CHAPTER 2

Estimation

In this chapter, we deal with problems involving point estimates. Section 2.1 covers the estimation of the bias of an estimator by the bootstrap technique. After showing you how to use the bootstrap to estimate bias in general, we will focus on the important application to the estimation of error rates in the classification problem.

This will require that we first provide you with an introduction to the classification problem and the difficulties with the classical estimation procedures when the training set is small. Another application to classification problems, the determination of a subset of features to be included in the classification rule, will be discussed in Section 8.2.

Section 2.2 explains how to bootstrap to obtain point estimates of location and dispersion parameters. When the distributions have finite second moments, the mean and the standard deviation are the common measures. However, we sometimes have to deal with distributions that do not even have first moments (the Cauchy distribution is one such example).

Such distributions come up in practice when taking ratios or reciprocals of random variables where the random variable in the denominator can take on the value zero or values close to zero. The commonly used location parameter is the median, and the interquartile range R is a common measure of dispersion where $R = L_{75} - L_{25}$ for L_{75} the 75th percentile of the distribution and L_{25} the 25th percentile of the distribution.

2.1. ESTIMATING BIAS

2.1.1. How to Do It by Bootstrapping

Let $E(X)$ denote the expected (or mean) value of a random variable X. For an estimator $\hat{\theta}$ of a parameter θ, we consider the random variable $\hat{\theta} - \theta$ for our X. The bias of an estimator $\hat{\theta}$ for θ is defined to be $b = E(\hat{\theta} - \theta)$. As an example, the sample variance,

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

based on a sample of n independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ from a population distribution with a finite variance, is an unbiased estimator for σ^2, the population variance where

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}.$$

On the other hand, for Gaussian populations the maximum likelihood estimator for σ^2 is equal to

$$(n-1)S^2/n.$$

It is a biased estimator with the bias equal to

$$-\sigma^2/n \qquad \text{since} \qquad E[(n-1)S^2/n] = (n-1)\sigma^2/n.$$

The bootstrap estimator B^* of b is then $E(\theta^* - \hat{\theta})$, where θ^* is an estimate of θ based on a bootstrap sample. A Monte Carlo approximation to B^* is obtained by doing k bootstrap replications as described in Section 1.1.

For the ith bootstrap replication, we denote the estimate of θ by θ_i^*. The Monte Carlo approximation to B^* is the average of the differences between the bootstrap sample estimates $\theta_i^*$of θ and the original sample estimate $\hat{\theta}$,

$$B_{\text{Monte}} = \sum_{i=1}^{k} (\theta_i^* - \hat{\theta})/k.$$

Generally, the purpose of estimating bias is to improve a biased estimator by subtracting an estimate of its bias from it. In Section 2.1.2, we shall see that Efron's definition of the bias is given by the negative of the definition given here [i.e., $B^* = E(\hat{\theta} - \theta^*)$], and consequently we will add the bias to the estimator rather than subtract it.

Bias correction was the original idea that led to a related resampling method, the jackknife [dating back to Quenouille (1949) and Tukey (1958)]. In the next section, we find an example of an estimator which in small samples has a large bias but not a very large variance. For this problem, the estimation of the prediction error rate in linear discriminant analysis, the bootstrap bias correction approach to the estimating the error rate is a spectacular success!

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

2.1.2. Error Rate Estimation in Discrimination

First you'll be given a brief description of the two-class discrimination problem. Then, some of the traditional procedures for estimating the expected conditional error rate (i.e., the expected error rate given a training set) will be described. Ncxt we will provide a description some of the various bootstrap-type estimators that have been applied. Finally, results are summarized for some of the simulation studies that compared the bootstrap estimators with the resubstitution and leave-one-out (or cross-validation) estimators.

I again emphasize that this particular example is one of the big success stories for the bootstrap. It is a case where there is strong empirical evidence for the superiority of bootstrap estimates over traditional methods, particularly when the sample sizes are small!

In the two-class discrimination problem you are given two classes of objects. A common example is the case of a target and some decoys that are made to look like the target. The data consist of a set of values for variables which are usually referred to as features.

We hope that the values of the features for the decoys will be different from the values for the targets. We shall also assume that we have a training set (i.e., a sample of features for decoys and a separate sample of features for targets where we know which values correspond to targets and which correspond to decoys). We need the training set in order to learn something about the unknown feature distributions for the target and the decoy.

We shall briefly mention some of the theory for the two-class problem. The interested reader may want to consult Duda and Hart (1973), Srivastava and Carter (1983, pp. 231–253), Fukunaga (1990), or McLachlan (1992) for more details.

Before considering the use of training data, for simplicity, let us suppose that we know exactly the probability density of the feature vector for the decoys and also for the targets. These densities shall be referred to as the class-conditional densities.

Now suppose someone discovers a new object and does not know whether it is a target or a decoy but does have measured or derived values for that object's features. Based on the features, we want to decide whether it is a target or a decoy.

This is a classical multivariate hypothesis testing problem. There are two possible decisions: (1) to classify the object as a decoy and (2) to classify the object as a target. Associated with each possible decision is a possible error: We can decide (1) when the object is a target or we can decide (2) when the object is a decoy.

Generally, there are costs associated with making the wrong decisions. These costs need not be equal. If the costs are equal, Bayes' theorem provides us with the decision rule that minimizes the cost.

For the reader who is not familiar with Bayes' theorem, it will be presented in the context of this problem, after we define all the necessary terms. Even with unequal costs, we can use Bayes' theorem to construct the decision rule which minimizes the expected cost. This rule is called the Bayes rule and it follows our intuition.

For equal costs, we classify the object as a decoy if the a posteriori probability of a decoy given that we observe the feature vector **x** is higher for the decoy than the a posteriori probability of a target given that we observe feature **x**. We classify it as a target otherwise.

Bayes' theorem gives us a way to compute these a posteriori probabilities. If our a priori probabilities are equal (i.e., before collecting the data we assume that the object is as likely to be a target as it is to be a decoy), the Bayes' rule is equivalent to the likelihood ratio test.

The likelihood ratio test classifies the object as the type which has the greater likelihood for **x** (i.e., the larger class conditional density). For more discussion see Duda and Hart (1973, p. 16).

Many real problems have unequal a priori probabilities; sometimes we can determine these probabilities. In the target versus decoy example, we may have intelligence information that the enemy will put out nine decoys for every real target. In that case, the a priori probability for a target is .1, whereas the a priori probability for a decoy is 0.9.

Let $P_D(\mathbf{x})$ be the class conditional density for decoys and let $P_T(\mathbf{x})$ be the class conditional density for targets. Let C_1 be cost of classifying a decoy as a target, C_2 the cost of classifying a target as a decoy, P_1 the a priori probability for a target and P_2 the a priori probability for a decoy.

Let $P(D|\mathbf{x})$ and $P(T|\mathbf{x})$ denote, respectively, the probability that an object with feature vector **x** is a decoy and the probability that an object with feature vector **x** is a target. For the two-class problem it is obvious that $P(T|\mathbf{x}) = 1 - P(D|\mathbf{x})$ since the object must be one of these two types. By the same argument, $P_1 = 1 - P_2$ for the two-class problem. Bayes' theorem states that

$$P(D|\mathbf{x}) = P_D(\mathbf{x})P_2/[P_D(\mathbf{x})P_2 + P_2(\mathbf{x})P_1] = P_D(\mathbf{x})P_2/[P_D(\mathbf{x})P_2 + P_T(\mathbf{x})(1 - P_2)].$$

The Bayes rule, which minimizes expected cost, is defined as follows:

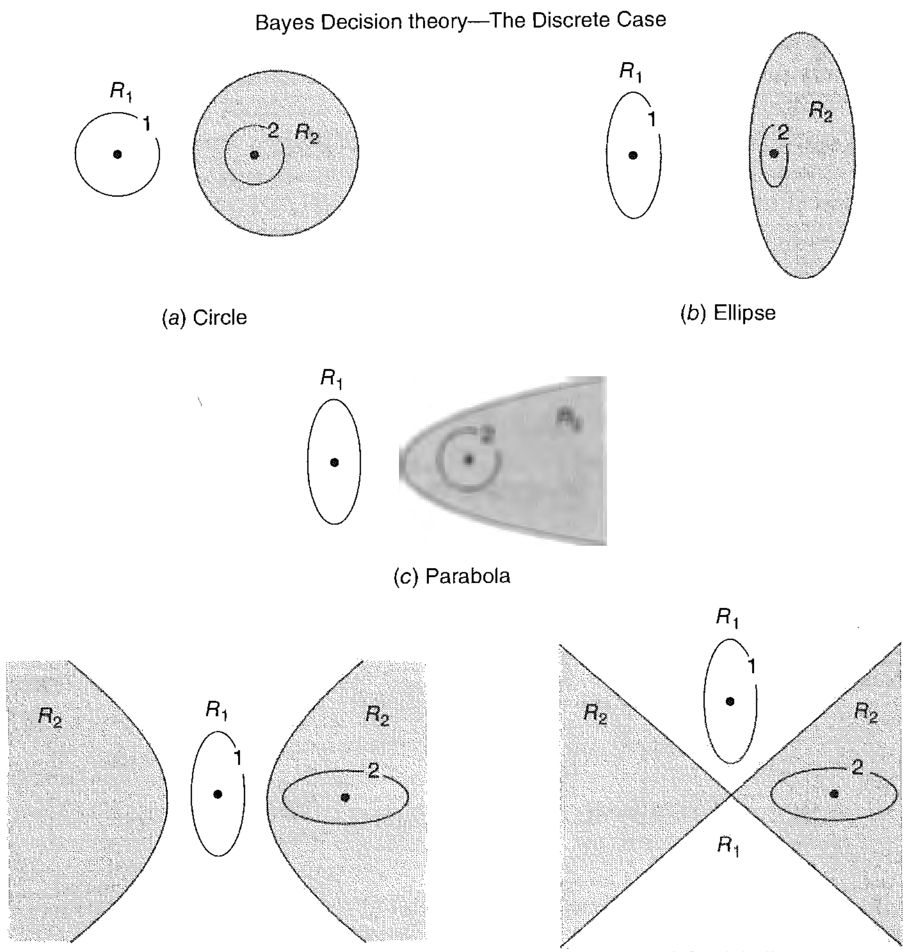
$$\text{Classify the object as a decoy if } \frac{P_D(\mathbf{x})}{P_T(\mathbf{x})} > K,$$

$$\text{Classify the object as a target if } \frac{P_D(\mathbf{x})}{P_T(\mathbf{x})} \le K,$$

where $K = (C_2P_1)/(C_1P_2)$. See Duda and Hart (1973, pp. 10–15) for a derivation of this result.

Notice that we have made no assumptions about the form of the class-conditional densities. The Bayes rule works for any probability densities. Of course, the form of the decision boundary and the associated error rates depend on these *known* densities. If we make the further assumption that the densities are both multivariate Gaussian with different covariance matrices, then Bayes' rule has a quadratic decision boundary (i.e., the boundary is a quadratic function of **x**).

If the densities are Gaussian and the covariance matrices are equal, then Bayes' rule has a linear boundary (i.e., the boundary is a linear function of **x**). Both of these results are derived in Duda and Hart (1973, pp. 22–31). The possible decision boundaries for Gaussian distributions with unequal covariances and two-dimensional feature vectors are illustrated in Figure 2.1, which was taken from Duda and Hart (1973, p. 31).



(*a*) Circle

(*b*) Ellipse

(*c*) Parabola

(*d*) Hyperbola

(*e*) Straight lines

Figure 2.1 Forms for decision boundaries for the general bivariate normal case. [From Duda and Hart (1973), p. 31, with permission from John Wiley and Sons, Inc.]

The circles and ellipses in the figure represent, say, the one sigma equal probability contours corresponding to the covariances. These covariances are taken to be diagonal without any loss of generality. The shaded region R_2 is the region in which class 2 is accepted.

In many practical problems the class-conditional densities are not known. If we assume the densities to be Gaussian, the training samples can be used to estimate the mean vectors and covariance matrices (i.e., the parameters required to determine the densities). If we have no knowledge of the form of the underlying densities, we may use available data whose classes are known (such data are referred to as the training data) to obtain density estimates.

One common approach is to use the kernel density estimation procedure. The rule used in practice replaces the Bayes rule (which is not known) with an approximation to it based on the replacement of the class-conditional densities in the Bayes rule with the estimated densities.

Although the resulting rule does not have the optimal properties of the Bayes rule, we argue that it is an appropriately optimal rule since as the training set gets larger and larger for both classes the estimated densities come closer and closer to the true densities and the rule comes closer and closer to the Bayes rule. For small sample sizes, it at least appears to be a reasonable approach. We shall call this procedure the estimated decision rule. To learn more about kernel discrimination, consult Hand (1981, 1982).

For known class-conditional densities, the Bayes rule can be applied and the error rates calculated by integrating these densities in the region in which a misclassification would occur. In parametric problems, so-called "plug-in" methods compute these integrals using the estimated densities obtained by plugging in the parameter estimates for their unknown values. These plug-in estimates of the error rates are known to be optimistically biased [i.e., they tend to underestimate the actual expected error rates; see Hills (1966)].

When we are unable to make any parametric assumptions, a naive approach is to take the estimated decision rule, apply it to the training data, and then count how many errors of each type would be made. We divide the number of misclassified objects in each class by their respective number of training samples to get our estimates of the error rates. This procedure is referred to as the resubstitution method, since we are substituting training samples for possible future cases and these training samples were already used to construct the decision rule.

In small to moderate sample sizes the resubstitution estimator is generally a poor estimator because it also tends to have a large optimistic bias (actually the magnitude of bias depends on the true error rate). Intuitively, the optimistic bias of the plug-in and resubstitution estimators is due to the fact that in both cases the training data are used to construct the rule and then reused to estimate the error rates.

Ideally, it would be better to estimate the error rates based on an independent set of data with known classes. This, however, creates a dilemma. It is wasteful to throw away the information in the independent set, since these data could be used to enlarge the training set and hence provide better estimates of the class-conditional densities. On the other hand, the holdout estimator, obtained by the separation of this independent data set for error rate estimation from the training set, eliminates the optimistic bias of resubstitution.

Lachenbruch (1967) [see also Lachenbruch, and Mickey (1968)] provided the leave-one-out estimate to overcome the dilemma. Each training vector is used in the construction of the rule. To estimate the error rate, the rule is reconstructed n times where n is the total number of training vectors. In the ith reconstruction, the ith training vector is left out of the construction.

We then count the ith vector as misclassified if the reconstructed rule would misclassify it. We take the total number misclassified in each class and divide by the number in the respective class to obtain the error rates.

This procedure is referred to as leave-one-out or cross-validation and the estimators are called the leave-one-out estimates or U estimates. Because the observations are left out one at a time, some have referred to it as the jackknife estimator, but Efron (1982a, pp. 53–58) defines another bias correction estimator to be the jackknife estimator [see also Efron (1983)].

Now, you'll be shown how to bootstrap in this application. Essentially, we will apply the bootstrap bias correction procedure that we learned about in Section 2.1.1 to the resubstitution estimator.

The resubstitution estimator, although generally poor in small samples, has a large bias that can be estimated by bootstrapping [see, for example, Chernick, Murthy, and Nealy (1985, 1986)]. Cross-validation (i.e., the leave-one-out estimator) suffers from a large variance for small training sample sizes. Despite this large variance, cross-validation has been traditionally the method of choice.

Glick (1978) was one of the first to recognize the problem of large variance with the leave-one-out estimate, and he proposed certain "smooth" estimators as an alternative. Glick's approach has since been followed up by Snapinn and Knoke (1984, 1985a).

Efron (1982a, 1983) showed that the bootstrap bias correction can produce an estimator that is nearly unbiased (the bias is small though not quite as small as for the leave-one-out estimator) and has a far smaller variance than the leave-one-out estimator. Consequently, the bootstrap is superior in terms of mean square error (a common measure of statistical accuracy).

As a guideline to the practitioner, I believe that the simulation studies to date indicate that for most applications, the .632 estimator is to be preferred. What follows is a description of the research studies to date that provide the evidence to support this general guideline. We shall now describe the various bootstrap estimators that were studied in Efron (1983) and in Chernick, Murthy, and Nealy (1985, 1986, 1988a,b).

It is important to clarify here what error rate we are estimating. It was pointed out by Sorum (1972) that when training data are involved, there are

at least three possible error rates to consider [see also Page (1985) for a more recent account].

In the simulation studies that we review here, only one error rate is considered. It is the expected error rate conditioned on the training set of size n. This averages the two error rates (weighing each equally). It is the natural estimator to consider since in the classification problem the training set is fixed and we need to predict the class for new objects based solely on our prior knowledge and the particular training set at hand.

A slightly different and less appropriate error rate would be the one obtained by averaging these conditional error rates over the distribution of possible training sets of size n. Without carefully defining the error rate to be estimated, confusion can arise and some comparisons may be inappropriate.

The resubstitution estimator and cross-validation have already been defined. The standard bootstrap (obtained using 100–200 bootstrap samples in the simulations of Efron and Chernick, Murthy, and Nealy) uses the bootstrap sample analog to Equation 2.10 of Efron (1983, p. 317) to correct the bias. Define the estimated bias as

$$\omega_h = E_*[\Sigma_i(n^{-1} - P_i^*)Q[y_i, \eta(t_i, X^*)]],$$

where E_* denotes the expectation under the bootstrap random sampling mechanism (i.e., sampling with replacement from the empirical distribution), $Q[y_i, \eta(t_i, X^*)]$ is the indicator function defined to be equal to one if $y_i = \eta(t_i, X^*)$ and zero if $y_i \neq \eta(t_i, X^*)$, y_i is the ith observation of the response, t_i is the vector of predictor variables, and η is the prediction rule.

$\mathbf{X}^*$ is the vector for a bootstrap sample a (of length n) and P_i^* is the ith repetition frequency (i.e. the proportion of cases in particular where the ith sample value occurs). The bootstrap estimate is then $e_{\text{boot}} = \text{err}_{\text{app}} + \omega_h$, where ω_h is the bootstrap estimate of the bias as define above.

This is technically slightly different from the simple bias correction procedure described in Section 2.1.1 but is essentially the same. Using the convention given in Efron (1983), this bias estimate is then added to the apparent error rate to produce the bootstrap estimate.

To be more explicit, let $X_1, X_2, \ldots, X_n$ denote the n training vectors where, say for convenience, $n = 2m$ for m an integer $X_1, X_2, \ldots, X_m$ come from class 1 and $X_{m+1}, X_{m+2}, \ldots, X_n$ come from class 2. A bootstrap sample is generated by sampling with replacement from the empirical distribution for the pooled data $X_1, X_2 \ldots, X_n$.

Although different, this is almost the same as taking m samples with replacement from $X_1, X_2, \ldots, X_m$ and another m samples with replacement from $X_{m+1}, X_{m+2}, \ldots, X_n$. In the latter case, each bootstrap sample contains m vectors from each class, whereas in the former case the number in each class varies according to a binomial distribution where N_1, the number from class 1, is binomial with parameters n and p (with $p = 1/2$) and N_2, the number from class 2, equals $n - N_1$. $E(N_1) = E(N_2) = n/2 = m$.

The approach used in Efron (1983) and Chernick, Murthy, and Nealy (1985, 1986, 1988a,b) is essentially the former approach except that the original training set itself is also selected in the same way as the bootstrap samples. So, for example, when $n = 14$, it is possible to have 7 training vectors from class 1 and 7 from class 2, but also we may have 6 from class 1 and 8 from class 2, and so on.

Once a bootstrap sample has been selected, we treat the bootstrap sample as though it were the training set. We construct the discriminant rule (linear for the simulations under discussion, but the procedure can apply to other forms such as quadratic) based on the bootstrap sample and subtracting the fraction of the observations in the bootstrap sample that would be misclassified by the same rule (where each observation is counted as many times as it occurs in the bootstrap sample).

The first term is a bootstrap sample estimate of the "true" error rate, while the second term is a bootstrap sample estimate of the apparent error rate. The difference is a bootstrap sample estimate of the optimistic bias in the apparent error rate. Averaging these estimates over the k Monte Carlo replications provides a Monte Carlo approximation to the bootstrap estimator.

An explicit formula for the bootstrap estimator and its Monte Carlo approximation is given on p. 317 of Efron (1983). Although the formulas are explicit, the notation is complicated. Nevertheless, the Monte Carlo approximation is simple to describe as we have done above.

The e_0 estimator was introduced as a variant to the bootstrap in Chatterjee and Chatterjee (1983), although the name e_0 came later in Efron (1983). For the e_0 estimate we simply count the total number of training vectors misclassified in each bootstrap sample. The estimate is then obtained by summing over all bootstrap samples and dividing by the total number of training vectors not included in the bootstrap samples.

The .632 estimator is obtained by the formula

$$\text{err}_{632} = 0.368\text{err}_{\text{app}} + 0.632e_0,$$

where err_{app} denotes the apparent error rate and e_0 is as defined in the previous paragraph. With only the exception of the very heavy-tailed distributions, the .632 estimator is the clear-cut winner over the other variants.

Some heuristic justification for this is given in Efron (1983) [see also Chernick and Murthy (1985)]. Basically, the .632 estimator appropriately balances the optimistic bias of the apparent error rate with the pessimistic bias of e_0. The reason for this weighting is that 0.368 is a decimal approximation to $1/e$, which is the asymptotic expected percentage of training vectors that are not included in a bootstrap sample.

Chernick, Murthy, and Nealy (1985) devised a variant called the MC estimator. This estimator is obtained just as the standard bootstrap. The difference is that a controlled bootstrap sample is generated in place of the ordinary bootstrap sample. In this procedure, the sample is restricted to include observations with replication frequencies as close as possible to the asymptotic expected replication frequency.

Another variant, also due to Chernick, Murthy, and Nealy (1985), is the convex bootstrap. In the convex bootstrap, the bootstrap sample contains linear combinations of the observation vectors. This smoothes out the sampling distribution for the bootstrap estimate by allowing a continuum of possible observations instead of just the original discrete set.

A theoretical difficulty with the convex bootstrap is that the bootstrap distribution does not converge to the true distribution since the observations are weighting according to λ which is chosen uniformly on [0,1]. This means that the "resamples" will not behave in large samples exactly like the original samples from the class-conditional densities. We can therefore not expect the estimated error rates to be correct for the given classification rule.

To avoid the inconsistency problem, Chernick, Murthy, and Nealy (1988b) introduced a modified convex bootstrap that concentrates the weight closer and closer to one of the samples, as the training sample size n increases. They also introduced a modification to the .632 estimator which they called the adaptive 632.

It was hoped that the modification of adapting the weights would improve the .632 estimator and increase its applicability, but results were disappointing. Efron and Tibshirani (1997a) introduce .632+, which also modifies the .632 estimator so that it works well for an even wider class of classification problems and a variety of class-conditional densities.

In Efron (1983) other variants—the double bootstrap, the randomized bootstrap, and the randomized double bootstrap—are also considered. The reader is referred to Efron (1983) for the formal definitions of these estimators. Of these, only the randomized bootstrap showed significant improvement over the ordinary bootstrap, and so these other variants were not considered. Follow-up studies did not include the randomized bootstrap.

The randomized bootstrap applies only to the two-class problem. The idea behind the randomized bootstrap is the modification of the empirical distributions for each class by allowing for the possibility that the observed training vectors for class 1 come from class 2 and vice versa.

Efron allowed a probability of occurrence of .1 to the opposite class in the simple version. After modifying the empirical distributions, bootstrap sampling is applied to the modified distributions rather than the empirical distributions and the bias is estimated and then corrected for, just as the standard bootstrap. In a way the randomized bootstrap smoothes the empirical distributions, an idea similar in spirit to the convex bootstrap.

Implementation of the randomized bootstrap by Monte Carlo is straightforward. We sample at random from the pooled training set (i.e., training data from both classes are mixed together) and then choose a uniform random number U. If $U \leq .9$, we assign the observation vector to its correct class. If not, we assign it to the opposite class. To learn more about the randomized bootstrap and other variations, see Efron (1983, p. 320).

For Gaussian populations and small training sample sizes (14–29) the .632 estimator is clearly superior in all the studies in which it was considered, namely Efron (1983), Chernick, Murthy, and Nealy (1985, 1986) and Jain, Dubes, and Chen (1987).

A paper by Efron and Tibshirani (1997), which we have already mentioned, looks at the .632 estimator and a variant called .632+. They treat more general classification problems as compared to just the linear (equal covariances) case that we focus on here.

Chernick, Murthy, and Nealy (1988a,b) consider multivariate (two-dimensional, three-dimensional, and five-dimensional) distributions. Uniform, exponential, and Cauchy distributions are considered. The uniform provides shorter than Gaussian tails to the distribution, and the bivariate exponential provides an example of skewness and the autoregressive family of Cauchy distributions provides for heavier-than-Gaussian tails to the distribution.

They found that for the uniform and exponential distributions the .632 estimator is again superior. As long as the tails are not heavy, the .632 estimator provides an appropriate weighting to balance the opposite biases of e_0 and the apparent error rate.

However, for the Cauchy distribution the e_0 no longer has a pessimistic bias and both the e_0 and the convex bootstrap outperform the .632 estimator. They conjectured that the result would generalize to any distributions with heavy tails. They also believe that skewness and other properties of the distribution which cause it to depart from the Gaussian distribution would have little effect on the relative performance of the estimators.

In Chernick, Murthy, and Nealy (1988b), the Pearson VII family of distributions was simulated for a variety of values of the parameter m. The probability density function is defined as

$$f(x) = \frac{\Gamma(M)|\Sigma|^{1/2}}{\Gamma(m-(p/2))\pi^{p/2}[1+(x-\mu)'(x-\mu)]m},$$

where μ is a location vector, Σ is a scaling matrix, m is a parameter that affects the dependence and controls the tail behavior, p is the dimension, and Γ is the gamma function. The symbol $|\ |$ denotes the determinant of the matrix.

The Pearson VII distributions are all elliptically contoured (i.e., contours of constant probability density are ellipses). An elliptic contoured density is a property the Pearson VII family shares with the Gaussian family of distributions. Only $p = 2$ was considered in Chernick, Murthy, and Nealy (1988b). The parameter m was varied from 1.3 to 3.0. For $p = 2$, second moments exist only for m greater than 2.5 and first moments exist only for m greater than 1.5.

Chernick, Murthy, and Nealy (1988b) found that when $m \leq 1.6$, the pattern observed for the Cauchy distributions in Chernick, Murthy, and Nealy (1988a) pertained; that is, the e_0 and the convex bootstrap were the best. As m decreases from 2.0 to 1.5, the bias of the e_0 estimator decreases and eventually it changes sign (i.e., goes from a pessimistic to an optimistic bias). For m greater than 2.0, the results are similar to the Gaussian and the light-tailed distributions where the .632 estimator is the clear winner.

Table 2.1 is taken from Chernick, Murthy, and Nealy (1988b). It summarizes for various values of m the relative performance of the estimators. The totals represent the number of cases for which the estimators ranked first, second, and third among the seven considered. The cases vary over the range of the "true" error rates that varied from about .05 to .50.

Table 2.2 is a similar summary taken from Chernick, Murthy, and Nealy (1986) which summarizes the results for the various Gaussian cases considered.

Again, we point out that for most applications the .632 estimator is preferred. It is not yet clear whether or not the smoothed estimators are as good as the best bootstrap estimates.

Snapinn and Knoke (1985b) claim that their estimator is better than the .632 estimator. Their study simulated both Gaussian distributions and a few non-Gaussian distributions.

They also show that the bias correction applied to the smoothed estimators by resampling procedures may be as good as their own smoothed estimators. This has not yet been confirmed in the published literature. Some results comparing the Snapinn and Knoke estimators with the .632 bootstrap and some other estimates in two-class cases are found in Hirst (1996).

For very heavy-tailed distributions, our recommendation would be to use the ordinary bootstrap or the convex bootstrap. But how does the practitioner know that the distributions are heavy-tailed? It may sometimes be possible to make such an assessment from knowledge as to how the data are generated for the practitioner to determine something about the nature of the tails of the distribution. One example would be when the data is ratios where the denominator can be close to zero. But in many practical cases it may not be possible.

To be explicit, consider the case where a feature is the ratio of two random variables and the denominator is known to be approximately Gaussian with zero mean; we will know that the feature has a distribution with tails like the Cauchy. This is because such cases are generalizations of the standard Cauchy distribution. It is a known result that the ratio of two independent Gaussian random variables with zero mean and the same variance has the standard Cauchy distribution. The Cauchy distribution is very heavy-tailed, and even the first moment or mean does not exist.

As the sample size becomes larger, it makes little difference which estimator is used, as the various bootstrap estimates and cross-validation are asymptotically equivalent (with the exception of the convex bootstrap). Even the apparent error rate may work well in very large samples where its bias is much reduced, although never zero. Exactly how large is large is difficult to say

Table 2.1 Summary Comparison of Estimators Using Root Mean Square Error (Number of Simulations on Which Estimator Attained Top Three Ranks)

	.632	MC	e_0	Boot	Conv	U	App	Total
				$M = 1.3$				
First	0	0	2	0	10	0	0	12
Second	3	0	0	9	0	0	0	12
Third	0	9	0	1	2	0	0	12
Total	3	9	2	10	12	0	0	36
				$M = 1.5$				
First	6	1	8	5	12	0	1	33
Second	8	4	0	14	7	0	0	33
Third	3	15	2	4	8	0	1	33
Total	17	20	10	23	27	0	2	99
				$M = 1.6$				
First	1	1	2	1	5	0	2	12
Second	4	3	0	5	0	0	0	12
Third	0	4	0	4	4	0	0	12
Total	5	8	2	10	9	0	2	36
				$M = 1.7$				
First	2	1	2	1	2	1	3	12
Second	3	3	1	4	1	0	0	12
Third	4	2	0	3	2	0	1	12
Total	9	6	3	8	5	0	1	36
				$M = 2.0$				
First	18	1	3	0	1	0	7	30
Second	10	4	4	2	5	2	3	30
Third	1	9	3	8	5	0	3	30
Total	29	14	10	10	11	2	13	90
				$M = 2.5$				
First	21	0	8	1	0	0	3	33
Second	10	3	4	5	4	2	5	33
Third	1	13	1	6	10	0	2	33
Total	32	16	13	12	14	2	10	99
				$M = 3.0$				
First	21	0	6	0	0	0	3	30
Second	9	3	5	3	2	2	6	30
Third	0	8	1	8	11	1	1	30
Total	30	11	12	11	13	3	10	90

Source: Chernick, Murthy, and Nealy (1988b).

Table 2.2 Summary Comparison

Rank	.632	MC	E_0	Boot	Conv	U	App	Total
First	72	1	29	6	0	0	1	109
Second	21	13	27	23	11	1	13	109
Third	7	20	8	25	37	7	5	109
Total	100	34	64	54	48	8	19	

Source: Chernick, Murthy, and Nealy (1986).

because the known studies have not yet adequately varied the size of the training sample.

2.1.3. Error Rate Estimation: An Illustrative Problem

In this problem, we have five bivariate normal training vectors from class 1 and have 5 from class 2. For class 1, the mean vector is $\begin{pmatrix}0\\0\end{pmatrix}$ and the covariance matrix is

$$\begin{bmatrix}10\\01\end{bmatrix}.$$

For class 2, the mean vector is $\begin{pmatrix}1\\1\end{pmatrix}$ and the covariance matrix is also

$$\begin{bmatrix}10\\01\end{bmatrix}.$$

The training vectors generated by random sampling from the above distributions are as follows:

For Class 1

$$\begin{pmatrix}2.052\\0.339\end{pmatrix}, \begin{pmatrix}1.083\\-1.320\end{pmatrix}, \begin{pmatrix}0.083\\-1.524\end{pmatrix}, \begin{pmatrix}1.278\\-0.459\end{pmatrix}, \begin{pmatrix}-1.226\\-0.606\end{pmatrix}.$$

For Class 2

$$\begin{pmatrix}1.307\\2.268\end{pmatrix}, \begin{pmatrix}-0.548\\1.741\end{pmatrix}, \begin{pmatrix}2.498\\0.813\end{pmatrix}, \begin{pmatrix}0.832\\1.409\end{pmatrix}, \begin{pmatrix}1.498\\2.063\end{pmatrix}.$$

We generate four bootstrap samples of size 10 and calculate the standard bootstrap estimate of the error rate. We also calculate e_0 and the apparent error rate in order to compute the .632 estimator. We denote by the indices 1, 2, 3, 4, and 5 the respective five bivariate vectors from class 1 and denote by the indices 6, 7, 8, 9, and 10 the respective five bivariate vectors from class 2. A bootstrap sample can be represented by a random set of 10 indices sampled with replacement from the integers 1 to 10. In this instance, our four bootstrap samples are **[9, 3, 10, 8, 1, 9, 3, 5, 2, 6]**, **[1, 5, 7, 9, 9, 9, 2, 3, 3, 9]**, **[6, 4, 3, 9, 2, 8, 7, 6, 7, 5]**, and **[5, 5, 2, 7, 4, 3, 6, 9, 10, 1]**.

Bootstrap sample numbers 1 and 2 have five observations from class 1 and five from class 2, bootstrap sample number 3 has four observations from class 1 and six from class 2, and bootstrap sample number 4 has six observations from class 1 and four from class 2. We also observe that in bootstrap sample number 1, indices 3 and 9 repeat once and indices 4 and 7 do not occur. In bootstrap sample number 2, index 9 occurs three times and index 3 twice while indices 4, 6, and 10 do not appear. In bootstrap sample number 3, indices 6 and 7 are repeated once while 1 and 10 do not appear.

Finally, in bootstrap sample number 4, only index 5 is repeated and index 8 is the only one not to appear. These samples are fairly typical of the behavior of bootstrap samples (i.e., sampling with replacement from a given sample), and they indicate how the bootstrap samples can mimic the variability due to sampling (i.e., the sample-to-sample variability).

Table 2.3 shows how the observations in the bootstrap sample were classified by the classification rule obtained using the bootstrap sample. We see that only in bootstrap samples 1 and 2 were any of the bootstrap observations misclassified. So for bootstrap samples 3 and 4 the bootstrap sample estimate of the apparent error rate is zero. In both bootstrap sample 1 and sample 2, only observation number 1 was misclassified and in each sample, observation number 1 appeared one time. So for these two bootstrap samples the estimate of apparent error is 0.1.

Table 2.4 shows the resubstitution counts for the original sample. Since none of the observations were misclassified, the apparent error rate or resubstitution estimate is also zero.

Table 2.3 Truth Table for the Four Bootstrap Samples

	Sample #1 Classified As		Sample #2 Classified As	
True Class	Class 1	Class 2	Class 1	Class 2
Class 1	4	1	4	1
Class 2	0	5	0	5
	Sample #3 Classified As		Sample #4 Classified As	
True Class	Class 1	Class 2	Class 1	Class 2
Class 1	4	0	6	0
Class 2	0	6	0	4

Table 2.4 Resubstitution Truth Table for Original Data

	Sample #1 Classified As	
True Class	Class 1	Class 2
Class 1	5	0
Class 2	0	5

In the first bootstrap sample, observation number 1 was the one misclassified. Observation numbers 4 and 7 did no appear. They both would have been correctly classified since their discriminant function values were 0.030724 for class 1 and –1.101133 for class 2 for observation 4 and –5.765286 for class 1 and 0.842643 for class 2 for observation 7. Observation 4 is correctly classified as coming from class 1 since its class 1 discriminant function value is larger than its class 2 discriminant function value. Similarly observation 7 is correctly classified as coming from class 2.

In the second bootstrap sample, observation number 1 was misclassified, and observation numbers 4, 6, and 10 were missing. Observation 3 was correctly classified as coming from class 1, and observations 60 and 10 were correctly classified as coming from class 2. Table 2.6 provides the coefficients of the linear discriminant functions for each of the four bootstrap samples.

It is an exercise for the reader to calculate the discriminant function values for observation numbers 4, 6, and 10 to see that the correct classifications would be made with bootstrap sample number 2.

In the third bootstrap sample, none of the bootstrap sample observations were misclassified but observation numbers 1 and 10 were missing. Using Table 2.6, we see that for class 1, observation number 1 has a discriminant function value of –3.8587, whereas for class 2 it has a discriminant function value of 2.6268.

Consequently, observation 1 would have been misclassified by the discrimination rule based on bootstrap sample number 3. The reader may easily check this and also may check that observation 10 would be correctly classified as coming from class 2 since its discriminant function value for class 1 is –9.6767 and 13.1749 for class 2.

In the fourth bootstrap sample, none of the bootstrap sample observations are misclassified and only observation number 8 is missing from the bootstrap sample. We see, however, by again computing the discriminant functions, that observation 8 would be misclassified as coming from class 1 since its class 1 discriminant function value is –2.1756 while its class 2 discriminant function value is –2.4171.

Another interesting point to notice from Table 2.5 is the variability of the coefficients of the linear discriminants. This variability in the estimated coefficients is due to the small sample size. Compare these coefficients with the ones given in Table 2.6 for the original data.

Table 2.5 Linear Discriminant Function Coefficients for Bootstrap Samples

True Class	Constant Term	Variable No. 1	Variable No. 2
	Bootstrap Sample No. 1		
Class 1	−1.793	0.685	−2.066
Class 2	−3.781	1.027	2.979
	Bootstrap Sample No. 2		
Class 1	−1.919	0.367	−2.481
Class 2	−3.353	0.584	3.540
	Bootstrap Sample No. 3		
Class 1	−2.343	0.172	−3.430
Class 2	−6.823	1.340	6.549
	Bootstrap Sample No. 4		
Class 1	−1.707	0.656	−2.592
Class 2	−6.130	0.469	6.008

Table 2.6 Linear Discriminant Function Coefficients for the Origiual Sample

Class Number	Constant Term	Variable No. 1	Variable No. 2
1	−1.493	0.563	−1.726
2	−4.044	0.574	3.653

The bootstrap samples give us an indication of the variability of the rule. This would otherwise be difficult to see. It also indicates that we can expect a large optimistic bias for resubstitution.

We can now compute the bootstrap estimate of bias:

$$\omega_{\text{boot}} = \frac{(0.1-0.1)+(0.1-0.1)+(0.1-0.)+(0.1-0)}{4} = 0.2/4 = 0.05.$$

Since the apparent error rate is zero, the bootstrap estimate of the error rate is also 0.05.

The e_0 estimate is the average of the four estimates obtained by counting in each bootstrap sample the fraction of the observations that do not appear in the bootstrap sample and that would be misclassified. We see from the results above that these estimates are 0.0, 0.0, 0.5, and 1.0 for bootstrap samples 1, 2, 3, and 4, respectively. This yields an estimated value of 0.375.

Another estimate similar to e_0 but distinctly different is obtained by counting all the observations left out of the bootstrap samples that would have been misclassified by the bootstrap sample rule and dividing by the total number of observations left out of the bootstrap samples. Since only two of the left-out observations were misclassified and only a total of eight observations were left out, this would give us an estimate of 0.250. This amounts to giving more weight to those bootstrap samples with more observations left out.

For the leave-one-out method, observation 1 would be misclassified as coming from class 2 and observation 8 would be misclassified as coming from class 1. This leads to a leave-one-out estimate of 0.200.

Now the .632 estimator is simply 0.368 × (apparent error rate) + 0.632 × (e_0). Since the apparent error rate is zero, the .632 estimate is 0.237.

Since the data were taken from independent Gaussian distributions, each with variance one and with the mean equal to zero for population 1 and with mean equal to one for population 2, the expected error rate for the optimal rule based on the distributions being known is easily calculated to approximately 0.240.

The actual error rate for the classifier based on a training set of size 10 can be expected to be even higher. We note that in this example, the apparent error rate and the bootstrap both underestimate the true error rate, whereas the e_0 overestimates it.

The .632 estimator comes surprisingly close to the optimal error rate and gives clearly a better estimate of the conditional error rate (0.295, discussed below) than the others. The number of bootstrap replications is so small in this numerical example that it should not be taken too seriously. It is simply one numerical illustration of the computations involved. Many more simulations are required to draw conclusions, and thus simulation studies such as the ones already discussed are what we should rely on.

The true conditional error rate given the training set can be calculated by integrating the appropriate Gaussian densities over the regions defined by the discriminant rule based on the original 10 sample observations. An approximation based on Monte Carlo generation of new observations from the two classes, classified by the given rule, yields for a sample size of 1000 new observations (500 from each class) an estimate of 0.295 for this true conditional error rate.

Since (for equal error rates) this Monte Carlo estimator is based on a binomial distribution with parameters $n = 1000$ and $p =$ the true conditional error rate, using $p = .3$, we have that the standard error of this estimate is approximately 0.0145 and an approximate 95% confidence interval for p is [0.266, 0.324]. So our estimate of the true conditional error rate is not very accurate.

If we are really interested in comparing these estimators to the true conditional error rate, we probably should have taken 50,000 Monte Carlo replications to better approximate it. By increasing the sample size by a factor of 50, we decrease the standard error by $\sqrt{50}$, which is a factor slightly greater than 7. Hence, the standard error of the estimate would be about 0.002 and the confidence interval would be $[p_h - 0.004, p_h + 0.004]$, where p_h is the point estimate of the true conditional error rate based on 50,000 Monte Carlo replications. We get 0.004 as the interval half-width since a 95% confidence interval requires a half-width of 1.96 standard errors (close to 2 standard errors).

The width of the interval would then be less than 0.01 and would be useful for comparison. Again, we should caution the reader that even if the true conditional error rate were close to the .632 estimate, we could not draw a strong conclusion from it because we would be looking at only one .632 estimate, one e_0 estimate, one apparent error rate estimate, and so on. It really takes simulation studies to account for the variability of the estimates for us to make valid comparisons.

2.1.4. Efron's Patch Data Example

Sometimes in making comparisons we are interested in computing the ratio of the two quantities. We are given a set of data that enables us to estimate both quantities, and we are interested in estimating the ratio of two quantities, What is the best way to do this? The natural inclination is to take the ratio of the two estimates. Such estimators are called ratio estimators.

However, statisticians know quite well that if both estimates are unbiased, the ratio estimate will be biased (except for special degenerate cases). To see why this is so, suppose that X is unbiased for $E(Y)$ and that Y is unbiased for μ. Since X is unbiased for θ, $E(X) = \theta$ and since Y is unbiased for μ, $E(Y) = \mu$. Then $\theta/\mu = E(X)/E(Y)$, but this is not $E(X/Y)$, which is the quantity that we are interested in. Let us further suppose that X and Y are statistically independent; then we have

$$E(X/Y) = E(X)E(1/Y) = \theta E(1/Y).$$

The reciprocal function $f(z) = 1/z$ is a convex function and therefore by Jensen's inequality (see Ferguson, 1967, pp. 76–78) implies that $\mathrm{f}(E(Y)) = \mathrm{f}(\mu) = 1/\mu \leq E(f(Y)) = E(1/Y)$. Consequently, $E(X/Y) = \theta\, E(1/Y) \geq \theta/\mu$. The only instance where equality holds is when Y equals a constant. Otherwise $E(X/Y) > \theta/\mu$ and the bias $B = E(X/Y) - \theta/\mu$ is positive. This bias can be large, and it is natural to try to improve the estimate of the ratio by adjusting for the bias. Ratio estimators are also common in survey sampling [see Cochran (1977) for some examples].

In Efron and Tibshirani (1993) an example of ratio estimator is given in Section 10.3 on pages 126–133. This was a small clinical trial used to show the FDA that a product produced at a new plant is equivalent to the product produced at the old plant where the agency had previously approved the product. In this example the product is a patch that infuses a certain natural hormone into the patient's bloodstream. The trial was a crossover trial involving eight subjects. Each subject was given three different patches: one patch that was manufactured at the old plant containing the hormone, one manufactured at the new plant containing the hormone, and a third patch (placebo) that contained no hormone.

The purpose of the placebo is to establish a baseline level to compare with the hormone. Presumably the subjects were treated in random order with regard to treatment, and between each treatment an appropriate wash-out period is applied to make sure that there is no lingering effect from the previous treatment.

The FDA has a well-defined criterion for establishing bioequivalence in such trials. They require that the new patch produces hormone levels that are within 20% of the amount produced by the old patch to the placebo. Mathematically, we express this as

$$\theta = [E(\text{new}) - E(\text{old})]/[E(\text{old}) - E(\text{placebo})]$$

and require that

$$|\theta| = [|E(\text{new}) - E(\text{old})|]/[|E(\text{old}) - E(\text{placebo})|] \leq 0.20.$$

So, for the FDA, the pharmaceutical company must show equivalence by rejecting the "null" hypothesis of non-equivalence in favor of the alternative of equivalence. So the null hypothesis is $|\theta| \leq 0.20$, versus the alternative that $|\theta| > 0.20$. This is most commonly done by applying Schurmann's two one-sided t tests. In recent years a two-stage group sequential test can be used with the hope of requiring a smaller total sample size than for the fixed sample size test.

For the ith subject we define z_i = (old patch blood level − placebo blood level) and y_i = (new patch blood level − old patch blood level). The natural estimate of θ is the plug-in estimate y_b/z_b, where y_b is the average of the eight y_i and z_b is the average of the eight z_i. As we have already seen, such a ratio estimator will be biased.

Table 2.7 shows the y and z values. Based on these data, we find that the plug-in estimate for θ is −0.0713, which is considerably less than the 0.20 in absolute value. However, the estimate is considerably biased and we might be able to improve our estimate with an adjustment for bias. The bootstrap can be used to estimate this bias as you have seen previously in the error rate estimation problem. The real problem is one of confidence interval estimation or hypothesis testing, and so the methods presented in Chapter 3 might be more appropriate. Nevertheless, we can see if the bootstrap can provide a better point estimate of the ratio. Efron and Tibshirani (1993) generated 400 bootstrap samples and estimated the bias to be 0.0043. They also estimated the standard error of the estimate, and the ratio of the bias estimate divided by the estimated standard error is only 0.041. This is small enough to indicate that the bias adjustment will not be important.

The patch data example is a case of equivalence of a product as it is manufactured in two different plants. It is also common for pharmaceutical

Table 2.7 Patch Data Summary

Subject	Old – Placebo (z)	New – Old (y)
1	8,406	–1,200
2	2,342	2,601
3	8,187	–2,705
4	8,459	1,982
5	4,795	–1,290
6	3,516	351
7	4,796	–638
8	10,238	–2,719
Average	6,342	–452.3

Source: Efron and Tibshirani (1993, p. 373), with permission from CRC Press, LLC.

companies to make minor changes in approved products since the change may improve the marketability of the product. To get the new product approved, the manufacturer must design a small bioequivalence trial much like the one shown in the patch data example. Recently, bootstrap methods have been developed to test for bioequivalence. There are actually three forms of bioequivalence defined. They are individual bioequivalence, average bioequivalence, and population bioequivalence. Depending on the application, one type may be more appropriate to demonstrate than another. We will give the formal definitions of these forms of bioequivalence and show examples of bootstrap methods for demonstrating individual and population bioequivalence in Chapter 8. The approach to individual bioequivalence was so successful that it has become a recommended approach in an FDA guidance document.

It is important to recognize that although the bootstrap adjustment will reduce the bias of the estimator and can do so substantially when the bias is large, it is not clear whether or not it improves the accuracy of the estimate. If we define the accuracy to be the root mean square error (rms), then since the rms error is the square root of the bias squared plus the variance, there is the possibility that although we decrease the bias, we could also be increasing the variance. If the increase in variance is larger than the decrease in the squared bias, the rms will actually increase. This tradeoff between bias and variance is common in a number of statistical problems including kernel smoothing, kernel density estimation, and the error rate estimation problem that we have seen. Efron and Tibshirani (1993, p. 138) caution about the hazards of bias correction methods.

2.2. ESTIMATING LOCATION AND DISPERSION

In this section, we consider point estimates of location parameters. For distributions with finite first and second moments the population mean is a

natural location parameter. The sample mean is the "best" estimate, and bootstrapping adds nothing to the parametric approach. We shall discuss this briefly.

For distributions without first moments, the median is a more natural parameter to estimate the location of the center of the distribution. Again, the bootstrap adds nothing to the point estimation, but we see in Section 2.2.2 that the bootstrap is useful in estimating standard errors and percentiles, which provide measures of the dispersion and measures of the accuracy of the estimates.

2.2.1. Means and Medians

For population distributions with finite first moments, the mean is a natural measure of central tendency. If the first moment does not exist, sample estimates can still be calculated but they tend to be unstable and they lose their meaning (i.e., the sample mean no longer converges to a population mean as the sample size increases).

One common example that illustrates this point is the standard Cauchy distribution. Given a sample size n from a standard Cauchy distribution, the sample mean is also standard Cauchy. So no matter how large we take n to be, we cannot reduce the variability of the sample mean.

Unlike the Gaussian or exponential distributions that have finite first and second moments and have sample means that converge in probability to the population mean, the Cauchy has a sample mean that does not converge in probability.

For distributions like the Cauchy, the sample median does converge to the population median as the sample size tends to infinity. Hence for such cases the sample median is a more useful estimator of the center of the distribution since the population median of the Cauchy and other heavy-tailed symmetric distributions best represents the "center" of the distribution.

If we know nothing about the population distribution at all, we may want to estimate the median since the population median always exists and is consistently estimated by the sample median regardless of whether or not the mean exists.

How does the bootstrap fit in when estimating a location parameter of a population distribution? In the case of Gaussian or the exponential distributions, the sample mean is the maximum likelihood estimate, is consistent for the population mean, and is the minimum variance unbiased estimate. How can the bootstrap top that?

In fact it cannot. In these cases the bootstrap could be used to estimate the mean but we would find that the bootstrap estimate is nothing but the sample mean itself, which is the average of all bootstrap samples, and the Monte Carlo estimate is just an approximation to the sample mean. It would be silly to bootstrap is such a case.

Nevertheless, for the purpose of developing a statistical theory for the bootstrap, the first asymptotic results were derived for the estimate of the mean when the variance is finite (Singh, 1981; Bickel and Freedman, 1981).

Bootstrapping was designed to estimate the accuracy of estimators. This is accomplished by using the bootstrap samples to estimate the standard deviation and possibly the bias of a particular estimator for problems where such estimates are not easily derived from the sample. In general, bootstrapping is not used to produce a better point estimate.

A notable exception was given in Section 2.1, where bias correction to the apparent error rate actually produced a better point estimate of the error rate. This is, however, an exception to the rule.

In the remainder of the book, we will learn about examples for which estimators are given, but we need to estimate their standard errors or construct confidence regions or test hypotheses about the corresponding population parameters.

For the case of distributions with heavy tails, we may be interested in robust estimates of location (the sample median being one such example). The robust estimators are given (e.g., Winsorized mean, trimmed mean, or sample median).

However, the bootstrap becomes useful as an approach to estimating the standard errors and to obtain confidence intervals for the location parameters based on these robust estimators. Some of the excellent texts that deal with robust statistical procedures are Chatterjee and Hadi (1988), Hampel, Ronchetti, Rousseeuw, and Stahel (1986), and Huber (1981).

2.2.2. Standard Errors and Quartiles

The standard deviation of an estimator (also referred to as the standard error for unbiased estimators) is a commonly used estimate of an estimator's variability. This estimate only has meaning if the distribution of the estimator of interest has a finite second moment. In examples for which the estimator's distribution does not have a finite second moment, the interquartile range (the 75th percentile minus the 25th percentile of the estimator's distribution) is often used as a measure of the variability.

Staudte and Sheather (1990, pp. 83–85) provide an exact calculation for the bootstrap estimate of the standard error of the median [originally derived by Maritz and Jarrett (1978)] and compare it to the Monte Carlo approximation for cell lifetime data (obtained as the absolute differences of seven pairs of independent identically distributed exponential random variables).

We shall review Staudte and Sheather's development and present their results here. For the median, they assume for convenience that the sample size n is odd (i.e., $n = 2m + 1$, for m an integer). This makes the exposition easier but is not a requirement.

Maritz and Jarrett (1978) actually provide explicit results for any n. It is just that the median is defined as the average of the two "middle" values when n is even and as the unique "middle" observation $m + 1$ when n is odd.

The functional representing the median is just $T(F) = F^{-1}(1/2)$, where F is the population cumulative distribution and F^{-1} is its inverse function. The sample median is just $X_{(m+1)}$, where $X_{(i)}$ denotes the ith-order statistic (i.e., the ith observation when ordered from smallest to largest).

An explicit expression for the variance of median of the bootstrap distribution can then be derived based on well-known results about order statistics. Let $X^*_{(1)}, \ldots, X^*_{(n)}$ denote the ordered observations from a bootstrap sample taken from $X_1, \ldots, X_n$. Let $x_{(i)}$ denote the ith smallest observation from the original sample. Let $N^*_i = \#\{j\colon X^*_j = x_{(i)}\}, i = 1, \ldots, n\}$.

Then it can be shown that $\sum_{i=1}^{k} N^*_i$ has the binomial distribution with parameters n and p, where $p = k/n$. Let P^* denote the probability under bootstrap sampling. It follows that

$$P^*\{X^*_{(m+1)} > x_{(k)}\} = P^*\left\{\sum_{i=1}^{k} N^*_i \le n\right\} = \sum_{j-0}^{n} \binom{n}{j} \binom{k}{n}^j \left(\frac{n-k}{n}\right)^{n-j}.$$

Using well-known relationships between binomial sums and the incomplete beta function, Staudte and Sheather (1990) find, letting $w_k = P^*\{X^*_{(m)} = x_{(k)}\}$, that

$$w_k = \frac{n!}{(m!)^2} \int_{(k-1)/n}^{k/n} (1-y)^m y^m dy$$

and then by simple probability calculations the bootstrap variance of $X^*_{(m+1)}$ is

$$\sum_{k-1}^{n} w_k x_{(k)} - \left(\sum_{k-1}^{n} w_k x_{(k)}\right)^2.$$

This result was first obtained by Maritz and Jarrett (1978) and later independently by Efron (1978). Taking the square root of the above expression, we have explicitly obtained, using properties of the bootstrap distribution for the median, the bootstrap estimate of the standard deviation of the sample median without doing any Monte Carlo approximation.

Table 2.8, taken from Staudte and Sheather (1990, p. 85), shows the results required to compute the standard error for the "sister cell" data set. In the table, p_k plays the role of w_k and the above equation using p_k gives.

$\mathrm{SE}_{\mathrm{BOOT}} = 0.173$. However, if we replace p_k with $\hat{p}_k$, we get 0.167 for a Monte Carlo approximation based on 500 bootstrap samples.

Table 2.8 Comparison of Exact and Monte Carlo Bootstrap Distributions for the Ordered Absolute Differences of Sister Cell Lifetimes

K	1	2	3	4	5	6	7
p_k	0.0102	0.0981	0.2386	0.3062	0.2386	0.0981	0.0102
$\hat{p}_k$	0.01	0.128	0.548		0.208	0.098	0.008
$x_{(k)}$	0.3	0.4	0.5	0.5	0.6	0.9	1.7

Source: Staudte and Sheather (1990, p. 85), with permission from John Wiley & Sons, Inc.

For other estimation problems the Monte Carlo approximation to the bootstrap may be required, since we may not be able to provide explicit calculations as we have just done for the median. The Monte Carlo approximation is straightforward. Let $\hat{\theta}$ be the sample estimate of θ and let $\hat{\theta}_i^*$ be the bootstrap estimate of for the ith bootstrap sample. Given k bootstrap samples, the bootstrap estimate of the standard deviation of the estimator $\hat{\theta}$ is, according to Efron (1982a),

$$SD_b\left\{\frac{1}{k-1}\sum_{j=1}^{k}\left[\theta_i^* - \bar{\theta}*\right]^2\right\}^{1/2},$$

where $\overline{\theta *}$ is the average of the bootstrap samples. Instead of $\overline{\theta *}$, one could equally well use $\hat{\theta}$ itself. The choice of $k-1$ in the denominator was made as the analog to the unbiased estimate of the standard deviation for a sample. There is no compelling argument for using $k-1$ instead of k in the formula.

For the interquartile range, one straightforward approach is to order the bootstrap sample estimates from smallest to largest. The bootstrap sample observation that equals the 25th percentile (or an appropriate average of the two bootstrap sample estimates closest to the 25th percentile) is subtracted from the bootstrap sample observation that equals the 75th percentile (or an appropriate average of the two bootstrap sample observations closest to the 75th percentile). Once these bootstrap sample estimates are obtained, bootstrap standard error estimates or other measures of spread for the interquartile range can be determined.

Other estimates of percentiles from a bootstrap distribution can be used to obtain bootstrap confidence intervals and test hypotheses as will be discussed in Chapter 3. Such methods could be applied to get approximate confidence intervals for standard errors, interquartile ranges, or any other parameters that can be estimated from a bootstrap sample (e.g., medians, trimmed means, Winsorized means, M-estimates, or other robust location estimates).

2.3. HISTORICAL NOTES

For the error rate estimation problem there is a great deal of literature. For developments up to 1974 see the survey article by Kanal (1974) and see the extensive bibliography by Toussaint (1974). In addition, for multivariate Gaussian features McLachlan has derived the asymptotic bias of the apparent error rate (i.e., the resubstitution estimate) in McLachlan (1976) and it is not zero!

The bias of plug-in rules under parametric assumptions is discussed in Hills (1966). A collection of articles including some bootstrap work can be found in Choi (1986).

There have been a number of simulation studies showing the superiority of versions of the bootstrap over cross-validation when the training sample size is small. Most of the studies have considered linear discriminant functions (although Jain, Dubes, and Chen consider quadratic discriminants). Most consider the two-class problem with two-dimensional feature vectors.

However, Efron (1982a, 1983) and Chernick, Murthy, and Nealy (1985, 1986, and 1988a) considered five-dimensional feature vectors as well. Also, in Chernick, Murthy, and Nealy (1985, 1986, 1988a) some three-class problems were considered. Chernick, Murthy, and Nealy (1988a,b) were the first to simulate the performance of these bootstrap estimators for linear discriminant functions when the populations were not Gaussian. Hirst (1996) proposes a smoothed estimator (a generalization of the Snapinn and Knoke approach) for cases with three or more classes and provides detailed simulation studies showing the superiority of his method. He also compares .632 with the smoothed estimator of Snapinn and Knoke (1985) in two-class problems.

Chatterjee and Chatterjee (1983) considered only the two-class problem, doing only one-dimensional Gaussian simulations with equal variance. They were, however, the first to consider a variant of the bootstrap which Efron later refers to as e_0 in Efron (1983). They also provided an estimated standard error for their bootstrap error rate estimation.

The smoothed estimators have also been compared with cross-validation by Snapinn and Knoke (1984, 1985a). They show that their estimators have smaller mean square error than cross-validation for small training samples sizes, but unfortunately not much has been published comparing the smoothed estimates with the bootstrap estimates. We are aware of one unpublished study, Snapinn and Knoke (1985b), and some results in Hirst (1996).

In the simulation studies of Efron (1983), Chernick, Murthy, and Nealy (1985, 1986), Chatterjee and Chatterjee (1983), and Jain, Dubes, and Chen (1987), only Gaussian populations were considered.

Only Jain, Dubes, and Chen (1987) considered classifiers other than linear discriminants. They looked at quadratic and nearest-neighbor rules. Performance was measured by mean square error of the conditional expected error rate.

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