Summer 2024

# Accounting For Variability Due To Resampling Using Bootstrapping

Dipendra Phuyal

ACCOUNTING FOR VARIABILITY DUE TO RESAMPLING USING

BOOTSTRAPPING

by

DIPENDRA PHUYAL

(Under the Direction of Charles W .Champ)

ABSTRACT

Bradley Efron (1979) introduced bootrapping. Typically a researcher is interested in studying a process which generates individuals. The collection of individuals the process has (actual) or could have (conceptual) generated is the population. The collection of conceptual members of the population is an uncountable collection. Hence, the population is an uncountable collection of individuals. The collection of individuals the process has generated (actual individuals) is representative of what the process can generate and will be referred to as the representative sample. The size of this sample is a nonnegative integer valued random variable $N$ which may be a constant random variable such as in statistically designed experiments in which the researcher decides the value of $N$ before collecting the data. In general, the variable $N$ is a random variable $N$ whose value is generated by the process. We note that in the design of experiments the researcher becomes a part of the process that generates "actual" individuals in the population. There is variability that must be accounted for among the measurement on the individuals in the representative sample. From the representative sample, the researcher using a sampling method will select a sample referred to as the researcher's sample. There is variability of among the measurements that must be accounted for in the researcher's sample. A bootstrap sample introduces further variability that must be accounted for. We will study bootstrapping in which we account for the variability in the bootstrap sample.

INDEX WORDS: bootstrap sample mean, bootstrap sample variance, independent samples, random sample, variability

ACCOUNTING FOR VARIABILITY DUE TO RESAMPLING USING

BOOTSTRAPPING

by

DIPENDRA PHUYAL

B.A., Tehrathum Multiple Campus, 2010

M.A., Tribhuvan University, Kathmandu Nepal, 2014

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial

Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

# ACCOUNTING FOR VARIABILITY DUE TO RESAMPLING USING BOOTSTRAPPING

by

DIPENDRA PHUYAL

Major Professor: Charles W .Champ
Committee: Andrew V. Sills
Divine Wanduku

Electronic Version Approved:
July 2024

# DEDICATION

To my beloved mother, Lila Devi Phuyal. Your unwavering love, strength, and sacrifices have been the foundation of all my achievements. After the untimely passing of Dad, you carried the weight of our family on your shoulders, embodying resilience and determination. You have been my greatest inspiration, teaching me the values of hard work, perseverance, and compassion

Every step of this journey, every late night and early morning, has been fueled by the desire to make you proud. This accomplishment is not mine alone; it is a testament to your endless support, encouragement, and belief in me.

ACKNOWLEDGMENTS

I would like to express my special thanks to my supervisor, Dr. Charles W. Champ, for his consistent support throughout this entire project. Your expertise in the field and valuable suggestions have shaped my research journey in a positive way. I will be forever grateful for your guidance, which helped me complete my academic journey here at Georgia Southern University.

I am extremely grateful to my family for their unconditional support through the highs and lows of my academic journey. Also, I would like to thank all my friends for sharing their experiences with me, which helped me stay calm during this challenging time.

Finally, I would like to extend my heartfelt gratitude to my thesis committee members, Dr. Divine Wanduku and Dr. Andrew V. Sills, for their invaluable suggestions.

# TABLE OF CONTENTS

# LIST OF TABLES

Table Page

# CHAPTER 1

# INTRODUCTION

## 1.1 INTRODUCTION

Typically a researcher is interested in studying a process which generates individuals. The collection of individuals the process has or could have generated is the population. The collection of conceptual members of the population is an uncountable collection. Hence, the population is an uncountable collection of individuals. The collection of individuals the process has generated (actual individuals) is representative of what the process can generate and will be referred to as the representative sample. The size of this sample is a non-negative integer valued random variable $\boldsymbol{N}$ which may be a constant random variable such as in statistically designed experiments in which the researcher decides the value of $\boldsymbol{N}$ before collecting the data. In general, the variable $\boldsymbol{N}$ is a random variable $\boldsymbol{N}$ whose value is generated by the process. We note that in the design of experiments the researcher becomes a part of the process that generates ”actual” individuals in the population.

On each individual, let's assume we are interested in taking the measurement $X$ (as opposed to some vector of measurements). We assume here that $X$ is either a discrete or continuous measurement (random variable). If $X$ is discrete (continuous), the distribution of $X$ can be described by the probability mass (density) function $P(X = x)(f_X(x))$. The mean $\mu_X$ and variance $\sigma_X^2$ of the distribution of $X$ are defined by

$$\mu_X = \begin{cases} \sum_x xP(X=x), & \text{if X is discrete;} \\ \int_{-\infty}^{\infty} xf_X(x)dx, & \text{if X is continuous;} \end{cases}$$
$$\sigma_X^2 = \begin{cases} \sum_x (x-\mu_x)^2 P(X=x), & \text{if X is discrete;} \\ \int_{-\infty}^{\infty} (x-\mu_x)^2 f_X(x)dx, & \text{if X is continuous.} \end{cases} \tag{1.1}$$

Other parameters of the distribution of $X$ that are of interest to researchers are the quantiles of the distribution. The $100\gamma\, th$ quantile $x_{1-\gamma}$ of the distribution satisfies the inequalities

$$P(X \leq x_{1-\gamma}) \geq \gamma \quad and \quad P(X < x_{1-\gamma}) \leq \gamma$$

In order to gain information about the distribution of $X$, assume that the researcher is allowed to select n individuals from the N individuals in the representative sample with $1 \leq n \leq N$. Note that it is possible for n=N. We refer to this case as a census. As we have noted, the researcher may be part of the process that generates individuals which occurs in the statistical design of experiments. The value of n may be fixed by the researcher, but in some applications the size n of the researcher's sample is a random variable. However, in our discussion here, we will assume that n has a known value. For convenience of discussion, we represent the measurements on the N individuals in the representative sample by $X_1, ..., X_n, X_n + 1, ..., X_N$ with the first n of these denoting the $X$ measurements on the individuals in the researcher's sample. Assume that the process is generating individuals such that one may assume that the measurements $X_1, ..., X_n, X_n + 1, ..., X_N$ are independent and identically distributed with common distribution described by $P(X = x)\,(f_X(x))$. (Note that if X is a continuous measurement, we are not considering any process in which the probability density function $f_X(x)$ does not exist.

One commonly recommended method for the researcher generating the researcher's sample is a method known as a simple random sampling method. This sampling method gives each of the $\binom{N}{n}$ possible researcher's samples of size n with the equal chance of being selected. The resulting sample is called a simple random sample (SRS). If the researcher can argue that $X$ measurements in the representative sample are stochastically independent and have the same distribution, then it follows that the $X$ measurements to be obtained on the individuals in a simple random sample would be stochastically independent and have the same distribution. This will be stated more concisely by stating that the sample is a random sample. In what follows, we require that the $X$ measurements to be taken on the n individuals in the researcher's sample are stochastically independent and have the same distribution. That is, the collection of the N random variables $X_1, ...X_n, X_n + 1, ...X_N$ is a random sample and the collection of the n random variables $X_1, ...X_n$ is also a random sample.

The following schematic is used to summarize the previous paragraph to this point in our discussion.

$$\underset{X \sim P(X=x) \text{ or } (f_X(x)}{\text{Population}} \xrightarrow{\text{Process Generates from the Population}} \underset{\text{Measurements } X_1,...X_n,X_n+1,...X_N}{\text{Representative Sample}} \subset \text{Population}$$

$$\xrightarrow{ResearcherGenerates from the Representative Sample} \text{Researcher's Sample} \subseteq \underset{\text{Measurements } X_1,...X_n}{\text{Representative Sample}}$$

Note that,

Researcher's Sample $\subseteq$ Representative Sample $\subset$ Population.

The probability mass (density ) function $P(X = x)(f_X(x))$ is typically not known. It is usually modelled by a probability mass (density) function $P^{Model}(X = x) \quad (f_X^{Model}(x))$. One of the most commonly used family of models for the distribution of a continuous measurements $X$ is the family of Normal distributions. A Normal distribution is described by a probability density function of the form

$$f_X^{Model}(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$$

Where $-\infty < \infty$ and $\sigma > 0$ are distributional parameters. One can show that

$$\begin{aligned} &(1) \quad f_X^{\text{Model}}(x) > 0 \\ &(2) \quad \int_{-\infty}^{\infty} f_X^{\text{Model}}(x)\, dx = 1 \end{aligned}$$

and when used as a model it describes(models) the distribution of $X$ It can be shown the mean and variance of the distribution of $X$ under this model are respectively,

$$\mu_X = \mu \quad \text{and} \quad \sigma_X^2 = \sigma^2$$

The following statistics are of interest. The representative samples mean and variance are

$$\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i \quad and \quad S_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X}_N)^2$$

The sample mean and sample variance of the researcher's sample are

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad and \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Also, we will be interested in the following statistics.

$$\overline{X}_{N-n} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - n\overline{X}) \quad and \quad S_{N-n}^2 = \frac{1}{N-n-1}\sum_{i=n+1}^{N}(X_i - \overline{X}_{N-n})^2$$

The statistics the researcher will be allowed to observe are $\overline{X}_n$ ,$S_n^2$. To make inferences about such population values as $\mu_X$ and $\sigma_X^2$ using the numerical information $X_1, ...X_n$ in the researcher's sample, are estimation methods. One can also use this same numerical information to "predict" the values of statistics such as $\overline{X}_N$ and $S_N^2$ describing the representative sample. For example,the observed value of $\overline{X}_n$ is often used as a point estimator of the parameter $\mu_X$ and the observed value of the random interval

$$\left(\overline{X}_n - t_{n-1}\frac{\alpha}{2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n-1}\frac{\alpha}{2}\frac{S_n}{\sqrt{n}}\right)$$

as an interval estimate (confidence interval) of $\mu_X$, where $t_{n-1}, \frac{\alpha}{2}$ is the $100(1-\frac{\alpha}{2})th$ percentile of a t-distribution with n-1 degree of freedom. The margin of error $(MOE_{CI})$ and $(L_{CI})$ of this random interval are, respectively,

$$MOE_{CI} = t_{n-1}\frac{\alpha}{2}\frac{S_n}{\sqrt{n}} \quad L_{CI} = 2t_{n-1}\frac{\alpha}{2}\frac{S_n}{\sqrt{n}}$$

Note that both $(MOE_{CI})$ and $(L_{CI})$ are random variables.

The statistics $\overline{X}_n$ can also be used as a point predictor for statistics $\bar{X_N}$. One can show that $E(\overline{X}_n|\overline{X}_N, N) = \overline{X}_N$. A $100(1-\alpha)th$ prediction interval for the statistics $\overline{X}_N$ conditioned on size N of representative sample is the observed value of the random interval

$$\left(\overline{X}_n - t_{n-1}\frac{\alpha}{2}\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n-1}\frac{\alpha}{2}\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}}\right)$$

If N is not given this interval cannot be observed. However, since the factor $\sqrt{\frac{N-n}{N}} < 1$ then this interval is contained in the interval

$$\left(\overline{X}_n - t_{n-1}\frac{\alpha}{2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n-1}\frac{\alpha}{2}\frac{S_n}{\sqrt{n}}\right)$$

So for large value of N, removing the factor $\sqrt{\frac{N-n}{N}}$ gives the prediction interval for $\overline{X}_N$ in which one can be more confident than $100(1-\alpha)\%$ that $\overline{X}_N$ is in the interval. The margin of error $(MOE_{PI})$ and length $(L_{PI})$ of this interval are, respectively,

$$MOE_{CI} = t_{n-1}\frac{\alpha}{2}\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}} \leq t_{n-1}\frac{\alpha}{2}\frac{S_n}{\sqrt{n}} \quad L_{CI} = 2t_{n-1}\frac{\alpha}{2}\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}} \leq 2t_{n-1}\frac{\alpha}{2}\frac{S_n}{\sqrt{n}}$$

We see that $L_{CI} - L_{PI} = 2t_{n-1}\frac{\alpha}{2}(1 - \sqrt{\frac{N-n}{N}})\frac{S_n}{\sqrt{n}} > 0$

But as n approaches N, $L_{PI}$ approaches 0. Hence , in this case, the confidence interval is wider. It is also interesting to note that a prediction interval for the statistics $\overline{X}_{N-n}$ is

$$\left(\overline{X}_n - t_{n-1}\frac{\alpha}{2}\sqrt{\frac{N}{N-N}}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n-1}\frac{\alpha}{2}\sqrt{\frac{N}{N-n}}\frac{S_n}{\sqrt{n}}\right)$$

Where $\overline{X}_n$ and $S_n$ are observed values of researcher's sample mean and standard deviation, respectively. One can see the confidence interval for $\mu_X$ is wider than the prediction interval $\overline{X}_N$

## 1.2 Conclusion

We have discussed a meaning for population, representative sample, and the researcher's sample. In what follows, we examine the concept of bootstrapping and its use in point and interval estimation and prediction methods as well as test of significance and hypothesis. It will be shown that 'Bootstrapping' adds extra unneeded variability to these methods.

# CHAPTER 2

# BOOTSTRAPPING

## 2.1 Introduction

The bootstrap method has been discussed by several authors which include among others Efron (1979), Efron (1981), Efron (1982), Schenker (1985), Wu (1986), Efron and Tibshirani (1993), Mooney and Duval (1993), Shao and Tu (1995), Davison and Hinkely (1997), Moore, McCabe, Duckworth, and Sclove (2003), Hesterberg, Moore, Monaghan, Clipson, and Epstein (2003), and Good (2005). We examine the concept of bootstrapping and its use in point and interval estimation and prediction methods. It will be shown that "bootstrapping" adds extra, unneeded variability to these methods. Schenker (1985) states "The percentile method and bias-corrected method of Efron (1981, 1982) are discussed. When these methods are used to construct nonparameteric confidence intervals for the variance of a normal distribution, the coverage probabilities are substantially below the nominal level for small to moderate samples. This is due to the inapplicability of assumptions underlying the methods. These assumptions are difficult or impossible to check in complicated situations for which the bootstrap is intended. Therefore, bootstrap confidence intervals should be used with caution in complex problems." Efron and Tibshirani (1993) states "Dissussions of some of the issues concerning bootstrap confidence intervals appear in Schenker (1985), Robinson (1986, 1987), Peters (and Freedman (1987), Hinkley (1988), and in the psychology literature, Lunneborg (1985), Rasmussen (1987), and Efron (1988)."

After observing the $n$ measurements $X_1, \ldots, X_n$ on the $n$ individuals in the sample, the researcher decides to select $m$ groups of $n$ numbers as follows. For the $i$th sample, an individual is randomly selected from the $n$ individuals in the researher's sample and the $X$-measurement $X^*$ on the individual is recorded. Again from the $n$ individuals in the researcher's sample (the first individual selected has been placed back into the group), an individual is randomly selected and the individual's $X$-measurement $X^*$ is recorded, and so forth. The $n$ $X$-measurements are $X_1^*, \ldots, X_n^*$, to be more compactly denoted by $\mathbf{X}^* = [X_1^*, \ldots, X_n^*]^{\mathbf{T}}$. This is a bootstrap sample. We summarize as follows.

Population

$X \sim P(X = x)$ or $f_X(x)$

$\xrightarrow{\text{Process Generates from the Population}}$

Representative Sample $\subset$ Population

Measurements: $X_1, X_2, \ldots, X_N$

$\xrightarrow{\text{Researcher Generates from the Representative Sample}}$

Researcher's Sample $\subseteq$ Representative Sample

Measurements: $X_1, \ldots, X_n$

$\xrightarrow{\text{Researcher Generates from the Researcher's sample}}$

Bootstrap Sample $\subseteq$ Researcher's Sample

Measurements: $X_1^*, \ldots, X_n^*$

Note that

Bootstrap Sample$(n)$ $\subseteq$ Researcher's Sample$(n)$ $\subseteq$ Representative Sample$(N)$ $\subset$ Population$(\infty)$

It is interesting to note that there are only a finite number of different possible bootstrap samples of size n that can be generated from the n measurements on a sample of individuals. There are two ways to count the bootstrap samples. For example for $n = 3$, we may count the two samples as

$$\{X^*_{i,1} = x_1, X^*_{i,2} = x_1, X^*_{i,3} = x_2\} \quad \text{and} \quad \{X^*_{i,1} = x_1, X^*_{i,2} = x_2, X^*_{i,3} = x_1\}$$

as distinct samples. If this is the case then there are $n^n$ possible bootstrap samples for a sample of size n. On the other hand, we could view both of these samples as the same sample,$\{x_1, x_1, x_2\}$. In this case, there is 1 bootstrap sample for $n = 1$ and there are

$$m = \sum_{k=0}^{n} k \binom{2n-k-2}{n-k} = \binom{2n-1}{n}$$

bootstrap samples for $n > 1$. This can be found in literature. It is not difficult see that $m \leq n^n$ for all positive integers n. Some example of *m* are given in the following table.

Table 2.1: Values of *m*

| n | $m = \binom{2n-1}{n}$ |
|---|---|
| 4 | 35 |
| 5 | 126 |
| 10 | 92,378 |
| 15 | 77,558,760 |
| 20 | 68,923,264,410 |
| 25 | 63,205,303,218,876 |
| 30 | 59,132,290,782,430,712 |
| 35 | 56,093,138,908,331,422,716 |
| 40 | 53,753,604,366,668,088,230,810 |
| 50 | 50,445,672,272,782,096,667,406,248,628 |

For example if $n = 20$, then there are 68,923,264,410 possible bootstrap samples.

It is also worth noting that it is possible to obtain a sample in which all the elements are the same value $x_i$, where $i = 1, 2, ...n$. Thus , there are

$$m - n = \binom{2n-1}{n} - n$$

bootstrap samples that have at least two distinct values from the values $x_1, ...x_n$. Also note, the researcher's sample is one of the $\mathbf{m}$ bootstrap samples.

A bootstrap sample containing $f_1$ value(s) of $x_1$, $f_2$ value(s) of $x_2$,..., $f_n$ value(s) of $x_n$, has probability

$$\frac{\frac{n!}{f_1!\,f_2!\,...f_n!}}{n^n}$$

of occurring. For example, for the case in which $n = 3$, we have

$$P(x_1, x_1, x_3) = \frac{\frac{3!}{2!\,1!\,0!}}{3^3} = \frac{1}{9}$$

Whereas,

$$P(x_1, x_2, x_3) = \frac{\frac{3!}{1!\,1!\,1!}}{3^3} = \frac{2}{9}$$

Given that $X_1 = x_1, ..., X_n = x_n$, the measurements $X^*_{i,1}, ..., X^*_{i,n}$ to be selected from the n measurements in the sample is a set of n independent discrete random variables. Furthermore, the measurements to be taken in each of the m sample are independent with common distribution expressed in tabular form by

$$\begin{bmatrix} X^* : & X_1 & X_2 \ldots & X_n \\ P(X^*|X_1, ..., X_n) : & \frac{1}{n} & \frac{1}{n} \ldots & \frac{1}{n} \end{bmatrix}$$

and in functional form by

$$P(X^*|X_1, \ldots, X_n) = \frac{1}{n} I_{\{X_1, ..., X_n\}}(X^*)$$

, Where $I_A(w) = 1$ if $w \in A$ and zero otherwise. It is important to note that $X^*$ is a discrete random variable whose conditional distribution given $X_1 = X_1, \ldots, X_n = X_n$ describes the variability introduced by researcher's bootstrap sampling method. The conditional mean and variance of this distribution are

$$\mu_{X^*}|X_1, \ldots, X_n = \sum_{i=1}^{n} X_i \times \frac{1}{n} = \bar{X}_n$$

$$\sigma^2_{X^*}|X_1, \ldots, X_n = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \times \frac{1}{n} = \frac{n-1}{n} S_n^2 = (1 - \frac{1}{n}) S_n^2$$

Before the measurements $X_1, \ldots, X_n$ on the individuals in the researcher's sample are made the values

$$\mu_{X^*|X_1,\ldots,X_n} = \overline{X}_n \text{ and } \sigma^2_{X^*|X_1,\ldots,X_n} = \left(1 - \frac{1}{n}\right) S_n^2$$

are random variables.

## 2.2 Bootstrap Samples

The bootstrap sampling method used in our example results in the random variables $X_1^*, \ldots, X_{n_B}^*$ being conditionally stochastically independent with identical distribution that of the distribution of $X^*$ given $X_1, \ldots, X_n$. We let $\overline{X}^*$ and $S^{2*}$ represent the mean and variance of the bootstrap sample of size $n_B$. Define

$$\overline{X}^*_{n_B} = \frac{1}{n_B} \sum\nolimits_{i=1}^{n_B} \left(X_i^* \,|X_1, \ldots, X_n\right) \text{ and } S^{2*}_{n_B} = \frac{1}{n_B - 1} \sum\nolimits_{i=1}^{n_B} \left[ \left(X_i^* - \overline{X}^*_{n_B}\right)^2 |X_1, \ldots, X_n \right].$$

One can show that

$$E\left(\overline{X}^*_{n_B} \,|X_1, \ldots, X_n\right) = \overline{X}_n \text{ and } E\left(S^{2*}_{n_B} \,|X_1, \ldots, X_n\right) = \left(1 - \frac{1}{n}\right) S_n^2.$$

Further, we have

$$V\left(\overline{X}^*_{n_B} \,|X_1, \ldots, X_n\right) = \frac{V\left(X^* \,|X_1, \ldots, X_n\right)}{n_B} = \frac{n-1}{n} \frac{S_n^2}{n_B}.$$

Using the law of total expection, we have the unconditional expectation of $\overline{X}^*_{n_B}$

$$E\left(\overline{X}^*_{n_B}\right) = E\left[E\left(\overline{X}^*_{n_B}|X_1,\ldots,X_n\right)\right] = E\left(\overline{X}_n\right) = \mu.$$

and using the total law of variance the unconditional variance of $\overline{X}^*_{n_B}$

$$\begin{aligned}
V\left(\overline{X}^*_{n_B}\right) &= E\left[V\left(\overline{X}^*_{n_B}|X_1,\ldots,X_n\right)\right] + V\left[E\left(\overline{X}^*_{n_B}|X_1,\ldots,X_n\right)\right] \\
&= \frac{n-1}{n\left(n_B\right)} E\left(S_n^2\right) + V\left(\overline{X}_n\right) = \frac{n-1}{n_B}\frac{\sigma^2}{n} + \frac{\sigma^2}{n} \\
&= \left(\frac{n-1}{n_B} + \frac{n_B}{n_B}\right)\frac{\sigma^2}{n} = \frac{n_B+n-1}{n_B}\frac{\sigma^2}{n}.
\end{aligned}$$

## 2.3 Duplication When Bootstrapping

For $n_B > n$, we would expect duplicate bootstrap values for $X^*$. Let $X_{1:n},\ldots,X_{n:n}$ be the order statistics of the random sample $X_1,\ldots,X_n$. Here $n_B$ is the bootstrap sample size. Let $f_{i:n}$ represent the frequency that $X_{i:n}$ appears in the sample. Further, let $a$ be a positive integer for a given $0 < \alpha \leq 0.5$ such that

$$\frac{\sum_{i=1}^{a} f_{i:n}}{n_B} \geq \alpha/2.$$

This would suggest that $X_{a:n}$ is the estimated $100\left(\alpha/2\right)$th percentile using bootstrapping. Now let $b$ be a positive integer such that

$$\frac{\sum_{i=b}^{n} f_{i:n}}{n_B} \geq 1-\alpha/2.$$

This would suggest that $X_{b:n}$ is the estimated $100\left(1-\alpha/2\right)$th percentile using bootstrapping.

## 2.4 CONCLUSION

Bootstrapping was introduced. We have discussed the distribution of a bootstrap random variable $X^*$ and its distribution. We have shown that the mean and variance of the distribution of $X^*$ are, respectively, the sample mean $\overline{X}_n$ of the researcher's sample and $(n-1)\,S_n^2/n$ with $S_n^2$ the variance of the researcher's sample.

# CHAPTER 3

# DISTRIBUTION OF THE BOOTSTRAP SAMPLE MEAN $\overline{X}^*$

## 3.1 INTRODUCTION

A random sample of size $n_B$ with $X^*$ measurement $X_1^*, \ldots, X_{n_B}^*$ is to be taken from the distribution of $X^*$ given by

$$\begin{bmatrix} X^*: & X_1 & X_2 & \ldots & X_n \\ P\left(X^* \left| X_1, \ldots, X_n\right.\right): & \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{bmatrix}.$$

The bootstrap samples mean and variance of the distribution of $X_1^*, \ldots, X_{n_B}^*$ are, respectively, of this sample are

$$\overline{X}^* = \frac{1}{n_B} \sum\nolimits_{i=1}^{n_B} X_i^* \text{ and } S^{2*} = \frac{1}{n_B - 1} \sum\nolimits_{i=1}^{n_B} \left(X_i^* - \overline{X}^*\right)^2$$

Since the bootstrap sample is a random sample, we have

$$E\left(\overline{X}^* \left| X_1, \ldots, X_n\right.\right) = \overline{X}_n \text{ and } V\left(\overline{X}^* \left| X_1, \ldots, X_n\right.\right) = \frac{V\left(X^*\right)}{n},$$

where $V\left(X^*\right) = (n-1) S^2/n$. It follows that the unconditional mean and variance are

$$E\left[E\left(\overline{X}^* \left| X_1, \ldots, X_n\right.\right)\right] = E\left(\overline{X}\right) = \mu$$

using the law of total expectations and

$$\begin{aligned} V\left(\overline{X}^*\right) &= E\left[V\left(\overline{X}^* \left| X_1, \ldots, X_n\right.\right)\right] + V\left[E\left(\overline{X}^* \left| X_1, \ldots, X_n\right.\right)\right] \\ &= E\left[\frac{(n-1) S^2}{n^2}\right] + V\left(\overline{X}\right) = \left(\frac{1}{n} - \frac{1}{n^2}\right) E\left(S^2\right) + V\left(\overline{X}\right) \\ &= \left(\frac{\sigma^2}{n} - \frac{\sigma^2}{n^2}\right) + \frac{\sigma^2}{n} = \left(2 - \frac{1}{n}\right) \frac{\sigma^2}{n} \\ &= \left(2 - \frac{1}{n}\right) V\left(\overline{X}\right) \end{aligned}$$

using the law total of variance. Thus, as the sample size $n$ increases, the variance $V\left(\overline{X}^*\right)$ of the bootstrap sample mean becomes almost twice the variance of the sample mean with

$$\lim_{n\to\infty} V\left(\overline{X}^*\right) = \lim_{n\to\infty}\left(2-\frac{1}{n}\right) V\left(\overline{X}\right) = 2V\left(\overline{X}\right).$$

### 3.2 Standard Error of the Sample Mean

For a sample $X_1,\ldots,X_n$ of measurements, the standard error of the sample mean $\overline{X}_n$ is $se_{\overline{X}} = \sigma/\sqrt{n}$. An estimator for $se_{\overline{X}}$ is $\widehat{se}_{\overline{X}} = S_n/\sqrt{n}$, where $S_n$ is the standard deviation of the sample. On page 42 of Efron and Tibsharani (1993), the authors suggest estimating $\sigma$ by

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2}.$$

Efron and Tibshirani (1993) also suggest using

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum\nolimits_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2}.$$

For the $i$th bootstrap sample $\mathbf{X}_i^*$, an estimator for $\sigma$ is

$$\widehat{\sigma}_i^* = \sqrt{\frac{1}{n}\sum\nolimits_{j=1}^{n}\left(X_{i,j}^*-\overline{X}_i^*\right)^2} \text{ or } \widehat{\sigma}_i^* = \sqrt{\frac{1}{n-1}\sum\nolimits_{j=1}^{n}\left(X_{i,j}^*-\overline{X}_i^*\right)^2}$$

where

$$\overline{X}^* = \frac{1}{n}\sum\nolimits_{j=1}^{n} X_{i,j}^*.$$

The statistic $\widehat{\sigma}$ under the independent Normal model is the maximum likelihood estimator for $\sigma$. It follows using $\widehat{\sigma}$ as an estimator of $\sigma$ that

$$\widehat{se}_{\overline{X}} = \sqrt{\frac{1}{n^2}\sum\nolimits_{i=1}^{n}\left(X_j-\overline{X}_n\right)^2} \text{ or } \widehat{se}_{\overline{X}} = \sqrt{\frac{1}{n(n-1)}\sum\nolimits_{i=1}^{n}\left(X_j-\overline{X}_n\right)^2}.$$

The $i$th bootstrap estimator $\widehat{se}_{i,\overline{X}}^*$ of $se_{\overline{X}}$ is as follows. We have

$$\widehat{se}_{i,\overline{X}}^* = \sqrt{\frac{1}{n^2}\sum\nolimits_{j=1}^{n}\left(X_{i,j}^*-\overline{X}_i^*\right)^2} \text{ or } \widehat{se}_{i,\overline{X}}^* = \sqrt{\frac{1}{n(n-1)}\sum\nolimits_{j=1}^{n}\left(X_{i,j}^*-\overline{X}_i^*\right)^2},$$

where

$$\overline{X}_i^* = \frac{1}{n} \sum\nolimits_{j=1}^{n} X_{i,j}^*.$$

A bootstrap estimator for the $se_{\overline{X}}$ is based on $n_B$ bootstrap samples $\mathbf{X}_1^*, \ldots, \mathbf{X}_{n_B}^*$. For the $i$th bootstrap sample $\mathbf{X}_i^*$, we calculate $\overline{X}_i^*$. The bootstrap estimator of the standard error of the sample mean according to Efron and Tibsharani (1993) on page 47 is

$$\widehat{se}_{\overline{X}}^* = \sqrt{\frac{1}{n_B - 1} \sum\nolimits_{i=1}^{n_B} \left(\overline{X}_i^* - \overline{\overline{X}}^*\right)^2},$$

where

$$\overline{\overline{X}}^* = \frac{1}{n_B} \sum\nolimits_{i=1}^{n_B} \overline{X}_i^*.$$

## 3.3 Possible Bootstrap Samples

For a given sample of size $n$ as previously shown, there are

$$m = \binom{2n-1}{n}$$

possible bootstrap samples. The $i$th bootstrap sample mean $\overline{X}_i^*$ can now be expressed as

$$\overline{X}_i^* = \frac{1}{n} \sum\nolimits_{j=1}^{n} X_{i,j}^* = \frac{1}{n} \sum\nolimits_{j=1}^{n} L_{i,j}^* X_j,$$

where $X_{i,j}^*$ is the $j$th $X$ value selected for the $i$th bootstrap sample and $L_{i,j}^*$ is the number of times $X_j$ appears in the $i$th sample with

$$\sum\nolimits_{j=1}^{n} L_{i,j}^* = n.$$

The matrix $\mathbf{L}^*$ has $i$th row $L^*_{i,1}, \ldots, L^*_{i,n}$, where $i = 1, \ldots, \binom{2n-1}{n}$. Define the $\binom{2n-1}{n} \times n$ matrix $\mathbf{L}^*$ by

$$\mathbf{L}^* = \begin{bmatrix} L^*_{1,1} & L^*_{1,2} & \ldots & L^*_{1,n} \\ L^*_{2,1} & L^*_{2,2} & \ldots & L^*_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ L^*_{m,1} & L^*_{m,2} & \ldots & L^*_{m,n} \end{bmatrix}.$$

The rows of $\mathbf{L}^*$ are the possible bootstrap samples of size $n$. For example, if $n = 3$, we have

$$\mathbf{L}^* = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \\ 0 & 3 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \\ 2 & 0 & 1 \\ 2 & 1 & 0 \\ 3 & 0 & 0 \end{bmatrix}.$$

For the case in which $n = 3$, we have

$$\frac{1}{3}\mathbf{L}^* \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 0 & 0 & 3 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \\ 0 & 3 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \\ 2 & 0 & 1 \\ 2 & 1 & 0 \\ 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} \overline{X}_1^* \\ \overline{X}_2^* \\ \overline{X}_3^* \\ \overline{X}_4^* \\ \overline{X}_5^* \\ \overline{X}_6^* \\ \overline{X}_7^* \\ \overline{X}_8^* \\ \overline{X}_9^* \\ \overline{X}_{10}^* \end{bmatrix} = \begin{bmatrix} X_3 \\ (X_2 + 2X_3)/3 \\ (2X_2 + X_3)/3 \\ X_2 \\ (X_1 + X_2 + X_3)/3 \\ (X_1 + 2X_3)/3 \\ (X_1 + 2X_2)/3 \\ (2X_1 + X_3)/3 \\ (2X_1 + X_2)/3 \\ X_1 \end{bmatrix}.$$

It follows that, for $n = 3$, the possible values of bootstrap sample mean $\overline{X}^*$ are $X_3$, $(X_2 + 2X_3)/3$, $(2X_2 + X_3)/3$, $X_2$, $(X_1 + X_2 + X_3)/3$, $(X_1 + 2X_3)/3$, $(X_1 + 2X_2)/3$, $(2X_1 + X_3)/3$, $(2X_1 + X_2)/3$, and $X_1$. Each of these values are equally likely to occur with probability $1/\binom{2(3)-1}{3} = 1/10$. This gives us the distribution of the bootstrap sample mean as follows.

$$\begin{bmatrix} \overline{X}^* : & X_3 & \frac{X_2+2X_3}{3} & \frac{2X_2+X_3}{3} & X_2 & \frac{X_1+X_2+X_3}{3} & \frac{X_1+2X_3}{3} & \frac{X_1+2X_2}{3} & \frac{2X_1+X_3}{3} & \frac{2X_1+X_2}{3} & X_1 \\ P\left(\overline{X}^*\right) : & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \end{bmatrix}$$

The mean of this discrete distribution is determined as follows.

$$\begin{aligned} E\left(\overline{X}^*\right) &= \frac{1}{10}[X_3 + \frac{X_2 + 2X_3}{3} + \frac{2X_2 + X_3}{3} + X_2 + \frac{X_1 + X_2 + X_3}{3} + \frac{X_1 + 2X_3}{3} \\ &\quad + \frac{X_1 + 2X_2}{3} + \frac{2X_1 + X_3}{3} + \frac{2X_1 + X_2}{3} + X_1] \\ &= \frac{10X_1 + 10X_2 + 10X_3}{3(10)} \\ &= \frac{X_1 + X_2 + X_3}{3} = \overline{X}_3. \end{aligned}$$

The variance of the distribution of $\overline{X}^*$ is determined as follows.

$$V\left(\overline{X}^*\right) = \frac{1}{10}\sum_{i=1}^{10}\left(\overline{X}_i^* - \overline{X}_3\right)^2.$$

Theorem 3.1: Each of the $X_i$'s appears $\binom{2n-1}{n-1}$ times in the matrix $\mathbf{L}^*$.

Proof: The proof can be found in Wilf and Zeilberger (1990). See the Appendix for a proof of this theorem.

Theorem 3.2: For a sample of size $n$, the average of the possible bootstrap sample means is $\overline{X}_n$.

Proof: We see that

$$\frac{1}{\binom{2n-1}{n-1}}\sum_{i=1}^{\binom{2n-1}{n-1}}\overline{X}_i^* = \frac{\binom{2n-1}{n-1}X_1 + \ldots + \binom{2n-1}{n-1}X_n}{n\binom{2n-1}{n-1}} = \overline{X}_n.$$

In general for a given sample size $n$, we have the distribution of $\overline{X}^*$ described as follows.

$$\begin{bmatrix} \overline{X}^*: & \overline{X}_1^* & \overline{X}_2^* & \ldots & \overline{X}_m^* \\ P\left(\overline{X}^*\right): & \frac{1}{m} & \frac{1}{m} & \ldots & \frac{1}{m} \end{bmatrix},$$

where $m = \binom{2n-1}{n-1}$. The mean of this distribution is $\overline{X}_n$ and the variance of the distribution is expressed as

$$V\left(\overline{X}^*\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\overline{X}_i^* - \overline{X}_n\right)^2.$$

## 3.4 Conclusion

We have discussed the distribution of the bootstrap sample mean $\overline{X}^*$. It was shown that the variance of the distribution of the bootstrap sample mean is almost twice the variance of the mean of the researcher's sample.

# CHAPTER 4

## DISTRIBUTION OF THE BOOTSTRAP SAMPLE VARIANCE $S^{2*}$

A bootstrap sample $X_1^*, \ldots, X_{n_B}^*$ of size $n_B$ is taken from the distribution of $X^*$. The bootstrap sample variance is defined by

$$S^{2*} = \frac{1}{n_B - 1} \sum\nolimits_{i=1}^{n_B} \left( X_i^* - \overline{X}^* \right)^2,$$

where

$$\overline{X}^* = \frac{1}{n_B} \sum\nolimits_{i=1}^{n_B} X_i^*$$

is the bootsrap sample mean. The expectation $E\left(S^{2*}\right)$ of $S^{2*}$ is determined as follows.

$$E\left(S^{2*}\right) = \frac{1}{n_B - 1} \sum\nolimits_{i=1}^{n_B} E\left[ \left( X_i^* - \overline{X}^* \right)^2 \right].$$

Note that

$$E\left[ \left( X_i^* - \overline{X}^* \right)^2 \right] = V\left( X_i^* - \overline{X}^* \right) + \left[ E\left( X_i^* - \overline{X}^* \right) \right]^2.$$

One can easily show that $E\left( X_i^* - \overline{X}^* \right) = 0$. Hence,

$$E\left[ \left( X_i^* - \overline{X}^* \right)^2 \right] = V\left( X_i^* - \overline{X}^* \right).$$

We can write

$$X_i^* - \overline{X}^* = \frac{n_B - 1}{n_B} X_i^* - \frac{1}{n_B} \sum\nolimits_{j=1, j \neq i}^{n_B} X_j^*.$$

It follows that

$$\begin{aligned} V\left( X_i^* - \overline{X}^* \right) &= \left( \frac{n_B - 1}{n_B} \right)^2 V\left( X_i^* \right) + \frac{1}{n_B^2} \sum\nolimits_{j=1, j \neq i}^{n} V\left( X_j^* \right) \\ &= \left( \frac{n_B - 1}{n_B} \right)^2 V\left( X^* \right) = \left( \frac{n_B - 1}{n_B} \right)^2 V\left( X^* \right) + \frac{n_B - 1}{n_B^2} V\left( X^* \right) \\ &= \left[ \left( \frac{n_B - 1}{n_B} \right)^2 + \frac{n_B - 1}{n_B^2} \right] V\left( X^* \right) = \frac{n_B - 1}{n_B} V\left( X^* \right). \end{aligned}$$

Therefore,

$$E\left(S^{2*}\right)=\frac{1}{n_B-1}\sum\nolimits_{i=1}^{n_B}\frac{n_B-1}{n_B}V\left(X^*\right)=V\left(X^*\right)=\frac{n-1}{n}S^2.$$

Hence, the bootstrap sample variance $S^{2*}$ is unbiased in estimating variance of the distribution of $X^*$.

The sample variance $S^2$ of the $n>1$ measurements $X_1,\ldots,X_n$ is defined by

$$S^2=\frac{1}{n-1}\sum\nolimits_{i=1}^{n}\left(X_i-\overline{X}\right)^2,$$

where

$$\overline{X}=\frac{1}{n}\sum\nolimits_{i=1}^{n}X_i.$$

The sample standard deviation $S$ is the principal square root of the sample variance, where

$$S=\sqrt{\frac{1}{n-1}\sum\nolimits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}.$$

One can show that

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

Hence,

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

Using the aforementioned results, we can express the $i$th bootstrap sample variance as

$$S_i^{2*}=\frac{n\sum_{j=1}^{n}L_{i,j}^*X_j^2-\left(\sum_{j=1}^{n}L_{i,j}^*X_j\right)^2}{n\left(n-1\right)},$$

where $\mathbf{L}_i^*$ is the $i$th row of the $\binom{2n-1}{n}\times n$ matrix $\mathbf{L}^*$ of possible bootstrap samples. The $i$th bootstrap sample standard deviation is

$$S_i^*=\sqrt{\frac{n\sum_{j=1}^{n}L_{i,j}^*X_j^2-\left(\sum_{j=1}^{n}L_{i,j}^*X_j\right)^2}{n\left(n-1\right)}}.$$

There are $m = \binom{2n-1}{n}$ possible bootstrap samples. Hence, there are $m$ possible bootstrap sample variances (sample standard deviations) denoted by $S_1^{2*}, \ldots, S_m^{2*}$ ($S_1^*, \ldots, S_m^*$). We can express the distribution of the bootstrap sample variances by

$$\begin{bmatrix} S_1^{2*}: & S_1^{2*} & S_2^{2*} & \ldots & S_m^{2*} \\ P\left(S^{2*} \left| X_1, \ldots, X_n\right.\right): & \frac{1}{m} & \frac{1}{m} & \ldots & \frac{1}{m} \end{bmatrix}.$$

The mean and variance of this distribution are

$$E\left(S^{2*}\right) = \frac{1}{m}\sum\nolimits_{i=1}^{n} S_i^{2*} = \overline{S}^{2*} \text{ and } V\left(S^{2*}\right) = \frac{1}{m}\sum\nolimits_{i=1}^{n}\left(S_i^{2*} - \overline{S}^{2*}\right)^2,$$

with

$$\overline{S}^{2*} = \frac{1}{m}\sum\nolimits_{j=1}^{m} S_i^{*2} \text{ and } \overline{S}^{*} = \frac{1}{m}\sum\nolimits_{j=1}^{m} S_i^{*}.$$

The $i$th row of $\mathbf{L}^*$ are the $n$ frequencies $f_j$ that indicate the number of times $X_j$ appears in the $i$th bootstrap sample. This is the average of the possible bootstrap sample variances and standard deviations, where $m = \binom{2n-1}{n}$. The expectation and variance of $\overline{S}^*$ are

$$E\left(\overline{S}^{*2}\right) = \frac{1}{m}\sum\nolimits_{j=1}^{m} E\left(S_i^{*2}\right) \text{ and } V\left(\overline{S}^{*2}\right) = \frac{1}{m^2}\sum\nolimits_{j=1}^{m} V\left(S_i^{*2}\right) \text{ and}$$
$$E\left(\overline{S}^{*}\right) = \frac{1}{m}\sum\nolimits_{j=1}^{m} E\left(S_i^{*}\right) \text{ and } V\left(\overline{S}^{*}\right) = \frac{1}{m^2}\sum\nolimits_{j=1}^{m} V\left(S_i^{*}\right).$$

We do not know in general how to obtain $E\left(\overline{S}^{*2}\right)$, $V\left(\overline{S}^{*2}\right)$, $E\left(\overline{S}^{*}\right)$, and $V\left(\overline{S}^{*}\right)$ analytically. However, we can obtain simulated estimates of these parameters for a given model. We assume the data is generated from a $N(10, 4)$ distribution. The following table gives the estimates of $E\left(\overline{S}^{*}\right)$ and $V\left(\overline{S}^{*}\right)$ for $n = 7, 8$.

Table of Estimates of $E\left(\overline{S}^{*}\right)$ and $V\left(\overline{S}^{*}\right)$

| $n$ | $\widehat{E}\left(\overline{S}^{*}\right)$ | $\widehat{V}\left(\overline{S}^{*}\right)$ |
|---|---|---|
| 7 | 0.77785793 | 0.052265 |
| 8 | 0.8056029 | 0.04762375 |

Under the independent Normal model, we have for the variance and standard deviation of the researcher's sample,

$$
\begin{aligned}
E\left(S_n^2\right) = \sigma^2, V\left(S_n^2\right) &= \frac{2\sigma^4}{n-1}, V\left(S_n\right) = c_4\sigma, \\
V\left(S\right) = V\left(\sigma W^{1/2}/\sqrt{n-1}\right) &= \frac{\sigma^2}{n-1} V\left(W_i^{1/2}\right) \\
&= \frac{\sigma^2}{n-1}\left[E\left(W_i\right) - E^2\left(W_i\right)\right] \\
&= \frac{\sigma^2}{n-1}\left[\sigma^2 - c_4^2\sigma^2\right] \\
&= \frac{\left(1-c_4^2\right)\sigma^4}{n-1},
\end{aligned}
$$

with $W = (n-1)\,S_n^2/\sigma^2 \sim \chi_{n-1}^2$ and

$$
c_4 = \frac{\sqrt{2}\Gamma\left(\frac{n-2}{2}\right)}{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)}.
$$

The function $c_4$ of $n$ is an unbiasing constant for the researcher's sample standard deviation.

The distribution of the bootstrap sample standard deviation $S^*$ is the discrete distribution

$$
\begin{bmatrix} S^*: & S_1^* & S_2^* & \ldots & S_m^* \\ P\left(S^*\right): & \frac{1}{m} & \frac{1}{m} & \ldots & \frac{1}{m} \end{bmatrix},
$$

where $m = \binom{2n-1}{n}$. The mean and variance of the distribution of $S^*$ are

$$
\mu_{S^*} = \frac{1}{m}\sum\nolimits_{i=1}^{m} S_i^* = \overline{S}^* \text{ and } \sigma_{S^*}^2 = \frac{1}{m}\sum\nolimits_{i=1}^{m}\left(S_i^* - \overline{S}^*\right)^2.
$$

The question that arises from the aforementioned analysis is "why bootstrap?" It has been stated that bootstrapping causes the estimation method to be more "stable." At this point, we do not have a definition of what is meant by a method being "stable." Our example shows that "bootstrapping" does not provide the researcher with *more information* about the process that generated the data than the information found in the sample. It simply adds more variability to the statistics used to summarize the information in the sample.

## 4.1 Conclusion

We have discussed the distribution of the bootstrap sample variance and the boostrap sample standard deviation.

# CHAPTER 5

# CONFIDENCE AND PREDICTION INTERVALS FOR $\mu$ AND $\overline{X}_N$

## 5.1 INTRODUCTION

Suppose we generate a (pseudo) random sample $X_1, \ldots, X_n$ of size $n$ from a $N(\mu, \sigma^2)$. One method for constructing a $100(1-\alpha)\%$ confidence interval for $\mu$ is the observed value of the random interval

$$\left(\overline{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right) \quad (1).$$

This the typically recommend $t$-interval. Suppose we generated a random sample of size $n = 5$ from a $N(10, 4)$ distribution and obtain the following data: $7.800671$, $9.33206464$, $8.15044618$, $12.96949046$, and $10.42013007$. A $95\%$ confidence interval for $\mu = 10$ using this method is $(7.149, 12.320)$ rounded to three decimal places.

On page 160 in Efron and Tibshirani (1993), the authors present a method for obtaining a bootstrap confidence interval for the population mean $\mu$. Suppose we generate $n_B$ independent bootstrap samples $\mathbf{X}_1^*, \ldots, \mathbf{X}_{n_B}^*$ and calculate for each one

$$T_i^* = \frac{\overline{X}_i^* - \overline{X}_n}{\widehat{se}_{i,\overline{X}}^*},$$

where $\overline{X}_i^*$ is the bootstrap sample mean the $i$th bootstrap sample $\mathbf{X}_i^* = [X_{i1}^*, \ldots, X_{in}^*]^{\mathbf{T}}$ and

$$\widehat{se}_{i,\overline{X}}^* = \sqrt{\frac{1}{n(n-1)}\sum\nolimits_{j=1}^{n}\left(X_{i,j}^* - \overline{X}_i^*\right)^2} = \frac{S_i^*}{\sqrt{n}},$$

for $i = 1, \ldots, n_B$. The $n_B$ values $T_i^*$ values are ordered $T_{1:n_B}^*, \ldots, T_{n_B:n_B}^*$. A method is used to obtain the a bootstrap estimator $\widehat{t}_{n-1,\alpha/2}^*$ of $t_{n-1,\alpha/2}$. The authors suggest a method when $n_B(\alpha/2)$ is not an integer. For $\alpha \leq 0.5$, define

$$k = \lceil (n_B + 1)(\alpha/2) \rceil .$$

The bootstrap estimators for $-t_{n-1,\alpha/2}$ and $t_{n-1,\alpha/2}$ are, respectively, $-\widehat{t}^*_{n-1,\alpha/2} = T^*_{k:n_B}$ and $\widehat{t}^*_{n-1,\alpha/2} = T^*_{n_B+1-k:n_B}$. The $100\,(1-\alpha)\,\%$ bootstrap confidence interval of the population mean $\mu$ is

$$\left(\overline{X}_n - T^*_{k:n_B}\frac{S_n}{\sqrt{n}}, \overline{X}_n - T^*_{B+1-k:n_B}\frac{S_n}{\sqrt{n}}\right) = \left(\overline{X}_n - \widehat{t}^*_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + \widehat{t}^*_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right).$$

For $n_B = 1000$, we obtained a $95\%$ bootstrap confidence interval for $\mu$ of $(7.973, 11.152)$ rounded to three decimal places. This gives us a interval of length smaller than the standard $t$-interval of $(7.148877245, 12.32024388)$.

In Chapter 13 of Efron and Tibshirini (1993), they discuss confidence intervals based on bootstrap-$t$, bootstrap percentiles, bias-corrected (BC), and the approximate bias-corrected methods (ABC). For the data, $7.800671, 9.33206464, 8.15044618, 12.96949046,$ and $10.42013007$, a $95\%$ bootstrap confidence interval based on $n_B = 1000$ for $\mu$ is $(8.179, 11.515)$. This interval is included in the bootstrap $t$-interval which is included in the standard $t$-interval. The authors also discuss the bias-corrected method (BC) and the approximate bootstrap confidence (ABC) method. For $n_B = 1000$ and these data, the BC method yields a $95\%$ confidence interval for $\mu$ of $(8.317, 11.946)$. The ABC method will give the same results as the BC method. The ABC method requires less computation than the BC method. As pointed out, Schenker (1985) states it is for confidence intervals “difficult or impossible to check in the complicated situations for which the bootstrap is intended.”

## 5.2 Bootstrap Prediction Interval for $\overline{X}_N$

There is a process that generates individuals in a researcher’s study. Let $X_1, \ldots, X_N$ be the $X$ measurements on these $N$ individuals. Assume this sample is a random sample. We refer to this sample as the representative sample of size $N$. The mean of these measurements is

$$\overline{X}_N = \frac{1}{N}\sum\nolimits_{i=1}^{N} X_i.$$

Many believe that the representative sample is the "population." Hence, the "population" mean would be $\overline{X}_N$. From this sample, the researcher is allowed to select a sample of size $1 \leq n \leq N$ of individuals with $X$ measurements $X_1, \ldots, X_n$. We refer to this sample as the researcher's sample. Assuming the representative sample is a random sample, then the researcher's sample is a random sample. If $n = N$, the researcher's sample is called a census. The mean of the researcher's sample is defined by

$$\overline{X}_n = \frac{1}{n}\sum\nolimits_{i=1}^{n} X_i.$$

We represent the mean of the remaining $X$ measuements by $\overline{X}_{N-n}$. Under the assumption that the representative sample is a random sample, the means $\overline{X}_n$ and $\overline{X}_{N-n}$ are independent.

Observe that

$$\overline{X}_N = \frac{n}{N}\overline{X}_n + \frac{N-n}{N}\overline{X}_{N-n} = \frac{n\overline{X}_n + (N-n)\,\overline{X}_{N-n}}{n}.$$

It follows that

$$\overline{X}_n - \overline{X}_N = \frac{N-n}{N}\left(\overline{X}_n + \overline{X}_{N-n}\right).$$

If we assume $X_i \sim N\left(\mu, \sigma^2\right)$, then it can be shown that $\overline{X}_n - \overline{X}_N \sim N\left(0, \frac{N-n}{N}\sigma^2/n\right)$. One can show that

$$Z = \frac{\overline{X}_n - \overline{X}_N}{\sqrt{\frac{N-n}{N}}\frac{\sigma}{\sqrt{n}}} \sim N\left(0, 1\right).$$

It follows that

$$T = \frac{\overline{X}_n - \overline{X}_N}{\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}}} = \frac{\frac{\overline{X}_n - \overline{X}_N}{\sqrt{\frac{N-n}{N}}\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2} / (n-1)}}$$
$$= \frac{Z}{\sqrt{\chi^2_{n-1} / (n-1)}} \sim t_{n-1},$$

where the random variables $Z$ and $\chi^2_{n-1}$ are independent. Hence, we have

$$-t_{n-1,\alpha/2} < \frac{\overline{X}_n - \overline{X}_N}{\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}}} < t_{n-1,\alpha/2} \text{ or}$$
$$\overline{X}_n - t_{n-1,\alpha/2}\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}} < \overline{X}_N < \overline{X}_n + t_{n-1,\alpha/2}\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}}.$$

The observed value of the random interval

$$\left(\overline{X}_n - t_{n=1,\alpha/2}\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n=1,\alpha/2}\sqrt{\frac{N-n}{N}}\frac{S_n}{\sqrt{n}}\right)$$

is a $100\,(1-\alpha)\,\%$ prediction interval for $\overline{X}_N$. Here we are assuming that $N$ is known. If $N$ is not known, one can assume that $(N-n)\,/N$ is approximately equal to $1$ if $N$ is "large" relative to $n$, we have the observed value of the random interval

$$\left(\overline{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right)$$

is an approximate $100\,(1-\alpha)\,\%$ prediction interval for $\overline{X}_N$.

The following data was simulated from a $N\,(10, 4)$ distribution: $7.80067147$, $9.33206464$, $8.15044618$, $12.96949046$, and $10.42013007$. A $95\%$ prediction interval for $\overline{X}_N$ for our given sample of size $n = 5$ is

$$(7.149, 12.320)$$

rounded to three decimal places.

A bootstrap prediction interval would have the form

$$\left(\overline{X}_n - T^*_{k:n_B}\frac{S_n}{\sqrt{n}}, \overline{X}_n - T^*_{n_B+1-k:n_B}\frac{S_n}{\sqrt{n}}\right).$$

For data, we have a $95\%$ prediction interval of

$$\left(9.73456 - 2.082765\frac{2.082434}{\sqrt{5}}, 9.73456 + 1.676555\frac{2.082434}{\sqrt{5}}\right)$$
$$= (7.795, 11.296)$$

rounded to three decimal places. We see that the $95\%$ bootstrap prediction interval $(7.795, 11.296)$ for $\overline{X}_N$ is narrower than the $95\%$ prediction interval (7.149,12.320) for $\overline{X}_N$.

## 5.3 CONCLUSION

In this chapter, we have looked at examples of methods for calculating confidence intervals for the population mean $\mu$. Also, we examined prediction intervals for the representative sample mean $\overline{X}_N$.

# CHAPTER 6

# TEST OF HYPOTHESIS

## 6.1 INTRODUCTION

A statistical hypothesis is a statement about the distribution of a random variable $X$. For example, a researcher may believe that the population mean $\mu$ is greater than a give value $\mu_0$. This hypothesis is referred to as the researcher's (also the althernative) hypothesis and it is expressed by $H_a : \mu > \mu_0$. In a test of hypothesis, one assumes that the opposite in truth value to the researcher's hypothesis is true. This hypothesis is referred to as the null hypothesis which is commonly expressed as $H_0$. In our example, $H_0 : \mu \leq \mu_0$.A random sample of size $n$ is to be taken whose $X$-values are $X_1, \ldots, X_n$. The collection of all possible samples of size $n$ is the sample space $\mathbf{S}$. The researcher selects a subcollection $\mathbf{C}$ of the sample space and uses the decision rule to reject $H_0$ if the observed sample is in $\mathbf{C}$. The subcollection $\mathbf{C}$ is selected such that the maximum probability the random sample $X_1, \ldots, X_n$ is in $\mathbf{C}$ is $\alpha$. The value $\alpha$ is called the size of the test. The possible null and alternative hypotheses about the population mean are given in the following table.

Table 6.1: Hypotheses about the Population Mean

| | Null Hypothesis | Researcher's (Alternative) Hypothesis |
|---|---|---|
| Case 1 | $H_0 : \mu \leq \mu_0$ | $H_a : \mu > \mu_0$ |
| Case 2 | $H_0 : \mu = \mu_0$ | $H_a : \mu \neq \mu_0$ |
| Case 3 | $H_0 : \mu \geq \mu_0$ | $H_a : \mu < \mu_0$ |

The researcher is allowed to take a random sample $X_1, \ldots, X_n$ of a fixed sample size $n$. The $t$-test is a likelihood ratio test. The following decision rules are based on the test statistic

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

where $\overline{X}$ and $S$ are respectively the mean and standard deviation of the sample. The null hypothesis is rejected if the observed value of $T$

Table 6.2: Decision Rules

| **Case** | **Decision Rule** |
|---|---|
| Case 1 | $T \geq t_{n-1,\alpha}$ |
| Case 2 | $\lvert T \rvert \geq t_{n-1,\alpha/2}$ |
| Case 3 | $T \leq -t_{n-1,\alpha}$ |

Note that $T$ has a non-central $t$-distribution with $n-1$ degrees of freedom and non-centrality parameter $\theta = \frac{(\mu - \mu_0)}{\frac{\sigma}{\sqrt{n}}}$, where $\mu$ and $\sigma$ are respectively the population mean and standard deviation.

The significance levels ($SL$) for each of the three case are given in the following table

Table 6.3: Decision Rules

| **Case** | **Decision Rule** |
|---|---|
| Case 1 | $1 - F_{t_{n-1,0}}(T)$ |
| Case 2 | $2\left[1 - F_{t_{n-1,0}}(\lvert T \rvert)\right]$ |
| Case 3 | $F_{t_{n-1,0}}(T)$ |

where $F_{t_{n-1,0}}(t)$ is the cumulative distribution function of a $t$-distribution with $n-1$ degrees of freedom. The significance level and its observed value are statistics. On the other hand, the $p$-value is a probability. Assuming the null hypothesis is true, it is the probability that the observed significance level of another sample of size $n$ will be less than or equal to the observed significance level of the researcher's sample. The magnitude of the $p$-value is equal to the observed significance level of the researcher's sample.

A major car manufacturer wants to test a new engine to determine whether it meets new air-pollution standards. The mean emission $\mu$ of all engines of this type must be less than $20$ parts per million of carbon. Five engines are manufactured for testing purposes, and the emission level of each are determined. The data (in parts per million) are

$$\begin{bmatrix} 19.4 & 16.6 & 17.9 & 12.7 & 13.9 \end{bmatrix}$$

Do the data supply sufficient evidence to allow the manufacturer to conclude that this type of engine meets the pollution standard? Assume that the manufacturer is willing to risk a Type I error with maximum probability $\alpha = 0.05$.

A Normal probability plot of these data suggest that one can assume that the process that generated these data has an approximate Normal distribution. We will use a $t$-test to analyze these data. The null and alternative hypotheses are

$$H_0 : \mu \geq 20 \text{ and } H_a : \mu < 20.$$

The test statistic we will use is

$$T = \frac{\overline{X} - 20}{S/\sqrt{n}}.$$

The observed significance level ($OSL$) is $0.01736$ (rounded to five decimal places). The $p$-value is $0.01736$. Since the $OSL \leq \alpha$, implies the engine manufacture is meeting omission standards. Further, we note that Monnu (2024) showed that as $n$ increases the $OSL$ decreases.

In the two sample case, the sample space **S** is the collection of all possible combined of the two samples of the form

$$X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2},$$

where $X_{i,j}$ is the $X$ measurement on the $j$th individual for $j = 1, \ldots, n_i$ and population $i = 1.2$.

Possible null and researcher's hypotheses are

Table 6.4: Hypotheses about the Population Means

| | Null Hypothesis | Researcher's (Alternative) Hypothesis |
|---|---|---|
| Case 1 | $H_0 : \mu_1 \leq \mu_2$ | $H_a : \mu_1 > \mu_2$ |
| Case 2 | $H_0 : \mu_1 = \mu_2$ | $H_a : \mu_1 \neq \mu_2$ |
| Case 3 | $H_0 : \mu_1 \geq \mu_2$ | $H_a : \mu_1 < \mu_2$ |

The researcher selects a subcollection **C** of the sample space and uses the decision rule to reject $H_0$ if the observed combined sample is in **C**. The subcollection **C** is selected such that the maximum probability the combined random sample $X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2}$ is in **C** is $\alpha$.

If one assume that $X_{1,1}, \ldots, X_{1,n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$, $X_{2,1}, \ldots, X_{2,n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$, and $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the recommended test statistic

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}},$$

where $\overline{X}_i$ and $S_i^2$ are respectively the sample mean and sample variance with

$$S_p = \sqrt{\frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{(n_1 - 1) + (n_2 - 1)}}.$$

Under our model, the statistic $T$ has a non-central $t$-distribution with $n_1 + n_2 - 2$ degrees of freedom and non-centrality parameter $\theta$, where

$$\theta = \frac{(\mu_1 - \mu_2)/\sigma}{\sqrt{1/n_1 + 1/n_2}}.$$

The significance level ($SL$) is for each case is given in the following table.

Table 6.5: Estimated Approximate Significance Level

| **Case** | **Significance Level** |
|---|---|
| Case 1 | $1 - F_{t_{n_1+n_2-2,0}}(T)$ |
| Case 2 | $2\left[1 - F_{t_{n_1+n_2-2,0}}(\lvert T\rvert)\right]$ |
| Case 3 | $F_{t_{n_1+n_2-2,0}}(T)$ |

where $F_{t_{n_1+n_2-2,0}}(t)$ is the cumulative distribution function of a central $t$-distribution with  degrees of freedom.

Welch (1938) developed an estimated, approximate method for testing the aforementioned hypothesis based on the test statistic

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}},$$

where $\overline{X}_i$ and $S_i^2$ are the mean and variance of the sample from population $i = 1, 2$. The distribution of $T$ is has an estimated, approximate $t$-distribution with estimated degrees of freedom

$$\widehat{\nu} = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}.$$

Note that $T$ has an estimated approximat non-central $t$-distribution with $\widehat{\nu}$ degrees of freedom and noncentrality parameter

$$\theta = \frac{(\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{\delta}{\sqrt{\lambda^2/n_1 + 1/n_2}},$$

where $\delta = (\mu_1 - \mu_2)/\sigma_2$ and $\lambda^2 = \sigma_1^2/\sigma_2^2$. The estimated approximate significance level ($SL$) is for each case is given in the following table.

Table 6.6: Estimated Approximate Significance Level

| | **Estimated Approximate Significance Level** |
|---|---|
| **Case 1:** | $1 - F_{t_{\widehat{\nu},0}}(T)$ |
| **Case 2:** | $2\left[1 - F_{t_{\widehat{\nu},0}}(\lvert T\rvert)\right]$ |
| **Case 3:** | $F_{t_{\widehat{\nu},0}}(T)$ |

Champ and Hu (2024) derived the exact distribution of

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

They recommended an estimation exact method for computing a confidence interval for $\mu_1 - \mu_2$. Their method will not be examined further in our research.

Efron and Tibshironi (1993) give an example in which $16$ mice where randomly assigned to a treatment and control groups. Their survival times, in days, following a test surgery are given in the following table.

Table 6.7: Treatment vs. Control Data

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|
| **Treatment** | 94 | 38 | 23 | 197 | 99 | 16 | 141 | | |
| **Control** | 52 | 10 | 40 | 104 | 51 | 27 | 146 | 30 | 46 |

The question they posed "did the treatment prolong survival?" They assume that The observed value of $T$ is $t = 1.121$ (rounded to three decimal places), with $14$ degrees of freedom. The observed significance level ($OSL$) is $0.14061$ (rounded to five decimal places).

The $p$-value is equal in magnitude to $0.14061$. If we do not assume the population variances are equal, then using Welch (1938) method we have the observed value of $T$ is $t = 1.059$ (rounded to three decimal places). The observe value of $\widehat{\nu} = 9.654$ (rounded to three decimal places). The estimated approximate observed significance level ($OSL$) is $0.15775$ (rounded to five decimal places). The estimated approximate $p$-value is equal in magnitude the $OSL$ of $0.15775$.

## 6.2 Bootstrap Test of Hypothesis

Let $\mathbf{X}^* = [X_1, \ldots, X_n]^{\mathbf{T}}$ be a bootstrap random sample. Our interest is testing $H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$. Suppose we generate $n_B$ independent bootstrap samples $\mathbf{X}_1^*, \ldots, \mathbf{X}_{n_B}^*$ and calculate for each one

$$T_i^* = \frac{\overline{X}_i^* - \overline{X}_n}{\widehat{se}_i^*},$$

where $\overline{X}_i^*$ is the bootstrap sample mean the $i$th bootstrap sample $\mathbf{X}_i^* = [X_1^*, \ldots, X_n^*]^{\mathbf{T}}$, for $i = 1, \ldots, n_B$. On page 43 of Efron and Tibsharoni (1993), the authors define the estimated standard error by

$$\widehat{se} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n^2}}.$$

It follows that the bootstrap estimated standard error for the $i$th bootstrap sample is

$$\widehat{se}_i^* = \sqrt{\frac{\sum_{j=1}^{n}\left(X_{i,j}^* - \overline{X}_i^*\right)^2}{n\,(n-1)}} = S_i^*/\sqrt{n}.$$

The $n_B$ values $T_i^*$ values are ordered $T_{1:n_B}^*, \ldots, T_{n_B:n_B}^*$. A method is used to obtain the bootstrap estimator $\widehat{t}_{n-1,\alpha/2}^*$ of $t_{n-1,\alpha/2}$. The authors suggest a method when $n_B\,(\alpha/2)$ is not an integer. For $\alpha \leq 0.5$, define

$$k = \lceil (n_B + 1)\,\alpha \rceil\,.$$

The bootstrap estimators for $-t_{n-1,\alpha/2}$ is $-\widehat{t}^*_{n-1,\alpha/2} = T^*_{k:n_B}$. Our decision rule is to reject $H_0$ if the observed value of

$$T = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \leq -\widehat{t}^*_{n-1,\alpha} = T^*_{k:n_B}.$$

Similarly, for the other two cases with

$$T = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}},$$

we reject $H_0$ if the observed value of $T$ is where $k$ in Case 2 is $k = \lceil (n_B + 1)(\alpha/2) \rceil$.

Table 6.8: Hypotheses about the Population Mean

| | **Decision Rule** | $\widehat{ASL}^*$ |
|---|---|---|
| **Case 1:** | $T \geq \widehat{t}^*_{n-1,\alpha} = T^*_{n_B+1-k:n_B}$ | $P\left(T \geq \widehat{t}^*_{n-1,\alpha}\right)$ |
| **Case 2:** | $\lvert T\rvert \geq \widehat{t}^*_{n-1,\alpha/2} = T^*_{n_B+1-k:n_B}$ | $2P\left(\lvert T\rvert \geq \widehat{t}^*_{n-1,\alpha/2}\right)$ |

For our example, we have $H_0 : \mu \geq 20$ versus $H_a : \mu < 20$. The available data is

$$\begin{bmatrix} 19.4 & 16.6 & 17.9 & 12.7 & 13.9 \end{bmatrix}.$$

Base on $n_B = 1000$, we approximate $-\widehat{t}^*_{4,05} = -1.961742$ and the observed statistics $T = -3.14373$.

The

$$\widehat{ASL}^* = P(T_i \leq T) = P(T_i \leq -3.14373) = 0.018.$$

rounded to five decimal places. The $\widehat{ASL}^*$ is akin to the observed significance level. For these data, the $OSL = 0.0173614$

On page 203 of Efron and Tibshirani (1993) having observed $\widehat{\theta}$, the authors define the achieved sginificance level ($ASL$) as the probability of observing at least that large a value when the null hypothesis is true. That is,

$$ASL = P\left(\widehat{\theta}^* \geq \widehat{\theta}\right).$$

The $ASL$ appears to be akin to the $p$-value. The smaller the value of $ASL$, the stronger the evidence against the null hypothesis. Using bootstrapping, the $ASL$ is estimated by

$$\widehat{ASL}^* = \#\left(\widehat{\theta}^* \geq \widehat{\theta}\right)/n_B,$$

where $\widehat{\theta}^*$ is based on a bootstrap sample. This is a statistic since it can be observed from the data and is akin to the significance level ($SL$).

In the two sample case, Efron and Tibshirani (1993) compute an estimate of the $ASL$ as follows. Compute

$$T_j^* = \frac{\overline{X}^*_{1,j} - \overline{X}^*_{2,i}}{\sqrt{S^{2*}_{1,j}/n_1 + S^{2*}_{1,j}/n_2}},$$

where $\overline{X}^*_{i,j}$ and $S^{2*}_{i,j}$ are the mean and variance of the $i$th bootstrap sample, $i = 1, \ldots, n_B$. They approximate $ASL$ by

$$\widehat{ASL} = \#\left(T_i^* \geq t\right)/n_B,$$

where $t$ is the observed value of

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_1^2/n_1 + S_1^2/n_2}}$$

from the researcher's samples $X_{1,1}, \ldots, X_{1,n_1}$ and $X_{2,1}, \ldots, X_{2,n_2}$.

For the treatment and control data, the observed treatment mean is $\overline{x}_1 = 86.857$ (rounded to three decimal places) and the observed control mean is $\overline{x}_2 = 56.222$ (rounded to three decimal places). The mean of the combined samples is $\overline{y} = 143.079$ (rounded to three decimal places). On page 224 of Efron and Tibshirani (1993), the authors define

$$\widetilde{X}_{1i} = X_{1i} - \overline{X}_1 + \overline{Y} \text{ and } \widetilde{X}_{2i} = X_{2i} - \overline{X}_2 + \overline{Y}.$$

For the treatment and control data,

Table 6.9: Table of Transformed Values

| $i$ | $\widetilde{x}_{1i}$ | $\widetilde{x}_{2i}$ |
|---|---|---|
| 1 | 150.222 | 138.857 |
| 2 | 94.222 | 96.857 |
| 3 | 79.222 | 126.857 |
| 4 | 253.222 | 190.857 |
| 5 | 155.222 | 137.857 |
| 6 | 72.222 | 113.857 |
| 7 | 197.222 | 232.857 |
| 8 | | 116.857 |
| 9 | | 132.857 |

Now we form the $n_B$ bootstrap data sets $(\mathbf{X}_1^*, \mathbf{X}_2^*)$, where $\mathbf{X}_i^* = \left[\widetilde{X}_{i,1}, \ldots, \widetilde{X}_{i,n_i}\right]^{\mathbf{T}}$. Next evaluate $T_i^*$ defined by

$$T_i^* = \frac{\overline{X}_{1i}^* - \overline{X}_{2i}^*}{\sqrt{S_{1i}^{2*}/n_1 + S_{2i}^{2*}/n_2}}$$

for $i = 1, \ldots, n_B$. Approximate $\widehat{ASL}^*$ by

$$\widehat{ASL}^* = \#\left(T_i^* \geq t\right)/n_B,$$

where

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_1^2/n_1 + s_1^2/n_2}}.$$

For these data, $t = 1.059$ (rounded to three decimal places). For these data and $n_B = 1000$, we have $\widehat{ASL}^* = 0.129$. The $OSL = 0.158$ (rounded to three decimal places. How does one compare $\widehat{ASL}^*$ and the $OSL$?. Following Schenker (1985), it would be difficult to compare $\widehat{ASL}^*$ and the $OSL$.

## 6.3 Conclusion

We have examined parametric and bootstrap confidence intervals. Also, we examined parametric and bootstrap test of hypothesis. It seems that it would be difficult to compare parametric and bootstrap methods.

# CHAPTER 7

# CONCLUSION

## 7.1 General Conclusion

As was our intent, we have accounted for the added variability when taking a bootstrap sample from the researcher's sample. We have shown that the bootstrap sample mean is unbiased in estimating the population mean. However, the variability of the estimator is almost twice that of the sample mean.

## 7.2 Areas for Further Research

We are interested in studying bootstrap regression methods, bootstrap multivariate methods, and bootstrap methods in quality control. One of the measurements of interest in studying control charts is the run length distribution. It will be of interest to see how one could effectively use bootstrapping to study the performance measures of a control chart such as the average run length and standard deviation of the run length as well as percentiles of the run length distribution.

# REFERENCES


Champ, C.W. and Hu, F. (2024) "An Exact Expression of the Behrens-Fisher Distribution with Applications," *Applied Mathematical Analysis and Computation 1st SGMC*, Statesboro, GA, USA, April 2-3, 2021.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics* 7, 1-26.

Efron, B (1981), "Nonparametric Estimates of Standard Error: the Jackknife, the Bootstrap and Other Methods," *Biometrika* 68, 589-599.

Efron, B. (1982), "The Jackknife, the Bootstrap, and Other Resampling Plans," In *Society of Industrial and Applied Mathematics* CBMS-NSF Monographs, 38.

Efron, B. (1988), "Bootstrap Confidence Intervals: Good or Bad?," (with discussion) *Psychology Bulletin* 04, 293-296.

Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Davison, A.C. and Hinkley, D. V. (1997), *Bootstrap Methods and their Application*, Cambridge University Press.

Diaconis, P. and Efron, B. (1983), "Computer-intensive Methods in Statistics," *Scientific American*, May 116-130.

Fisher, R.A. (1958), *Statistical Methods for Research Workers*, Hafner Publishing Company Inc.: New York.

Good, P.I. (2005), *Permutation, Parametric and Bootstrap Tests of Hypotheses*, Third Edition, Springer.

Hesterberg, T.C., Moore, D.S., Monaghan, S., Clipson, A. and Epstein, R (2003), *Bootstrap Methods and Permutation Tests*, W.H. Freeman: New York.

Hinkley, D.V. (1988), "Bootstrap Methods (with discussion)," *Journal of the Royal Statistical Society B* 50, 321-337.

Lunnelborg, C.E. (1985), "Estimating the Correlation Coefficient: The Bootstrap Approach," *Psychology Bulletin* 98, 209-215.

Monnu, P.O. (2024), *The Distribution of the Significance Level*, Masters Thesis, Department of Mathematical Sciences, Georgia Southern University.

Mooney, C.Z. and Duval, R.D. (1993), "Bootstrapping, A Nonparametric Approach to Statistical Inference," Sage University Paper series on Quantitative Applications in the Social Sciences, 07-095, Sage: Newbury Park, CA.

Peters, S.C. and Freedman, D.A. (1987), "Balm for Bootstrap Confidence Intervals," *Journal of the American Statistical Association* 82, 186-187.

Robinson, G.K. (1986), "Bootstrap and Randomization Confidence Intervals," Proceedings of the Pacific Statistical Congress, 20-24 May 1985, Auckland (Francis, I.S., Manly, B.F.J., and Lam, F.C., editors) North Holland, Groningen, 49-50.

Robinson, G.K. (1986), "Nonparametric Confidence Intervals in Regression: The Bootstrap and Randomization Methods," New Perspetives in Theoretical and Assplied Statistics (Puri, M.L., Vilaplana, J.P., and Wertz, W., editors), 243-256, Wiley: New York.

Schenker, N. (1985), "Qualms about Bootstrap Confidence Intervals," *Journal of the American Statistical Association* 80(390), 360-361.

Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag, Inc.

Welch, B.L. (1938), "The Significance of the Difference Between Two Means When the Population Variances Are Unequal," *Biometrika* 29, 350–362.

Wilf, H.S. and Zeilberger, D. (1990), "Rational Functions Certify Combinatorial Identities," *Journal of the American Mathematical Society* 3, 147-158

Wu, C.F.J. (1986), “Jackknife, Bootstrap and other Resampling Methods in Regression Analysis,” *The Annals of Statistics* 14(4), 1261–1295.

# APPENDIX

Definition: A *weak-composition* of $n$ into $k$ parts is a $k$-tuple of non-negative integers that sum to $n$.

Definition: Let $c(n,k)$ denote the number of weak-compositions of $n$ into $k$ parts.

The next theorem requires the following result, which we will present as a lemma.

Lemma: For $n \geq 1$,

$$\sum\nolimits_{k=0}^{n} k\binom{2n-k-2}{n-k} = \binom{2n-1}{n-1}.$$

Proof: We will instead prove the summation in the equivalent form

$$\sum\nolimits_{k=0}^{n} k\binom{2n-k-2}{n-k} \Big/ \binom{2n-1}{n-1} = 1 \tag{7.1}$$

for $n = 1, 2, 3, \ldots$. The method of proof is that of Wilf and Zeilberger (1990). Let

$$F(n,k) = k\frac{\binom{2n-k-2}{n-k}}{\binom{2n-1}{n}} = k\frac{n!(n-1)!(2n-k-2)!}{(n-k)!(2n-1)!},$$

$$R(n,k) = -\frac{(nk-k+n)(2n-k-1)}{kn(n-1)},$$

$$G(n,k) = F(n,k)R(n,k) = -\frac{(2n-k-1)!(n-k+kn)(n-1)!}{(2n-1)!(n-k)!}.$$

Notice that $F(n,k)$ is the summand in (7.1). First, we wish to show that $F(n,k) = G(n,k+1) - G(n,k)$.

$$F(n,k) = G(n,k+1) - G(n,k) \quad \text{iff}$$

$$\frac{F(n,k)}{F(n,k)} = \frac{G(n,k+1)}{F(n,k)} - \frac{G(n,k)}{F(n,k)} \quad \text{iff}$$

$$1 = \frac{F(n,k+1)R(n,k+1)}{F(n,k)} - R(n,k) \quad \text{iff}$$

$$1 = \frac{(k-n)(nk+2n-k-1)}{kn(n-1)} + \frac{(nk-k+n)(2n-k-1)}{kn(n-1)} \quad \text{iff}$$

$$1 = 1,$$

which is clearly true. Now that we have established that $F(n,k) = G(n,k+1) - G(n,k)$, we will sum both sides of this equation for $k = 1, 2, 3, \ldots, n$. Thus we obtain

$$\begin{aligned}\sum\nolimits_{k=1}^{n} F(n,k) &= \sum\nolimits_{k=1}^{n} [G(n,k+1) - G(n,k)] \\ &= G(n,n+1) - G(n,1) \\ &= 0 - (-1) \\ &= 1\end{aligned}$$

where the penultimate equality follows because the summation telescopes.

Noting the $k = 0$ term in (7.1) is $0$, we obtain

$$\sum\nolimits_{k=1}^{n} F(n,k) = \sum\nolimits_{k=0}^{n} F(n,k) = 1$$
$$\sum\nolimits_{k=0}^{n} k\binom{2n-k-2}{n-k} / \binom{2n-1}{n} = 1$$

and the result follows.