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.

Now we demonstrate how the bootstrap works in 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 example, 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.

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.

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.

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

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

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 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 a further remark, the empirical distribution F_n is called the nonparametric maximum likelihood estimate (MLE) of F.

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.)

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,

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

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*

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

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

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

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*

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

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*

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

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* $AN(A_n, B_n^2)$.

It is of both theoretical and practical interest to characterize the error of approximation in the CLT. For the i.i.d. case, an exact bound on the error of approximation is provided by the following theorem due to Berry (1941) and Esseen (1945).

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,*

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

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 0.7975. For x is sufficiently large, while n remains fixed, the quantities $P[(S_n - ES_n)/\sqrt{Var(S_n)} \leq x]$ and $\Phi(x)$ each 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

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

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

3.2 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.

3.3 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

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

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

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

Now we give a fairly brief discussion of some 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 $\lambda_n(\hat{F}_n) - \lambda_n(F)$ 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),

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

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

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

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

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

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 sympton 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 Thinking in terms of the Asymptotics

Efron (1990) tried to give an explanation on *Why do maximum likelihood estimated distributions tend to be short-tailed*? To answer that question, he tried to use *asymptotic analysis* to address that question. In this section, we will present his argument on demonstrating the usefulness and pitfall of asymptotic analysis.

Again $X_1, X_2, \cdots, X_n$ are i.i.d. samples from an unknown distribution F. Quite often, we want to form a confidence interval for the expectation $\mu = E_F(X)$, say of the form

$$\bar{x} \pm z^{(\alpha)} \hat{se}, \quad (13)$$

where $z^{(\alpha)}$ is the 100αth normal percentile, e.g. $z^{(.95)} = 1.645$, and $\hat{se} = \hat{\sigma}/\sqrt{n}$, the estimated standard error of $\bar{x}$.

The standard interval (13) is based on taking literally the asymptotic normal approximation

$$\frac{\bar{x} - \mu}{\hat{se}} \xrightarrow{d} N(0,1)$$

with the estimated standard error $\hat{se}$ considered to be a fixed constant. In certain cases it is well known that both convergence to normality and constant of σ can be dramatically improved by considering instead of $\bar{x}$ and μ a monotone transformation $g(\bar{x})$ and $g(\mu)$. The classical example is that of the correlation coefficient from a bivariate normal sample, for which Fisher's inverse hyperbolic tangent approximation works beautifully.

Recall that the empirical distribution $\hat{F}_n$ is the nonparametric maximum likelihood estimate of F. The empirical distribution assigns probability

$$Prob_{\hat{F}}(A) = \frac{\#\{X_i \in A\}}{n} \tag{14}$$

to any set A in the sample space of the X's. This is an unbiased estimate of the true probability $Prob_F(A)$,

$$E\left[Prob_{\hat{F}_n}(A)\right] = Prob_F(A). \tag{15}$$

However the same unbiasedness does not apply to the variance functional: $Var_{\hat{F}}\{X\} = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$ has expectation

$$E\left[Var_{\hat{F}_n}(X)\right] = \frac{n-1}{n}Var(X). \tag{16}$$

We see that the variance function is underestimated by maximum likelihood, albeit mildly so. Elementary statistics courses recommen estimating the variance by

$$\frac{n}{n-1}Var_{\hat{F}_n}(X) = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1), \tag{17}$$

rather than by $Var_{\hat{F}_n}(X)$ itself. This removes the bias, but does not explain it.

The underestimating effect seen in (16) is quite small from the point of view of asymptotic theory. The choice of the MLE versus the unbiased estimate of σ^2 makes a difference of magnitude only $O(\hat{se}/n)$ in (13). This is a *third order* correction in the usual parlance. By comparison, ignoring the skewness of F in forming interval (13) makes a difference of $O(\hat{se}/\sqrt{n})$, a **second order** error; and using $\bar{x}$ instead of an efficient estimate $\hat{\mu}$, makes a

difference of $O(\hat{se})$, a **first order** error. Third order notwithstanding, the underestimation effect in (16) can cause substantial problems when n is small.

Under regularity conditions, it can be shown that the exact confidence limits are typically of the form

$$\bar{X}_n \pm \hat{se}\left(z^{(\alpha)} + \frac{A_n^{(\alpha)}}{\sqrt{n}} + \frac{B_n^{(\alpha)}}{n} + \cdots\right) \tag{18}$$

where n is the sample size. The standard interval (13) is **first-order correct** in the sense that the term $\bar{X}_n \pm \hat{se}A_n^{(\alpha)}/\sqrt{n}$ asymptotically dominates (18). However, the second-order term $\hat{se}z^{(\alpha)}$ can have a major effect in small-sample situations.

To make it precise, we consider the following problem. Let $X_1, \ldots, X_n$ be independent, with a common $N(\mu, \sigma^2)$ distribution. Let s^2 be the sample variance. Consider the pivot s^2/σ^2. This is distributed, of course, as $\chi^2_{n-1}/(n-1)$. Consider the bootstrap approximation, namely the distribution of s^{*2}/s^2, where s^{*2} is the variance of the resampled data. For $n = 20$, the bootstrap approximation is not good.

One source of this problem is in the tails of the normal: about 30% of the variance is contributed by 5% of the distribution and will be missed bytypical samples of size 20. In other words, s^2 is mean unbiased for σ^2, but quite skewed for moderate n: in most samples, s^2 is somewhat too small, counterbalanced by the few samples where s^2 is huge. The variance of s^{*2}/s^2 is largely controlled by the sample fourth moment, which is even more skewed. For large n, these problems go away: after all, s^2 and the bootstrap are consistent.

6 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

$$\begin{aligned} E\{\theta(F_n)|F_0\} &= E\{\mu + n^{-1}\sum_{i=1}^{n}(X_i - \mu)\}^3 \\ &= \mu^3 + n^{-1}3\mu\sigma^2 + n^{-2}\gamma, \end{aligned} \tag{19}$$

where $\gamma = E(X_1 - \mu)^3$ denotes population skewness. Using the nonparametric bootstrap, we obtain in direct analogy to (19)

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

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.

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

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

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

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

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

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

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

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