Chapter 1. Bootstrap Method

1 Introduction

1.1 The Practice of Statistics

Statistics is the science of learning from experience, especially experience that arrives a little bit at a time. Most people are not natural-born statisticians. Left to our own devices we are not very good at picking out patterns from a sea of noisy data. To put it another way, we all are too good at picking out non-existent patterns that happen to suit our purposes? Statistical theory attacks the problem from both ends. It provides optimal methods for finding a real signal in a noisy background, and also provides strict checks against the overinterpretation of random patterns.

Statistical theory attempts to answer three basic questions:

1. Data Collection: How should I collect my data?
2. Summary: How should I analyze and summarize the data that I've collected?
3. Statistical Inference: How accurate are my data summaries?

The bootstrap is a recently developed technique for making certain kinds of statistical inferences. It is only recently developed because it requires modern computer power to simplify the often intricate calculations of traditional statistical theory.

1.2 Motivated Example

We now illustrate the just mentioned three basic statistical concepts using a front-page news from the New York Times of January 27, 1987. A study was done to see if small aspirin doses would prevent heart attacks in healthy middle-aged men. The data for the aspirin study were collected in a particularly efficient way: by a controlled, randomized, double-blind study. One half of the subjects received aspirin and the other half received a control substance, or placebo, with no active ingredients. The subjects were randomly assigned to the aspirin or placebo groups. Both the subjects and the supervising physicians were blind to the assignments, with the statisticians keeping a secret code of who received which substance. Scientists, like everyone else, want the subject they are working on to succeed. The elaborate precautions of a controlled, randomized, blinded experiment guard against seeing benefits that don't exist, while maximizing the chance of detecting a genuine positive effect.

The summary statistics in the study are very simple:

	heart attacks (fatal plus non-fatal)	subjects
aspirin group:	104	11,037
placebo group:	189	11,034

What strikes the eye here is the lower rate of heart attacks in the aspirin group. The ratio of the two rates is

$$\hat{\theta} = \frac{104/11037}{189/11034} = 0.55.$$

It suggests that the aspirin-takers only have 55% as many as heart attacks as placebo-takers.

Of course we are not interested in $\hat{\theta}$. What we would like to know is θ, the true ratio, that is the ratio we would see if we could treat all subjects, and not just a sample of them. The tough question is how do we know that $\hat{\theta}$ might not come out much less favorably if the experiment were run again?

This is where statistical inference comes in. Statistical theory allows us to make the following inference: the true value of θ lies in the interval $0.43 < \theta < 0.70$ with 95% confidence. Note that

$$\theta = \hat{\theta} + (\theta - \hat{\theta}) = 0.55 + [\theta - \hat{\theta}(\omega_0)],$$

where θ and $\hat{\theta}(\omega_0)$ $(= 0.55)$ are two numbers. In statistics, we use $\theta - \hat{\theta}(\omega)$ to describe $\theta - \hat{\theta}(\omega_0)$. Since ω cannot be observed exactly, we instead study the fluctuation of $\theta - \hat{\theta}(\omega)$ among all ω. If, for most ω, $\theta - \hat{\theta}(\omega)$ is around zero, we can conclude statistically that θ is close to 0.55 $(= \hat{\theta}(\omega_0)$. (Recall the definition of consistency.) If $P(\omega : |\theta - \hat{\theta}(\omega)| < 0.1) = 0.95$, we claim that with 95% confidence that $\theta - 0.55$ is no more than 0.1.

In the aspirin study, it also track strokes. The results are presented as the following:

	strokes	subjects
aspirin group:	119	11,037
placebo group:	98	11,034

For strokes, the ratio of the two rates is

$$\hat{\theta} = \frac{119/11037}{98/11034} = 1.21.$$

It now looks like taking aspirin is actually harmful. However, the interval for the true stroke ratio θ turns out to be $0.93 < \theta < 1.59$ with 95% confidence. This includes the neutral value $\theta = 1$, at which aspirin would be no better or worse than placebo. In the language of statistical hypothesis testing, aspirin was found to be significantly beneficial for preventing heart attacks, but not significantly harmful for causing strokes.

In the above discussion, we use the sampling distribution of $\hat{\theta}(\omega)$ to develop intervals in which the true value of θ lies on with a high confidence level. The task of data analyst is to find the sampling distribution of the chosen estimator $\hat{\theta}$. Turn it into practice, we are quite often on finding right statistical table to look up.

Quite often, these tables are constructed based on the model-based sampling theory approach to statistical inference. In this approach, it starts with the assumption that the data arise as a sample from some conceptual probability distribution, f. When f is completely specified, we derive the distribution of $\hat{\theta}$. Recall that $\hat{\theta}$ is a function of the observed data. In deriving its distribution, those data will be viewed as random variables (why??). Uncertainties of our inferences can then be measured. The traditional parametric inference utilizes a priori assumptions about the shape of f. For the above example, we rely on the binomial distribution, large sample approximation of the binomial distribution, and the estimate of θ.

However, we sometimes need to figure out f intelligently. Consider a sample of weights of 27 rats ($n = 27$); the data are

$$57, 60, 52, 49, 56, 46, 51, 63, 49, 57, 59, 54, 56, 59, 57, 52, 52, 61, 59, 53, 59, 51, 51, 56, 58, 46, 53.$$

The sample mean of these data $= 54.6667$, standard deviation $= 4.5064$ with $cv = 0.0824$. For illustration, what if we wanted an estimate of the standard error of cv. Clearly, this would be a nonstandard problem. First, we may need to start with a parametric assumption on f. (How will you do it?) We may construct a nonparametric f estimator of (in essence) from the sample data. Then we can invoke either Monte Carlo method or large sample method to give an approximation on it.

Here, we will provide an alternative to the above approach. Consider the following nonparametric bootstrap method which relies on the empirical distribution function. As a demonstration, we apply the bootstrap method works to the stroke example.

1. Create two populations: the first consisting of 119 ones and $11037 - 119 = 10918$ zeros, and the second consisting of 98 ones and $11034 - 98 = 10936$ zeros.
2. (Monte Carlo Resampling) Draw with replacement a sample of 11037 items from the first population, and a sample of 11034 items from the second population. Each of these is called a *bootstrap sample*.
3. Derive the bootstrap replicate of $\hat{\theta}$:
$$\hat{\theta}^* = \frac{\text{prop. of ones in bootstrap sample \#1}}{\text{prop. of ones in bootstrap sample \#2}}.$$
4. Repeat this process (1-3) a large number of times, say 1000 times, and obtain 1000 *bootstrap replicates* $\hat{\theta}^*$.

As an illustration, the standard deviation turned out to be 0.17 in a batch of 1000 replicates that we generated. Also a rough 95% confidence interval is $(0.93, 1.60)$ which is derived by taking the 25th and 975th largest of the 1000 replicates.

Remark:

1. Initiated by Efron in 1979, the basic bootstrap approach uses Monte Carlo sampling to generate an empirical estimate of the $\hat{\theta}$'s sampling distribution.
2. Monte Carlo sampling builds an estimate of the sampling distribution by randomly drawing a large number of samples of size n from a population, and calculating for each one the associated value of the statistic $\hat{\theta}$. The relative frequency distribution of these $\hat{\theta}$ values is an estimate of the sampling distribution for that statistic. The larger the number of samples of size n will be, the more accurate the relative frequency distribution of these estimates will be.
3. With the bootstrap method, the basic sample is treated as the population and a Monte Carlo-style procedure is conducted on it. This is done by randomly drawing a large number of *resamples* of size n from this original sample (of size n either) with replacement. So, although each resample will have the same number of elements as the original sample, it could include some of the original data points more than once, and some not included. Therefore, each of these resamples will randomly depart from the original sample. And because the elements in these resamples vary slightly, the statistic $\hat{\theta}$, calculated from one of these resample will take on slightly different values.
4. The central assertion of the bootstrap method is that the relative frequency distribution of these $\hat{\theta}_{F_n}$'s is an estimate of the sampling distribution of $\hat{\theta}$.
5. How do we determine the number of *bootstrap replicates*?

Assignment 1. Do a small computer experiment to repeat the above process a few times and check whether you get the identical answers every time (with different random seeds).

Assignment 2. Read Ch. 11.4 of Rice's book. Comment on randomization, placebo effect, observational studies and fishing expedition.

Assignment 3. Do problems 1, 19 and 28 in Section 11.6 of Rice's book.

Now we come back to the cv example. First, we draw a random subsample of size 27 *with replacement*. Thus, while a weight of 63 appears in the actual sample, perhaps it would not appear in the subsample; or is could appear more than once. Similarly, there are 3 occurrences of the weight 57 in the actual sample, perhaps the resample would have, by chance, no values of 57. The point here is that a random sample of size 27 is taken from the original 27 data values. This is the first bootstrap resample with replacement ($b = 1$). From this resample, one computes $\hat{\mu}$, the $\hat{se}(\hat{\mu})$ and the cv and stores this in memory. Second, the whole process is repeated B times (where we will let $B = 1,000$ reps for this example).

Thus, we generate 1000 resample data sets ($b = 1, 2, 3, \ldots, 1000$) and from of each these we compute $\hat{\mu}$, $\hat{se}(\hat{\mu})$ and the cv and store these values. Third, we obtain the standard error of the cv by taking the standard deviation of the 1000 cv values (corresponding to the 1000 bootstrap samples). The process is simple. In this case, the standard error is 0.00917.

1.3 Odds Ratio

If an event has probability $P(A)$ of occurring, the **odds** of A occurring are defined to be

$$odds(A) = \frac{P(A)}{1 - P(A)}.$$

Now suppose that X denotes the event that an individual is exposed to a potentially harmful agent and that D denotes the event that the individual becomes diseased. We denote the complementary events as $\bar{X}$ and $\bar{D}$. The odds of an individual contracting the disease given that he is exposed are

$$odds(D|X) = \frac{P(D|X)}{1 - P(D|X)}$$

and the odds of contracting the disease given that he is not exposed are

$$odds(D|\bar{X}) = \frac{P(D|\bar{X})}{1 - P(D|\bar{X})}.$$

The **odds ratio** $\Delta = \frac{odds(D|X)}{odds(D|\bar{X})}$ is a measure of the influence of exposure on subsequent disease.

We will consider how the odds and odds ratio could be estimated by sampling from a population with joint and marginal probabilities defined as in the following table:

	$\bar{D}$	D	
$\bar{X}$	π_{00}	π_{01}	$\pi_{0.}$
X	π_{10}	π_{11}	$\pi_{1.}$
	$\pi_{.0}$	$\pi_{.1}$	1

With this notation,

$$P(D|X) = \frac{\pi_{11}}{\pi_{10} + \pi_{11}} \quad P(D|\bar{X}) = \frac{\pi_{01}}{\pi_{00} + \pi_{01}}$$

so that

$$odds(D|X) = \frac{\pi_{11}}{\pi_{10}} \quad odds(D|\bar{X}) = \frac{\pi_{01}}{\pi_{00}}$$

and the odds ratio is

$$\Delta = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}$$

the product of the diagonal probabilities in the preceding table divided by the product of the off-diagonal probabilities.

Now we will consider three possible ways to sample this population to study the relationship of disease and exposure.

- Random sample: From such a sample, we could estimate all the probabilities directly. However, if the disease is rare, the total sample size would have to be quite large to guarantee that a substantial number of diseased individuals was included.

- Prospective study: A fixed number of exposed and nonexposed individuals are sampled and then followed through time. The incidences of disease in those two groups are compared. In this case the data allow us to estimate and compare $P(D|X)$ and $P(D|\bar{X})$ and, hence, the odds ratio. The aspirin study described in the previous section can be viewed as this type of study.

- Retrospective study: A fixed number of diseased and undiseased individuals are sampled and the incidences of exposure in the two groups are compared. From such data we can directly estimate $P(X|D)$ and $P(X|\bar{D})$. Since the marginal counts of diseased and nondiseased are fixed, we cannot estimate the joint probabilities or the important conditional probabilities $P(D|X)$ and $P(D|\bar{X})$. Observe that

$$\begin{aligned} P(X|D) &= \frac{\pi_{11}}{\pi_{01}+\pi_{11}}, \quad 1-P(X|D) = \frac{\pi_{01}}{\pi_{01}+\pi_{11}}, \\ odds(X|D) &= \frac{\pi_{11}}{\pi_{01}}, \quad odds(X|\bar{D}) = \frac{\pi_{10}}{\pi_{00}}. \end{aligned}$$

  We thus see that the odds ratio can also be expressed as $odds(X|D)/odds(X|\bar{D})$.

Now we describe the study of Vianna, Greenwald, and Davies (1971) to illustrate the retrospective study. In this study they collected data comparing the percentages of tonsillectomies for a group of patients suffering from Hodgkin's disease and a comparable control group:

	Tonsillectomy	No Tonsillectomy
Hodgkin's	67	34
Control	43	64

Recall that the odds ratio can be expressed as $odds(X|D)/odds(X|\bar{D})$ and an estimate of it is $n_{00}n_{11}/(n_{01}n_{10})$, the product of the diagonal counts divided by the product of the off-diagonal counts. The data of Vianna, Greenwald, and Davies gives an estimate of odds ratio is

$$\frac{67 \times 64}{43 \times 34} = 2.93.$$

According to this study, the odds of contracting Hodgkin's disease is increased by about a factor of three by undergoing a tonsillectomy.

As well as having a point estimate 2.93, it would be useful to attach an approximate standard error to the estimate to indicate its uncertainty. We will use simulation (parametric bootstrap) to approximate the distribution of Δ. To do so, we need to generate random numbers according to a statistical model for the counts in the table of Vianna, Greenwald, and Davies. The model is that the count in the first row and first column, N_{11}, is binomially distributed with $n = 101$ and probability π_{11}. The count in the second row and second column, N_{22}, is binomially distributed with $n = 107$ and probability π_{22}. The distribution of the random variable

$$\hat{\Delta} = \frac{N_{11}N_{22}}{(101 - N_{11})(107 - N_{22})}$$

is thus determined by the two binomial distributions, and we could approximate it arbitrarily well by drawing a large number of samples from them. Since the probabilities π_{11} and π_{22} are unknown, they are estimated from the observed counts by $\hat{\pi}_{11} = 67/101 = 0.663$ and $\pi_{22} = 64/107 = 0.598$. A one thousand realizations generated on a computer gives the standard deviation 0.89.

2 Bootstrap Method

The bootstrap method introduced in Efron (1979) is a very general resampling procedure for estimating the distributions of statistics based on independent observations. The bootstrap method is shown to be successful in many situations, which is being accepted as an alternative to the asymptotic methods. In fact, it is better than some other asymptotic methods, such as the traditional normal approximation and the Edgeworth expansion. However, there are some counterexamples that show the bootstrap produces wrong solutions, i.e., it provides some inconsistent estimators.

Consider the problem of estimating variability of location estimates by the Bootstrap method. If we view the observations $x_1, x_2, \ldots, x_n$ as realizations of independent random variables with common distribution function F, it is appropriate to investigate the variability and sampling distribution of a location estimate calculated from a sample of size n. Suppose we denote the location estimate as $\hat{\theta}$. Note that $\hat{\theta}$ is a function of the random variables $X_1, X_2, \ldots, X_n$ and hence has a probability distribution, its sampling distribution, which is determined by n and F. We would like to know this sampling distribution, but we are faced with two problems:

1. we don't know F, and
2. even if we knew F, $\hat{\theta}$ may be such a complicated function of $X_1, X_2, \ldots, X_n$ that finding its distribution would exceed our analytic abilities.

First we address the second problem. Suppose we knew F. How could we find the probability distribution of $\hat{\theta}$ without going through incredibly complicated analytic calculations? The computer comes to our rescue-we can do it by simulation. We generate many, many samples, say B in number, of size n from F; from each sample we calculate the value of $\hat{\theta}$. The empirical distribution of the resulting values $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B$ is an approximation to the distribution function of $\hat{\theta}$, which is good if B is very large. If we wish to know the standard deviation of $\hat{\theta}$, we can find a good approximation to it by calculating the standard deviation of the collection of values $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$. We can make these approximations arbitrarily accurate by taking B to be arbitrarily large.

Assignment 4. Explain or prove that the simulation we just described will give a good approximation of the distribution function of θ.

All this would be well and good if we knew F, but we don't. So what do we do? We will consider two different cases. In the first case, F is unknown up to an unknown parameter η, i.e. $F(x|\eta)$. Without knowing η, the above approximation cannot be used. The idea of the **parametric bootstrap** is to simulate data from $F(x|\hat{\eta})$ where $\hat{\eta}$ should be a good estimate of η. Then it utilize the structure of F.

In the second case, F is completely unknown. The idea of the **nonparametric bootstrap** is to simulate data from the empirical cdf F_n. Here F_n is a discrete probability distribution that gives probability $1/n$ to each observed value $x_1, \cdots, x_n$. A sample of size n from F_n is thus a sample of size n drawn *with replacement* from the collection $x_1, \cdots, x_n$. The standard deviation of $\hat{\theta}$ is then estimated by

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B}\sum_{i=1}^{B}(\theta_i^* - \bar{\theta}^*)^2}$$

where $\theta_1^*, \ldots, \theta_B^*$ are produced from B sample of size n from the collection $x_1, \cdots, x_n$.

Now we use a simple example to illustrate this idea. Suppose $n = 2$ and observe $X_{(1)} = c < X_{(2)} = d$. Then X_1^*, X_2^* are independently distributed with

$$P(X_i^* = c) = P(X_i^* = d) = 1/2, \quad i = 1, 2.$$

The pairs (X_1^*, X_2^*) therefore takes on the four possible pairs of values

$$(c, c), (c, d), (d, c), (d, d),$$

each with probability 1/4. Thus $\theta^* = (X_1^* + X_2^*)/2$ takes on the values c, $(c+d)/2$, d with probabilities 1/4, 1/2, 1/4, respectively, so that $\theta^* - (c+d)/2$ takes on the values $(c-d)/2$, 0, $(d-c)/2$ with probabilities 1/4, 1/2, 1/4, respectively.

For the above example, we can easily calculate its bootstrap distribution. We can easily imagine that the above computation becomes too complicated to compute directly if n is large. Therefore, simple random sampling was proposed to generate bootstrap distribution. In the bootstrap literature, a variety alternatives are suggested other than simple random sampling.

Now we rewrite the above (generic) nonparametric bootstrap procedure into the following steps as follows. Refer to Efron and Tibshirani (1993) for detailed discussions. Consider the case where a random sample of size n is drawn from an unspecified probability distribution, F. The basic steps in the bootstrap procedure are

Step 1. Construct an empirical probability distribution, F_n, from the sample by placing a probability of $1/n$ at each point, $x_1, x_2, \cdots, x_n$ of the sample. This is the empirical distribution function of the sample, which is the nonparametric maximum likelihood estimate of the population distribution, F.

Step 2. From the empirical distribution function, F_n, draw a random sample of size n with replacement. This is a *resample.*

Step 3. Calculate the statistic of interest, T_n, for this resample, yielding T_n^*.

Step 4. Repeat steps 2 and 3 B times, where B is a large number, in order to create B resamples. The practical size of B depends on the tests to be run on the data. Typically, B is at least equal to 1000 when an estimate of confidence interval around T_n is required.

Step 5. Construct the relative frequency histogram from the B number of T_n^*'s by placing a probability of $1/B$ at each point, $T_n^{*1}, T_n^{*2}, \ldots, T_n^{*B}$. The distribution obtained is the bootstrapped estimate of the sampling distribution of T_n. This distribution can now be used to make inferences about the parameter θ, which is to be estimated by T_n.

We now introduce notations to illustrate the bootstrap method. Assumed the data $X_1, \cdots, X_n$, are independent and identically distributed (iid) samples from a k-dimensional population distribution F and the problem of estimating the distribution

$$H_n(x) = P\{R_n \leq x\},$$

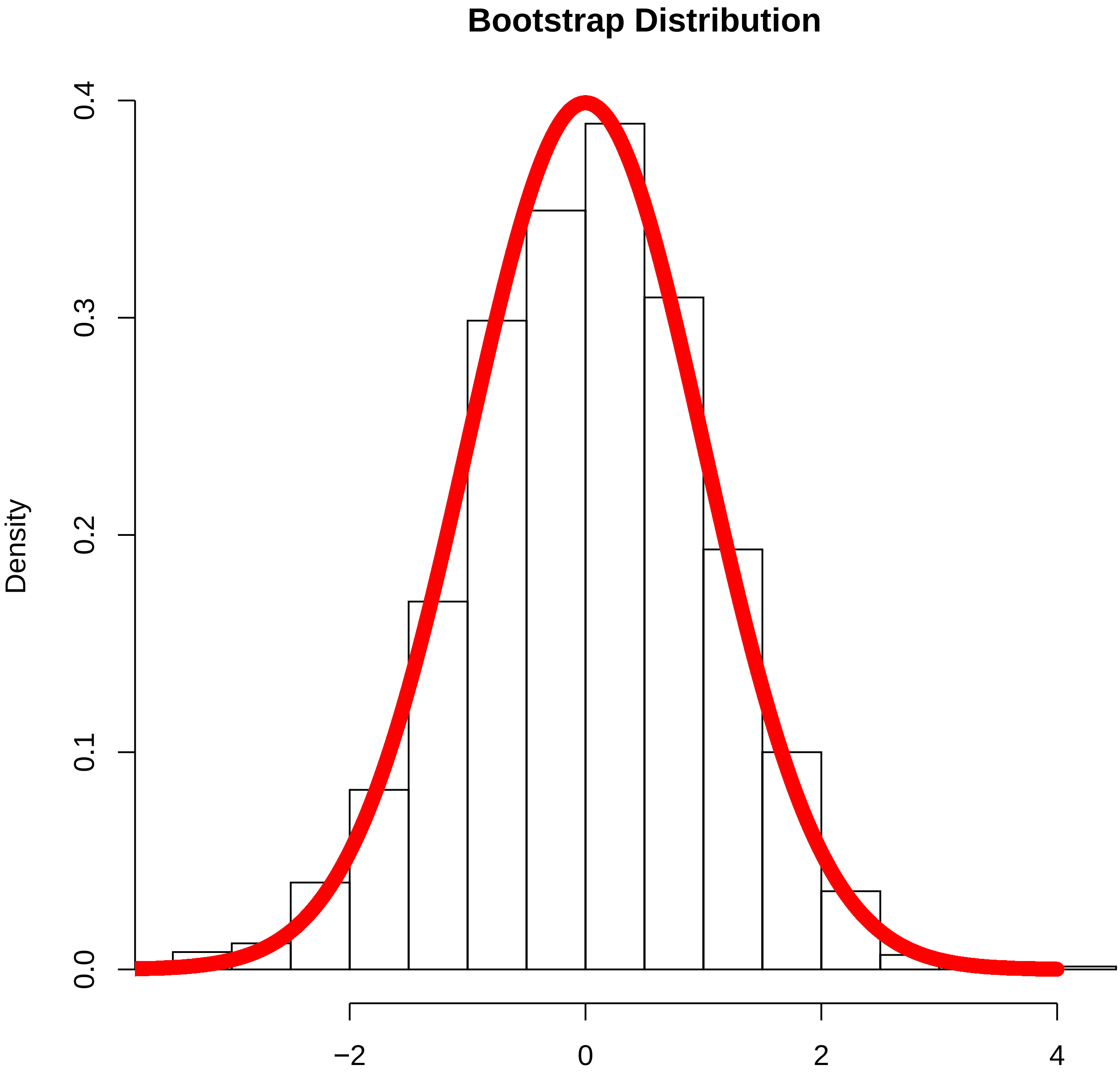
where $R_n = R_n(T_n, F)$ is a real-valued functional of F and $T_n = T_n(X_1, \cdots, X_n)$, a statistic of interest. Let $X_1^*, \cdots, X_n^*$ be a "bootstrap" samples iid from F_n, the empirical distribution based on $X_1, \cdots, X_n$, $T_n^* = T_n(X_1^*, \cdots, X_n^*)$, and $R_n^* = R_n(T_n^*, F_n)$. F_n is constructed by placing at each observation X_i a mass $1/n$. Thus F_n may be represented as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \qquad\qquad -\infty < x < \infty.$$

A bootstrap estimator of H_n is

$$\hat{H}_n(x) = P_*\{R_n^* \leq x\},$$

where for given $X_1, \cdots, X_n$, P_* is the conditional probability with respect to the random generator of bootstrap samples. Since the bootstrap samples are generated from F_n, this method is called the nonparametric bootstrap. Note that $\hat{H}_n(x)$ will depend on F_n and hence


Bootstrap Distribution
Density
0.0
0.1
0.2
0.3
0.4
−2
0
2
4

itself is a random variable. To be specific, $\hat{H}_n(x)$ will change as the data $\{x_1, \cdots, x_n\}$ changes. Recall that a bootstrap analysis is run to assess the accuracy of some primary statistical results. This produces bootstrap statistics, like standard errors or confidence intervals, which are assessments of error for the primary results.

As illustration, we consider the following three examples.

Example 1. Suppose that $X_1, \cdots, X_n \sim N(\mu, 1)$ and $R_n = \sqrt{n}(\bar{X}_n - \mu)$. Consider the estimation of

$$P(a) = P\{R_n > a | N(\mu, 1)\}.$$

The nonparametric bootstrap method will estimate $P(a)$ by

$$P_{NB}(a) = P\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) > a | F_n\}.$$

To be specific, we observe data $x_1, \cdots, x_n$ with mean $\bar{x}_n$. Let $Y_1, \ldots, Y_n$ denote a bootstrap sample of n observations drawn independently from F_n and let $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$. Then $P(a)$ is estimated by

$$P_{NB}(a) = P\{\sqrt{n}(\bar{Y}_n - \bar{x}_n) > a | F_n\}.$$

In principle, $P_{NB}(a)$ can be found by considering all n^n possible bootstrap sample. If all X_i's are distinct, then the number of different possible resamples equals the number of distinct ways of placing n indistinguishable objects into n numbered boxes, the boxes being allowed to contain any number of objects. It is known that it is equal to $C(2n-1, n) \approx (n\pi)^{-1/2} 2^{2n-1}$. When $n = 10$(20, respect.), $C(2n-1, n) \approx 92375(6.9 \times 10^{10}$, respect.). For small value of n, it is often feasible to calculate a bootstrap estimate exactly. However, for large samples, say $n \geq 10$, this becomes infeasible even at today's computer technology. Natural questions to ask are as follows:

- What are computationally efficient ways to bootstrap?
- Can we get bootstrap-like answers without Monte Carlo?

Moreover, we need to address the question of "evaluating" the performance of bootstrap method. For the above particular problem, we need to estimate $P_{NB}(a) - P(a)$ or $\sup_a |P_{NB}(a) - P(a)|$. As a remark, $P_{NB}(a)$ is a random variable since F_n is random. Efron (1992) proposed to use jackknife to give the error estimates for bootstrap quantities.

Suppose that additional information on F is available. Then it is reasonable to utilize this information in the bootstrap method. For example, F known to be normally distributed with unknown mean μ and variance 1. It is natural to use $\bar{x}_n$ to estimate μ and then estimate $P(a) = P\{R_n > a | N(\mu, 1)\}$ by

$$P_{PB}(a) = P\{\sqrt{n}(\bar{Y}_n - \bar{x}_n) > a | N(\bar{x}_n, 1)\}.$$

Since the bootstrap samples are generated from $N(\bar{x}_n, 1)$ which utilizes the information from a parametric form of F, this method is called the parametric bootstrap. In this case, it can be shown that $P_{PB}(a) = P(a)$ for all realization of $\bar{X}_n$. However, if F is known to be normally distributed with unknown mean and variance μ and variance σ^2 respectively, $P_{PB}(a)$ is no longer equal to $P(a)$.

Assignment 5. (a) Show that $P_{PB}(a) = \Phi(a/s_n)$ where $s_n^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$.
(b) Prove that $P_{PB}(a)$ is a consistent estimate of $P(a)$ for fixed a.
(c) Prove that $\sup_a |P_{PB}(a) - P(a)| \xrightarrow{P} 0$.

For the question of finding $P_{NB}(a)$, we can in principle write down the characteristic function and then apply the inversion formula. However, it is a nontrivial job. Therefore, Efron (1979) suggested to approximate $P_{NB}(a)$ by Monte Carlo resampling. (i.e., Sample-size resamples may be drawn repeatedly from the original sample, the value of a statistic computed for each individual resample, and the bootstrap statistic approximated by taking an average of an appropriate function of these numbers.)

Example 1. (cont.) Let us consider a sample containing two hundred values generated randomly from a standard normal population $N(0, 1)$. This is the original sample. In this example, the sampling distribution of the arithmetic mean is approximately normal with a mean roughly equal to 0 and a standard deviation approximately equal to $1/\sqrt{200}$. Now, let us apply the nonparametric bootstrap method to infer the result. One thousand and five hundred resamples are drawn from the original sample, and the arithmetic mean is calculated for each resample. These calculations are performed by using R functions as follows

Step 1. Randomly draw two hundred points from a standard normal population

$gauss < -rnorm(200, 0, 1)$

Step 2. Perform the nonparametric bootstrap study (1500 resamples)

$bootmean < -1 : 1500$

$for(i \; in \; 1 : 1500) \; bootmean[i] < -mean(sample(gauss, replace = T))$

Step 3. Do the normalization and comparison with $N(0, 1)$.

$bootdistribution < -sqrt(200) * (bootmean - mean(gauss))$

$hist(bootdistribution, freq = FALSE, main =$ "Bootstrap Distribution"$, xlab =$ "")

$x < -seq(-4, 4, 0.001); y < -(1/(sqrt(2 * pi))) * exp(-x^2/2)$

$points(x, y, col = 2)$

Now we state Levy's Inversion Formula which is taken from Chapter 6.2 of Chung (1974).

Theorem 1 *If $x_1 < x_2$ and x_1 and x_2 are points of continuity of F, then we have*

$$F(x_2) - F(x_1) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t) dt,$$

where $f(t)$ is the characteristic function.

Example 2. (Estimating the probability of success) Consider a probability distribution F putting all of its mass at zero or one. Let $\theta(F) = P(X = 1) = p$. Consider $R(\mathbf{X}, F) = \bar{X} - \theta(F) = \hat{p} - p$. Observed $\mathbf{X} = \mathbf{x}$, the bootstrap sample

$$X_1^*, \cdots, X_n^* \sim Bin(1, \theta(F_n)) = Bin(1, \bar{x}_n).$$

Note that

$$\begin{aligned} R(\mathbf{X}^*, F_n) &= \bar{X}_n^* - \bar{x}_n, \\ E_*(\bar{X}_n^* - \bar{x}_n) &= 0, \\ Var_*(\bar{X}_n^* - \bar{x}_n) &= \frac{\bar{x}_n(1-\bar{x}_n)}{n}. \end{aligned}$$

Recall that $n\bar{X}_n^* \sim Bin(n, \bar{x})$ and $n\bar{X}_n \sim Bin(n, p)$. It is known that if $\min\{n\bar{x}_n, n(1-\bar{x}_n)\} \geq 5$,

$$\frac{n\bar{X}_n^* - n\bar{x}_n}{\sqrt{n\bar{x}_n(1-\bar{x}_n)}} = \frac{\sqrt{n}(\bar{X}_n^* - \bar{x}_n)}{\sqrt{\bar{x}_n(1-\bar{x}_n)}} \sim N(0,1);$$

and if $\min\{np, n(1-p)\} \geq 5$,

$$\frac{n\bar{X}_n - np}{\sqrt{n\theta(1-p)}} = \frac{\sqrt{n}(\bar{X}_n - p)}{\sqrt{p(1-p)}} \sim N(0,1).$$

Based on the above approximation results, we conclude that the bootstrap method works if $\min\{n\bar{x}_n, n(1-\bar{x}_n)\} \geq 5$. The question remained to be studied is whether

$$P\{\min(n\bar{X}_n, n(1-\bar{X}_n)) \geq 5\} \to 0?$$

Example 3. (Estimating the median) Suppose we are interested in finding the distribution of $n^{1/2}\{F_n^{-1}(1/2) - F^{-1}(1/2)\}$ where $F_n^{-1}(1/2)$ and $F^{-1}(1/2)$ are the sample and population median respectively. Set $\theta(F) = F^{-1}(1/2)$. The normal approximation for this distribution will be discussed in Chapter 2. In this section, we consider the bootstrap approximation of the above distribution.

Consider $n = 2m - 1$. Then the sample median $F_n^{-1}(1/2) = X_{(m)}$ where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Let N_i^* denote the number of times x_i is selected in the bootstrap sampling procedure.

Set $\mathbf{N}^* = (N_1^*, \cdots, N_n^*)$. It follows easily that $\mathbf{N}^*$ follows a multinomial distribution with n trials and the probability of selection is $(n^{-1}, \cdots, n^{-1})$. Denote the order statistics of $x_1, \ldots, x_n$ by $x_{(1)} \leq \cdots \leq x_{(n)}$. Set $N_{[i]}^*$ to be the number of times of choosing $x_{(i)}$. Then for $1 \leq \ell < n$, we have

$$\begin{aligned} Prob_*(X_{(m)}^* > x_{(\ell)}) &= Prob_*\{N_{[1]}^* + \cdots + N_{[\ell]}^* \leq m-1\} \\ &= Prob\left\{Bin\left(n, \frac{\ell}{n}\right) \leq m-1\right\} = \sum_{j=0}^{m-1} C(n,j)\left(\frac{\ell}{n}\right)^j \left(1 - \frac{\ell}{n}\right)^{n-j}. \end{aligned}$$

Or,

$$Prob_*(T^* = x_{(\ell)} - x_{(m)}) = Prob\left\{Bin\left(n, \frac{\ell-1}{n}\right) \le m-1\right\} - Prob\left\{Bin\left(n, \frac{\ell}{n}\right) \le m-1\right\}.$$

When $n = 13$, we have

ℓ	2 or 12	3 or 11	4 or 10	5 or 9	6 or 8	7
probability	0.0015	0.0142	0.0550	0.1242	0.4136	0.2230

Quite often we use the mean square error to measure the performance of an estimator, $t(X)$, of $\theta(F)$. Or, $E_F T^2 = E_F(t(X) - \theta(F))^2$. We then can use bootstrap to estimate $E_F T^2$. Then the bootstrap estimate of $E_F T^2$ is

$$E_*(T^*)^2 = \sum_{\ell=1}^{13} [x_{(\ell)} - x_{(7)}]^2 Prob_*\{T^* = x_{(\ell)} - x_{(7)}\}.$$

It is known that $E_F T^2 \to [4nf^2(\theta)]^{-1}$ as n tends to infinity when F has a bounded continuous density. A natural question to ask is whether $E_*(T^*)^2$ is close to $E_F T^2$?

3 Validity of the Bootstrap Method

We now give a brief discussion on the validity of the bootstrap method. First, we state central limit theorems and its approximation error bound which will be used in proving that the bootstrap can provide a good approximation of distribution of $n^{1/2}(\hat{p} - p)$.

3.1 Central Limit Theorem

Perhaps the most widely known version of the CLT is

Theorem 2 *(Lindeberg-Levy) Let $\{X_i\}$ be iid with mean μ and finite variance σ^2. Then*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} N(0, \sigma^2).$$

The above theorem can be generalized to independent random variables which are not necessarily identically distributed.

Theorem 3 *(Lindeberg-Feller) Let $\{X_i\}$ be independent with mean $\{\mu_i\}$, finite variances $\{\sigma_i^2\}$, and distribution functions $\{F_i\}$. Suppose that $B_n^2 = \sum_{i=1}^n \sigma_i^2$ satisfies*

$$\frac{\sigma_n^2}{B_n^2} \to 0, \quad B_n \to \infty \qquad \text{as } n \to \infty.$$

Then $n^{-1}\sum_{i=1}^n X_i$ is $N(n^{-1}\sum_{i=1}^n \mu_i, n^{-2}B_n^2)$ if and only if the following Lindeberg condition satisfied

$$B_n^{-2}\sum_{i=1}^{n}\int_{|t-\mu_i|>\epsilon B_n} (t-\mu_i)^2 dF_i(t) \to 0, \quad n \to \infty \qquad \text{each } \epsilon > 0.$$

In the theorems previously considered, asymptotic normality was asserted for a sequence of sums $\sum_1^n X_i$ generated by a single sequence $X_1, X_2, \ldots$ of random variables. More generally, we may consider a *double array* of random variables

$$\begin{array}{cccc} X_{11}, & X_{12}, & \cdots, & X_{1K_1}; \\ X_{21}, & X_{22}, & \cdots, & X_{2K_2}; \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1}, & X_{n2}, & \cdots, & X_{nK_n}; \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

For each $n \geq 1$, there are K_n random variables $\{X_{nj}, 1 \leq j \leq K_n\}$. It is assumed that $K_n \to \infty$. The case $K_n = n$ is called a "triangular" array.

Denote by F_{nj} the distribution function of X_{nj}. Also, put

$$\begin{aligned} \mu_{nj} &= EX_{nj}, \\ A_n &= E\sum_{j=1}^{K_n} X_{nj} = \sum_{j=1}^{K_n} \mu_{nj}, \\ B_n^2 &= Var\left(\sum_{j=1}^{K_n} X_{nj}\right). \end{aligned}$$

We then have the following theorem.

Theorem 4 *(Lindeberg-Feller) Let $\{X_{nj} : 1 \leq j \leq K_n; n = 1, 2, \ldots\}$ be a double array with independent random variables within rows. Then the "uniform asymptotic negligibility" condition*

$$\max_{1 \leq j \leq K_n} P(|X_{nj} - \mu_{nj}| > \tau B_n) \to 0, \quad n \to \infty, \textit{each } \tau > 0,$$

and the asymptotic normality condition $\sum_{j=1}^{K_n} X_{nj}$ is $AN(A_n, B_n^2)$ together hold if and only if the Lindberg condition

$$B_n^{-2} \sum_{i=1}^{n} \int_{|t-\mu_i| > \epsilon B_n} (t - \mu_i)^2 dF_i(t) \to 0, \quad n \to \infty \textit{each } \epsilon > 0$$

is satisfied.

As a note, the independence is assumed only it within rows, which themselves may be arbitrarily dependent.

Corollary 1 *Suppose that, for some $v > 2$, $\sum_{j=1}^{K_n} E|X_{nj} - \mu_{nj}|^v = o(B_n^v)$, $n \to \infty$. Then $\sum_{j=1}^{K_n} X_{nj}$ is $AN(A_n, B_n^2)$.*

3.2 Approximation Error of CLT

It is of both theoretical and practical interest to characterize the error of approximation in the CLT. In this section, we just consider the i.i.d. case. The convergence in the Central

Limit Theorem is not uniform in the underlying distribution. For any fixed sample size n, there are distributions for which the normal distribution approximation to the distribution function of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is arbitrarily poor. However, there is an upper bound, due to Berry (1941) and Esseen (1942), to the error of the Central Limit Theorem approximation that shows the convergence is uniform for the class of distributions for which $|X-\mu|^3/\sigma^3$ is bounded above by a finite bound. We state this theorem without proof in one dimension.

Theorem 5 *If $X_1, \ldots, X_n$ are i.i.d. with distribution F and if $S_n = X_1 + \cdots + X_n$, then there exists a constant c (independent of F) such that for all x,*

$$\sup_x \left| P\left[\frac{S_n - ES_n}{\sqrt{Var(S_n)}} \le x\right] - \Phi(x)\right| \le \frac{c}{\sqrt{n}} \frac{E|X_1 - EX_1|^3}{[Var(X_1)]^{3/2}}$$

for all F with finite third moment.

Note that c in the above theorem is a universal constant. Various authors have thought to find the best constant c. Originally, c is set to be $33/4$ but it has been sharpened to be greater than 0.4097 and less than 0.7975. For x is sufficiently large, while n remains fixed, the quantity $P[(S_n - ES_n)/\sqrt{Var(S_n)} \le x]$ become so close to 1 that the bound given by above is too crude. The problem in this case may be characterized as one of approximation of *large deviation* probabilities, with the object of attention becoming the relative error in approximation of

$$1 - P[(S_n - ES_n)/\sqrt{Var(S_n)} \le x]$$

by $1 - \Phi(x)$ when $x \to \infty$.

When we have information about the third and higher moments of the underlying distribution, we may often improve on the normal approximation by considering higher-order terms in the expansion of the characteristic function. This leads to asymptotic expansions known as Edgeworth Expansions. We present without proof the two next terms in the Edgeworth Expansion.

$$\Phi(x) - \frac{\beta_1(x^2-1)}{6\sqrt{n}}\phi(x) - \left[\frac{\beta_2(x^3-3x)}{24n} + \frac{\beta_1^2(x^5 - 10x^3 + 15x)}{72n}\right]\phi(x).$$

where $\beta_1 = E(X-\mu)^3/\sigma^3$ and $\beta_2 = E(X-\mu)^4/\sigma^4 - 3$ are the coefficient of skewness and the coefficient of kurtosis, respectively, and where $\phi(x)$ represents the density of the standard normal distribution. This approximation is to be understood in the sense that the difference of the two sides when multiplied by n tends to zero as $n \to \infty$. Assuming the fourth moment exists, it is valid under the condition that

$$\limsup_{|t|\to\infty} |E(\exp\{itX\})| < 1.$$

This condition is known as *Cramer's Condition.* It holds, in particular, if the underlying distribution has a nonzero absolutely continuous component. The expansion to the term involving $1/\sqrt{n}$ is valid if the third moment exists, provided only that the underlying distribution is nonlattice, and even for lattice distributions it is valid provided a correction for continuity is made. See Feller (Vol. 2, Chap. XVI.4) for details.

Let us inspect this approximation. If we stop at the first term, we have the approximation given by the Central Limit Theorem. The next term is of order $n^{-1/2}$ and represents a correction for skewness, since this term is zero if $\beta_1 = 0$. In particular, if the underlying distribution is symmetric, the Central Limit Theorem approximation is accurate up to terms of order $1/n$. The remaining term is a correction for kurtosis (and skewness) or order $1/n$.

The Edgeworth Expansion is an asymptotic expansion, which means that continuing with further terms in the expansion with n fixed may not converge. In particular, expanding to further terms for fixed n may make the accuracy worse. There are a number of books treating the more advanced theory of Edgeworth and allied expansions. The review by Bhattacharya (1990), treats the more mathematical aspects of the theory and the book of Barndorff-Nielsen and Cox (1989) the more statistical. Hall (1992) is concerned with the application of Edgeworth Expansion to the bootstrap.

3.3 Estimation of the Probability of Success

We now discuss whether bootstrap method will give a consistent estimate of the distribution of $n^{1/2}(\hat{p}_n - p)$. For simplicity, we use an asymptotic analysis to evaluate it. Note that as $n \to \infty$, F_n will change accordingly. This is different from some asymptotic analysis in which the underlying distribution F never change with n.

Two different approaches are used to address this question. Since that the underlying distribution F_n changes with n, the first approach is to use the double array CLT to handle the case and the second approach is to use approximation result, Berry-Esseen bound.

Proof 1: Note that the bootstrap samples at sample size n as $Y_{n1}, \cdots, Y_{nn}$ which come from $Bin(1, \bar{x}_n)$. Note that $Bin(1, \bar{x}_n)$ is the so-called F_{nj} in the double array CLT. Then $\mu_{nj} = \bar{x}_n$, $A_n = n\bar{x}_n$, $K_n = n$, and $B_n^2 = n\bar{x}_n(1-\bar{x}_n)$.

1. Check UAN condition.

$$P(|Y_{nj} - \mu_{nj}| > \tau\sqrt{n[\bar{x}_n(1-\bar{x}_n)]}) = 0.$$

2. Check Lindberg condition.

$$\int_{|t-\mu_{nj}|>\epsilon B_n} (t-\mu_{nj})^2 dF_{nj}(t) = 0.$$

These imply that $\sum_{j=1}^{n} Y_{nj}$ is $AN(n\bar{x}_n, n\bar{x}_n(1-\bar{x}_n))$.

It is well known that $\sqrt{n}(\hat{p} - p)$ is asymptotically normally distributed with mean 0 and variance $p(1-p)$. If, for all realizations, $\bar{X}_n$ converges to p with probability 1, we then conclude that the bootstrap distribution of $\hat{p}$ will also converge to normal with mean 0 and variance $p(1-p)$. This gives a justification that the bootstrap method is consistent. As a remark, the bootstrap method is most powerful when we don't know how to do asymptotic analysis. In such a case, how do we justify that the bootstrap method is consistent is a challenging problem.

Proof 2: Using the Berry-Esseen bound, we have

$$\sup_x \left| P\left[\frac{n\bar{Y}_n - n\bar{x}_n}{\sqrt{\bar{x}_n\,(1-\bar{x}_n)}} \le x \right] - \Phi(x) \right| \le \frac{c}{\sqrt{n}} \frac{E(Y-\bar{x}_n)^3}{[\bar{x}_n(1-\bar{x}_n)]^{3/2}}.$$

If $\bar{x}_n(1-\bar{x}_n)$ is bounded away from zero, the right hand side will tend to zero. This gives a justification that the bootstrap method is consistent.

3.4 Statistical Functionals

Many statistics including the sample mean, the sample median and the sample variance, are consistent estimators of their corresponding population quantity: the sample mean $\bar{X}$ of the expectation $E(X)$, the p^{th} sample quantile of the p^{th} population quantile $F^{-1}(p)$, the k^{th} sample moment $\sum(X_i-\bar{X})^k/n$ of the k^{th} population moment $E[X-E(X)]^k$, etc. Any such population quantity is a function of the distribution F of the X_i and can therefore be written as $h(F)$, where h is a real-valued function defined over a collection $\mathcal{F}$ of distributions F. The mean $h(F) = E_F(X)$, for example, is defined over the class $\mathcal{F}$ of all F with finite expectation. Statistics which are representable as functionals $h(F)$ are so-called statistical functionals.

To establish the connection between the sequence of sample statistics and functional $h(F)$ that it estimates, define the *sample cdf* $\hat{F}_n$ by

$$\hat{F}_n(x) = \frac{\text{Number of } X_i \le x}{n}.$$

This is the cdf of a distribution that assigns probability $1/n$ to each of the n sample values $X_1, X_2, \ldots, X_n$. For the examples mentioned so far and many others, it turns out that the standard estimator of $h(F)$ based on n observations is equal to $h(\hat{F}_n)$, the plug-in estimator of $h(F)$. When $h(F) = E[X-E(X)]^k$, it is seen that

$$h(\hat{F}_n) = \frac{1}{n}(X_1-\bar{X})^k + \cdots + \frac{1}{n}(X_n-\bar{X})^k.$$

Note that $\sum(X_i-\bar{X})^k/n$ can be viewed as a function of n variables or as a function of $\hat{F}_n$.

Suppose we want to evaluate the performance of an estimator $\hat{\theta}_n$ of some parameter θ or functional $h(F)$. As an example, the sample median $\hat{\theta}_n$ as an estimator of the population median. We can use the following as a measure

$$\lambda_n(F) = P_F\left\{\sqrt{n}[\hat{\theta}_n - h(F)] \le a\right\}.$$

Note that population median can be written as $F^{-1}(0.5)$ and then the sample median can be viewed as a plug-in estimate $\hat{F}_n^{-1}(0.5)$. Again, we can estimate $\lambda_n(F)$ by the plug-in estimator $\lambda_n(\hat{F}_n)$ in which the distribution F of the X's by the distribution $\hat{F}_n$. In addition, the subscript F, which governs the distribution of $\hat{\theta}_n$, must also be changed to $\hat{F}_n$. To see what this last step means, write

$$\hat{\theta}_n = \theta(X_1, \ldots, X_n), \tag{1}$$

that is, express $\hat{\theta}_n$ not as a function of $\hat{F}_n$ but directly as a function of the sample $(X_1, \ldots, X_n)$. The dependence of the distribution of $\hat{\theta}_n$ on F results from the fact that $X_1, \ldots, X_n$ is a sample from F. To replace F by $\hat{F}_n$ in the distribution governing $\hat{\theta}_n$, we must therefore replace (1) by

$$\hat{\theta}_n^* = \theta(X_1^*, \ldots, X_n^*), \tag{2}$$

where $X_1^*, \ldots, X_n^*$ is a sample from $\hat{F}_n$. With this notation, $\lambda_n(\hat{F}_n)$ can now be written formally as

$$\lambda_n(\hat{F}_n) = P_{\hat{F}_n}\left\{\sqrt{n}[\hat{\theta}_n^* - h(\hat{F}_n)] \leq a\right\}. \tag{3}$$

When $\lambda_n(\hat{F}_n)$ is too complicated to compute directly, we can approximate $\lambda_n(\hat{F}_n)$ by $\lambda_{B,n}^*$ as suggested in Efron (1979). Here

$$\lambda_{B,n}^* = \frac{1}{n}\sum_1^B E_{\hat{F}_n}\hat{\theta}_n - h(\hat{F}_n). \tag{4}$$

Here we don't give any discussion of theoretical properties of the plug-in estimator $\lambda_n(\hat{F}_n)$, such as consistency and asymptotic normality. In much of the bootstrap literature, the bootstrap is said *to work* if $\lambda_n(\hat{F}_n)$ is consistent for estimating $\lambda_n(F)$ in the sense that $\lambda_n(\hat{F}_n) \to \lambda_n(F)$ for all F under consideration.

4 Inconsistent Bootstrap Estimator

Bickel and Freedman (1981) and Loh (1984) showed that the bootstrap estimators of the distributions of the extreme-order statistics are inconsistent. Let $X_{(n)}$ be the maximum of i.i.d. random variables $X_1, \ldots, X_n$ from F with $F(\theta) = 1$ for some θ, and let $X_{(m)}^*$ be the maximum of $X_{(1)}^*, \ldots, X_{(m)}^*$ which are i.i.d. from the empirical distribution F_n. Although $X_{(n)} \to \theta$, it never equals θ. But

$$P_*\{X_{(n)}^* = X_{(n)}\} = 1 - (1 - n^{-1})^n \to 1 - e^{-1},$$

which leads to the inconsistency of the bootstrap estimator.

The reason for the inconsistency of the bootstrap is that the bootstrap samples are drawn from F_n which is not exactly F. Therefore, the bootstrap may fail due to the lack of "continuity." We now illustrate that the bootstrap can produce wrong solutions, i.e., it provides some inconsistent estimators. Refer to Shao (1994) for further references. We focus on the case where the data $X_1, \ldots, X_n$ are i.i.d. samples from a k-dimensional population distribution F and the problem of estimating the distribution

$$H_n(x) = P\{R_n \leq x\}, \tag{5}$$

where $R_n = R_n(T_n, F)$ is a real-valued functional of F and $T_n = T_n(X_1, \ldots, T_n)$, a statistic of interest. A bootstrap estimator of H_n is

$$\hat{H}_n(x) = P_*\{R_n^* \leq x\}, \tag{6}$$

where $R_n^* = R_n(T_n^*, F_n)$. Let $\mu = EX_1$; $\theta = g(\mu)$, where g is a function from R^k to R; $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ be the sample mean; and $T_n = g(\bar{X}_n)$. Under the conditions that $Var(X_1) = \Sigma < \infty$ and g is first-order continuously differentiable at μ,

$$\sup_x \left| P\left\{\sqrt{n}(T_n^* - T_n) \leq x\right\} - P\left\{\sqrt{n}(T_n - \theta) \leq x\right\}\right| \to 0 \quad \text{a.s.}, \tag{7}$$

where $T_n^* = g(\bar{X}_n^*)$ and $\bar{X}_n^* = n^{-1}\sum_{i=1}^n X_i^*$.

Consider the situation where g is second-order continuously differentiable at μ with $\bigtriangledown^2 g(\mu) \neq 0$ but $\bigtriangledown g(\mu) = 0$. Using the Taylor expansion and $\bigtriangledown g(\mu) = 0$, we obtain that

$$T_n - \theta = \frac{1}{2}(\bar{X}_n - \mu)' \bigtriangledown^2 g(\mu)(\bar{X}_n - \mu) + o_P(n^{-1}). \tag{8}$$

This implies

$$n(T_n - \theta) \xrightarrow{d} \frac{1}{2} Z_\Sigma' \bigtriangledown^2 g(\mu) Z_\Sigma, \tag{9}$$

where Z_Σ is a random k-vector having normal distribution with mean 0 and covariance matrix Σ. From (9), $\sqrt{n}(T_n - \theta) \xrightarrow{P} 0$, and, therefore, result (7) is not useful when $\bigtriangledown g(\mu) = 0$ and we need to consider the bootstrap estimator of the distribution of $n(T_n - \theta)$ in this case.

Let $R_n = n(T_n - \theta)$, $R_n^* = n(T_n^* - T_n)$, and H_n and $\hat{H}_n$ be given by (5) and (6), respectively. Babu (1984) pointed out that $\hat{H}_n$ is inconsistent in this case. Similar to (8),

$$T_n^* - T_n = \bigtriangledown g(\bar{X}_n)'(\bar{X}_n^* - \bar{X}_n) + \frac{1}{2}(\bar{X}_n^* - \bar{X}_n)' \bigtriangledown^2 g(\bar{X}_n)(\bar{X}_n^* - \bar{X}_n) + o_P(n^{-1}) \quad \text{a.s.} \tag{10}$$

B the continuity of $\bigtriangledown^2 g$ and Theorem 2.1 of Bickel and Freedman (1981), for almost all given sequences $X_1, X_2, \ldots$,

$$\frac{n}{2}(\bar{X}_n^* - \bar{X}_n)' \bigtriangledown^2 g(\bar{X}_n)(\bar{X}_n^* - \bar{X}_n) \xrightarrow{d} \frac{1}{2} Z_\Sigma' \bigtriangledown^2 g(\mu) Z_\Sigma. \tag{11}$$

From $\bigtriangledown g(\mu) = 0$,

$$\sqrt{n} \bigtriangledown g(\bar{X}_n) = \sqrt{n} \bigtriangledown^2 g(\mu)(\bar{X}_n - \mu) + o_P(1) \xrightarrow{d} \bigtriangledown^2 g(\mu) Z_\Sigma. \tag{12}$$

Hence, for almost all given $X_1, X_2, \ldots$, the conditional distribution of $n \bigtriangledown g(\bar{X}_n)'(\bar{X}_n^* - \bar{X}_n)$ does not have a limit. It follows from (10) and (11) that for almost all given $X_1, X_2, \ldots$, the conditional distribution of $n(T_n^* - T_n)$ does not have a limit. Therefore, $\hat{H}_n$ is inconsistent as an estimator of H_n.

The symptom of this problem in the present case is that $\bigtriangledown g(\bar{X}_n)$ is not necessarily equal to zero when $\bigtriangledown g(\mu) = 0$. As a result, the expansion in (10), compared with the expansion in (8), has an extra nonzero term $\bigtriangledown g(\bar{X}_n)'(\bar{X}_n^* - \bar{X}_n)$ which does not converge to zero fast enough, and, therefore, $\hat{H}_n$ cannot mimic H_n.

5 Bias Reduction via the Bootstrap Principle

In this section, we will use an example to illustrate the bias reduction via the Bootstrap principle. Consider $\theta_0 = \theta(F_0) = \mu^3$, where $\mu = \int x dF_0(x)$. Set $\hat{\theta} = \theta(F_n) = \bar{X}^3$. Elementary calculations show that

$$\begin{aligned} E\{\theta(F_n)|F_0\} &= E\{\mu + n^{-1}\sum_{i=1}^{n}(X_i - \mu)\}^3 \qquad (13) \\ &= \mu^3 + n^{-1}3\mu\sigma^2 + n^{-2}\gamma, \end{aligned}$$

where $\gamma = E(X_1 - \mu)^3$ denotes population skewness. Using the nonparametric bootstrap, we obtain in direct analogy to (13)

$$E\{\theta(F_n^*)|F_n\} = \bar{X}^3 + n^{-1}3\bar{X}\hat{\sigma}^2 + n^{-2}\hat{\gamma},$$

where $\hat{\sigma}^2 = n^{-1}\sum(X_i - \bar{X})^2$ and $\hat{\gamma} = n^{-1}\sum(X_i - \bar{X})^3$ denote sample variance and skewness respectively. Using the bootstrap principle, $E\{\theta(F_n^*)|F_n\} - \theta(F_n)$ is used to estimate $\theta(F_n) - \theta(F_0)$. Note that $\theta_0 = \theta(F_n) - (\theta(F_n) - \theta_0)$. Or, θ_0 can be estimated by $\theta(F_n) - [E\{\theta(F_n^*)|F_n\} - \theta(F_n)]$ or $2\theta(F_n) - E\{\theta(F_n^*)|F_n\}$. Therefore the bootstrap bias-reduced estimate is $2\bar{X}^3 - (\bar{X}^3 + n^{-1}3\bar{X}\hat{\sigma}^2 + n^{-2}\hat{\gamma})$. Or, $\hat{\theta}_{NB} = \bar{X}^3 - n^{-1}3\bar{X}\hat{\sigma}^2 - n^{-2}\hat{\gamma}$.

Now we check whether $\hat{\theta}_{NB}$ really reduces bias. Observe that for general distributions with finite third moments,

$$\begin{aligned} E(\bar{X}^3) &= \mu^3 + n^{-1}3\mu\sigma^2 + n^{-2}\gamma, \\ E(\bar{X}\hat{\sigma}^2) &= \mu\sigma^2 + n^{-1}(\gamma - \mu\sigma^2) - n^{-2}\gamma, \\ E(\hat{\gamma}) &= \gamma(1 - 3n^{-1} + 2n^{-2}). \end{aligned}$$

It follows that

$$E(\theta(F_n)) - \theta_0 = n^{-2}3(\mu\sigma^2 - \gamma) + n^{-3}6\gamma - n^{-4}2\gamma$$

for general distributions.

6 Jackknife

One of the central goals of data analysis is an estimate of the uncertainties in fit parameters. Sometimes standard methods for getting these errors are unavailable or inconvenient. In that case we may resort to a couple of useful statistical tools that have become popular since the advent of fast computers. One is called the “jackknife” (because one should always have this tool handy) and the other the “bootstrap”. One of the earliest techniques to obtain reliable statistical estimators is the jackknife technique. Here we describe the jackknife method, which was invented in 1949 by Quenouille and developed further by Tukey in 1958. As the father of EDA, John Tukey attempted to use Jackknife to explore how a model is influenced by subsets of observations when outliers are present. The name Jackknife was coined by Tukey to imply that the method is an all-purpose statistical tool.

Quenouille (1949) introduced a technique for reducing the bias of a serial correlation estimator based on splitting the sample into two half-samples. In his 1967 paper he generalized this idea into splitting the sample into g group of size h each, $n = gh$, and explore its general applicability. It requires less computational power than more recent techniques such as bootstrap method.

Suppose we have a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and an estimator $\hat{\theta} = s(\mathbf{x})$. The jackknife focuses on the samples that *leaves out one observation at a time*:

$$\mathbf{x}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

for $i = 1, 2, \ldots, n$, called *jackknife samples.* The ith jackknife sample consists of the data set with the ith observation removed. Let $\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)})$ be the ith jackknife replication of $\hat{\theta}$.

The jackknife estimate of standard error defined by

$$\hat{se}_{jack} = \left[\frac{n-1}{n} \sum_i (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2 \right]^{1/2},$$

where $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$.

The jackknife only works well for linear statistics (e.g., mean). It fails to give accurate estimation for non-smooth (e.g., median) and nonlinear (e.g., correlation coefficient) cases. Thus improvements to this technique were developed. Now we consider Delete-d jackknife. Instead of leaving out one observation at a time, we leave out d observations. Therefore, the size of a delete-d jackknife sample is $n - d$, and there are $C(n, d)$ jackknife samples. Let $\hat{\theta}_{(s)}$ denote $\hat{\theta}$ applied to the data set with subset s removed. The formula for the delete-d jackknife estimate of s.e. is

$$\left[\frac{n-d}{dC(n,d)} \sum_i (\hat{\theta}_{(s)} - \hat{\theta}_{(\cdot)})^2 \right]^{1/2},$$

where $\hat{\theta}_{(\cdot)} = \sum_s \hat{\theta}_{(s)}/C(n, d)$ and the sum is over all subsets s of size $n - d$ chosen without replacement for $x_1, x_2, \ldots, x_n$. It can be shown that the delete-d jackknife is consistent for the median if $\sqrt{n}/d \to 0$ and $n - d \to \infty$. Roughly speaking, it is preferrable to choose a d such that $\sqrt{n} < d < n$ for the delete-d jackknife estimation of standard error.

We just describe how to obtain standard error estimate of an estimator $\hat{\theta}$ based on the sample of size n. Now we demonstrate that the jackknife can be used to reduce the biasd estimate $\hat{\theta}$. If the bias of $\hat{\theta}$ is of the order n^{-1}, we write

$$E(\hat{\theta}) - \theta = \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots .$$

Hence,

$$E_F(\hat{\theta}_{(\cdot)}) = \theta + \frac{a_1(F)}{n-1} + \frac{a_2(F)}{(n-1)^2} + \cdots .$$

To estimating the bias, the jackknife gives

$$B\hat{i}as_{jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}).$$

Then the jackknife estimate of θ is

$$\tilde{\theta} = \hat{\theta} - B\hat{i}as_{jack}.$$

Note that

$$\tilde{\theta} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}.$$

We can show easily that the bias of jackknife estimate $\tilde{\theta}$ is of the order n^{-2}.

7 Resampling Methods

The term "resampling" has been applied to a variety of techniques for statistical inference, among which stochastic permutation and the bootstrap are the most characteristic. There are at least four major types of resampling methods which include the randomization exact test, cross-validation, jackknife, and bootstrap. Although today they are unified under a common theme, it is important to note that these four techniques were developed by different people at different periods of time for different purposes. In this chapter, we already discuss two of them. In this section, we will describe the randomization exact test and cross-validation.

For the randomization exact test, it is also known as the permutation test. This test was developed by R.A. Fisher (1935), the founder of classical statistical testing. Both noted that with a large sample the *exact* Fisher test is not feasible because of the computational difficulty (before the age of powerful computers). Hence, in his later years Fisher lost interest in the permutation method because there were no computers in his days to automate such a laborious method. As a remedy, people suggested that a randomly-generated selected subset of the possible permutations could provide the benefits of the permutation test without excessive computational cost.

Randomization exact test is a test procedure in which data are randomly re-assigned so that an exact p-value is calculated based on the permutated data. Let's look at the following example. Assume that in an experiment comparing Web-based and text-based instructional methods, subjects obtained the following scores:

$$\{99, 90, 93, \ldots\} \quad \text{versus} \quad \{87, 89, 97, \ldots\}.$$

After the researcher has run a two-sample t-test, the test returns a t-value of 1.55. If the classical procedure is employed, the researcher can check this $t_{observed}$ against the $t_{critical}$ in the t-distribution to determine whether the group difference is significant. However, in resampling, instead of consulting a theoretical t-distribution, the researcher keeps swapping observations across the two groups, many more t-values will be returned. The purpose of this procedure is to artificially simulate "chance." Sometimes the t is large, but other times it is small. After exhausting every possibility, say 100, the inquirer can put these t-values together to plot an empirical distribution curve, which is built on the empirical sample data. When the t-value of 1.55 occurs only 5 times out of 100 times, the researcher can conclude that the exact p-value (the probability that this difference happens by chances alone) is 0.05.

The underlying idea was to use the power of sampling, in a fashion similar to the way it is used in empirical samples from large universes of data, in order to approximate the ideal test based on the complete set of permutations. So the idea was to gain the benefits of the classical array of methods - though not a parametric test in this case - by the technical device of simulation sampling. It raises a challenge question on whether the approximation would be satisfactory.

For cross-validation, it separates the available data into two or more segments, and then tests the model generated in one segment against the data in the other segment(s). The goal is to find out whether the result is replicable or just a matter of random fluctuations. Take regression as an example. In the process of implementing a cross-validation, the first sub-sample is usually used for deriving the regression equation while another sub-sample is used for generating predicted scores from the first regression equation. Next, the cross-validity coefficient is computed by correlating the predicted scores and the observed scores on the outcome variable. Refer to Stone (1974) on using cross-validation to assess statistical predictions. Clearly there is no re-use of the same data, nor is there any use of repeated simulation trials.

Bootstrap means that one available sample gives rise to many others by resampling (a concept reminiscent of pulling yourself up by your own bootstrap). While the original objective of cross-validation is to verify replicability of results and that of Jackknife is to detect outliers, Efron developed bootstrap with inferential purposes. The principles of cross-validation, Jackknife, and bootstrap are very similar, but bootstrap overshadows the others for it is a more thorough procedure in the sense that it draws many more sub-samples than the others. Through simulations, it is found that the bootstrap technique provides less biased and more consistent results than the Jackknife method does. Nevertheless, Jackknife is still useful in EDA for assessing how each sub-sample affects the model.

Supporters of resampling have raised a number of reasons to justify the aforementioned techniques.

- Empirical: Classical procedures rely on theoretical distributions, which requires strong assumptions of both the sample and the population. For example, we may need to start with the assumption that *the data conform to a bell-shaped curve and the need to focus on statistical measures whose theoretical properties can be analyzed mathematically.* But the inferential leap form the sample to the population may be problematic, especially when the population is ill-defined. If one is skeptical of the use of theoretical distributions, empirical-based resampling is a good alternative (Diaconis and Efron, 1983).

- Clarity: Conceptually speaking, resampling is clean and simple. Sophisticated mathematical background is not required to comprehend resampling. Thus, the researchers can pay attention to the content of their research rather than worrying about which test procedure could reduce the family-wise error from 0.06 to 0.05 in multiple comparison.

- Distribution: Classical procedures require distributional assumptions, which are usually met by a large sample size. When the sample size is small and does not conform to the parametric assumptions, resampling is recommended as a remedy (Diaconis and Efron, 1983). However, Good (2000) stated that permutation tests are still subject to the Behrens-Fisher problem, in which estimation is problematic when population variances are unknown. To be specific, permutation tests still assume equal variances as what is required in classical tests.

- Small sample size: Even if the data structure meets the parametric assumptions, a study with small sample size will be crippled by the low power level. Bootstrapping could treat a small sample as the virtual population to " “generate” more observations.

- Large sample size: Usually resampling is a remedy for small sample size, however, the same technique can also be applied to the situation of overpowering, in which there are too many subjects. Given a very large sample size, one can reject virtually any null hypothesis. When the researcher obtains a large sample size, he/she could divide the sample into subsets, and then apply a cross-validation.

- Replications: Classical procedures do not inform researchers of how likely the results are to be replicated. Repeated experiments in resampling such as cross-validation and bootstrap can be used as internal replications. Replications are essential to certain classical procedures such as multiple regression.

Despite these justifications, some methodologists are skeptical of resampling for the following reasons:

- Assumption: Stephen E. Fienberg mocked resampling by saying, “You’re trying to get something for nothing. You use the same numbers over and over again until you get an answer that you can’t get any other way. In order to do that, you have to assume something, and you may live to regret that hidden assumption later on.” Every theory and procedure is built on certain assumptions and requires a leap of faith to some degree. Indeed, the classical statistics framework requires more assumptions than resampling does.

- Generalization: Some critics argued that resampling is based on one sample and therefore the generalization cannot go beyond that particular sample. One critic even went further to say, “I do wonder, though, why one would call this (resampling) inference?”

- Bias and bad data: It asserted that confidence intervals obtained by simple bootstrapping are always biased though the bias decreases with sample size. If the sample comes from a normal population, the bias in the size of the confidence interval is at least $n/(n-1)$, where n is the sample size. Nevertheless, one can reduce the bias by more complex bootstrap procedures. Some critics challenged that when the collected data are biased, resampling would just repeat and magnify the same mistake. However, if one asserts that the data are biased, one must know the attributes of the underlying population. As a matter of fact, usually the population is infinite in size and unknown in distribution. Hence, it is difficult to judge whether the data are bad. Further, if the data were biased, classical procedures face the same problem as resampling. While replications in resampling could partially alleviate the problem, classical procedures do not provide any remedy.

- Accuracy: Some critics question the accuracy of resampling estimates. If the researcher doesn’t conduct enough experimental trials, resampling may be less accurate than conventional parametric methods. However, this doesn’t seem to be a convincing argument because today’s high-power computers are capable of running billions of simulations. For example, in StatXact, a software program for exact tests, the user could configure the resampling process to run with maximum RAM for 1 billion samples with no time limit.

References

[1] Babu, G.J. (1984). Bootstrapping statistics with linear combinations of chi-squares as weak limit. *Sankhya* Ser. A **46** 85-93.

[2] Berry, A.C. (1941). The acciracy of the Gausian approximation to the sum of independent variables. *Trans. Amer. Math. Soc.* **49** 122-136.

[3] Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196-1217.

[4] Chung, K.L. (1974). *A Course in Probability Theory.* 2nd ed., Academic Press, New York.

[5] Diaconis, P., and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, May, 116-130.

[6] Edgington, E.S. (1995). *Randomization tests.* New York: M. Dekker.

[7] Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *Ann. Statist.* **7** 1-26.

[8] Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 63, 589-599.

[9] Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* SIAM, Philadelphia.

[10] Efron, B. (1990). Six questions raised by the bootstrap. Technical Report No. 350, Department of Statistics, Stanford University.

[11] Efron, B. (1992). Jackknife after bootstrap Bootstrap methods: (with discussions). *J. R. Statist. Soc. B* **54** 1-127.

[12] Efron, B, and Tibshirani, R.J. (1993). *An introduction to the bootstrap.* Chapman and Hall, London.

[13] Fisher, R.A. (1956). *Statistical methods and scientific inference.* Edinburgh: Oliver and Boyd.

[14] Esseen, C.G. (1945). Fourier analysis of distribution functions. *Acta Math.* **77** 1-125.

[15] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13-30.

[16] Loh, W.Y. (1984). Estimating an endpoint of a distribution with resampling methods. *Ann. Statist.* **12** 1543-1550.

[17] Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Soc. Series B*, **11**, 18-84.

[18] Shao, J. (1994). Bootstrap sample size in nonregular cases. *Proc. of the Amer. Math. Soc.* **122**1251-1262.

[19] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B*, **36** 111-147.

[20] Tukey, J.W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, **29**, 614.