $\theta_1^2 + \theta_2^2$.

**(c)** Modify the test of part (a) to obtain a procedure that is level $\alpha$ for $H : \theta_1 = \theta_1^0, \theta_2 = \theta_2^0$ and exhibit the corresponding family of confidence circles for $(\theta_1, \theta_2)$.

*Hint:* (c) $X_1 - \theta_1^0, X_2 - \theta_2^0$ are independent $\mathcal{N}(\theta_1 - \theta_1^0, \sigma^2), \mathcal{N}(\theta_2 - \theta_2^0, \sigma^2)$, respectively.

**6.** Let $X_1, \ldots, X_n$ be a sample from a population with density $f(t - \theta)$ where $\theta$ and $f$ are unknown, but $f(t) = f(-t)$ for all $t$, and $f$ is continuous and positive. Thus, we have a location parameter family.

**(a)** Show that testing $H : \theta \le 0$ versus $K : \theta > 0$ is equivalent to testing

$$H' : P[X_1 \ge 0] \le \frac{1}{2} \text{ versus } K' : P[X_1 \ge 0] > \frac{1}{2}.$$

**(b)** The *sign test* of $H$ versus $K$ is given by,

$$\begin{aligned} \delta_k(X) &= 1 \text{ if } \left(\sum_{i=1}^n 1[X_i \ge 0]\right) \ge k \\ &= 0 \text{ otherwise.} \end{aligned}$$

Determine the smallest value $k = k(\alpha)$ such that $\delta_{k(\alpha)}$ is level $\alpha$ for $H$ and show that for $n$ large, $k \cong \frac{1}{2}n + \frac{1}{2}z(1-\alpha)\sqrt{n}$.

**(c)** Show that $\delta_{k(\alpha)}(X_1 - \theta_0, \ldots, X_n - \theta_0)$ is a level $\alpha$ test of $H : \theta \le \theta_0$ versus $K : \theta > \theta_0$.

**(d)** Deduce that $X_{(n-k(\alpha)+1)}$ (where $X_{(j)}$ is the $j$th order statistic of the sample) is a level $(1-\alpha)$ LCB for $\theta$ *whatever be $f$ satisfying our conditions.*

**(e)** Show directly that $P_\theta[X_{(j)} \le \theta]$ and $P_\theta[X_{(j)} \le \theta \le X_{(k)}]$ do not depend on $f$ or $\theta$.

**(f)** Suppose that $\alpha = 2^{-(n-1)} \sum_{j=0}^{k-1} \binom{n}{j}$. Show that $P[X_{(k)} \le \theta \le X_{(n-k+1)}] = 1 - \alpha$.

**(g)** Suppose that we drop the assumption that $f(t) = f(-t)$ for all $t$ and replace our assumptions by: $X_1, \ldots, X_n$ is a sample from a population with density $f$ and median $\nu$. Show that if we replace $\theta$ by $\nu$, the conclusions of (a)–(f) still hold.

**7.** Suppose $\theta = (\eta, \tau)$ where $\eta$ is a parameter of interest and $\tau$ is a nuisance parameter. We are given for each possible value $\eta_0$ of $\eta$ a level $\alpha$ test $\delta(\mathbf{X}, \eta_0)$ of the composite hypothesis $H : \eta = \eta_0$. Let $C(\mathbf{X}) = \{\eta : \delta(\mathbf{X}, \eta) = 0\}$.

**(a)** Show that $C(\mathbf{X})$ is a level $(1-\alpha)$ confidence region for the parameter $\eta$ and conversely that any level $(1-\alpha)$ confidence region for $\eta$ is equivalent to a family of level $\alpha$ tests of these composite hypotheses.

**(b)** Find the family of tests corresponding to the level $(1-\alpha)$ confidence interval for $\mu$ of Example 4.4.1 when $\sigma^2$ is unknown.

**8.** Suppose $X, Y$ are independent and $X \sim \mathcal{N}(\nu, 1), Y \sim \mathcal{N}(\eta, 1)$. Let $\rho = \nu/\eta, \theta = (\rho, \eta)$. Define

$$\begin{aligned}\delta(X, Y, \rho) &= 0 \text{ if } |X - \rho Y| \leq (1+\rho^2)^{\frac{1}{2}} z(1 - \frac{1}{2}\alpha) \\ &= 1 \text{ otherwise.}\end{aligned}$$

**(a)** Show that $\delta(X, Y, \rho_0)$ is a size $\alpha$ test of $H : \rho = \rho_0$.

**(b)** Describe the confidence region obtained by inverting the family $\{\delta(X, Y, \rho)\}$ as in Problem 4.5.7. Note that the region is not necessarily an interval or ray. This problem is a simplified version of that encountered in putting a confidence interval on the zero of a regression line.

**9.** Let $X \sim \mathcal{N}(\theta, 1)$ and $q(\theta) = \theta^2$.

**(a)** Show that the lower confidence bound for $q(\theta)$ obtained from the image under $q$ of the ray $(X - z(1-\alpha), \infty)$ is

$$\begin{aligned}\underline{q}(X) &= (X - z(1-\alpha))^2 \text{ if } X \geq z(1-\alpha) \\ &= 0 \text{ if } X < z(1-\alpha).\end{aligned}$$

**(b)** Show that

$$\begin{aligned}P_\theta[\underline{q}(X) \leq \theta^2] &= 1 - \alpha \text{ if } \theta \geq 0 \\ &= \Phi(z(1-\alpha) - 2\theta) \text{ if } \theta < 0\end{aligned}$$

and, hence, that $\sup_\theta P_\theta[\underline{q}(X) \leq \theta^2] = 1$.

**10.** Let $\alpha(S, \theta_0)$ denote the $p$-value of the test of $H : \theta = \theta_0$ versus $K : \theta > \theta_0$ in Example 4.1.3 and let $[\underline{\theta}(S), \bar{\theta}(S)]$ be the exact level $(1 - 2\alpha)$ confidence interval for $\theta$ of Example 4.5.2. Show that as $\theta$ ranges from $\underline{\theta}(S)$ to $\bar{\theta}(S)$, $\alpha(S, \theta)$ ranges from $\alpha$ to a value no smaller than $1 - \alpha$. Thus, if $\theta_0 < \underline{\theta}(S)$ ($S$ is inconsistent with $H : \theta = \theta_0$), the quantity $\Delta = \bar{\theta}(S) - \theta_0$ indicates how far we have to go from $\theta_0$ before the value $S$ is not at all surprising under $H$.

**11.** Establish (iii) and (iv) of Example 4.5.2.

**12.** Let $\eta$ denote a parameter of interest, let $\boldsymbol{\tau}$ denote a nuisance parameter, and let $\boldsymbol{\theta} = (\eta, \boldsymbol{\tau})$. Then the level $(1 - \alpha)$ confidence interval $[\underline{\eta}(\mathbf{x}), \bar{\eta}(\mathbf{x})]$ for $\eta$ is said to be *unbiased* if

$$P_{\boldsymbol{\theta}}[\underline{\eta}(\mathbf{X}) \leq \eta' \leq \bar{\eta}(\mathbf{X})] \leq 1 - \alpha \text{ for all } \eta' \neq \eta, \text{ all } \boldsymbol{\theta}.$$

That is, the interval is unbiased if it has larger probability of covering the true value $\eta$ than the wrong value $\eta'$. Show that the Student $t$ interval $(4.5.1)$ is unbiased.

*Hint:* You may use the result of Problem 4.5.7.

**13.** *Distribution-free Confidence Regions for Quantiles.* Let $X_1, \ldots, X_n$ be a sample from a population with continuous and increasing distribution $F$. Let $x_p = F^{-1}(p)$, $0 < p < 1$, be the $p$th quantile of $F$. (See Section 3.5.) Suppose that $p$ is specified. Thus, $x_{.95}$ could be the 95th percentile of the salaries in a certain profession, or $x_{.05}$ could be the fifth percentile of the duration time for a certain disease.

**(a)** Show that testing $H : x_p \leq 0$ versus $K : x_p > 0$ is equivalent to testing $H' : P(X \geq 0) \leq (1-p)$ versus $K' : P(X \geq 0) > (1-p)$.

**(b)** The *quantile sign test* $\delta_k$ of $H$ versus $K$ has critical region $\{\mathbf{x} : \sum_{i=1}^n 1[X_i \geq 0] \geq k\}$. Determine the smallest value $k = k(\alpha)$ such that $\delta_{k(\alpha)}$ has level $\alpha$ for $H$ and show that for $n$ large, $k(\alpha) \cong h(\alpha)$, where

$$h(\alpha) \cong n(1-p) + z_{1-\alpha}\sqrt{np(1-p)}.$$

**(c)** Let $x^*$ be a specified number with $0 < F(x^*) < 1$. Show that $\delta_k(X_1 - x^*, \ldots, X_n - x^*)$ is a level $\alpha$ test for testing $H : x_p \leq x^*$ versus $K : x_p > x^*$.

**(d)** Deduce that $X_{(n-k(\alpha)+1)}$ ($X_{(j)}$ is the $j$th order statistic of the sample) is a level $(1-\alpha)$ LCB for $x_p$ whatever be $f$ satisfying our conditions.

**(e)** Let $S$ denote a $\mathcal{B}(n,p)$ variable and choose $k$ and $l$ such that $1 - \alpha = P(k \leq S \leq n - l + 1) = \sum_{j=k}^{n-l+1} p^j(1-p)^{n-j}$. Show that $P(X_{(k)} \leq x_p \leq X_{(n-l)}) = 1 - \alpha$. That is, $(X_{(k)}, X_{(n-l)})$ is a level $(1-\alpha)$ confidence interval for $x_p$ whatever be $F$ satisfying our conditions. That is, it is distribution free.

**(f)** Show that $k$ and $l$ in part (e) can be approximated by $h\left(\frac{1}{2}\alpha\right)$ and $h\left(1 - \frac{1}{2}\alpha\right)$ where $h(\alpha)$ is given in part (b).

**(g)** Let $\widehat{F}(x)$ denote the empirical distribution. Show that the interval in parts (e) and (f) can be derived from the pivot

$$T(x_p) = \frac{\sqrt{n}[\widehat{F}(x_p) - F(x_p)]}{\sqrt{F(x_p)[1 - F(x_p)]}}.$$

*Hint:* Note that $F(x_p) = p$.

**14.** *Simultaneous Confidence Regions for Quantiles.* In Problem 13 preceding we gave a distribution-free confidence interval for the $p$th quantile $x_p$ for $p$ fixed. Suppose we want a distribution-free confidence region for $x_p$ valid for all $0 < p < 1$. We can proceed as follows. Let $F$, $\widehat{F}^-(x)$, and $\widehat{F}^+(x)$ be as in Examples 4.4.6 and 4.4.7. Then

$$P(\widehat{F}^-(x) \leq F(x) \leq \widehat{F}^+(x) \text{ for all } x \in (a,b)) = 1 - \alpha.$$

**(a)** Show that this statement is equivalent to

$$P(\underline{x}_p \leq x_p \leq \bar{x}_p \text{ for all } p \in (0,1)) = 1 - \alpha$$

where $\underline{x}_p = \sup\{x : a < x < b,\ \widehat{F}^+(x) \leq p\}$ and $\bar{x}_p = \inf\{x : a < x < b,\ \widehat{F}^-(x) \geq p\}$. That is, the desired confidence region is the band consisting of the collection of intervals $\{[\underline{x}_p, \bar{x}_p] : 0 < p < 1\}$. Note that $\underline{x}_p = -\infty$ for $p < d_\alpha$ and $\bar{x}_p = \infty$ for $p > 1 - d_\alpha$.

**(b)** Express $\underline{x}_p$ and $\bar{x}_p$ in terms of the critical value of the Kolmogorov statistic and the order statistics.

**(c)** Show how the statistic $A_n(F)$ of Problem 4.4.17(a) and (c) can be used to give another distribution-free simultaneous confidence band for $x_p$. Express the band in terms of critical values for $A_n(F)$ and the order statistics. Note the similarity to the interval in Problem 4.4.13(g) preceding.

**15.** Suppose $X$ denotes the difference between responses after a subject has been given treatments $A$ and $B$, where $A$ is a placebo. Suppose that $X$ has the continuous distribution $F$. We will write $F_X$ for $F$ when we need to distinguish it from the distribution $F_{-X}$ of $-X$. The hypothesis that $A$ and $B$ are equally effective can be expressed as $H : F_{-X}(t) = F_X(t)$ for all $t \in R$. The alternative is that $F_{-X}(t) \neq F_X(t)$ for some $t \in R$. Let $\widehat{F}_X$ and $\widehat{F}_{-X}$ be the empirical distributions based on the i.i.d. $X_1, \dots, X_n$ and $-X_1, \dots, -X_n$.

**(a)** Consider the test statistic

$$D(\widehat{F}_X, \widehat{F}_{-X}) = \max\{|\widehat{F}_X(t) - \widehat{F}_{-X}(t)| : t \in R\}.$$

Show that if $F_X$ is continuous and $H$ holds, then $D(\widehat{F}_X, \widehat{F}_{-X})$ has the same distribution as $D(\widehat{F}_U, \widehat{F}_{1-U})$, where $\widehat{F}_U$ and $\widehat{F}_{1-U}$ are the empirical distributions of $U$ and $1 - U$ with $U = F(X) \sim \mathcal{U}(0, 1)$.

*Hint:* $n\widehat{F}_X(x) = \sum_{i=1}^n 1[F_X(X_i) \leq F_X(x)] = n\widehat{F}_U(F(x))$ and

$$\begin{aligned} n\widehat{F}_{-X}(x) &= \sum_{i=1}^n 1[-X_i \leq x] = \sum_{i=1}^n 1[F_{-X}(-X_i) \leq F_{-X}(x)] \\ &= n\widehat{F}_{1-U}(F_{-X}(x)) = n\widehat{F}_{1-U}(F(x)) \text{ under } H. \end{aligned}$$

See also Example 4.1.5.

**(b)** Suppose we measure the difference between the effects of $A$ and $B$ by $\frac{1}{2}$ the difference between the quantiles of $X$ and $-X$, that is, $\nu_F(p) = \frac{1}{2}[x_p + x_{1-p}]$, where $p = F(x)$. Give a distribution-free level $(1 - \alpha)$ simultaneous confidence band for the curve $\{\nu_F(p) : 0 < p < 1\}$.

*Hint:* Let $\Delta(x) = F_{-X}^{-1}(F_X(x)) - x$, then

$$\begin{aligned} n\widehat{F}_{-X}(x + \Delta(x)) &= \sum_{i=1}^n 1[-X_i \leq x + \Delta(x)] \\ &= \sum_{i=1}^n 1[F_{-X}(-X_i) \leq F_X(x)] = n\widehat{F}_{1-U}(F_X(x)). \end{aligned}$$

Moreover, $n\widehat{F}_X(x) = \sum_{i=1}^n 1[F_X(X_i) \leq F_X(x)] = n\widehat{F}_U(F_X(x))$. It follows that if we set $F^*_{-X,\Delta}(x) = \widehat{F}_{-X}(x + \Delta(x))$, then $D(\widehat{F}_X, F^*_{-X,\Delta}) \overset{\mathcal{L}}{=} D(\widehat{F}_U, \widehat{F}_{1-U})$, and by solving $D(\widehat{F}_X, \widehat{F}_{-X,\Delta}) \leq d_\alpha$ for $\Delta$, where $d_\alpha$ is the $\alpha$th quantile of the distribution of

$D(\widehat{F}_U, \widehat{F}_{1-U})$, we get a distribution-free level $(1-\alpha)$ simultaneous confidence band for $\Delta(x) = F_{-X}^{-1}(F_X(x)) - x = -2\nu_F(F(x))$. Properties of this and other bands are given by Doksum, Fenstad and Aaberge (1977).

**(c)** *A Distribution and Parameter-Free Confidence Interval.* Let $\theta(\cdot) : \mathcal{F} \to R$, where $\mathcal{F}$ is the class of distribution functions with finite support, be a location parameter as defined in Problem 3.5.17. Let

$$\widehat{\nu}_F^- = \inf_{0<p<1} \widehat{\nu}_F^-(p),\ \widehat{\nu}_F^+ = \sup_{0<p<1} \widehat{\nu}_F^+(p)$$

where $[\widehat{\nu}_F^-(p), \widehat{\nu}_F^+(p)]$ is the band in part (b). Show that for given $F \in \mathcal{F}$, the probability is $(1-\alpha)$ that the interval $[\widehat{\nu}_F^-, \widehat{\nu}_F^+]$ contains the location set $L_F = \{\theta(F) : \theta(\cdot)$ is a location parameter$\}$ of *all* location parameter values at $F$.

*Hint:* Define $H$ by $H^{-1}(p) = \frac{1}{2}[F_X^{-1}(p) - F_X^{-1}(1-p)] = \frac{1}{2}[F_X^{-1}(p) + F_{-X}^{-1}(p)]$. Then $H$ is symmetric about zero. Also note that

$$x = H^{-1}(F(x)) - \frac{1}{2}\Delta(x) = H^{-1}(F(x)) + \nu_F(F(x)).$$

It follows that $X$ is stochastically between $X_S - \underline{\nu}_F$ and $X_S + \bar{\nu}_F$ where $X_S \equiv H^{-1}(F(X))$ has the symmetric distribution $H$. The result now follows from the properties of $\theta(\cdot)$.

**16.** As in Example 1.1.3, let $X_1, \dots, X_n$ be i.i.d. treatment $A$ (placebo) responses and let $Y_1, \dots, Y_n$ be i.i.d. treatment $B$ responses. We assume that the $X$'s and $Y$'s are independent and that they have respective continuous distributions $F_X$ and $F_Y$. To test the hypothesis $H$ that the two treatments are equally effective, we test $H : F_X(t) = F_Y(t)$ for all $t$ versus $K : F_X(t) \neq F_Y(t)$ for some $t \in R$. Let $\widehat{F}_X$ and $\widehat{F}_Y$ denote the $X$ and $Y$ empirical distributions and consider the test statistic

$$D(\widehat{F}_X, \widehat{F}_Y) = \max_{t \in R} |\widehat{F}_Y(t) - \widehat{F}_X(t)|.$$

**(a)** Show that if $H$ holds, then $D(\widehat{F}_X, \widehat{F}_Y)$ has the same distribution as $D(\widehat{F}_U, \widehat{F}_V)$, where $\widehat{F}_U$ and $\widehat{F}_V$ are independent $\mathcal{U}(0,1)$ empirical distributions.

*Hint:* $n\widehat{F}_X(t) = \sum_{i=1}^n 1[F_X(X_i) \leq F_X(t)] = n\widehat{F}_U(F_X(t))$; $n\widehat{F}_Y(t) = \sum_{i=1}^n 1[F_Y(Y_i) \leq F_Y(x_p)] = nF_V(F_X(t))$ under $H$.

**(b)** Consider the parameter $\delta_p(F_X, F_Y) = y_p - x_p$, where $x_p$ and $y_p$ are the $p$th quantiles of $F_X$ and $F_Y$. Give a distribution-free level $(1-\alpha)$ simultaneous confidence band $[\widehat{\delta}_p^-, \widehat{\delta}_p^+ : 0 < p < 1]$ for the curve $\{\delta_p(F_X, F_Y) : 0 < p < 1\}$.

*Hint:* Let $\Delta(x) = F_Y^{-1}(F_X(x)) - x$, then

$$n\widehat{F}_Y(x + \Delta(x)) = \sum_{i=1}^n 1[Y_i \leq F_Y^{-1}(F_X(x))] = \sum_{i=1}^n 1[F_Y(Y_i) \leq F_X(x)] = nF_V(F_X(x)).$$

Moreover, $n\widehat{F}_X(x) = n\widehat{F}_U(F_X(x))$. It follows that if we set $F^*_{Y,\Delta}(x) = \widehat{F}_Y(x + \Delta(x))$, then $D(\widehat{F}_X, F^*_{Y,\Delta}) \overset{\mathcal{L}}{=} D(\widehat{F}_U, \widehat{F}_V)$. Let $d_\alpha$ denote a size $\alpha$ critical value for $D(\widehat{F}_U, \widehat{F}_V)$, then by solving $D(\widehat{F}_X, F^*_{Y,\Delta}) \leq d_\alpha$ for $\Delta$, we find a distribution-free level $(1-\alpha)$ simultaneous confidence band for $\Delta(x_p) = F_X^{-1}(p) - F_Y^{-1}(p) = \delta_p(F_X, F_Y)$. Properties of this and other bands are given by Doksum and Sievers (1976).

**(c)** A parameter $\theta = \delta(\cdot,\cdot) : \mathcal{F} \times \mathcal{F} \to R$, where $\mathcal{F}$ is the class of distributions with finite support, is called a *shift parameter* if $\theta(F_X, F_{X+a}) = \theta(F_{X-a}, F_X) = a$ and

$$Y_1 \overset{st}{\geq} Y,\ X_1 \overset{st}{\geq} X \Rightarrow \theta(F_X, F_Y) \leq \theta(F_X, F_{Y_1}),\ \theta(F_X, F_Y) \geq \theta(F_{X_1}, F_Y).$$

Let $\underline{\delta} = \min_{0<p<1} \delta_p(F_X, F_Y)$ and $\bar{\delta} = \max_{0<p<1} \delta_p(F_X, F_Y)$. Show that if $\theta(\cdot,\cdot)$ is a shift parameter, then $\theta(F_X, F_Y)$ is in $[\underline{\delta}, \bar{\delta}]$.

*Hint:* Set $Y^* = X + \Delta(X)$, then $Y^* \overset{\mathcal{L}}{=} Y$, moreover $X + \underline{\delta} \leq Y^* \leq X + \bar{\delta}$. Now apply the axioms.

**(d)** Show that $E(Y) - E(X)$, $\delta_p(\cdot,\cdot)$, $0 < p < 1$, $\underline{\delta}$, and $\bar{\delta}$ are shift parameters.

**(e)** *A Distribution and Parameter-Free Confidence Interval.* Let $\widehat{\delta}^- = \min_{0<p<1} \widehat{\delta}^-(p)$, $\widehat{\delta}^+ = \max_{0<p<1} \widehat{\delta}^+(p)$. Show that for given $(F_X, F_Y) \in \mathcal{F} \times \mathcal{F}$, the probability is $(1-\alpha)$ that the interval $[\widehat{\delta}^-, \widehat{\delta}^+]$ contains the *shift parameter set* $\{\theta(F_X, F_Y) : \theta(\cdot,\cdot)$ is a shift parameter$\}$ of the values of *all* shift parameters at $(F_X, F_Y)$.

## Problems for Section 4.6

**1.** Suppose $X_1, \ldots, X_n$ is a sample from a $\Gamma\left(p, \frac{1}{\theta}\right)$ distribution, where $p$ is known and $\theta$ is unknown. Exhibit the UMA level $(1-\alpha)$ UCB for $\theta$.

**2. (a)** Consider the model of Problem 4.4.2. Show that

$$\underline{\theta}^* = (2\sum_{i=1}^n t_i^2 X_i)/\sum_{i=1}^n t_i^4 - 2z(1-\alpha)\sigma[\sum_{i=1}^n t_i^4]^{-\frac{1}{2}}$$

is a uniformly most accurate lower confidence bound for $\theta$.

**(b)** Consider the unbiased estimate of $\theta$, $T = (2\sum_{i=1}^n X_i)/\sum_{i=1}^n t_i^2$. Show that

$$\underline{\theta} = (2\sum_{i=1}^n X_i)/\sum_{i=1}^n t_i^2 - 2\sigma\sqrt{n}z(1-\alpha)/\sum_{i=1}^n t_i^2$$

is also a level $(1-\alpha)$ confidence bound for $\theta$.

**(c)** Show that the statement that $\underline{\theta}^*$ is more accurate than $\underline{\theta}$ is equivalent to the assertion that $S = (2\sum_{i=1}^n t_i^2 X_i)/\sum_{i=1}^n t_i^4$ has uniformly smaller variance than $T$.

*Hint:* Both $\underline{\theta}$ and $\underline{\theta}^*$ are normally distributed.

**3.** Show that for the model of Problem 4.3.4, if $\mu = 1/\lambda$, then $\bar{\mu} = 2\sum_{i=1}^n X_i^c/x_{2n}(\alpha)$ is a uniformly most accurate level $1-\alpha$ UCB for $\mu$.

**4.** Construct uniformly most accurate level $1-\alpha$ upper and lower confidence bounds for $\mu$ in the model of Problem 4.3.6 for $c$ fixed, $n = 1$.

**5.** Establish the following result due to Pratt (1961). Suppose $[\underline{\theta}^*, \bar{\theta}^*]$, $[\underline{\theta}, \bar{\theta}]$ are two level $(1-\alpha)$ confidence intervals such that

$$P_\theta[\underline{\theta}^* \le \theta' \le \bar{\theta}^*] \le P_\theta[\underline{\theta} \le \theta' \le \bar{\theta}] \text{ for all } \theta' \ne \theta.$$

Show that if $(\underline{\theta}, \bar{\theta})$, $(\underline{\theta}^*, \bar{\theta}^*)$ have joint densities, then $E_\theta(\bar{\theta}^* - \underline{\theta}^*) \le E_\theta(\bar{\theta} - \underline{\theta})$.

*Hint:* $E_\theta(\bar{\theta} - \underline{\theta}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\int_s^t du\right)p(s,t)dsdt = \int_{-\infty}^{\infty} P_\theta[\underline{\theta} \le u \le \bar{\theta}]du$, where $p(s,t)$ is the joint density of $(\underline{\theta}, \bar{\theta})$.

**6.** Let $U, V$ be random variables with d.f.'s $F, G$ corresponding to densities $f, g$, respectively, satisfying the conditions of Problem B.2.12 so that $F^{-1}, G^{-1}$ are well defined and strictly increasing. Show that if $F(x) \le G(x)$ for all $x$ and $E(U), E(V)$ are finite, then $E(U) \ge E(V)$.

*Hint:* By Problem B.2.12(b), $E(U) = \int_0^1 F^{-1}(t)dt$.

**7.** Suppose that $\underline{\theta}^*$ is a uniformly most accurate level $(1-\alpha)$ LCB such that $P_\theta[\underline{\theta}^* \le \theta] = 1-\alpha$. Prove Corollary 4.6.1.

*Hint:* Apply Problem 4.6.6 to $V = (\theta - \underline{\theta}^*)^+$, $U = (\theta - \underline{\theta})^+$.

**8.** In Example 4.6.2, establish that the UMP test has acceptance region (4.6.3).

*Hint:* Use Examples 4.3.3 and 4.4.4.

## Problems for Section 4.7

**1. (a)** Show that if $\boldsymbol{\theta}$ has a beta, $\beta(r,s)$, distribution with $r$ and $s$ positive integers, then $\boldsymbol{\lambda} = s\boldsymbol{\theta}/r(1-\boldsymbol{\theta})$ has the $F$ distribution $\mathcal{F}_{2r,2s}$.

*Hint:* See Sections B.2 and B.3.

**(b)** Suppose that given $\boldsymbol{\theta} = \theta$, $X$ has a binomial, $\mathcal{B}(n,\theta)$, distribution and that $\boldsymbol{\theta}$ has beta, $\beta(r,s)$ distribution with $r$ and $s$ integers. Show how the quantiles of the $F$ distribution can be used to find upper and lower credible bounds for $\lambda$ and for $\theta$.

**2.** Suppose that given $\boldsymbol{\lambda} = \lambda$, $X_1, \dots, X_n$ are i.i.d. Poisson, $\mathcal{P}(\lambda)$ and that $\boldsymbol{\lambda}$ is distributed as $V/s_0$, where $s_0$ is some constant and $V \sim \chi_k^2$. Let $T = \sum_{i=1}^n X_i$.

**(a)** Show that $(\boldsymbol{\lambda} \mid T = t)$ is distributed as $W/s$, where $s = s_0 + 2n$ and $W \sim \chi_m^2$ with $m = k + 2t$.

**(b)** Show how quantiles of the $\chi^2$ distribution can be used to determine level $(1-\alpha)$ upper and lower credible bounds for $\lambda$.

**3.** Suppose that given $\boldsymbol{\theta} = \theta$, $X_1, \dots, X_n$ are i.i.d. uniform, $\mathcal{U}(0,\theta)$, and that $\boldsymbol{\theta}$ has the Pareto, $Pa(c,s)$, density

$$\pi(t) = sc^s/t^{s-1},\ t > c,\ s > 0,\ c > 0.$$

**(a)** Let $M = \max\{X_1, \dots, X_n\}$. Show that $(\boldsymbol{\theta} \mid M = m) \sim Pa(c', s')$ with $c' = \max\{c, m\}$ and $s' = s + n$.

**(b)** Find level $(1 - \alpha)$ upper and lower credible bounds for $\theta$.

**(c)** Give a level $(1 - \alpha)$ confidence interval for $\theta$.

**(d)** Compare the level $(1-\alpha)$ upper and lower credible bounds for $\theta$ to the level $(1-\alpha)$ upper and lower confidence bounds for $\theta$. In particular consider the credible bounds as $n \to \infty$.

**4.** Suppose that given $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\tau}) = (\mu_1, \mu_2, \tau)$, $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ are two independent $\mathcal{N}(\mu_1, \tau)$ and $\mathcal{N}(\mu_2, \tau)$ samples, respectively. Suppose $\boldsymbol{\theta}$ has the improper prior $\pi(\theta) = 1/\tau$, $\tau > 0$.

**(a)** Let $s_0 = \Sigma(x_i - \bar{x})^2 + \Sigma(y_j - \bar{y})^2$. Show formally that the posterior $\pi(\theta \mid \mathbf{x}, \mathbf{y})$ is proportional to

$$\pi(\tau \mid s_0)\pi(\mu_1 \mid \tau, \bar{x})\pi(\mu_2 \mid \tau, \bar{y})$$

where $\pi(\tau \mid s_0)$ is the density of $s_0/V$ with $V \sim \chi_{m+n-2}$, $\pi(\mu_1 \mid \tau, \bar{x})$ is a $\mathcal{N}(\bar{x}, \tau/m)$ density and $\pi(\mu_2 \mid \tau, \bar{y})$ is a $\mathcal{N}(\bar{y}, \tau/n)$ density.

*Hint:* $p(\theta \mid \mathbf{x}, \mathbf{y})$ is proportional to

$$p(\boldsymbol{\theta})p(\mathbf{x} \mid \mu_1, \tau)p(\mathbf{y} \mid \mu_2, \tau).$$

**(b)** Show that given $\tau$, $\mu_1$ and $\mu_2$ are independent in the posterior distribution $p(\theta \mid \mathbf{x}, \mathbf{y})$ and that the joint density of $\Delta = \mu_1 - \mu_2$ and $\tau$ is

$$\pi(\Delta, \tau \mid \mathbf{x}, \mathbf{y}) = \pi(\tau \mid s_0)\pi(\Delta \mid \bar{x} - \bar{y}, \tau)$$

where $\pi(\Delta \mid \bar{x} - \bar{y}, \varphi)$ is the $\mathcal{N}(\bar{x} - \bar{y}, \tau(m^{-1} + n^{-1}))$ distribution.

**(c)** Set $s^2 = s_0/(m + n - 2)$. Show that the posterior distribution $\pi(t \mid \mathbf{x}, \mathbf{y})$ of

$$t = \frac{\Delta - (\bar{x} - \bar{y})}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

is (Student) $t$ with $m + n - 2$ degrees of freedom.

*Hint:* $\pi(\Delta \mid \mathbf{x}, \mathbf{y})$ is obtained by integrating out $\tau$ in $\pi(\Delta, \tau \mid \mathbf{x}, \mathbf{y})$.

**(d)** Use part (c) to give level $(1-\alpha)$ credible bounds and a level $(1-\alpha)$ credible interval for $\Delta$.

## Problems for Section 4.8

**1.** Let $X_1, \dots, X_{n+1}$ be i.i.d. as $X \sim \mathcal{N}(\mu, \sigma_0^2)$, where $\sigma_0^2$ is known. Here $X_1, \dots, X_n$ is observable and $X_{n+1}$ is to be predicted.

**(a)** Give a level $(1 - \alpha)$ prediction interval for $X_{n+1}$.

**(b)** Compare the interval in part (a) to the Bayesian prediction interval (4.8.3) by doing a frequentist computation of the probability of coverage. That is, suppose $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$. Take $\sigma_0^2 = \tau^2 = 1$, $n = 100$, $\eta_0 = 10$, and $\alpha = .05$. Then the level of the frequentist interval is 95%. Find the probability that the Bayesian interval covers $X_{n+1}$ for $\mu = 5, 8, 9, 9.5, 10, 10.5, 11, 12, 15$. Present the results in a table and a graph.

**2.** Let $X_1, \dots, X_{n+1}$ be i.i.d. as $X \sim F$, where $X_1, \dots, X_n$ are observable and $X_{n+1}$ is to be predicted. A level $(1-\alpha)$ *lower (upper) prediction* bound on $Y = X_{n+1}$ is defined to be a function $\underline{Y}(\bar{Y})$ of $X_1, \dots, X_n$ such that $P(\underline{Y} \le Y) \ge 1 - \alpha$ $(P(Y \le \bar{Y}) \ge 1 - \alpha)$.

**(a)** If $F$ is $\mathcal{N}(\mu, \sigma_0^2)$ with $\sigma_0^2$ known, give level $(1-\alpha)$ lower and upper prediction bounds for $X_{n+1}$.

**(b)** If $F$ is $\mathcal{N}(\mu, \sigma^2)$ with $\sigma^2$ unknown, give level $(1-\alpha)$ lower and upper prediction bounds for $X_{n+1}$.

**(c)** If $F$ is continuous with a positive density $f$ on $(a, b)$, $-\infty \le a < b \le \infty$, give level $(1-\alpha)$ distribution free lower and upper prediction bounds for $X_{n+1}$.

**3.** Suppose $X_1, \dots, X_{n+1}$ are i.i.d. as $X$ where $X$ has the exponential distribution

$$F(x \mid \theta) = 1 - e^{-x/\theta},\ x > 0,\ \theta > 0.$$

Suppose $X_1, \dots, X_n$ are observable and we want to predict $X_{n+1}$. Give a level $(1-\alpha)$ prediction interval for $X_{n+1}$.

*Hint:* $2X_i/\theta$ has a $\chi_2^2$ distribution and $nX_{n+1}/\sum_{i=1}^n X_i$ has an $\mathcal{F}_{2,2n}$ distribution.

**4.** Suppose that given $\boldsymbol{\theta} = \theta$, $X$ is a binomial, $\mathcal{B}(n, \theta)$, random variable, and that $\boldsymbol{\theta}$ has a beta, $\beta(r, s)$, distribution. Suppose that $Y$, which is not observable, has a $\mathcal{B}(m, \theta)$ distribution given $\boldsymbol{\theta} = \theta$. Show that the conditional (predictive) distribution of $Y$ given $X = x$ is

$$q(y \mid x) = \left( \begin{array}{c} m \\ y \end{array} \right) B(r + x + y, s + n - x + m - y)/B(r + x, s + n - x)$$

where $B(\cdot, \cdot)$ denotes the beta function. (This $q(y \mid x)$ is sometimes called the Pólya distribution.)

*Hint:* First show that

$$q(y \mid x) = \int p(y \mid \theta)\pi(\theta \mid x) d\theta.$$

**5.** In Example 4.8.2, let $U^{(1)} < \cdots < U^{(n+1)}$ denote $U_1, \dots, U_{n+1}$ ordered. Establish (4.8.2) by using the observation that $U_{n+1}$ is equally likely to be any of the values $U^{(1)}, \dots, U^{(n+1)}$.

### Problems for Section 4.9

**1.** Let $X$ have a binomial, $\mathcal{B}(n, \theta)$, distribution. Show that the likelihood ratio statistic for testing $H : \theta = \frac{1}{2}$ versus $K : \theta \ne \frac{1}{2}$ is equivalent to $|2X - n|$.

*Hint:* Show that for $x \le \frac{1}{2}n$, $\lambda(x)$ is an increasing function of $-(2x-n)$ and $\lambda(x) = \lambda(n-x)$.

*In Problems 2–4, let $X_1, \dots, X_n$ be a $\mathcal{N}(\mu, \sigma^2)$ sample with both $\mu$ and $\sigma^2$ unknown.*

**2.** In testing $H : \mu \le \mu_0$ versus $K : \mu > \mu_0$ show that the one-sided, one-sample $t$ test is the likelihood ratio test (for $\alpha < \frac{1}{2}$).

*Hint:* Note that $\widehat{\mu}_0 = \bar{X}$ if $\bar{X} \le \mu_0$ and $= \mu_0$ otherwise. Thus, $\log \lambda(\mathbf{x}) = 0$, if $T_n \le 0$ and $= (n/2)\log(1 + T_n^2/(n-1))$ for $T_n > 0$, where $T_n$ is the $t$ statistic.

**3.** *One-Sided Tests for Scale.* We want to test $H : \sigma^2 \le \sigma_0^2$ versus $K : \sigma^2 > \sigma_0^2$. Show that

**(a)** Likelihood ratio tests are of the form: Reject if, and only if,

$$\frac{n\widehat{\sigma}^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^n (X_i - \bar{X})^2 \ge c.$$

*Hint:* $\log \lambda(\mathbf{x}) = 0$, if $\widehat{\sigma}^2/\sigma_0^2 \le 1$ and $= (n/2)[\widehat{\sigma}^2/\sigma_0^2 - 1 - \log(\widehat{\sigma}^2/\sigma_0^2)]$ otherwise.

**(b)** To obtain size $\alpha$ for $H$ we should take $c = x_{n-1}(1-\alpha)$.
*Hint:* Recall Theorem B.3.3.

**(c)** These tests coincide with the tests obtained by inverting the family of level $(1-\alpha)$ lower confidence bounds for $\sigma^2$.

**4.** *Two-Sided Tests for Scale.* We want to test $H : \sigma = \sigma_0$ versus $K : \sigma \ne \sigma_0$.

**(a)** Show that the size $\alpha$ likelihood ratio test accepts if, and only if,

$$c_1 \le \frac{1}{\sigma_0^2}\sum_{i=1}^n (X_i - \bar{X})^2 \le c_2 \text{ where } c_1 \text{ and } c_2 \text{ satisfy,}$$

(i) $F(c_2) - F(c_1) = 1 - \alpha$, where $F$ is the d.f. of the $\chi^2_{n-1}$ distribution.

(ii) $c_1 - c_2 = n \log c_1/c_2$.

**(b)** Use the normal approximation to check that

$$\begin{aligned} c_{1n} &= n - \sqrt{2n}z(1 - \tfrac{1}{2}\alpha) \\ c_{2n} &= n + \sqrt{2n}z(1 - \tfrac{1}{2}\alpha) \end{aligned}$$

approximately satisfy (i) and also (ii) in the sense that the ratio

$$\frac{c_{1n} - c_{2n}}{n \log c_{1n}/c_{2n}} \to 1 \text{ as } n \to \infty.$$

**(c)** Deduce that the critical values of the commonly used equal-tailed test, $x_{n-1}(\frac{1}{2}\alpha)$, $x_{n-1}(1 - \frac{1}{2}\alpha)$ also approximately satisfy (i) and (ii) of part (a).

**5.** The following blood pressures were obtained in a sample of size $n = 5$ from a certain population: 124, 110, 114, 100, 190. Assume the one-sample normal model.

**(a)** Using the size $\alpha = 0.05$ one-sample $t$ test, can we conclude that the mean blood pressure in the population is significantly larger than 100?

**(b)** Compute a level $0.95$ confidence interval for $\sigma^2$ corresponding to inversion of the equal-tailed tests of Problem 4.9.4.

**(c)** Compute a level $0.90$ confidence interval for the mean blood pressure $\mu$.

**6.** Let $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ be two independent $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ samples, respectively.

**(a)** Show that the MLE of $\theta = (\mu_1, \mu_2, \sigma^2)$ is $(\bar{X}, \bar{Y}, \widetilde{\sigma}^2)$, where $\widetilde{\sigma}^2$ is as defined in Section 4.9.3.

**(b)** Consider the problem of testing $H : \mu_1 \leq \mu_2$ versus $K : \mu_1 > \mu_2$. Assume $\alpha \leq \frac{1}{2}$. Show that the likelihood ratio statistic is equivalent to the two-sample $t$ statistic $T$.

**(c)** Using the normal approximation $\Phi(z(\alpha) + \sqrt{n_1 n_2/n}(\mu_1 - \mu_2)/\sigma)$ to the power, find the sample size $n$ needed for the level $0.01$ test to have power $0.95$ when $n_1 = n_2 = \frac{1}{2}n$ and $(\mu_1 - \mu_2)/\sigma = \frac{1}{2}$.

**7.** The following data are from an experiment to study the relationship between forage production in the spring and mulch left on the ground the previous fall. The control measurements ($x$'s) correspond to 0 pounds of mulch per acre, whereas the treatment measurements ($y$'s) correspond to $500$ pounds of mulch per acre. Forage production is also measured in pounds per acre.

| $x$ | 794 | 1800 | 576 | 411 | 897 |
|---|---|---|---|---|---|
| $y$ | 2012 | 2477 | 3498 | 2092 | 1808 |

Assume the two-sample normal model with equal variances.

**(a)** Find a level $0.95$ confidence interval for $\mu_2 - \mu_1$.

**(b)** Can we conclude that leaving the indicated amount of mulch on the ground significantly improves forage production? Use $\alpha = 0.05$.

**(c)** Find a level $0.90$ confidence interval for $\sigma$ by using the pivot $s^2/\sigma^2$.

**8.** Suppose $\mathbf{X}$ has density $p(\mathbf{x}, \theta)$, $\theta \in \Theta$, and that $T$ is sufficient for $\theta$. Show that $\lambda(\mathbf{X}, \Theta_0, \Theta_1)$ depends on $\mathbf{X}$ only through $T$.

**9.** The normally distributed random variables $X_1, \dots, X_n$ are said to be *serially correlated* or to follow an autoregressive model if we can write

$$X_i = \theta X_{i-1} + \epsilon_i, \; i = 1, \dots, n,$$

where $X_0 = 0$ and $\epsilon_1, \dots, \epsilon_n$ are independent $\mathcal{N}(0, \sigma^2)$ random variables.

**(a)** Show that the density of $\mathbf{X} = (X_1, \dots, X_n)$ is

$$p(\mathbf{x}, \theta) = (2\pi\sigma^2)^{-\frac{1}{2}n} \exp\{-(1/2\sigma^2) \sum_{i=1}^{n} (x_i - \theta x_{i-1})^2\}$$

for $-\infty < x_i < \infty, i = 1, \dots, n, x_0 = 0$.

**(b)** Show that the likelihood ratio statistic of $H : \theta = 0$ (independence) versus $K : \theta \neq 0$ (serial correlation) is equivalent to $-(\sum_{i=2}^{n} X_i X_{i-1})^2 / \sum_{i=1}^{n-1} X_i^2$.

**10.** (An example due to C. Stein). Consider the following model. Fix $0 < \alpha < \frac{1}{2}$ and $\alpha/[2(1-\alpha)] < c < \alpha$. Let $\Theta$ consist of the point $-1$ and the interval $[0, 1]$. Define the frequency functions $p(x, \theta)$ by the following table.

| $\theta$ \ $x$ | $-2$ | $-1$ | $0$ | $1$ | $2$ |
|---|---|---|---|---|---|
| $-1$ | $\frac{1}{2}\alpha$ | $\frac{1}{2} - \alpha$ | $\alpha$ | $\frac{1}{2} - \alpha$ | $\frac{1}{2}\alpha$ |
| $\neq -1$ | $\theta c$ | $\left(\frac{1-c}{1-\alpha}\right)(\frac{1}{2} - \alpha)$ | $\left(\frac{1-c}{1-\alpha}\right)\alpha$ | $\left(\frac{1-c}{1-\alpha}\right)(\frac{1}{2} - \alpha)$ | $(1-\theta)c$ |

**(a)** What is the size $\alpha$ likelihood ratio test for testing $H : \theta = -1$ versus $K : \theta \neq -1$?

**(b)** Show that the test that rejects if, and only if, $X = 0$, has level $\alpha$ and is strictly more powerful whatever be $\theta$.

**11.** *The power functions of one- and two-sided t tests.* Suppose that $T$ has a noncentral $t$, $\mathcal{T}_{k,\delta}$, distribution. Show that,

**(a)** $P_\delta[T \geq t]$ is an increasing function of $\delta$.

**(b)** $P_\delta[|T| \geq t]$ is an increasing function of $|\delta|$.

*Hint:* Let $Z$ and $V$ be independent and have $\mathcal{N}(\delta, 1), \chi_k^2$ distributions respectively. Then, for each $v > 0$, $P_\delta[Z \geq t\sqrt{v/k}]$ is increasing in $\delta$, $P_\delta[|Z| \geq t\sqrt{v/k}]$ is increasing in $|\delta|$. Condition on $V$ and apply the double expectation theorem.

**12.** Show that the noncentral $t$ distribution, $\mathcal{T}_{k,\delta}$, has density

$$f_{k,\delta}(t) = \frac{1}{\sqrt{\pi k}(\frac{1}{2}k) 2^{\frac{1}{2}(k+1)}} \int_0^\infty x^{\frac{1}{2}(k-1)} e^{-\frac{1}{2}\{x + (t\sqrt{x/k} - \delta)^2\}} dx.$$

*Hint:* Let $Z$ and $V$ be as in the preceding hint. From the joint distribution of $Z$ and $V$, get the joint distribution of $Y_1 = Z/\sqrt{V/k}$ and $Y_2 = V$. Then use $p_{Y_1}(y_1) = \int p_{Y_1, Y_2}(y_1, y_2) dy_2$.

**13.** *The F Test for Equality of Scale.* Let $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$ be two independent samples from $\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)$, respectively, with all parameters assumed unknown.

**(a)** Show that the LR test of $H : \sigma_1^2 = \sigma_2^2$ versus $K : \sigma_2^2 > \sigma_1^2$ is of the form: Reject if, and only if, $F = [(n_1 - 1)/(n_2 - 1)]\Sigma(Y_i - \bar{Y})^2/\Sigma(X_i - \bar{X})^2 \geq C$.

**(b)** Show that $(\sigma_1^2/\sigma_2^2)F$ has an $\mathcal{F}_{n_2-1,n_1-1}$ distribution and that critical values can be obtained from the $\mathcal{F}$ table.

**(c)** Justify the two-sided $F$ test: Reject $H$ if, and only if, $F \geq f(1 - \alpha/2)$ or $F \leq f(\alpha/2)$, where $f(t)$ is the $t$th quantile of the $\mathcal{F}_{n_2-1,n_1-1}$ distribution, as an approximation to the LR test of $H : \sigma_1 = \sigma_2$ versus $K : \sigma_1 \neq \sigma_2$. Argue as in Problem 4.9.4.

**(d)** Relate the two-sided test of part (c) to the confidence intervals for $\sigma_2^2/\sigma_1^2$ obtained in Problem 4.4.10.

**14.** The following data are the blood cholesterol levels ($x$'s) and weight/height ratios ($y$'s) of 10 men involved in a heart study.

| $x$ | 254 | 240 | 279 | 284 | 315 | 250 | 298 | 384 | 310 | 337 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 2.71 | 2.96 | 2.62 | 2.19 | 2.68 | 2.64 | 2.37 | 2.61 | 2.12 | 1.94 |

Using the likelihood ratio test for the bivariate normal model, can you conclude at the 10% level of significance that blood cholesterol level is correlated with weight/height ratio?

**15.** Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample from a bivariate $\mathcal{N}(0, 0, \sigma_1^2, \sigma_2^2, \rho)$ distribution. Consider the problem of testing $H : \rho = 0$ versus $K : \rho \neq 0$.

**(a)** Show that the likelihood ratio statistic is equivalent to $|r|$ where

$$r = \sum_{i=1}^{n} X_i Y_i \Bigg/ \sqrt{\sum_{i=1}^{n} X_i^2 \sum_{j=1}^{n} Y_j^2}.$$

**(b)** Show that if we have a sample from a bivariate $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ distribution, then $P[\widehat{\rho} \geq c]$ is an increasing function of $\rho$ for fixed $c$.

*Hint:* Use the transformations and Problem B.4.7 to conclude that $\widehat{\rho}$ has the same distribution as $S_{12}/S_1 S_2$, where

$$S_1^2 = \sum_{i=2}^{n} U_i^2,\ S_2^2 = \sum_{i=2}^{n} V_i^2,\ S_{12} = \sum_{i=2}^{n} U_i V_i$$

and $(U_2, V_2), \ldots, (U_n, V_n)$ is a sample from a $\mathcal{N}(0, 0, 1, 1, \rho)$ distribution. Let $R = S_{12}/S_1 S_2$, $T = \sqrt{n-2}R/\sqrt{1 - R^2}$, and using the arguments of Problems B.4.7 and B.4.8, show that given $U_2 = u_2, \ldots, U_n = u_n$, $T$ has a noncentral $\mathcal{T}_{n-2}$ distribution with noncentrality parameter $\rho$. Because this conditional distribution does not depend on $(u_2, \ldots, u_n)$, the continuous version of (B.1.24) implies that this is also the unconditional distribution. Finally, note that $\widehat{\rho}$ has the same distribution as $R$, that $T$ is an increasing function of $R$, and use Problem 4.8.11(a).

**16.** Let $\lambda(\mathbf{X})$ denote the likelihood ratio statistic for testing $H : \rho = 0$ versus $K : \rho \neq 0$ in the bivariate normal model. Show, using (4.9.4) and (4.9.5) that $2 \log \lambda(\mathbf{X}) \xrightarrow{\mathcal{L}} V$, where $V$ has a $\chi_1^2$ distribution.

**17.** Consider the bioequivalence example in Problem 3.2.9.

**(a)** Find the level $\alpha$ LR test for testing $H : \theta \in [-\epsilon, \epsilon]$ versus $K : \theta \notin [-\epsilon, \epsilon]$.

**(b)** In this example it is reasonable to use a level $\alpha$ test of $H$: $\theta \notin [-\epsilon, \epsilon]$ versus $K$: $\theta \in [-\epsilon, \epsilon]$. Propose such a test and compare your solution to the Bayesian solution based on a continuous loss function given in Problem 3.2.9. Consider the cases $\eta_0 = 0, \tau_0^2 \to \infty$, and $\eta_0 = 0, n \to \infty$.

## 4.11 NOTES

### Notes for Section 4.1

(1) The point of view usually taken in science is that of Karl Popper [1968]. Acceptance of a hypothesis is only provisional as an adequate current approximation to what we are interested in understanding. Rejection is more definitive.

(2) We ignore at this time some real-life inadequacies of this experiment such as the placebo effect (see Example 1.1.3).

(3) A good approximation (Durbin, 1973; Stephens, 1974) to the critical value is $c_n(t) = t/(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$ where $t = 1.035, 0.895$ and $t = 0.819$ for $\alpha = 0.01$, .05 and 0.10, respectively.

### Notes for Section 4.3

(1) Such a class is sometimes called essentially complete. The term *complete* is then reserved for the class where strict inequality in (4.3.3) holds for some $\theta$ if $\varphi \notin \mathcal{D}$.

(2) The theory of complete and essentially complete families is developed in Wald (1950), see also Ferguson (1967). Essentially, if the parameter space is compact and loss functions are bounded, the class of Bayes procedures is complete. More generally the closure of the class of Bayes procedures (in a suitable metric) is complete.

### Notes for Section 4.4

(1) If the continuity correction discussed in Section A.15 is used here, $S$ in $\bar{\theta}(\mathbf{X})$ would be replaced by $S + \frac{1}{2}$, and $S$ in $\underline{\theta}(\mathbf{X})$ is replaced by $S - \frac{1}{2}$.

### Notes for Section 4.5

(1) In using $\underline{\theta}(S)$ as a confidence bound we are using the region $[\underline{\theta}(S), 1]$. Because the region contains $C(\mathbf{X})$, it also has confidence level $(1 - \alpha)$.

## 4.12 REFERENCES

Barlow, R. and F. Proschan, *Mathematical Theory of Reliability* New York: J. Wiley & Sons, 1965.

Bickel, P., "Inference and auditing: the Stringer bound." *Internat. Statist. Rev., 60*, 197–209 (1992).

BICKEL, P., E. HAMMEL, AND J. W. O'CONNELL, "Is there a sex bias in graduate admissions?" *Science, 187*, 398–404 (1975).

BOX, G. E. P., *Apology for Ecumenism in Statistics and Scientific Inference*, Data Analysis and Robustness, G. E. P. Box, T. Leonard, and C. F. Wu, Editors New York: Academic Press, 1983.

BROWN, L. D., T. CAI, AND A. DAS GUPTA, "Interval estimation for a binomial proportion," *Statistical Science*, 101–128 (2001).

DOKSUM, K. A. AND G. SIEVERS, "Plotting with confidence: Graphical comparisons of two populations," *Biometrika, 63*, 421–434 (1976).

DOKSUM, K. A., G. FENSTAD, AND R. AABERGE, "Plots and tests for symmetry," *Biometrika, 64*, 473–487 (1977).

DURBIN, J., "Distribution theory for tests based on the sample distribution function," *Regional Conference Series in Applied Math., 9*, SIAM, Philadelphia, Pennsylvania (1973).

FERGUSON, T., *Mathematical Statistics. A Decision Theoretic Approach* New York: Academic Press, 1967.

FISHER, R. A., *Statistical Methods for Research Workers*, 13th ed. New York: Hafner Publishing Company, 1958.

HALD, A., *Statistical Theory with Engineering Applications* New York: J. Wiley & Sons, 1952.

HEDGES, L. V. AND I. OLKIN, *Statistical Methods for Meta-Analysis* Orlando, FL: Academic Press, 1985.

JEFFREYS, H., *The Theory of Probability* Oxford: Oxford University Press, 1961.

LEHMANN, E. L., *Testing Statistical Hypotheses*, 2nd ed. New York: Springer, 1997.

POPPER, K. R., *Conjectures and Refutations; the Growth of Scientific Knowledge*, 3rd ed. New York: Harper and Row, 1968.

PRATT, J., "Length of confidence intervals," *J. Amer. Statist. Assoc., 56*, 549–567 (1961).

SACKROWITZ, H. AND E. SAMUEL–CAHN, "$P$ values as random variables—Expected $P$ values," *The American Statistician, 53*, 326–331 (1999).

STEPHENS, M., "EDF statistics for goodness of fit," *J. Amer. Statist., 69*, 730–737 (1974).

STEIN, C., "A two-sample test for a linear hypothesis whose power is independent of the variance," *Ann. Math. Statist., 16*, 243–258 (1945).

TATE, R. F. AND G. W. KLETT, "Optimal confidence intervals for the variance of a normal distribution," *J. Amer. Statist. Assoc., 54*, 674–682 (1959).

VAN ZWET, W. R. AND J. OSTERHOFF, "On the combination of independent test statistics," *Ann. Math. Statist., 38*, 659–680 (1967).

WALD, A., *Sequential Analysis* New York: Wiley, 1947.

WALD, A., *Statistical Decision Functions* New York: Wiley, 1950.

WANG, Y., "Probabilities of the type I errors of the Welch tests," *J. Amer. Statist. Assoc., 66*, 605–608 (1971).

WELCH, B., "Further notes on Mrs. Aspin's tables," *Biometrika, 36*, 243–246 (1949).

WETHERILL, G. B. AND K. D. GLAZEBROOK, *Sequential Methods in Statistics* New York: Chapman and Hall, 1986.