Jain, Dubes, and Chen (1987) and Chatterjee and Chatterjee (1983) also considered confidence intervals and the standard error of the estimators, respectively. Chernick, Murthy, and Nealy (1988a,b), Hirst (1996) and Snapinn and Knoke (1985b) considered certain non-Gaussian populations. The most recent results on the .632 estimator and an enhancement of it called .632+ are given in Efron and Tibshirani (1997a).

McLachlan has done a lot of research in discriminant analysis and particularly on error rate estimation. His survey article (McLachlan, 1986) provides a good review of the issues and the literature including bootstrap results up to 1986. Some of the developments discussed in this chapter appear in McLachlan (1992), where he devotes an entire chapter, (Chapter 10) to the estimation of error rates. It includes a section on bootstrap (pp. 346–360).

An early account of discriminant analysis methods is given in Lachenbruch (1975). Multivariate simulation methods such as those used in studies by Chernick, Murthy, and Nealy are covered in Johnson (1987).

The bootstrap distribution for the median is also discussed in Efron (1982a, Chapter 10, pp. 77–78). Mooney and Duval (1993) discuss the problem of estimating the difference between two medians.

Justification (consistency results) for the bootstrap approach to individual bioequivalence came in Shao, Kübler, and Pigeot (2000). The survey article by Pigeot (2001) is an excellent reference for the advantages and disadvantages of the bootstrap and the jackknife in biomedical research, and it includes coverage of the individual bioequivalence application.

CHAPTER 3

Confidence Sets and Hypothesis Testing

Because of the close relationship between tests of hypotheses and confidence intervals, we include both in this chapter. Section 3.1 deals with "nonparametric" bootstrap confidence intervals (i.e., little or no assumptions are made about the form of the distribution being sampled).

There has also been some work on parametric forms of bootstrap confidence intervals and on methods for reducing or eliminating the use of Monte Carlo replications. We shall not discuss these in this text but do include references to the most relevant work in the historical notes (Section 3.5). Also, the parametric bootstrap is discussed briefly in Chapter 6.

Section 3.1.2 considers the simplest technique, the percentile method. This method works well when the statistic used is a pivotal quantity and has a symmetric distribution [see Efron (1981c, and 1982a)].

The percentile method and various other bootstrap confidence interval estimates require a large number of Monte Carlo replications for the intervals to be both accurate (i.e., be as small as possible for the given confidence level) and nearly exact (i.e., if the procedure were repeated many times the percentage of intervals that would actually include the "true" parameter value is approximately the stated confidence levels).

This essentially states for exactness that the actual confidence level of the interval is approximately the stated level. So, for example, if we construct a 95% confidence interval, we would expect that our procedure would produce intervals that contain the true parameter in 95% of the cases. Such is the definition of a confidence interval.

Unfortunately for "nonparametric" intervals, we cannot generally do this. The best we can hope for is to have approximately the stated coverage. Such

Bootstrap Methods: A Guide for Practitioners and Researchers, Second Edition, by Michael R. Chernick

intervals will be called approximately correct or almost exact. As the sample size increases and the number of bootstrap Monte Carlo replications increases, we can expect the percentile method to be approximately correct and accurate.

Another method that Hall (1992a) refers to as the percentile method is also mentioned in Section 3.1.2. Hall refers to Efron's percentile method as the "other" percentile method.

For pivotal quantities that do not have symmetric distributions, the intervals can be improved by bias adjustment and acceleration constants. This is the approach taken in Efron (1987) and is the topic of Section 3.1.3.

Another approach that also provides better bootstrap confidence intervals is called bootstrap iteration (or double bootstrap). This approach has been studied in detail by Hall and Martin, among others, and is covered in Section 3.1.4. There we provide a review of research results and the developments from Martin (1990a) and Hall (1992a).

In each of the sections, examples are given to instruct the reader in the proper application of the methods, to illustrate their accuracy and correctness. Important asymptotic results will be mentioned, but we shall not delve into the asymptotic theory.

Section 3.1.5 deals with the bootstrap t method for generating bootstrap-type confidence intervals. In some problems, the bootstrap t method may be appropriate and has better accuracy and correctness than the percentile method. It is easier to implement than methods involving Efron's corrections. It is not as computer intensive as the iterated bootstrap. Consequently, it is popular in practice. We applied it in the Passive Plus DX clinical trial at Pacesetter. So Section 3.1.5 is intended to provide the definition of it so that the reader may apply it. The bootstrap t was introduced by Efron, in his monograph (Efron, 1982a).

In Section 3.2, the reader is shown the connection between confidence intervals and hypothesis tests. This close connection enables the reader to see how a confidence interval for a parameter can be reinterpreted in terms of the acceptance or rejection of a hypothesis test with a null hypothesis that the parameter is a specified value.

The confidence level is directly related to the significance level of the test. Knowing this, the reader will be able to test hypotheses by constructing bootstrap confidence intervals for the parameter.

In Section 3.3, we provide examples of hypothesis tests to illustrate the usefulness of the bootstrap approach. In some cases, we can compare the bootstrap tests with other nonparametric tests including the permutation tests from Good (1994) or Manly (1991, 1997).

Section 3.4 provides an historical perspective on the literature for confidence interval estimation and hypothesis testing using the bootstrap approach.

3.1. CONFIDENCE SETS

Before introducing the various bootstrap-type confidence intervals, we will review what a confidence set or region is and then, in Section 3.1.1, present Hartigan's typical value theorem in order to motivate the percentile method of Section 3.1.2. Section 3.1.3 then explains how refinements can be made to handle asymmetric cases where the percentile method does not work well.

Section 3.1.4 presents bootstrap interation. Bootstrap iteration or double bootstrapping is another approach to confidence intervals that overcomes the deficiencies of the percentile method. In Section 3.1.5, we present the bootstrap t method that also overcomes deficiencies of the percentile method but is simpler and more commonly used in practice than the iterated bootstrap and other bootstrap modifications to the percentile method.

What is a confidence set for a parameter vector? Suppose we have a parameter vector ν that belongs to an n-dimensional Euclidean space (denoted by R^n). A confidence set with confidence coefficient $1 - \alpha$ is a set in R^n determined on the basis of a random sample and having the property that if the random sampling were repeated infinitely many times with a new region generated each time, then $100^* (1 - \alpha)\%$ of the time the region will contain ν.

In the simplest case where the parameter is one-dimensional, the confidence region will be an interval or the union of two or more disjoint intervals.

In parametric families of population distributions involving nuisance parameters (parameters required to uniquely specify the distribution but which are not of interest to the investigator) or when very little is specified about the population distribution, it may not be possible to construct confidence sets which have a confidence coefficient that is exactly $1 - \alpha$ for all possible ν and all possible values of the nuisance parameters [see Bahadur and Savage (1956), for example]. We shall see that the bootstrap percentile method will at least provide us with confidence intervals that have confidence coefficient approaching $1 - \alpha$ as the sample size becomes very large.

If we only assume that the population distribution is symmetric, then the typical value theorem of Hartigan (1969) tells us that subsampling methods (e.g., random subsampling) can provide confidence intervals that are exact (i.e., have confidence coefficient $1 - \alpha$ for finite sample sizes). We shall now describe these subsampling methods and present the typical value theorems.

3.1.1. Typical Value Theorems for M-Estimates

We shall consider the case of independent identically distributed observations from a symmetric distribution on the real line. We denote the n random variables by $X_1, X_2, \ldots, X_n$ and their distribution by F_θ. For any set A let $P_\theta(A)$ denote the probability that a random variable X with distribution F_θ has its

value in the set A. As in Efron (1982a, p. 69) we will assume that F_θ has a symmetric density function $f(\cdot)$ so that

$$P_\theta(A) = \int_{-A} f(x-\theta)\,dx,$$

where

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1, \qquad f(x) \geq 0, \qquad \text{and} \qquad f(-x) = f(x).$$

An M-estimate $\hat{\theta}(x_1, x_2, \ldots, x_n)$ for θ, is any solution to the equation

$$\sum_i \psi(x_i - t) = 0.$$

Here we assume that the observed data $X_i = x_i$ for $i = 1, 2, \ldots, n$ are fixed while t is the variable to solve for.

We note that in general M-estimates need not be unique. The function Ψ is called the kernel, and Ψ is assumed to be antisymmetric and strictly increasing [i.e., $\Psi(-z) = -\Psi(z)$ and $\Psi(z + h) > \Psi(z)$ for all z and for $h > 0$]. Examples of M-estimates are given in Efron (1982a). For an appropriately chosen functions, Ψ many familiar estimates can be shown to be M-estimates including the sample mean and the sample median.

Consider the set of integers $(1, 2, 3, \ldots, n)$. The number of nonempty subsets of this set is $2^n - 1$. Let S be any one of these non-empty subsets. Let $\hat{\theta}_s$ denote an M-estimate based on only those values x_i for i belonging to S.

Under our assumptions about Ψ these M-estimates will be different for differing choices of S. Now let $I_1, I_2, \ldots, I_2^n$ denote the following partition of the real line:

$$\{I_1 = (-\infty, a_1), I_2 = [a_1, a_2), I_3 = [a_2, a_3), \ldots, I_{2^n-1} = [a_{2^n-2}, a_{2^n-1})\},$$

and

$$I_{2^n} = [a_{2^n-1}, +\infty)$$

where a_1 is the smallest $\hat{\theta}_S$, a_2 is the second smallest $\hat{\theta}_S$, and so on. We now are able to state the first typical value theorem.

Theorem 3.1.1.1. The Typical Value Theorem (Hartigan, 1969). The true value of θ has probability $1/2^n$ of being in the interval I_i for $i = 1, 2, \ldots, 2^n$, where I_i is defined as above.

The proof of this theorem is given in Efron (1982a, pp. 70–71). He attributes the method of proof to the paper by Maritz (1979). The theorem came originally from Hartigan (1969), who attributes it to Tukey and Mallows.

We now define a procedure called random subsampling. Let $S_1, S_2, S_3, \ldots, S_{B-1}$ be $B - 1$ of the $2^n - 1$ non-empty subsets of $\{1, 2, \ldots, n\}$ selected at random without replacement and let $I_1, I_2, \ldots, I_B$ be the partition of the real line obtained by ordering the corresponding $\hat{\theta}_s$ values, We then have the following typical value theorem, which can be viewed as a corollary to the previous theorem.

Theorem 3.1.1.2. The true value of θ has probability $1/B$ of being in the interval I_i for $i = 1, 2, \ldots, B$ where I_i is defined as above.

For more details and discussion about these results see Efron (1982a). The important point here is that we know the probability that each interval contains θ.

We can then construct an exact $100(j/B)$ percent confidence region for $1 \leq j \leq B - 1$ by simply combining any j of the intervals. The most sensible approach would be to paste together the j intervals in the "middle" if a two-sided interval is desired.

3.1.2. Percentile Method

The percentile method is the most obvious way to construct a confidence interval for a parameter based on bootstrap estimates. Suppose that $\hat{\theta}_i^*$ is the ith bootstrap estimate from the ith bootstrap sample where each bootstrap sample is of size n. By analogy with the case of random subsampling, we would expect that if we ordered the observations from smallest to largest, we would expect an interval that contains 90% of the $\hat{\theta}_i^*$ to be a 90% confidence interval for θ. The most sensible way to choose the interval that excludes the lowest 5% and the highest 5%.

A bootstrap confidence interval generated this way is called a percentile method confidence interval or, more specifically, Efron's percentile method confidence interval. This result (the exact confidence level) would hold if the typical value theorem applied to bootstrap sample estimates just as it did to random subsample estimates. Remember, we also had the symmetry condition and the estimator had to be an M-estimator in Hartigan's theorem.

Unfortunately, even if the distribution is symmetric and the estimator is an M-estimator as is the case for the sample median of, say, a Cauchy distribution, the bootstrap percentile method would not be exact (i.e., the parameter is contained in the generated intervals in exactly the advertised proportion of intervals as the number of generated cases becomes large).

Efron (1982a, pp. 80–81) shows that for the median, the percentile method provides nearly the same confidence interval as the nonparametric interval based on the binomial distribution. So the percentile method works well in some cases even though it is not exact.

Really, the main difference between random subsampling and bootstrapping is that bootstrapping involves sampling with replacement from the origi-

nal sample whereas random subsampling selects without replacement from the set of all possible subsamples. As the sample size becomes large, the difference in the distribution of the bootstrap estimates and the subsample estimates becomes small. Therefore, we expect the bootstrap percentile interval to be almost the same as the random subsample interval. So the percentile intervals inherit the exactness property of the subsample interval asymptotically (i.e., as the sample size becomes infinitely large).

Unfortunately, in the case of small samples (especially for asymmetric distributions) the percentile method does not work well. But fortunately, there are modifications that will get around the difficulties as we shall see in the next section.

In Chapter 3 of Hall (1992a), several bootstrap confidence intervals are defined. In particular, see Section 3.2 of Hall (1992a). In Hall's notation, F_0 denotes the population distribution, F_1 the empirical distribution and F_2 denotes the distribution of the samples drawn at random and with replacement from F_1.

Let φ_0 be the unknown parameter of interest which is expressible as a functional of the distribution F_0. So $\varphi_0 = \varphi(F_0)$. A theoretical α-level percentile confidence interval for φ_0 (by Hall's definition) is the interval $I_1 = (-\infty, \psi + t_0)$, where t_0 is defined so that

$$P(\varphi_0 \le \psi + t_0) = \alpha.$$

Alternatively, if we define

$$f_1(F_0, F_1) = I\{\varphi(F_0) \le \varphi(F_1) + t\} - \alpha,$$

then t_0 is a value of t such that $f_1(F_0, F_1) = 0$.

By analogy, a bootstrap one-sided percentile interval for φ_0 would be obtained by solving the equation

$$f_1(F_1, F_2) = 0 \tag{3.1}$$

since in bootstrapping, F_1 replaces F_0 and F_2 replaces F_1. If $\hat{t}_0$ is a solution to Eq. (3.1), the interval $(-\infty, \varphi(F_2) + (\hat{t}_0)$ is a one-sided bootstrap percentile confidence interval for φ. Here $\varphi(F_2)$ is the bootstrap sample estimate for φ. This is a natural way to define a percentile confidence interval according to Hall. It can easily be approximated by Monte Carlo, but differs from Efron's percentile method. Hall refers to Efron's percentile as the "other" percentile method or the "backwards" percentile method.

3.1.3. Bias Correction and the Acceleration Constant

Efron and Tibshirani (1986, pp. 67–70) describe four methods for constructing approximate confidence intervals for a parameter θ. They provide the assumptions required for each method to work well. In going from the first method to the fourth, the assumptions become less restrictive while the methods become more complicated but more generally applicable.

The first method is referred to as the standard method. It is obtained by taking the estimator $\hat{\theta}$ of θ and an estimate of its standard deviation $\hat{\sigma}$. The interval $[\hat{\theta} - \hat{\sigma}z_\alpha, \hat{\theta} + \hat{\sigma}z_\alpha]$ is the standard $100(1 - \alpha)\%$ approximate confidence interval for θ. This method works well if $\hat{\theta}$ has an approximate Gaussian distribution with mean θ and standard deviation σ independent of θ.

The second method is the bootstrap percentile method (Efron's definition) described in Section 3.1.2. It works well, when there exists a monotone transformation $\phi = g(\theta)$, such that $\hat{\phi} = g(\hat{\theta})$ is approximately Gaussian with mean ϕ and standard deviation τ independent of ϕ.

The third method is the bias-corrected bootstrap interval, which we discuss in this section. It works well if the transformation $\hat{\phi} = g(\hat{\theta})$ is approximately Gaussian with mean $\phi - z_0\tau$, where z_0 is the bias correction and τ is the standard deviation of $\hat{\phi}$ that does not depend on ϕ.

The fourth method is the BC_a method, which incorporates an acceleration constant a. For it to work well, $\hat{\phi}$ is approximately Gaussian with mean $\phi - z_0\tau_\phi$, where z_0 is the bias correction and τ_ϕ is the standard deviation of $\hat{\phi}$, which does depend on ϕ as follows: $\tau_\phi = 1 + a\phi$, where a is the acceleration constant to be defined later in this section. These results are summarized in Table 6 of Efron and Tibshirani (1986) and are reproduced in Table 3.1.

Efron and Tibshirani (1986) claim that the percentile method automatically incorporates normalizing transformations. To illustrate the difficulties that can be encountered with the percentile method, they consider the case where θ is the bivariate correlation coefficient from a two-dimensional Gaussian distribution and the sample size is 15.

In this case, there is no monotone transformation g that maps $\hat{\theta}$ into $\hat{\phi}$ with $\hat{\phi}$ Gaussian with mean ϕ and constant variance τ^2 independent of ϕ. For a set of data referred to as the "law school data," Efron and Tibshirani (1986) show that the sample bivariate correlation is 0.776.

Assuming we have bivariate Gaussian data with a sample of size 15 and a sample correlation estimate equal to 0.776, we would find that for a bootstrap sample the probability that the correlation coefficient is less than 0.776 based on the bootstrap estimate is only 0.431.

For any monotone transformation, this would also be the probability that the transformed value of the bootstrap sample correlation is less than the transformed value of the original sample correlation [i.e., $g(0.776)$]. However, for the transformed values to be Gaussian or at least a good approximation to the Gaussian distribution and centered about $g(0.776)$, this probability would have to be 0.500 and not 0.431. Note that for symmetric distributions like the Gaussian, the mean is equal to the median. But we do not see that here for the correlation coefficient.

What we see here is that, at least for some values of θ different from zero, no such transformation will work well. Efron and Tibshirani remedy this

Table 3.1 Four Methods of Setting Approximate Confidence Intervals for a Real Valued Parameter θ

Method	Abbreviation	α-Level Endpoint	Correct if
1. Standard	$\theta_S[\alpha]$	$\hat{\theta}+\hat{\sigma}z^{(\alpha)}$	$\hat{\theta} \approx N(\theta, \sigma^2)$ σ is constant
2. Percentile	$\theta_P[\alpha]$	$\hat{G}^{-1}(\alpha)$	There exists a monotone transformation such that $\hat{\phi} = g(\hat{\theta})$, where, $\phi = g(\theta)$, $\hat{\phi} \approx N(\phi, \tau^2)$ and τ is constant.
3. Bias-corrected	$\theta_{BC}[\alpha]$	$\hat{G}^{-1}([\phi[2z_\alpha + z^{(\alpha)}])$	There exists a monotone transformation such that $\hat{\phi} \approx N(\phi - z_0\tau, \tau^2)$ and z_0 and τ are constant.
4. BC_a	$\theta_{BC_a}[\alpha]$	$\hat{G}^{-1}\left(\phi\left[z_0+\frac{[z_0+z^{(\alpha)}]}{1-a[z_0+z^{(\alpha)}]}\right]\right)$	There exists a monotone transformation such that $\hat{\phi} \approx N(\phi - z_0\tau_\phi, \tau_\phi^2)$, where $\tau\phi = 1 + a\phi$ and z_0 and a are constant.

Note: Each method is correct under more general assumptions than its predecessor. Methods 2, 3, and 4 are defined in terms of the percentile of G, the bootstrap distribution.

Source: Efron and Tibshirani (1986, Table 6) with permission from The Institute of Mathematical Statistics.

problem by making a bias correction to the percentile method. Basically, the percentile method works if exactly 50% of the bootstrap distribution for $\hat{\theta}$ is less than $\hat{\theta}$.

By applying the Monte Carlo approximation, we determine an approximation to the bootstrap distribution. We find the 50th percentile of this distribution and call it $\hat{\theta}^*_{50}$. Taking this bias B to be $\hat{\theta} - \hat{\theta}^*_{50}$, we see that $\hat{\theta} - B$ equals $\hat{\theta}^*_{50}$ and so B is called the bias correction.

Another way to look at it, which is explicit but may be somewhat confusing, is to define $z_0 = \Phi^{-1}\{\hat{G}(\hat{\theta})\}$ (where Φ^{-1} is the inverse of the cumulative Gaussian distribution and $\hat{G}$ is the cumulative bootstrap sample distribution for θ. For a central $100(1 - 2\alpha)\%$ confidence interval, we then take the lower endpoint to be $\hat{G}^{-1}(\Phi\{2z_0 + z^{(\alpha)}\})$ and the upper endpoint to be $\hat{G}^{-1}(\Phi\{2z_0 + z^{(1-\alpha)}\})$. This is how Efron defines the bias correction method in Efron (1982a) and Efron

and Tibshirani (1986), where $z^{(\alpha)}$ satisfies $\Phi(z^{(\alpha)}) = \alpha$. Note that we use the "hat" notation over the cumulative bootstrap distribution G to indicate that Monte Carlo estimate of it is used.

It turns out that in the case of the law school data (assuming that it is a sample from a bivariate Gaussian distribution) the exact central 90% confidence interval is [0.496, 0.898]. The percentile method gives an interval of [0.536, 0.911] and the bias-corrected method yields [0.488, 0.900]. Since the bias-corrected method comes closer to the exact interval, we can conclude, in this case, that it is better than percentile method for the correlation coefficient.

What is important here is that this bias-correction method will work no matter what the value of θ really is. This means that after the adjustment, the monotone transformation leads to a distribution that is approximately Gaussian and whose variance does not depend on the transformed value, ϕ. If the variance cannot be made independent of ϕ, then a further adjustment, referred to as the acceleration constant a, is required.

Schenker (1985) provides an example for which the bias-correct percentile method did not work very well. It involves a χ^2 random variable with 19 degrees of freedom. In Efron and Tibshirani (1986) and Efron (1987) it is shown that the use of an acceleration constant overcomes the difficulty. It turns out in examples like Schenker's that there is a monotone transformation that works after a bias correction. The problem is that the resulting Gaussian distribution has a standard deviation τ_ϕ that depends linearly on ϕ (i.e., $\tau_\phi = 1 + a\phi$, where a is called the acceleration constant). A difficulty in the application of this modification to the bootstrap is the determination of the acceleration constant, a.

Efron found that a good approximation to the constant is one-sixth of the skewness of the score statistic evaluated at $\hat{\theta}$. See Efron and Tibshirani (1986) for details and examples of the computations involved.

Although this method seems to work in very general cases, it is complicated and may not be necessary. Bootstrap iteration to be explained in Section 3.1.4 is an alternative, as is the bootstrap percentile t method of Section 3.1.5.

These methods have a drawback that they share with the bootstrap percentile t intervals, namely, that they are not monotone in the assumed level of coverage (i.e., one could decrease the confidence level and not necessarily get a shorter interval that is contained in the interval obtained at the higher confidence level). This is not a desirable property and goes counter to our intuition about how confidence intervals should behave.

3.1.4. Iterated Bootstrap

A number of authors have contributed to the literature on bootstrap iteration, and we mention many of these contributors in the historical notes (Section 3.4). Major contributions were made by Peter Hall and his graduate student

Michael Martin. Martin (1990a) provides a clear and up-to-date summary of these advances [see also Hall (1992a, Chapter 3)].

Under certain regularity conditions on the population distributions, there has developed an asymptotic theory for the degree of closeness of the bootstrap confidence intervals to their stated coverage probability. Details can be found in a number of papers [e.g., Hall (1988b), Martin (1990a)].

An approximate confidence interval is said to be first-order accurate if its coverage probability differs from its advertised coverage probability by terms which go to zero at a rate of $n^{-1/2}$. The standard intervals discussed in Section 3.1.3 are first-order accurate. The BC_a intervals of Section 3.1.3 and the iterated bootstrap intervals to be discussed in this section are both second-order accurate (i.e., the difference goes to zero at rate n^{-1}).

A more important property for a confidence interval than just being accurate would be for the interval to be as small as possible for the given coverage probability. It may be possible to construct a confidence interval using one method which has coverage probability of 0.95, and yet it may be possible to find another method to use which will also provide a confidence interval with coverage probability 0.95, but the latter interval is actually shorter!

Confidence intervals that are "optimal" in the sense of being the shortest possible for the given coverage are said to be "correct." Efron (1990) provides a very good discussion of this issue along with some examples.

A nice property of these bootstrap intervals (i.e., the BC_a and the iterated bootstrap) is that in addition to being second-order accurate, they are also close to the ideal of "correct" interval in a number of problems where it makes sense to talk about "correct" intervals.

In fact the theory has gone further to show for certain broad parametric families of distributions that corrections can be made to get third-order accurate (i.e., with rate $n^{-3/2}$) intervals (Hall, 1988; Cox and Reid, 1987a and Welch and Peers, 1963).

Bootstrap iteration provides another way to improve the accuracy of bootstrap confidence intervals. Martin (1990a) discusses the approach of Beran (1987) and shows for one-sided confidence intervals that each bootstrap iteration improves the coverage by a factor of $n^{-1/2}$ and for two-sided intervals by n^{-1}.

What is a bootstrap iteration? Let us now describe the process. Suppose we have a random sample $\mathbf{X}$ of size n with observations denoted by $X_1, X_2, X_3, \ldots, X_n$. Let $X_1^*, X_2^*, X_3^*, \ldots, X_n^*$ denote a bootstrap sample obtained from this sample and let $\mathbf{X}^*$ denote this sample. Let I_0 denote a nominal $1-\alpha$ level confidence interval for a parameter ϕ of the population from which the original sample was taken. For example, I_0 could be a $1-\alpha$ level confidence interval for ϕ obtained by Efron's percentile method. To illustrate the dependence of I_0 on the original sample $\mathbf{X}$ and the level $1-\alpha$, we denote it as $I_0(\alpha|\mathbf{X})$. We then denote the actual coverage of the interval $I_0(\alpha|\mathbf{X})$ by $\pi_0(\alpha)$.

Let β_α be the solution to

$$\pi_0(\beta_\alpha) = P\{\theta \in I_0(\beta_\alpha|\mathbf{X}\} = 1-\alpha. \tag{3.2}$$

Now let $I_0(\beta_\alpha|\mathbf{X}^*)$ denote the version of I_0 computed using the resample in place of the original sample. The resampling principle of Hall and Martin (1988a) states that to obtain better coverage accuracy than given by the original interval I_0 we use $I_0(\beta_\alpha|\mathbf{X}^*)$ where

$$\hat{\beta}_\alpha \text{ is the estimate of } \beta_\alpha$$

in Equation (3.2) obtained by replacing ϕ with $\hat{\theta}$ and $\mathbf{X}$ with $\mathbf{X}^*$. To iterate again we just use the newly obtained interval in place of I_0 and apply the same procedure to it. An estimate based on a single iteration is called the double bootstrap and is the most common iterated estimate used in practice.

The algorithm just described is theoretically possible but in practice a Monte Carlo approximation must be used. In the Monte Carlo approximation B bootstrap resamples are generated. Details of the bootstrap iterated confidence interval are given in Martin (1990a, pp. 1113–1114). Although it is a complicated procedure to describe the basic idea is that by resampling from the B bootstrap resamples, we can estimate the point β_α and use that estimate to correct the percentile intervals. Results for particular examples using simulations are also given in Martin (1990a).

Clearly, the price paid for this added accuracy in the coverage of the confidence interval is an increase in the number of Monte Carlo replications. If we have an original sample size n and each bootstrap resample is of size n, then the number of replications will be nB_1B_2 where B_1 is the number of bootstrap samples taken from the original sample and B_2 is the number of bootstrap samples taken from each resample. In his example of two-sided intervals for the studentized mean from a folded normal distribution, Martin (1990a) uses $n = 10$, $B_1 = B_2 = 299$. The examples do seem to be in agreement with the asymptotic theory in that a single bootstrap iteration does improve the coverage in all cases considered.

Bootstrap iteration can be applied to any bootstrap confidence interval to improve the rate of convergence to the level $1-\alpha$. Hall (1992a) remarks that although his version of the percentile method may be more accurate than Efron's, bootstrap iteration works better on Efron's percentile method. The reason is not clear and the observation is based on empirical findings. A single bootstrap iteration provides the same type correction as BC_a does to Efron's percentile method. Using more than one bootstrap iteration is not common practice. This is due to the large increase in complexity and computation compared to the small potential gain in accuracy of the confidence interval.

3.1.5. Bootstrap Percentile t Confidence Intervals

The iterated bootstrap method and the BC_a confidence interval both provide improvements over Efron's percentile method, but both are complicated and the iterated bootstrap is even more computer-intensive than other bootstraps. The idea of the bootstrap percentile t method is found in Efron (1982a). A clearer presentation can be found in Efron and Tibshirani (1993, pp. 160–167). As a consequence of these attributes, it is popular in practice.

It is a simple method and has higher-order accuracy compared to Efron's percentile method. To be precise, bootstrap percentile t confidence intervals are second-order accurate (when they are appropriate). See Efron and Tibshirani (1993, pp. 322–325). Consequently, it is popular in practice. We used it in the Passive Plus DX clinical trial.

We shall now describe it briefly. Suppose that we have a parameter θ and an estimate θ_h for θ. Let θ^* be a nonparametric bootstrap estimate for θ based on a bootstrap sample and let S^* be an estimate of the standard deviation for θ_h based on the bootstrap samples. Define $T^* = (\theta^* - \theta_h)/S^*$. For each of the B bootstrap estimates θ^*, there is a corresponding T^*. We find the percentiles of T^*. For an approximate two-sided $100(1 - 2\alpha)\%$ confidence interval for θ, we take the interval $[\theta_h - t^*_{(1-\alpha)}S, \theta_h - t^*_{(\alpha)}]$, where $t^*_{(1-\alpha)}$ is the $100(1 - \alpha)$ percentile of the T^* values and $t^*_{(\alpha)}$ is the 100α percentile of the T^* values and S is the estimated standard deviation for θ_h. This we call the bootstrap t (or bootstrap percentile t as Hall refers to it) two-sided $100(1 - 2\alpha)\%$ confidence interval for .

A difficulty with the bootstrap t is the need for an estimate of the standard deviation S for θ_h and the corresponding bootstrap estimate S^*. In some problems there are obvious estimates, as in the simple case of a sample mean or the difference between the experimental group and control group means). For more complex parameters (e.g., C_{pk}) S may not be available.

3.2. RELATIONSHIP BETWEEN CONFIDENCE INTERVALS AND TESTS OF HYPOTHESES

In Section 3.1 of Good (1994), hypothesis testing for a single location parameter, θ, of a univariate distribution is introduced. In this it is shown how confidence intervals can be generated based on the hypothesis test. Namely for a $100(1 - \alpha)\%$ confidence interval, you include the values of θ at which you would not reject the null hypotheses at the level α. Conversely, if we have a $100(1 - \alpha)\%$ confidence interval for θ, we can construct an α level hypothesis test by simply accepting the hypothesis that $\theta = \theta_0$ if θ_0 is contained in the $100(1 - \alpha)\%$ confidence interval for θ and rejecting if it is outside of the interval.

In problems involving nuisance parameters, this procedure becomes more complicated. Consider the case of estimating the mean μ of a normal distribution when the variance σ^2 is unknown. The statistic $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ has Student's t distribution with $n - 1$ degrees of freedom where

$$\bar{x} = \sum_{i-1}^{n} x_i/n, s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}.$$

Here n is the sample size and x_i is the ith observed value. What is nice about the t statistic is that its distribution is independent of the nuisance parameter σ^2 and it is a pivotal quantity. Because its distribution does not depend on σ^2 or any other unknown quantities, we can use the tables of the t distribution to determine probabilities such as $P[a \le t \le b]$, where $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$.

Now t is also a pivotal quantity, which means that probability statements like the one above can be converted into confidence statements involving the unknown mean, μ. So if

$$P[a \le t \le b] = 1 - \alpha, \tag{3.3}$$

then the probability is also $1 - \alpha$ that the random interval

$$\left[\bar{x} - \frac{bs}{\sqrt{n}}, \bar{x} - \frac{as}{\sqrt{n}}\right] \tag{3.4}$$

includes the true value of the parameter μ, This random interval is then a $100(1 - \alpha)\%$ confidence interval for μ.

The interval (3.4) is a $100(1 - \alpha)\%$ confidence interval for μ, and we can start with Eq. (3.3) and get Eq. (3.4) or vice versa. If we are testing the hypothesis that $\mu = \mu_0$ versus the alternative that μ differs from μ_0, using (3.2), we replace μ with μ_0 in the t statistic and reject the hypothesis at the α level of significance if $t < a$ or if $t > b$.

We have seen earlier in this chapter how to construct various bootstrap confidence intervals with confidence level approximately $100(1 - \alpha)\%$. Using these bootstrap confidence intervals, we will be able to construct hypothesis tests by rejecting parameter values if and only if they fall outside the confidence interval. In the case of a translation family of distributions, the power of the test for the translation parameter is connected to the width of the confidence interval.

In the next section we shall illustrate the procedure by using a bootstrap confidence interval for the ratio of two variances in order to test the equality of the variances. This one example should suffice to illustrate how bootstrap tests can be obtained.

3.3. HYPOTHESIS TESTING PROBLEMS

In principle, we can use any bootstrap confidence interval for a parameter to construct a hypothesis test just as we have described it in the previous section (as long as we have a pivotal or asymptotically pivotal quantity or have no nuisance parameters). Bootstrap iteration and the use of bias correction with the acceleration constant are two ways by which we can provide more accuracy to the confidence interval by making the interval shorter without increasing the significance level. Consequently, the corresponding hypothesis test based on the iterated bootstrap or BC_a confidence interval will be more powerful than the test based on Efron's percentile interval, and it will more closely maintain the advertised level of the test.

Another key point that relates to accuracy is the choice of a test statistic that is asymptotically pivotal. Fisher and Hall (1990) pointed out that tests based on pivotal statistics often result in significance levels that differ from the advertised level by $O(n^{-2})$ as compared to $O(n^{-1})$ for tests based on non-pivotal statistics.

As an example, Fisher and Hall (1990) show that for the one-way analysis of variance, the F ratio is appropriate for testing equality of means when the variances are equal from group to group. For equal (homogeneous) variances the F ratio test is asymptotically pivotal.

However, when the variances differ (i.e., are heterogeneous) the F ratio depends on these variances, which are nuisance parameters. For the heterogeneous case the F ratio is not asymptotically pivotal. Fisher and Hall use a statistic first proposed by James (1951) which is asymptotically pivotal. Additional work on this topic can be found in James (1954).

In our example, we will be using an F ratio to test for equality of two variances. Under the null hypothesis that the two variances are equal, the F ratio will not depend on the common variance and is therefore pivotal.

In Section 3.3.2 of Good (1994), he points out that permutation tests had not been devised for this problem. On the other hand, there is no problem with bootstrapping. If we have n_1 samples from one population and n_2 from the second, we can independently resample with sample sizes of n_1 and n_2 from population one and population two, respectively.

We construct a bootstrap value for the F ratio by using a bootstrap sample of size n_1 from the sample from population one to calculate the numerator (a sample variance estimate for population one) and a bootstrap sample of size n_2 from the sample from population two to calculate the denominator (a sample variance estimate for population two). Since the two variances are equal under the null hypothesis, we expect the ratio to be close to one. By repeating this many times, we are able to get a Monte Carlo approximation to the bootstrap distribution for the F ratio. This distribution should be centered about one when the null hypothesis is true, and the extremes of the bootstrap distribution tell us how far from one we need to set our threshold

for the test. Since the F ratio is pivotal under the null hypothesis, we use the percentiles of the Monte Carlo approximation to the bootstrap distribution to get critical points from the hypothesis test. Alternatively, we could use the more sophisticated bootstrap confidence intervals, but in this case it is not crucial.

In the above example under the null hypothesis we assume $\sigma_1^2/\sigma_2^2 = 1$, and we would normally reject the null hypothesis in favor of the alternative that $\sigma_1^2/\sigma_2^2 \neq 1$, if the F ratio differs significantly from 1. However, in Hall (1992a, Section 3.12) he points out that the F ratio for the bootstrap sample should be compared or "centered" at the sample estimate rather than at the hypothesized value. Such an approach is known to generally lead to more powerful tests than the approach based on sampling at the hypothesized value. See Hall (1992a) or Hall and Wilson (1991) for more examples and a more detailed discussion of this point.

3.3.1. Tendril DX Lead Clinical Trial Analysis

In 1995 Pacesetter Inc., a St. Jude Medical Company that produces pacemakers and leads for patients with bradycardia, submitted a protocol to the United States Food and Drug Administration (FDA) for a clinical trial to demonstrate the safety and effectiveness of an active fixation steroid eluting lead. The study called for the comparison of the Tendril DX model 1388T with a concurrent control, the market-released Tendril model 1188T active fixation lead.

The two leads are almost identical, with the only differences being the use of titanium nitride on the tip of the 1388T lead and the steroid eluting plug also in the 1388T lead. Both leads were designed for implantation in either the atrial or the ventricular chambers of the heart, to be implanted with dual chamber pacemakers (most commonly Pacesetter's Trilogy DR+ pulse generator).

From the successful clinical trials of a competitor's steroid eluting leads and other research literature, it is known that the steroid drug reduces inflammation at the area of implantation. This inflammation results in an increase in the capture threshold for the pulse generator in the acute phase (usually considered to be the first six months post-implant).

Pacesetter statisticians (myself included) proposed as its primary endpoint for effectiveness a 0.5-volt or greater reduction in the mean capture threshold at the three-month follow-up for patients with 1388T leads implanted in the atrial chamber when they are compared to similar patients with 1188T leads implanted in the atrial chamber. The same hypothesis test was used for the ventricular chamber.

Patients entering the study were randomized as to whether they received the 1388T steroid lead or the 1188T lead. Since the effectiveness of steroid is well established from other studies in the literature, Pacesetter argued that it

would be unfair to patients in the study to give them only a 50–50 chance of receiving the 1388T lead (which is expected to provide less inflammation and discomfort and lower capture thresholds).

So Pacesetter designed the trial to have reasonable power to detect a 0.5-volt improvement and yet give the patient a 3-to-1 chance of receiving the 1388T lead. Such an unbalanced design required more patients for statistical conformation of the hypothesis (i.e., based on Gaussian assumptions, a balanced design required 50 patients in each group, whereas with the 3-to-1 randomization 99 patients were required in the experimental group and 33 in the control group to achieve the same power for the test at the 0.05 significance level), a total of 132 patients compared to the 100 for the balanced design.

The protocol was approved by the FDA and the trial proceeded. Interim reports and a pre-market approval report (PMA) were submitted to the FDA and the leads were approved for market release in June 1997.

Capture thresholds take on very discrete values due to the discrete programmed settings. Since the early data at three months was expected to be convincing but the sample size possibly relatively small, nonparametric approaches were taken as alternatives to the standard t tests based on Gaussian assumptions.

The parametric methods would only be approximately valid for large sample sizes due to the non-Gaussian nature of capture threshold distributions (possibly skewed, discrete and truncated). The Wilcoxon rank sum test was used as the nonparametric standard for showing improvement in the mean (or median) of the capture threshold distribution, and the bootstrap percentile method was also used to test the hypothesis.

Figures 3.1 and 3.3 show the distributions (i.e., histograms) of bipolar capture thresholds for 1188T and 1388T leads in the atrium and the ventricle, respectively, at the three-month follow-up visit. The variable, named "leadloc," refers to the chamber of the heart where the lead was implanted.

Figures 3.2 and 3.4 provide the bootstrap histogram of the difference in mean atrial capture threshold and mean ventricular capture threshold, respectively, for the 1388T leads versus the 1188T leads at the three-month follow-up.

The summary statistics in the box are N, the number of bootstrap replications; Mean, the mean of the sampling distribution; Std Deviation, the standard deviation of the bootstrap samples; Minimum, the smallest values out of the 5000 bootstrap estimates of the mean difference; and Maximum, the largest value out of the 5000 bootstrap estimates of the mean difference. Listed on the figures is the respective number of samples for the control (1188T) leads and for the investigational (1388T) leads in the original sample for which the comparison is made.

It also shows the mean difference of the original data that should be (and is) close in value to the bootstrap estimate of the sample mean. The estimate of the standard deviation for the mean difference is also given on the figures.

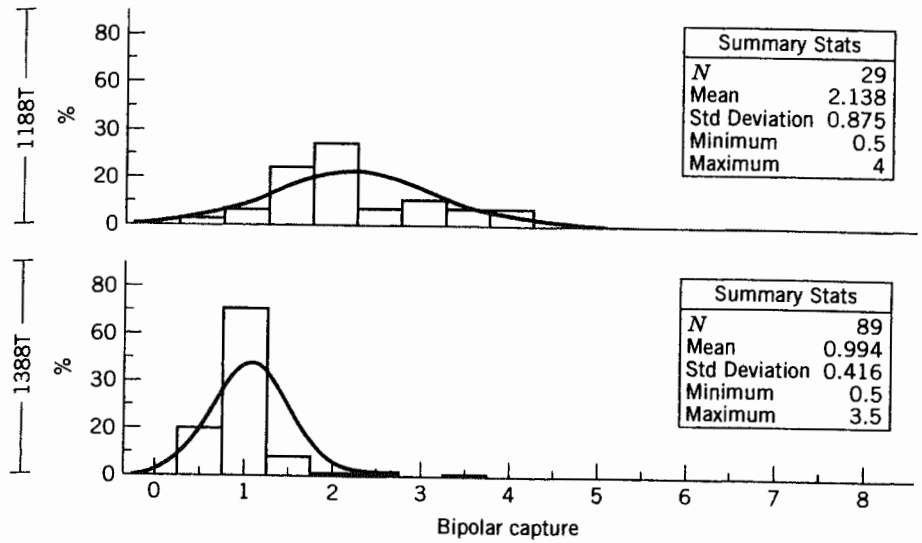

Figure 3.1 Capture threshold distributions for the three-month visit (leadloc; atrial chamber).

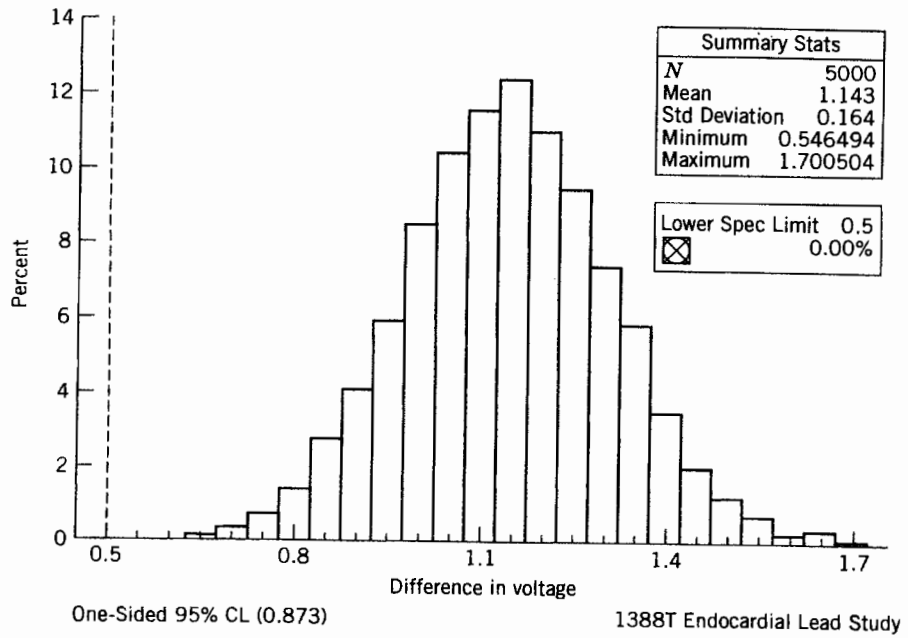

Figure 3.2 Distribution of bootstrapped data sets (atrium) bipolar three-month visit data as of March 15, 1996.

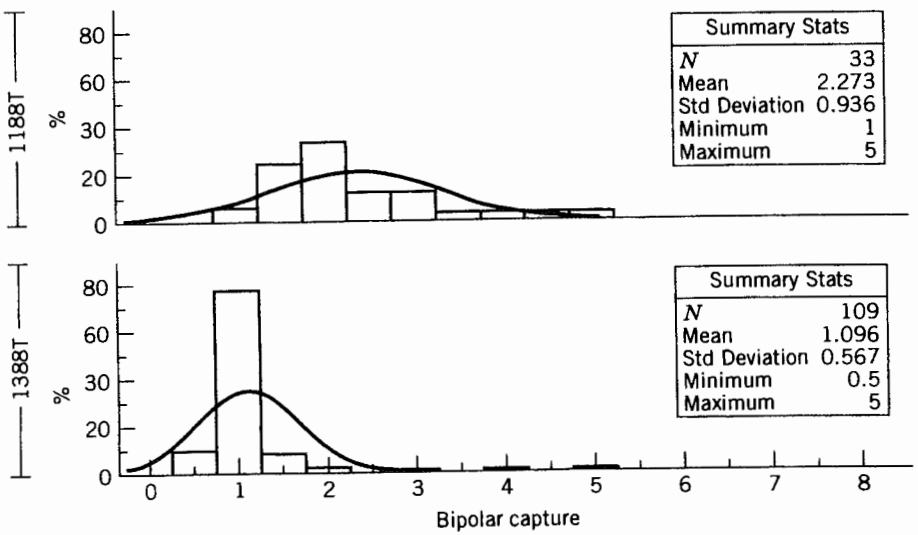

Figure 3.3 Capture threshold distributions for three-month visit (leadloc; ventricular chamber).

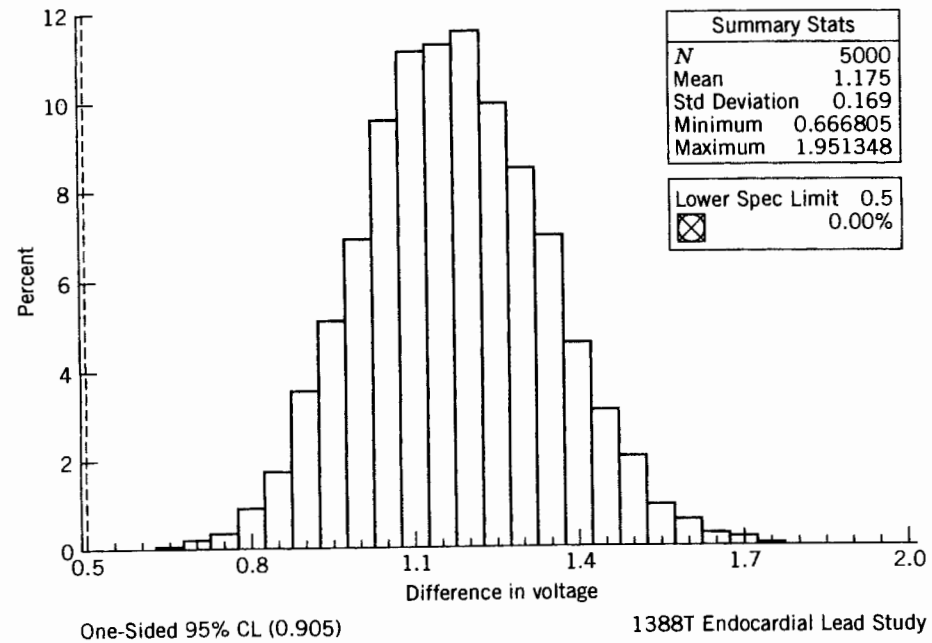

Figure 3.4 Distribution of bootstrapped data sets (ventricle) bipolar three-month data as of March 15, 1996.

We note that this too is very close in value to the bootstrap estimate for these data.

The histograms are based 5000 bootstrap replications on the mean differences. Also shown on the graph of the histogram is the lower 5th percentile (used in Efron's percentile method as the lower bound on the true difference for the hypothesis test). The proportion of the bootstrap distribution below zero provides a bootstrap percentile p-value for the hypothesis of no improvement versus a positive improvement in capture threshold.

Due to the slight skewness in the shape of the histogram that can be seen in Figures 3.1 and 3.3, the Pacesetter statisticians were concerned that the percentile method for determining the bootstrap lower confidence bound on the difference in the mean values might not be sufficiently accurate.

The bootstrap percentile t method was considered, but time did not permit the method to be developed in time for the submission. In a later clinical trial, Pacesetter took the same approach with the comparison of the control and treatment for the Passive Plus DX clinical trial.

The bootstrap percentile t method is a simple method to program and appears to overcome some of the shortcomings of Efron's percentile method without the complications of bias correction and acceleration constants. This technique was first presented by Efron as the bootstrap (Efron, 1982a, Section 10.10). Later, in Hall (1986a) asymptotic formulas were developed for the coverage error of the bootstrap percentile t method. This is the method discussed previously in Section 3.1.5.

The Passive Plus DX lead is a passive fixation steroid eluting lead that was compared with a non-steroid approved version of the lead. The 3:1 randomization of treatment group to control group was used in the Passive Plus study also.

In the Passive Plus study, the capture thresholds behaved similarly to those for the leads in the Tendril DX study. The main difference in the results was that the mean differences were not quite as large (i.e., close to the 0.5-volt improvement for the steroid lead over the non-steroid lead, whereas for Tendril DX the improvement was close to a 1.0-volt improvement).

In the Passive Plus study, both the bootstrap percentile method lower 95% and the bootstrap percentile t method lower 95% confidence bounds were determined.

3.4. AN APPLICATION OF BOOTSTRAP CONFIDENCE INTERVALS TO BINARY DOSE–RESPONSE MODELING

At pharmaceutical companies, a major part of the early phase II development is the establishment of a dose–response relationship for a drug that is being considered for marketing. At the same time estimation of doses that are minimally effective or maximally safe are important to determine what is the best dose or small set of doses to carry over into phase III trials. The following

example, Klingenberg (2007), was chosen because it addresses methods that are important in improving the phase 2 development process for new pharmaceuticals (an important application) and it provides an example where resampling methods are used in a routine fashion. Permutation methods are used for p-value adjustment due to multiplicity, and bootstrap confidence intervals are used to estimate the minimum effective dose after proof of concept.

In the spirit of faster development of drugs through adaptive design concepts, Klingenberg (2007) proposes a unified approach to determining proof of concept with a new drug followed by dose–response modeling and dose estimation. In this paper, Klingenberg describes some of the issues that have motivated this new statistical research. The purpose of the paper is to provide a unified approach to proof of concept (PoC) phase 2a clinical trials with the dose finding phase 2b trials in an efficient way when the responses are binary. The goal at the end of phase 2 is to find a dose for the drug that will be safe and effective and therefore will have a good chance for success in phase 3. Klingenberg cites the following statistics as an indication of the need to find different approaches that have better chances of achieving the phase 2 objectives.

He notes that the current failure rate for phase 3 trials is approaching 50%, largely attributed to improper target dose estimation/selection in phase II and incorrect or incomplete knowledge of the dose–response, and the FDA reports that 20% of the approved drugs between 1980 and 1989 had the initial dose changed by *more than 33%*, in most cases lowering it. So current approaches to phase 2 trials are doing a poor job of achieving the objectives since poor identification of dose is leading to the use of improper doses that lead to wasted phase 3 trials and even when the trials succeed, they often do so with a less than ideal choice of dose and in the post-marketing phase the dose is determined to be too high and reduced dramatically.

The idea of the approach is to use the following strategy: (1) Work with the clinical team to identify a reasonable class of potential dose–response models; (2) from this comprehensive set of models, choose the ones that best describe the dose–response data; (3) use model averaging to estimate a target dose; (4) decide which models, if any, significantly pick up the signal, establishing PoC; (5) use the permutation distribution of the maximum penalized deviance over the candidate set to determine the best model (s0); and (6) use the best model to estimate the minimum effective dose (MED). Important aspects of the approach are the use of permutation methods to determine adjusted p-values and control the error rate of declaring spurious signals as significant (due to the multiplicity of models considered). A thorough evaluation and comparison of the approach to popular contrast tests reveals that its power is as good or better in detecting a dose–response signal under a variety of situations, with many more additional benefits: It incorporates model uncertainty in proof of concept decisions and target dose estimation,

yields confidence intervals for target dose estimates (MED), allows for adjustments due to covariates, and extends to more complicated data structures. Klingenberg illustrates his method with the analysis of a Phase II clinical trial.

The bootstrap enters into this process as the procedure for determining confidence intervals for the dose. Permutation methods due to Westfall and Young (1993) were used for the p-value adjustment. Westfall and Young (1993) also devised a bootstrap method for p-value adjustment that is very similar to the permutation approach and could also have been used. We cover the bootstrap method for p-value adjustment with some applications in Chapter 8.

The unified approach that is used by Klingenberg is similar to the approach taken by Bretz, Pinheiro, and Branson (2005) for normally distributed data but applied to binomial distributed data. MED estimation in this paper follows closely the approach of Bretz, Pinheiro, and Branson (2005). A bootstrap percentile method confidence interval for MED is constructed using the fit to the chosen dose-response model. The confidence interval is constructed conditional on the establishment of PoC. Klingenberg illustrates the methodology by reanalyzing data from a phase 2 clinical trial using a unified approach.

In Klingenberg's example, the key variable is a binary indicator for the relief of symptoms from irritable bowl syndrome (IBS), a disorder that is reported to affect up to 30% of all Americans at sometime during their lives (American Society of Colon and Rectal Surgeons, www.fascrs.org). A phase II clinical trial investigated the effcacy of a compound against IBS in women at $k = 5$ dose levels ranging from placebo to 24 mg. Expert opinion was used to determine a target dose. Here, Klingenberg reanalyzes these data within the statistical framework of the unified approach.

Preliminary studies with only two doses indicated a placebo effect of roughly 30% and a maximal possible dose effect of 35%. However, prior to the trial, investigators were uncertain about the monotonicity and curvature of a possible dose effect. The first eight models and the zero effect model are pictured in Figure 3.5 for a particular prior choice of parameter values, cover a broad range of dose–response shapes deemed plausible for his particular compound, and were selected to form the candidate set. The candidate models had to be somewhat broad because the investigators could not rule out strongly concave or convex patterns or even a down-turn at higher doses, and hence the candidate set includes models to the see these possible effects. All models in Figure 3.5, most with fractional polynomial (Roystone and Altman, 1994) linear predictor form, are fit to the data by maximum likelihood, but some of the models might not converge for every possible data set.

The author is interested in models that pick up a potential signal observed in a dose–response study. To this end, he compared each of the eight models to the model of no dose effect via a (penalized) likelihood ratio test. A description of the models is given in Table 3.2.

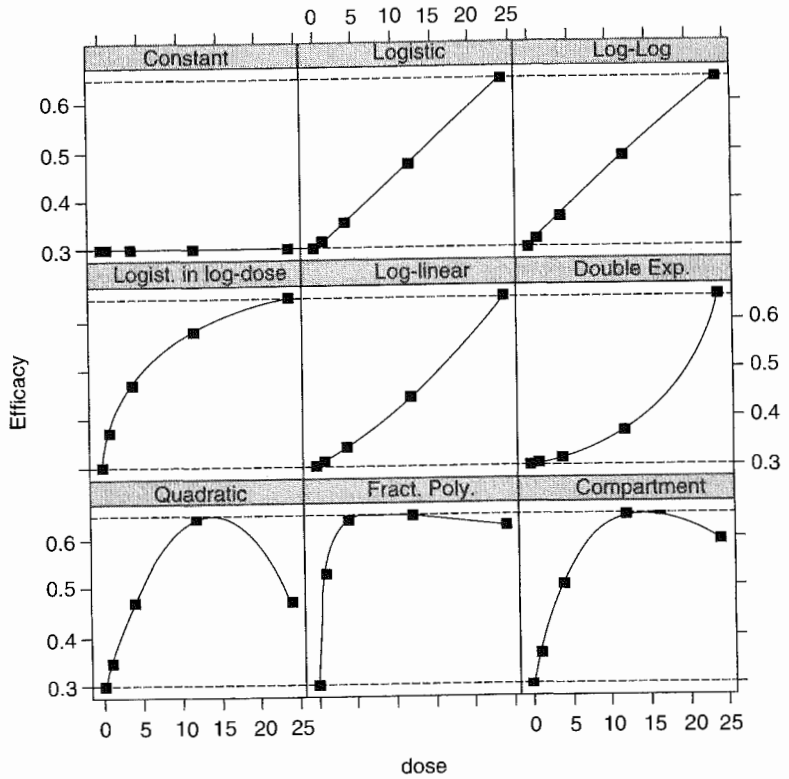

Figure 3.5 A zero effect model and eight candidate dose–response models. [Taken from Klingenberg (2007) with permission.]

Table 3.2 Dose–Response Models for the Efficacy of Irritable Bowel Syudrome Compound

M	Model	Link	Predictor	Number of Permutations
M_1:	Logistic	logit	$\beta_0 + \beta_1 d$	2
M_2:	Log-Log	log–log	$\beta_0 + \beta_1 d$	2
M_3:	Logistic in log-dose	logit	$\beta_0 + \beta_1 \log(d+1)$	2
M_4:	Log-linear	log	$\beta_0 + \beta_1 d$	2
M_5:	Double-exponential	identity	$\beta_0 + \beta_1 \exp(\exp(d/\max(d)))$	2
M_6:	Quadratic	identity	$\beta_0 + \beta_1 d + \beta_2 d^2$	3
M_7:	Fractional Poly	logit	$\beta_0 + \beta_1 \log(d+1) + \beta_2/(d+1)$	3
M_8:	Compartment	identity	$\beta_0 + \beta_1 d \exp(-d/\beta_2)$, $\beta_2 > 0$	3
M_9:	Square root	logit	$\beta_0 + \beta_1 d^{1/2}$	2
M_{10}:	Emax	logit	$\beta_0 + \beta_1 d/(\beta_2 + d)$, $\beta_2 > 0$	3

Source: Taken from Klingenberg (2007) with permission.

Table 3.3 G_s^2 -Statistics, p-Values, Target Dose Estimates and Model Weights

Model Number	Model Type	G_s^2	Raw p-Value	Adjusted p-Value	MED (mg)	Model Weight (%)
M_1	Logistic	3.68	0.017	0.026	N/A	0
M_2	Log–Log	3.85	0.015	0.024	N/A	0
M_3	Logistic in log-dose	10.53	$<10^{-3}$	0.001	7.9	6
M_4	Log-linear	3.25	0.022	0.032	N/A	0
M_5	Double-exponential	0.90	0.088	0.106	N/A	0
M_6	Quadratic	6.71	0.005	0.005	7.3	1
M_7	Fractional Polynomial	15.63	$<10^{-4}$	$<10^{-4}$	0.7	81
M_8	Compartment	11.79	$<10^{-3}$	$<10^{-3}$	2.5	12
		Critical value 2.40			MED (avg.) = 1.4	
					95% Conf. Int. = [0.4, 12.0]	

Source: Adapted from Klingenberg (2007) with permission.

Table 3.3 gives the results for the tests of the models including the raw and adjusted p-values. Also included are the weights used in the model averaging. For each model that was included the point estimate of the MED is given. Also we see that the weighted average of the four selected models is 1.4 and the 95% bootstrap percentile confidence interval is [0.4, 12.0]. The critical value for the null (permutation) distribution of the maximum penalized deviance is shown to be 2.4, and seven of the eight models (all but M_5) have a test statistic that exceeds the critical value. But only models (M_3, M_6, M_7, and M_8) were used in the final averaging.

3.5. HISTORICAL NOTES

Bootstrap confidence intervals were introduced in Efron (1982a, Chapter 10). Efron's percentile method and the bias corrected percentile method were introduced at that time. Efron also introduced in Efron (1982a) the bootstrap t intervals and illustrated these techniques with the median as the parameter.

It was recognized at that time that confidence interval estimation was a tougher problem than estimating standard errors and considerably more bootstrap samples would be required (i.e., 1000 bootstrap samples for confidence intervals where only 100 would be required for standard error estimates). For more discussion of this issue see Section 7.1.

In an important work, Schenker (1985) shows that bias adjustment to Efron's percentile method is not always sufficient to provide "good" confi-

dence intervals. Nat Schenker's examples motivated Efron to come up with the use of an acceleration constant as well as a bias correction in the modification of the confidence interval endpoints. This led to a significant improvement in the bootstrap confidence intervals and removed Schenker's objections.

The idea of bootstrap iteration to improve confidence interval estimation appears in Hall (1986a), Beran (1987), Loh (1987), Hall and Martin (1988a), and DiCiccio and Romano (1988). The methods of Hall, Beran, and Loh all differ in the way they correct the critical point(s). Loh refers to his approach as bootstrap calibration.

Hall (1986b) deals with sample size requirements. Specific application to the confidence interval estimation for the correlation coefficient is given in Hall, Martin, and Schucany (1989). For further developments in bootstrap iteration see Martin (1990a), Hall (1992a), or Davison and Hinkley (1997).

Some of the asymptotic theory is based on formal Edgeworth expansions that were rigorously developed in Bhattacharya and Ghosh (1978) [see Hall (1992a) for a detailed account with applications to the bootstrap]. Other asymptotic expansions such as saddlepoint approximations may provide comparable confidence intervals without the need for Monte Carlo [see the monograph by Field and Ronchetti (1990) and the papers by Davison and Hinkley (1988) and Tingley and Field (1990)].

DiCiccio and Efron (1992) also obtain very good confidence intervals without Monte Carlo for data from an exponential family of distributions. DiCiccio and Romano (1989a) also produce accurate confidence limits by making some parametric assumptions.

Some the research in the 1980s and late 1990s suggests that the Monte Carlo approximation may not be necessary (see Section 7.3 and the references above) or that the number of Monte Carlo replications can be considerably reduced by variance reduction techniques [see Section 7.2 and Davison, Hinkley, and Schechtman (1986), Therneau (1983), Hesterberg (1988), Johns (1988), and Hinkley and Shi 1989]. The most recent developments can be found in Hesterberg (1995a,b, 1996, 1997).

Discussions of bootstrap hypothesis tests appear in the early paper of Efron (1979a) and some work can be found in Beran (1988c), Hinkley (1988), Fisher and Hall (1990) and Hall and Wilson (1991). Specific applications and Monte Carlo studies of bootstrap hypothesis testing problems are given in Dielman and Pfaffenberger (1988), Rayner (1990a,b), and Rayner and Dielman (1990).

Fisher and Hall (1990) point out that even though there are close connections between bootstrap hypothesis tests and confidence intervals there are also important differences which lead to specialized treatment. They recommend the use of asymptotic pivotal quantities in order to maintain a close approximation to the advertised significance level for the test.

Ideas are illustrated using the analysis of variance problem with both real and simulated data sets. Results based on Edgeworth expansions and Cornish–Fisher expansions clearly demonstrate the advantage of bootstrapping pivotal

statistics for both hypothesis testing and confidence intervals [see Hall (1992a)]. Lehmann (1986) is the second edition of a classic reference on hypothesis testing and any reader wanting a rigorous treatment of the subject would be well advised to consult that text.

The first application of Edgeworth expansions to derive properties for the bootstrap is Singh (1981). The work of Bickel and Freedman (1981) is similar to that of Singh (1981) and also uses Edgeworth expansions. Their work shows how bootstrap methods correct for skewness.

Both papers applied one-term Edgeworth expansion corrections. Much of the development of Edgeworth expansions goes back to the determination of particular cumulants, as in James (1955, 1958).

The importance of asymptotically pivotal quantities was not brought out in the early papers because the authors considered a nonstudentized sample mean and assumed the population variance is known. Rather this result was first mentioned by Babu, and Singh in a series of papers (Babu and Singh, 1983, 1984a, and 1985). Another key paper on the use of Edgeworth expansions for hypothesis testing is Abramovitch and Singh (1985).

Hall (1986a, 1988b) wrote two key papers which demonstrate the value of asymptotically pivotal quantities in the accuracy of bootstrap confidence intervals.

Hall (1986a) derives asymptotic formulas for coverage error of the bootstrap percentile t confidence intervals and Hall (1988b) gives a general theory for bootstrap confidence intervals. Theoretical comparisons of variations on bootstrap percentile t confidence intervals are given in Bickel (1992). Other papers that support the use of pivotal statistics are Beran (1987) and Liu and Singh (1987).

Methods based on symmetric bootstrap confidence intervals are introduced in Hall (1988a). Hall also defines "short" bootstrap confidence intervals in Hall (1988b) [see also Hall (1992a) for some discussions]. The idea for the "short" bootstrap confidence intervals goes back to Buckland (1980, 1983).

Efron first proposed his version of the percentile method in Efron (1979a) [see also Efron (1982a) for detailed discussions]. The BC_a intervals were first given in Efron (1987). Buckland (1983, 1984, 1985) provide applications for Efron's bias correction intervals along with algorithms for their construction.

Bootstrap iteration in the context of confidence intervals is introduced in Hall (1986a) and Beran (1987). Hall and Martin (1988a) develop a general framework for bootstrap iteration. Loh (1987) introduced the notion of bootstrap calibration. When applied to bootstrap confidence intervals, calibration is equivalent to bootstrap iteration.

Other important works related to confidence intervals and hypothesis testing include Beran (1986, 1990a,b).