

PERFORMANCE OF BOOTSTRAP CONFIDENCE
REGION FOR BINOMIAL DISTRIBUTION WITH
UNKNOWN PARAMETERS p AND m

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

UNIVERSITY OF REGINA

By

Chengyu Gao

Regina, Saskatchewan

Fer 2020

© Copyright 2020: Chengyu Gao

Abstract

The goal of this research is to find out the Performance of the bootstrap confidence region of a binomial distribution with unknown parameters. The research is designed as follow:

The first step is to estimate unknown parameters m and p from binomial distribution. In my case, I focus on method of moment to estimate. Since these estimators do not have moments of all orders, I cannot obtain the mean, variance, and covariance for these estimators. Thus, the Delta Method is used to derive the asymptotic normality of the joint distribution of the Method of Moments estimators. After finding the estimators \hat{p} , and \hat{m} , I will work on the Asymptotic Normality of the Estimators.

For Asymptotic Normality of the Estimators by method of moment, I will find the sampling from the binomial distribution with sample mean $\bar{X} - mp$ and sample variance $S^2 - mp(1 - p)$ are asymptotically normal with zero mean vector and covariance matrix. I will apply the Delta-method consists of expansion of \hat{p} , and \hat{m} into two-dimensional Taylor series expansion, then using partial derivatives of these

functions, so I should have the covariance matrix.

Next section, I find out that if random vector X is normally distributed with the mean vector E and covariance matrix Σ is distributed as chi-square with 2 degrees of freedom. Therefore, the $100(1 - \alpha)\%$ confidence region should be $X_2^2(p, m) \leq X_2^2(\alpha)$ with 2 degree of freedom.

Some general steps of using the independent and dependent bootstrap sampling will be the next. I will give example of creating both independent and dependent bootstrap samples with different k , where k is the number of copies of original sample. I also will talk about the coverage probability of confidence regions and the areas of confidence regions.

Acknowledgements

I am very grateful to my supervisor, Dr. Andrei Volodin. He encouraged me with doing my master degree and helped me with research paper. Also, I want to appreciate my parents for supporting me. I want to thank Dr. Dianliang Deng, Dr. Michael Kozdron and all instructors of Statistics department at University of Winnipeg and University of Regina for helping me build solid background of statistics. I would like to thank Sarah Carnochan Naqvi and my friend Sichen Liu help me with software latex. Especially, I appreciate Jie Li who encouraged me to pursue my master degree and future study.

Contents

Abstract	i
Acknowledgements	iii
Table of Contents	iv
List of Figures	vii
List of Tables	ix
1 Introduction	1
1.1 Research Objectives	1
1.2 Research Hypothesis	2
1.3 Research Statement	3
1.4 The Significance of This Research	3
2 Literature Review	6
2.1 General Statistics Literature	6

2.2	Binomial Distribution	7
2.3	Point Estimation	8
2.3.1	Unbiased Estimator	8
2.3.2	Mean Square Error	10
2.4	Method of Moment	11
2.4.1	Sample Mean and Sample Variance Estimation	12
2.5	Maximum Likelihood Estimation	13
2.6	Delta Method	13
2.6.1	Convergence in Distribution	14
2.6.2	Slutsky's Theorem	15
2.7	Bootstrap	15
2.7.1	Independent Bootstrap	17
2.7.2	Dependent Bootstrap	18
2.8	Coverage Probability	20
2.9	Chi-Square Distribution	20
2.9.1	Chi-Squared Test	21
3	Research Methods	22
3.1	Introduction of Research Methods	22
3.2	Bivariate Normal Distribution	23
3.3	Method of Moment Estimator	24

3.4	The First Four Moments of Binomial Distribution	25
3.5	Delta Method	28
3.6	Confidence Region	30
3.6.1	Confidence Region for Binomial Distribution	31
3.6.2	Test Statistic of Confidence Region	32
3.7	The Performance of Confidence Region	33
3.7.1	Coverage Probability of Confidence Region	33
3.7.2	Area of Confidence Region	34
4	Statistical Simulation	36
4.1	Introduction	36
4.2	Simulation	37
4.3	Coverage Probability of 95% Confidence Region	37
4.4	Coverage Probability of 90% Confidence Region	45
4.5	Areas of Confidence Region	49
5	Conclusion and Future Work	59
5.1	Conclusion	59
5.2	Future Work	60
	References	62

List of Figures

2.1	Figure shows the general steps of bootstrapping from towards data science [16].	16
3.1	Figure shows the confidence region of mean of \hat{p} and \hat{m} from 2000 independent bootstrap sample follows $\text{Bin}(500, 0.55)$	35
4.1	The Areas 95% Confidence regions of $\text{Bin}(p = 0.4, m = 250)$ for independent bootstrap and dependent bootstrap with $k = 5, 50$	50
4.2	The Areas 95% Confidence regions of $\text{Bin}(p = \{0.4, 0.6, 0.8\}, m = 500)$ for independent bootstrap method	52
4.3	The Areas 95% Confidence regions of $\text{Bin}(p = 0.6, m = \{100, 250, 500\})$ for independent bootstrap method	54
4.4	The Areas 95% Confidence regions of $\text{Bin}(p = 0.6, m = \{100, 250, 500\})$ for dependent bootstrap method	56

4.5	The Areas 95% Confidence regions of $Bin(p = 0.6, m = \{100, 250, 500\})$	
	for dependent bootstrap method	58

List of Tables

4.1	Coverage probabilities of 95% confidence regions of Independent Bootstrap samples	38
4.2	Coverage probabilities of 95% confidence regions of dependent Bootstrap samples $k = 5$	39
4.3	Coverage probabilities of 95% confidence regions of dependent Bootstrap samples $k = 10$	40
4.4	Coverage probabilities of 95% confidence regions of dependent Bootstrap samples $k = 25$	41
4.5	Coverage probabilities of 95% confidence regions of dependent Bootstrap samples $k = 50$	42
4.6	Coverage probabilities of 95% confidence regions of dependent Bootstrap samples $k = 100$	43
4.7	Coverage probabilities of 95% confidence regions of dependent Bootstrap samples $k = 250$	44

4.8	Coverage probabilities of 95% confidence regions of dependent Bootstrap samples $k = 500$	45
4.9	Coverage probabilities of 90% confidence regions of independent Bootstrap samples	46
4.10	Coverage probabilities of 90% confidence regions of dependent Bootstrap samples $k = 5$	47
4.11	Coverage probabilities of 90% confidence regions of dependent Bootstrap samples $k = 50$	48

Chapter 1

Introduction

1.1 Research Objectives

The estimation of binomial distribution parameters m and p achieve in four common inference methods: Bayesian Method, the Maximum Likelihood, Method of Moments, Method of Least Squares. Method of Moments is the only method for this research.

The independent bootstrap sampling method will be applied for resampling the estimators \hat{m} and \hat{p} , in order to construct the $100(1 - \alpha)\%$ Confidence regions and calculate the convergence probabilities, as well as the dependent bootstrap sampling method with different k .

The asymptotic normality of the joint distribution of method of moment estimators \hat{m} and \hat{p} is proved by the delta method.

The confidence region of estimators \hat{m} and \hat{p} can be written as form which follows

the chi-square distribution with two degree of freedom.

1.2 Research Hypothesis

The goal is to see whether the bootstrap procedure can influence the confidence regions for $\text{Bin}(m,p)$, such as changing the coverage probability, and the areas of confidence regions. Furthermore, I will compare the results for both independent bootstrap samples and dependent bootstrap samples, in order to test the performance of bootstrap sampling. In my case, I focus on method of moment to estimate. Delta Method is used to derive the asymptotic normality of the joint distribution of the Method of Moments estimators. After finding the estimators \hat{p} , and \hat{m} , I will work on the Asymptotic Normality of the Estimators.

For the simulation part, I plan to generate N binomial samples with fixed m and p with each sample size n . Then I will generate the estimators for all m and p . Applying the independent and dependent bootstrap for re-sampling propose, I will have B sets of bootstrap samples of the estimators for m and p . A summary of all B

bootstrap samples of coverage probabilities can give some results of this research. Moreover, I can change the value of true m , p , B and k to compare all the results by plots and tables.

1.3 Research Statement

Using chi-square test, to define the rules for Confidence regions of Estimation of m and p from binomial distribution by method of moment on independent bootstrap method and dependent bootstrap method with different k . The coverage probabilities of confidence regions are not evidently different, but the Areas of confidence regions.

1.4 The Significance of This Research

The method of Sampling is the first step of statistical analysis. However, the samples which selected from the population are not perfect. Because the dataset of population must include every entry of the select topic, it is impossible to find all of them. Therefore, constructing samples in statistical way is one of the major goal for statisticians.

A good statistical analysis relying on precise samples. The statisticians and data analyst spend time, patience, and financial support ceaselessly, in order to construct accurate samples. The classical sampling method, such as simple random sampling is often costly and inefficient due to all elements with same probability of being selected (Yates, David, and Daren, 2008). This research will focus on using simple and accurate sampling method: the bootstrap sampling.

One of the superiority of the bootstrap sampling is that the bootstrap method does not rely on a highly completed sample data. The method of bootstrap sampling

is designed for replacing and controlling the statistical features (Efron and Tibshirani, 1993). The simple sampling methods, such as random sampling methods are suitable for applying any statistical problems (Varian, 2005). The bootstrap procedure can easily estimate the confidence regions and standard errors for Binomial distribution with two unknown parameters. This works even when we do not have large and perfect samples. Even though it is not possible to find the real confidence regions, the bootstrap can derive precise and stable results (DiCiccio, and Efron, 1996). A perfect data will be costly and difficult to find, sometimes it does not even exist. Therefore, the bootstrap allow us to get good statistical information, even with data set small or some sampling issue. The bootstrap sampling can be done in two ways (independent bootstrap and dependent bootstrap).

Confidence region (confidence ellipse) is a generalized form of confidence interval with two or more dimensions. To understand confidence region, a ground work of understanding confidence interval, which is one of the major tool in statistical analysis is important. For example, If a researcher gets the sample with known standard deviation, the $100(1 - \alpha)\%$ confidence interval can be written as

$$(\hat{\mu} - \Phi^{-1}(.95)\hat{\sigma}/\sqrt{n}, \hat{\mu} + \Phi^{-1}(.95)\hat{\sigma}/\sqrt{n})$$

where $\alpha = 0.05$. This is clear to see that when the researchers would like analyzing one estimator at one time, so in such a case it is one-dimensional intervals to analyze. Somehow, to analyze the two-dimension Binomial distribution with unknown

m (number of trials) and p (probability of success of each trials), the performance of Confidence region is crucial and essential.

Chapter 2

Literature Review

2.1 General Statistics Literature

This chapter is introduction of the statistical literature which related to my research. I will show the theorem definition and explanation.

In statistics, some researchers often find confidence intervals (CI) for the parameters of their interest. CI is a range of values of interest, for example the mean of random samples. CIs must be with particular probabilities also called confidence coefficient in statistics words. The confidence coefficient can be changed by the researcher; and CI relevant to the estimator of sampling distribution (Dekking, 2005). In other words, the exact value of the unknown parameter is inside of the CI by $\nu\%$ chance. In my thesis, I set up $\nu\%$ equal to 0.9 and 0.95.

Statistical median is useful method to measure the centre of sample data, which it can separate first and second half of data. The statistical mean is different to the

median, since the mean is the average number of data. To calculate the median, we can rank all numbers of the data. If the total number of data set is odd number, the median is the number with middle position of order. If the total number of data set is even number, the median can be calculate as the half of summation of the last number of first half of data and the first number of second half of data.

2.2 Binomial Distribution

In statistics, the binomial distribution is discrete distribution which has two unknown parameters m and p . In the experiment, a total number of m independent trials can have k success trials, and p is the probability of success. Because p is probability of success, so the interval of p is $0 \leq p \leq 1$; moreover, $k \leq m$. The probability mass function of binomial distribution can be written as

$$P(X = k) = \binom{m}{k} p^k (1 - p)^{m-k}, k = 0, 1, \dots, m.$$

Haldane was the first researcher that published statistically problems of binomial distribution(Haldane, 1941). According to his results, we can estimated parameter m and p by solving Method of Moments estimation. After solving the first two moments, it will be easy to find sample mean and sample variance which often denote as \bar{X} and S^2 . Solving the equations obtain the estimation of m and p (will show in ch3).

2.3 Point Estimation

Point Estimation is a significant content of statistical inference. There are two general methods of finding point estimators, the method of moment estimation and Maximum Likelihood Estimation. The method of moment estimation is the oldest and simplest way to obtain the point estimators, even though it sometimes is not the best option. The Maximum Likelihood method is the most prevalent estimation method. According to Casella, a measurable function $f = f(Y_1, \dots, Y_i)$ is called a statistic. Any estimator is a statistic. Estimate is a numerical value of an estimator [2]. It can be considerable to identify estimator and estimate. The estimator is a function of random sample and estimate is the accurate value of the random sample. Thus, the estimator is $f = f(Y_1, \dots, Y_i)$, and the estimate is accurate value $f = f(y_1, \dots, y_i)$, where Y_1, \dots, Y_i is iid random sample.

In statistics, we are really care to get a “good” estimators. The good estimators must Satisfy several standards, such as Unbiased estimator, and small value of mean squared error.

2.3.1 Unbiased Estimator

Before we talk about unbiased estimator, the definition of bias of an estimator is a requirement. If a estimator is biased, the estimate will be different than the exact

value of the parameter of interest. Also, it does not equal to the mathematical expectation of the random sample. In mathematical notation, suppose θ is an estimator of any random sample Y_1, \dots, Y_n . The bias of $\hat{\theta}$ can be written as

$$Bias_{\theta}[\hat{\theta}] = E_{y|\theta}[\hat{\theta}] - \theta = E_{y|\theta}[\hat{\theta} - \theta]$$

where $E_{y|\theta}$ means the mathematical expectation of any unknown distribution of y .

For unbiased estimator, the bias of estimator must equal to zero. Thus, the mathematical expectation of the estimator must equal to the true value of the population parameter. we can define as:

$$Bias_{\theta}[\hat{\theta}] = 0$$

$$E[\hat{\theta}] - \theta = 0$$

$$E[\hat{\theta}] = \theta$$

For example, the sample mean is unbiased estimator. Suppose Y_1, \dots, Y_m is iid random variables, we knew that $\bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_i$. it can be shown as:

$$\begin{aligned} E(\bar{Y}) &= E\left(\frac{1}{m} \sum_{i=1}^m Y_i\right) \\ &= E\left(\frac{1}{m}(Y_1 + Y_2 + \dots + Y_m)\right) \\ &= \frac{1}{m}(E(Y_1) + E(Y_2) + \dots + E(Y_m)) \\ &= \frac{1}{m}(\mu + \mu + \dots + \mu) \\ &= \frac{1}{m}m\mu \\ &= \mu \end{aligned}$$

where $E(Y_i) = \mu$. Therefore, we can see that sample mean is unbiased estimator.

2.3.2 Mean Square Error

The error is measurement of the difference between the estimate of sample parameter and the exact value of population parameter. The mean squared error (MSE) of an estimator is average of squared difference between the estimate and the expected value of parameter. MSE is most likely greater than zero due to sampling randomization. In other words, the random sample for our estimator can only carry the data which can improve the precision of estimator (lahmann, 2004).

In real world, If we are interesting in some topics (say the average income of citizen in Beijing), it can be very costly to get information of all citizen in Beijing. Over 20 millions data need to collect, that will cost too much money and time. The best solution is survey sampling. However, there is always some errors existed. In numerical case, the estimate of sample parameter minus the expected value of population parameter can be either positive or negative; after squared up the result, it would be positive.

Let Y_1, \dots, Y_m be a random sample, λ is an estimator of $G(Y_1, \dots, Y_m)$. The MSE can be defined by $MSE_\lambda(G) = E_\lambda[(G - \lambda^2)]$.

We can expand the square as:

$$\begin{aligned}
MSE_\lambda(G) &= E_\lambda[(G - \lambda)^2] \\
&= E_\lambda[(G - E_\lambda(G) + E_\lambda(G) - \lambda)^2] \\
&= E_\lambda[G - E_\lambda(G)]^2 + 2E_\lambda[G - E_\lambda(G)][E_\lambda(G) - \lambda] + [E_\lambda(G) - \lambda]^2 \\
&= E_\lambda[G - E_\lambda(G)]^2 + 2(E_\lambda(G) - \lambda)(E_\lambda(G) - E_\lambda(G)) + [E_\lambda(G) - \lambda]^2 \\
&= E_\lambda[G - E_\lambda(G)]^2 + [E_\lambda(G) - \lambda]^2 \\
&= Var_\lambda(G) + (Bias_\lambda(G))^2
\end{aligned}$$

where $Bias_\lambda = E_\lambda(G) - \lambda$.

If G is an unbiased estimator of λ , the MSE will be equal to variance of G . We can notate as:

$$MSE_\lambda(G) = Var_\lambda(G).$$

2.4 Method of Moment

The method of moment is the very first method to obtain population parameters. it was first introduced by Chebyshev. it generally set k unknown parameters $\theta_1, \theta_2, \dots, \theta_k$ is found by solving the system of k equations obtained by equating the first k sample moments with first population moments. so that is

$$M_i = \mu_i = \frac{1}{n} \sum_{j=1}^n X_j^i$$

sample i^{th} moment.

$$\mu_i = E(X^i)$$

shown as the population i^{th} moment. Thus, the method of moment estimator can be solving $M_i = \mu_i$ where $i=1,2,\dots,k$.

2.4.1 Sample Mean and Sample Variance Estimation

The sample mean and sample variance are very significant element to any statistically analysis. The sample mean The first sample moment $\acute{M}_1 = \frac{1}{n} \sum_{k=1}^n X_k$ is called the sample mean and usually denoted as \bar{X} . Then the sample variance S^2 can be obtained by second sample moment $M_2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2$. To have the value of method of moment estimators, setting equation system of

$$\mu = E(X) = \frac{1}{n} \sum_{k=1}^n X_k$$

and

$$\sigma^2 = E[(X - \mu)^2] = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2$$

then solve for parameters.

2.5 Maximum Likelihood Estimation

The Maximum Likelihood Method is another prevalent estimation method, and it estimates the population parameters by finding the maximum values of the likelihood equations. The value can be a point in the parameter space, so Maximum Likelihood Estimation is one of the major method of point estimation(Rossi, 2018). General steps of Maximum Likelihood Estimation: First, we need to make the the likelihood function is differentiable. The Second step is taking the first derivative of the likelihood function. Then, setting it equal to Zero and solving for the estimator. The last step is testing whether its second derivative is less than Zero.

Mathematically, the Maximum Likelihood Estimators can be written as:

Say, if we have a Random sample of (Y_1, Y_2, \dots, Y_i) , then the pdf or pmf can be shown as $f(x_j; \theta)$. The likelihood function is

$$\begin{aligned} L(\theta|y_1, y_2, \dots, y_i) &= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_i = y_i) \\ &= f(y_1; \theta) \cdot f(y_2; \theta) \cdots f(y_i; \theta) = \prod_{j=1}^i f(y_j; \theta). \end{aligned}$$

In this case, we denote $\hat{\theta}$ as the Maximum Likelihood Estimators of θ .

2.6 Delta Method

The Delta method is used as tool of driving the variance of asymptotic distribution. In some books, the author gives a general idea of Delta method,such as (Cramér,

2016). One of the simplest way can be defined as:

suppose \bar{Z}_n has an asymptotic normal distribution, so $\sqrt{n}(\bar{Z}_n - \alpha) \xrightarrow{d} N(0, \beta^2)$ as $n \rightarrow \infty$. If W is a continuous and differentiable function, and $W'(\alpha) \neq 0$, then

$$\sqrt{n}(W(\bar{Z}_n) - W(\alpha)) \xrightarrow{d} N(0, (W'(\alpha))^2 \beta^2).$$

Since $W'(\alpha) \neq 0$, we can derive it as:

$$\frac{\sqrt{n}(W(\bar{Z}_n) - W(\alpha))}{|W'(\alpha)|\beta} \xrightarrow{d} N(0, 1)$$

as $n \rightarrow \infty$.

2.6.1 Convergence in Distribution

Convergence in distribution is also named as converge weakly. It converges very slowly to a “better” distribution. In other words, when n small, it can be very different to the current distribution. And, as n goes up, it will be more and more closer to the distribution with more and more slower speed. Formly, it can be written as:

suppose $G_Y(y)$ is distribution function of any random variable Y , $\{Y_n, n \geq 1\}$ is a sequence of random variables and $-\infty < y < \infty$. $Y_n \xrightarrow{d} Y$ as $n \rightarrow \infty$. iff

$$\lim_{n \rightarrow \infty} F_{Y_n}(y) = F_Y(y), \text{ for all } y \text{ such that } F_Y \text{ is continues.}$$

2.6.2 Slutsky's Theorem

The Slutsky's Theorem gave idea of operation for convergent sequence of random variables (Goldberger, 1964). If $A_n \rightarrow A$ in distribution and $B_n \rightarrow c$ in probability, where c is constant, then

1. $A_n B_n \xrightarrow{d} Ac$
2. $A_n + B_n \xrightarrow{d} A + c$
3. $\frac{A_n}{B_n} \xrightarrow{d} A/c$, where $c \neq 0$

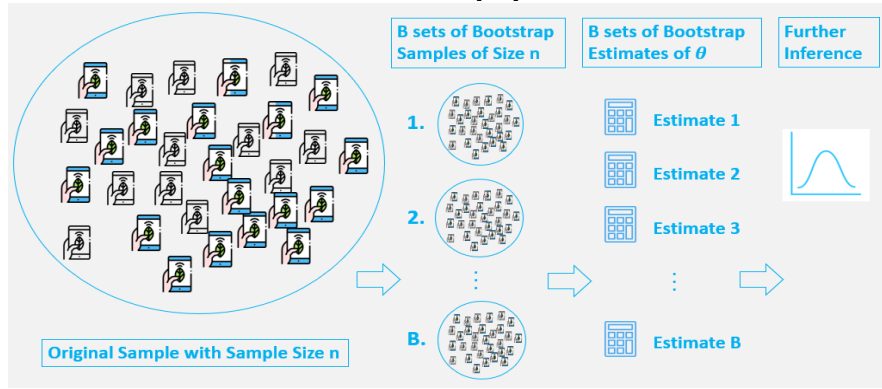
2.7 Bootstrap

Sampling often is the first step of a statistical analysis. The bootstrap method is an important method of sampling. The propose of applying bootstrap method is making scientific decision of the population parameters. The bootstrap method was first shown to the public in 1979 from a book "Bootstrap methods: another look at the jackknife" authored by Bradley Efron (Efron, 1979). The characteristic of bootstrap method is a re-sampling method rather than sampling method for the original sample, that is directly from the population. Because it is a re-sampling method, the MSE of estimator reduced manifestly and relevantly. Generally, some steps of bootstrap as follow:

1. Set the original sample with sample size n from population

2. Take B sets of bootstrap sample with sample size m
3. Find the estimate of all B bootstrap samples
4. Make statistical inference (Confidence intervals)

Figure 2.1: Figure shows the general steps of bootstrapping from towards data science [16].



I am going to apply the Bootstrap method for my thesis, because I am interesting in if the bootstrap sampling can reducing the errors of the estimate for both parameters m and p of binomial distribution. Furthermore, the bootstrap has two re-sampling methods, the independent bootstrap and dependent bootstrap. I will set the Confidence Regions for each independent bootstrap and dependent bootstrap samples. Moreover, the Probability of coverage of Confidence Region is a numerical standard to clarify the precision of estimation.

2.7.1 Independent Bootstrap

In statistics, we make inference and conclusion on sample, but the bootstrap re-sampling method play the same game on bootstrap samples. Bootstrap samples can be taking in two ways. The independent bootstrap method as a re-sampling method need to have a sample directly from population called original sample. Often, a simple random sampling works well in this situation. Mathematically, the independent bootstrap may written as follow:

Suppose, there exists an unknown distribution G , a simple random sample $Y = (y_1, \dots, y_n)$. For the Bootstrap sample $Y^* = (y_1^*, \dots, y_m^*)$, the value of estimate of sample parameter should be equal to the estimate of Bootstrap sample parameter $\hat{\alpha} = g(Y) = g(Y^*) = \hat{\alpha}^*$.

The independent bootstrap randomly selects variables with replacement from original sample of n iid random variables from population. All variables have $\frac{1}{n}$ probability to be selected, and we stop selecting once we have m random variables. An independent bootstrap sample with sample size m is completed. For example, suppose a sequence of 1, 2, 3, 4, 5, 6, 7, 8, 9, 0 as an original sample where $n = 10$. An independent bootstrap sample with $m = 20$ can be 1, 1, 2, 2, 4, 4, 5, 6, 0, 2, 7, 8, 4, 6, 7, 8, 9, 5, 2, 0. An extreme sample of 20 ones can be exist with probability of $(\frac{1}{10})^{20}$. In Statistical analysis, independent bootstrap sample size m is often greater than or equal to the

original sample size $n(m \geq n)$. The independent bootstrap CI can be written as

$$CI = \left(\hat{\theta} + \hat{\sigma}_{\theta} \cdot t_{(\frac{\alpha}{2})}^*, \hat{\theta} + \hat{\sigma}_{\theta} \cdot t_{(1-\frac{\alpha}{2})}^* \right).$$

Where

$$t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\hat{\sigma}_{\theta b}^*}.$$

The independent bootstrap sample mean and standard deviation as follow:

$$\hat{\theta}_b^* = \bar{y}_b^* = \frac{1}{m} \sum_{j=1}^m y_{jb}^*,$$

and

$$\hat{\sigma}_{\theta b}^* = \sqrt{\frac{\sum_{j=1}^m (y_{jb}^* - \bar{y}_b^*)^2}{m-1}}.$$

2.7.2 Dependent Bootstrap

The dependent bootstrap has same propose of reducing the variety of sample estimator, but it design in different re-sampling method. There are 3 differences as following:

1. selecting dependent bootstrap sample with replacement
2. selecting from k copies of original sample
3. the sample size of dependent bootstrap sample will less than or equal to k copies
time the original sample size($m \leq nk$)

Due to the differences, dependent bootstrap sample actually has lower MSE of the estimator. Especially, k is small, that decreases more variation compare with larger k . In other words, as k increases, dependent bootstrap sample is more and more closer to independent bootstrap sample. If $k \rightarrow \infty$, independent bootstrap and dependent bootstrap appear similar characteristic.

The same example for the independent bootstrap, suppose $k = 5$ and $m = 20$ as well. 1, 1, 2, 2, 4, 4, 5, 6, 0, 2, 7, 8, 4, 6, 7, 8, 9, 5, 2, 0 can also be a dependent bootstrap sample; however, the extreme sample of 20 ones does not Satisfy the nature of dependent bootstrap (Smith and Taylor, 2001).

The dependent bootstrap CI can be written as

$$CI = \left(\hat{\theta} + \hat{\sigma}_{\theta} \cdot t_{(\frac{\alpha}{2})}^{\bullet}, \hat{\theta} + \hat{\sigma}_{\theta} \cdot t_{(1-\frac{\alpha}{2})}^{\bullet} \right).$$

Where

$$t_b^{\bullet} = \frac{\hat{\theta}_b^{\bullet} - \hat{\theta}}{\hat{\sigma}_{\theta b}^{\bullet}}.$$

The dependent bootstrap sample mean and standard deviation as follow:

$$\hat{\theta}_b^{\bullet} = \bar{y}_b^{\bullet} = \frac{1}{m} \sum_{j=1}^m y_{jb}^{\bullet}$$

and

$$\hat{\sigma}_{\theta b}^{\bullet} = \sqrt{\frac{\sum_{j=1}^m (y_{jb}^{\bullet} - \bar{y}_b^{\bullet})^2}{m-1}} \sqrt{\frac{kn-m}{m(kn-1)}}$$

Where $\frac{kn-m}{kn-1}$ is the finite population correction factor.

2.8 Coverage Probability

The coverage probability is proportion of times of the exact value of estimates inside of the confidence interval. But we often calculate the number of times of estimates which are outside of the confidence interval, the coverage probability equal to one minus the number over the total sample size. Suppose the population mean $E(x)$ as the interest of any distribution, the coverage probability is

$$P(L < \mathbb{E}[X] < U) = 1 - \frac{\#\{\hat{L} > \mathbb{E}[X]\} + \#\{\hat{U} < \mathbb{E}[X]\}}{\#\{\hat{C}\}},$$

For the numerator, the number of $E(x)$ less than lower limit plus the the number of $E(x)$ greater than upper limit. The number of all CIs are the denominator. In general, the coverage probability is a standard of the design of experiment. The coverage probability approximately equal to the confidence level appears a good result of experiment. Even though coverage probability slightly greater or less than confidence level is acceptable, I am more desirable of less than confidence level.

2.9 Chi-Square Distribution

The Chi-square distribution has very close relation with normal distribution. When mean equal to 0 and variance equal to 1, a special case of normal distribution called the standard normal distribution. The Chi-square distribution contain the summation of ν iid random variables of standard normal distribution, where ν

equal to the degree of freedom of Chi-square distribution. The Chi-square distribution with ν degree of freedom has a pdf

$$f(y) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}y^{\frac{\nu}{2}-1}\exp(-\frac{y}{2})$$

where $y > 0$, otherwise, $f(y) = 0$.

2.9.1 Chi-Squared Test

The chi-squared test is normally used for goodness of fit or test of independency. In this paper, the confidence region asymptotically distributed as chi-square distribution with 2 degree of freedom, since the confidence region is two-dimensional (m and p). Similar to any other statistical test, whether the null hypothesis is current at α level of significant.

Some general steps for chi-squared test:

1. set the null hypothesis versus alternative hypothesis
2. calculate the test statistic
3. calculate the degree of freedom
4. compare the test statistic with value of chi-square with degree of freedom
5. conclude whether the null hypothesis is rejected

Reject region: $\chi^2 \geq \chi_{1-\alpha}^2(df)$.

Chapter 3

Research Methods

3.1 Introduction of Research Methods

There are several methodologies in this research, include Binomial distribution with unknown parameter (m, p) , method of moment estimation, Asymptotic Normality of the Estimators, Delta method, confidence regions, bootstrap sampling method. The theoretical results will be proved in this chapter. The aim of this chapter is comparing the confidence regions of the independent and dependent bootstrap with different k . In addition, the effect of probability of coverage of confidence regions with different value of population parameter m and p . The methods and general simulation results can be applies in this chapter, but I will collect all tables and plots of the simulation results in chapter 4.

3.2 Bivariate Normal Distribution

According to Colin Rose, suppose $\{X_1, \dots, X_n\}$ are iid random samples follow normal distribution (μ_X, σ_X^2) , and $\{Y_1, \dots, Y_n\}$ are iid random samples follow normal distribution (μ_Y, σ_Y^2) . The Bivariate Normal Distribution is distributed by vector $\begin{pmatrix} X \\ Y \end{pmatrix}$ (2011). The pdf can be written as:

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp \left\{ -\frac{1}{1-\rho^2} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \right] \right\}.$$

Some the mathematical results:

$$EX_i = \mu_X, EY_i = \mu_Y, \text{Var}(X_i) = \sigma_X^2, \text{Var}(Y_i) = \sigma_Y^2 \text{ and } \text{Cov}(X_i, Y_i) = \rho\sigma_X\sigma_Y,$$

Where the correlation coefficient $\rho = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$. Thus, the covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

and the mean vector is

$$\vec{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}.$$

According to delta method, $\sqrt{n}(\bar{Z}_n - \alpha) \xrightarrow{d} N(0, \beta^2)$ as $n \rightarrow \infty$, we can easily show that as $n \rightarrow \infty$.

$$\sqrt{n} \begin{pmatrix} \bar{X}_n - \mu_X \\ \bar{Y}_n - \mu_Y \end{pmatrix} \xrightarrow{d} N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right).$$

where $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ and $\bar{Y}_n = \frac{1}{n} \sum_{k=1}^n Y_k, n \geq 1$.

3.3 Method of Moment Estimator

Suppose $\{Y_1, \dots, Y_n\}$ are iid random samples follow binomial distribution (m, p) .

We can find the mean and variance as follow:

$$\bar{Y} = \frac{1}{n} \sum_{k=1}^n Y_k \text{ and } S^2 = \frac{1}{n-1} \sum_{k=1}^n (Y_k - \bar{Y})^2$$

. The population mean and variance of binomial distribution are:

$$E(y) = mp, Var(y) = mp(1 - p).$$

Theorem 3.3.1 *The estimators of the parameters m and p by the Method of Moments are*

$$\hat{p}_n = \frac{\bar{Y} - S^2}{\bar{Y}}, \quad \hat{m}_n = \frac{\bar{Y}^2}{\bar{Y} - S^2}.$$

Proof. By method of moment, $M_i = \mu_1 = \frac{1}{n} \sum_{k=1}^n X_k^i$. The i^{th} moment are:

$$\mu_i = E(X^i) = M_i.$$

Since $E(y) = mp, Var(y) = mp(1 - p)$.

$$\bar{Y} = mp$$

$$S^2 = mp(1 - p).$$

Solving these equations

$$S^2 = mp(1 - p) = \bar{Y}(1 - p)$$

$$S^2 = \bar{Y} - \bar{Y}p$$

$$\bar{Y}p = \bar{Y} - S^2$$

so $\hat{p} = \frac{\bar{Y} - S^2}{\bar{Y}}$. Then

$$\bar{Y} = mp$$

and $\hat{p} = \frac{\bar{Y} - S^2}{\bar{Y}}$

$$\hat{p} = m \left(\frac{\bar{Y} - S^2}{\bar{Y}} \right), \text{ then } \hat{m} = \frac{\bar{Y}^2}{\bar{Y} - S^2}.$$

QED

3.4 The First Four Moments of Binomial Distribution

The moment generating function is the common way to find the moments. Let $Y = \{Y_1, \dots, Y_n\}$ follows Binomial distribution, the MGF of Binomial distribution:

$$M_Y(t) = [(1 - p) + pe^t]^m.$$

Proof The First four moments of the Binomial distribution are

$$E(Y) = mp, \quad E(Y - EY)^2 = mp(1 - p), \quad E(Y - EY)^3 = mp(1 - p)(1 - 2p),$$

$$E(Y - EY)^4 = 3m^2p^2(1 - p)^2 + mp(1 - p)(1 - 6p(1 - p)),$$

Taking derivative of MGF:

$$M'_Y(t) = m[1 - p + pe^t]^{m-1}(pe^t)$$

setting $t = 0$

$$\begin{aligned} M'_Y(0) &= m[1 - p + pe^0]^{m-1}(pe^0) \\ &= m[1 - p + p]^{m-1}(p) \\ &= m(1)^{m-1}p \\ &= mp \end{aligned}$$

Thus, $\mu = M_1 = \mu'_1 = mp$.

Taking second derivative of MGF:

$$M''_X(t) = m[1 - p + pe^t]^{m-1}(pe^t) + (pe^t)m(m-1)[1 - p + pe^t]^{m-2}(pe^t)$$

setting $t = 0$

$$\begin{aligned} M''_Y(0) &= m[1 - p + pe^0]^{m-1}(pe^0) + (pe^0)m(m-1)[1 - p + pe^0]^{m-2}(pe^0) \\ &= m[1 - p + p]^{m-1}(p) + pm(m-1)[1 - p + p]^{m-2}p \\ &= m(1)^{m-1}(p) + pm(m-1)1^{m-2}p \\ &= mp + m(m-1)p^2 \end{aligned}$$

Since $E(Y)^2 = mp^2$, the variance is

$$Var(Y) = \mu_2 = E(Y^2) - (EY)^2 = mp + m(m-1)p^2 - (mp)^2 = mp + (mp)^2 - mp^2 - (mp)^2 = mp(1-p).$$

Taking second derivative of MGF:

$$M'''(t) = [1 - p + pe^t]^{m-1} m (pe^t) + m (pe^t) (m-1) [1 - p + pe^t]^{m-2} (pe^t) \\ + p^2 e^{2t} m (m-1) (m-2) [1 - p + pe^t]^{m-3} (pe^t) + [1 - p + pe^t]^{m-2} m (m-1) (2p^2 e^{2t})$$

again setting $t = 0$.

$$M'''(0) = mp(1 + 3mp - 3p - 3mp^2 + 2p^2 + (mp)^2)$$

Then $E(Y^3) = \mu'_3 = mp(1 + 3mp - 3p - 3mp^2 + 2p^2 + (mp)^2)$. The formula of third moment:

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2\mu'_3$$

then μ_3 is given as following :

$$E(Y - EY)^3 = \mu_3 = mp(1 - p)(1 - 2p).$$

Where $\mu'_1 = mp$ and $\mu'_2 = mp + m(m-1)p^2$.

Taking fourth derivative of MGF:

$$M^{iv}(t) = m(m-1)(m-2)(m-3)p^4 e^{4t} [pe^t + (1-p)]^{m-4} \\ + 3m(m-1)(m-2)e^{3t} p^3 [pe^t + (1-p)]^{m-3} + 3m(m-1)(m-2)p^2 e^{2t} [pe^t + (1-p)]^{m-2} \\ + m(m-1)p^2 e^{2t} [pe^t + (1-p)]^{m-2} + mpe^t [pe^t + (1-p)]^{m-1}$$

again setting $t = 0$

$$E(X^4) = \mu'_4 = mp [m^3 p^3 - 6m^2 p^3 + 11mp^3 - 6p^3 + 6m^2 p^2 - 18mp^2 + 12p^2 + 7mp - 7p + 1]$$

The formula of fourth moment:

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_1'^2\mu_2' - 3\mu_1'^4$$

then μ_4 is given as following :

$$E(X - EX)^4 = \mu_4 = 3m^2p^2(1-p)^2 + mp(1-p)[1 - 6p(1-p)].$$

QED

3.5 Delta Method

According salma's thesis, $\sqrt{n} \begin{pmatrix} \hat{p}_n - p \\ \hat{m}_n - m \end{pmatrix}$ for $n \rightarrow \infty$ converges to the bivariate normal distribution with the zero mean vector and covariance matrix $\Sigma = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}$ (Saad, 2019). Then, plug in the value of the First four moments: $\sigma^2 = \mu_2 = mp(1-p), \mu_3 = mp(1-p)(1-2p)$ For $\mu_4 - \sigma^4$ can be written as:

$$\begin{aligned} \mu_4 - \sigma^4 &= 3m^2p^2(1-p)^2 + mp(1-p)[1 - 6p(1-p)] - (mp(1-p))^2 \\ &= 2m^2p^2(1-p)^2 + mp(1-p)[1 - 6p(1-p)] \end{aligned}$$

therefore

$$\tilde{\Sigma} = \begin{pmatrix} mp(1-p) & mp(1-p)(1-2p) \\ mp(1-p)(1-2p) & 2m^2p^2(1-p)^2 + mp(1-p)(1-6p(1-p)) \end{pmatrix}.$$

Apply the of delta method of \hat{p}_n and \hat{m}_n , We chnage variables and set $k_1 = \bar{Y}$ and

$k_2 = S^2$. Therefore, in this term

$$\hat{p}_n = g_1(\bar{Y}, S^2) = \frac{\bar{Y} - S^2}{\bar{Y}}, \text{ that is, } g_1(k_1, k_2) = \frac{k_1 - k_2}{k_1}$$

and

$$\hat{m}_n = g_2(\bar{Y}, S^2) = \frac{\bar{Y}^2}{\bar{Y} - S^2}, \text{ that is, } g_2(k_1, k_2) = \frac{k_1^2}{k_1 - k_2}.$$

Partial derivatives:

$$\begin{aligned} \frac{\partial g_1(k_1, k_2)}{\partial k_1} &= \frac{k_2}{k_1^2} \\ \frac{\partial g_1(k_1, k_2)}{\partial k_2} &= -\frac{1}{k_1} \\ \frac{\partial g_2(k_1, k_2)}{\partial k_1} &= \frac{k_1(k_1 - 2k_2)}{(k_1 - k_2)^2} \\ \frac{\partial g_2(k_1, k_2)}{\partial k_2} &= \frac{k_1^2}{(k_1 - k_2)^2} \end{aligned}$$

Also taking into consideration that $\mu_1 = E(\bar{Y}) = E(Y) = mp$ and $\mu_2 = E(S^2) =$

$\text{Var}(Y) = mp(1 - p)$ we can write that

$$\begin{aligned} g_1(\mu_1, \mu_2) &= \frac{mp - mp(1 - p)}{mp} = p \\ g_2(\mu_1, \mu_2) &= \frac{(mp)^2}{mp - mp(1 - p)} = m \\ \frac{\partial g_1(\mu_1, \mu_2)}{\partial t_1} &= \frac{mp(1 - p)}{(mp)^2} = \frac{1 - p}{mp} \\ \frac{\partial g_1(\mu_1, \mu_2)}{\partial t_2} &= -\frac{1}{mp} \\ \frac{\partial g_2(\mu_1, \mu_2)}{\partial t_1} &= \frac{mp(mp - 2mp(1 - p))}{(mp - mp(1 - p))^2} = \frac{2p - 1}{p^2} \\ \frac{\partial g_2(\mu_1, \mu_2)}{\partial t_2} &= -\frac{(mp)^2}{(mp - mp(1 - p))^2} = -\frac{1}{p^2}. \end{aligned}$$

By the Delta method, the random vector $\sqrt{n} \begin{pmatrix} \hat{p}_n - p \\ \hat{m}_n - m \end{pmatrix}$ is asymptotically normal $N(\vec{0}, \Sigma)$ with zero mean $\vec{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix $\Sigma = B\tilde{\Sigma}B'$, where

$$B = \begin{pmatrix} \frac{1-p}{mp} & -\frac{1}{mp} \\ \frac{2p-1}{p^2} & \frac{1}{p^2} \end{pmatrix}$$

And

$$\tilde{\Sigma} = \begin{pmatrix} (mp(1-p))^2 & mp(1-p)(1-2p) \\ mp(1-p)(1-2p) & 2m^2p^2(1-p)^2 + mp(1-p)(1-6p(1-p)) \end{pmatrix}.$$

solving the matrix equation we can easily get:

$$\Sigma = \begin{pmatrix} \sigma_p^2 & \rho\sigma_p\sigma_m \\ \rho\sigma_p\sigma_m & \sigma_m^2 \end{pmatrix},$$

where $\sigma_p^2 = \frac{(1-p)(a+p)}{m}$, $\sigma_m^2 = \frac{m(1-p)a}{p^2}$,

$\rho = -\sqrt{\frac{a}{a+p}}$, and $a = 2(1-p)(m-1)$.

3.6 Confidence Region

The Confidence Region of Binomial distribution is the major part for research. Some the theorem was proved by lehmann of the statistical inference of parameters p and m (Lehmann, 2004).

3.6.1 Confidence Region for Binomial Distribution

Suppose $\vec{Y} = (Y_1, Y_2, \dots, Y_n)$ is a random vector follows normal distribution with mean $\vec{\nu} = (\nu_1, \nu_2, \dots, \nu_n)$, and covariance matrix Σ . Thus, $(\vec{Y} - \vec{\nu})^T \Sigma^{-1} (\vec{Y} - \vec{\nu})$ follows chi-square distribution with n degree freedom (Lehmann, 2004).

It is clear that the binomial distribution has two parameters, so n=2 for Method of moment estimates. It can be show that the mean vector $\vec{\nu} = \begin{pmatrix} p \\ m \end{pmatrix}$. From previous page, the covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_p^2 & \rho\sigma_p\sigma_m \\ \rho\sigma_p\sigma_m & \sigma_m^2 \end{pmatrix}$$

The inverse matrix:

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{\det[\Sigma]} \begin{pmatrix} \sigma_p^2 & \rho\sigma_p\sigma_m \\ \rho\sigma_p\sigma_m & \sigma_m^2 \end{pmatrix}^T \\ &= \frac{1}{\sigma_p^2\sigma_m^2 - \rho^2\sigma_p^2\sigma_m^2} \begin{pmatrix} \sigma_m^2 & -\rho\sigma_p\sigma_m \\ -\rho\sigma_p\sigma_m & \sigma_p^2 \end{pmatrix} \\ &= \frac{1}{1 - \rho^2} \begin{pmatrix} \frac{1}{\sigma_p^2} & -\frac{\rho}{\sigma_p\sigma_m} \\ -\frac{\rho}{\sigma_p\sigma_m} & \frac{1}{\sigma_m^2} \end{pmatrix} \end{aligned}$$

Now plug in the value of $\sigma_p^2 = \frac{(1-p)(a+p)}{m}$, $\sigma_m^2 = \frac{m(1-p)a}{p^2}$,

$$\rho = -\sqrt{\frac{a}{a+p}}, \text{ and } a = 2(1-p)(m-1).$$

$$\begin{aligned}
\Sigma^{-1} &= \frac{1}{1 - \frac{a}{a+p}} \begin{pmatrix} \frac{m}{(1-p)(a+p)} & -\frac{\sqrt{\frac{a}{a+p}}}{\sqrt{\frac{(1-p)(a+p)}{m}} \sqrt{\frac{m(1-p)a}{P^2}}} \\ -\frac{\sqrt{\frac{a}{a+p}}}{\sqrt{\frac{(1-p)(a+p)}{m}} \sqrt{\frac{m(1-p)a}{P^2}}} & \frac{p^2}{m(1-p)a} \end{pmatrix} \\
&= \frac{a+p}{p} \begin{pmatrix} \frac{m}{(1-p)(a+p)} & -\frac{\sqrt{\frac{a}{a+p}}}{\sqrt{\frac{(1-p)^2(a+p)a}{P^2}}} \\ -\frac{\sqrt{\frac{a}{a+p}}}{\sqrt{\frac{(1-p)^2(a+p)a}{P^2}}} & \frac{p^2}{m(1-p)a} \end{pmatrix} \\
&= \frac{a+p}{p} \begin{pmatrix} \frac{m}{(1-p)(a+p)} & -\sqrt{\frac{aP^2}{(a+p)^2 a (1-p)^2}} \\ -\sqrt{\frac{aP^2}{(a+p)^2 a (1-p)^2}} & \frac{p^2}{m(1-p)a} \end{pmatrix} \\
&= \frac{a+p}{p} \begin{pmatrix} \frac{m}{(1-p)(a+p)} & -\frac{P}{(a+p)(1-p)} \\ -\frac{P}{(a+p)(1-p)} & \frac{p^2}{m(1-p)a} \end{pmatrix} \\
&= \frac{1}{p(1-p)} (1-p)(p+a) \begin{pmatrix} \frac{m}{(1-p)(a+p)} & -\frac{P}{(a+p)(1-p)} \\ -\frac{P}{(a+p)(1-p)} & \frac{p^2}{m(1-p)a} \end{pmatrix} \\
&= \frac{1}{p(1-p)} \begin{pmatrix} m & -p \\ -p & \frac{p^2(p+a)}{ma} \end{pmatrix}
\end{aligned}$$

3.6.2 Test Statistic of Confidence Region

The statistics is asymptotically distributed to Chi-square distribution. Since there are two unknown parameters, degree of freedom is two. The mathematical form

follows:

$$X_2^2(p, m) = \frac{1}{p(1-p)} \begin{pmatrix} \hat{p} - p & \hat{m} - m \end{pmatrix} \begin{pmatrix} m & -p \\ -p & \frac{p^2(p+a)}{ma} \end{pmatrix} \begin{pmatrix} \hat{p} - p \\ \hat{m} - m \end{pmatrix}$$

simplify the matrix:

$$X_2^2(p, m) = \frac{1}{p(1-p)} \left(m(\hat{p} - p)^2 - 2p(\hat{p} - p)(\hat{m} - m) + \frac{p^2(p+a)}{ma}(\hat{m} - m)^2 \right)$$

follows chi-square distribution with 2 degree of free asymptotically.

For a α level of significance, the $100(1 - \alpha)\%$ confidence region shall inside of this ellipse $\{(p, m) | X_2^2(p, m) \leq X_2^2(\alpha)\}$.

3.7 The Performance of Confidence Region

There are total 1000 binomial samples as original sample. Then, $B = 2000$ independent bootstrap samples, and $B = 2000$ dependent bootstrap samples with $k = 5, 10, 25, 50, 100, 500, 1000$. To evaluate the performance, the coverage of probability of confidence region, δ , and area of confidence region, \hat{A} , will be applied.

3.7.1 Coverage Probability of Confidence Region

From section 2.8, the coverage probability were derived as

$$P(L < \mathbb{E}[X] < U) = 1 - \frac{\#\{\hat{L} > \mathbb{E}[X]\} + \#\{\hat{U} < \mathbb{E}[X]\}}{\#\{\hat{C}\}}$$

However, the coverage of probability of confidence region, δ approximately equal to the proportion which is the number of total confidence regions less or equal to critical value of chi-square distribution with 2 degree of freedom at significant level α , over the total number of confidence regions. Since each bootstrap sample can determine a confidence region, the total number of confidence regions will equal to total bootstrap samples. In this research, the total bootstrap samples $B = 2000$. We can write the mathematical form:

$$\delta \approx P(X_2^2(p, m) \leq X_2^2(\alpha)) = \frac{\#\{X_2^2(p, m) \leq X_2^2(\alpha)\}}{\#\{X_2^2(p, m)\}}.$$

Furthermore, $\delta \approx (1 - \alpha)$. It will accept that the coverage probability slightly less or greater than confidence level, but I am more desirable of less than confidence level. As a reason, the errors exist and they will change the proportion; $\delta > (1 - \alpha)$ will be considered as being perfect and perfection does not exist in real the world.

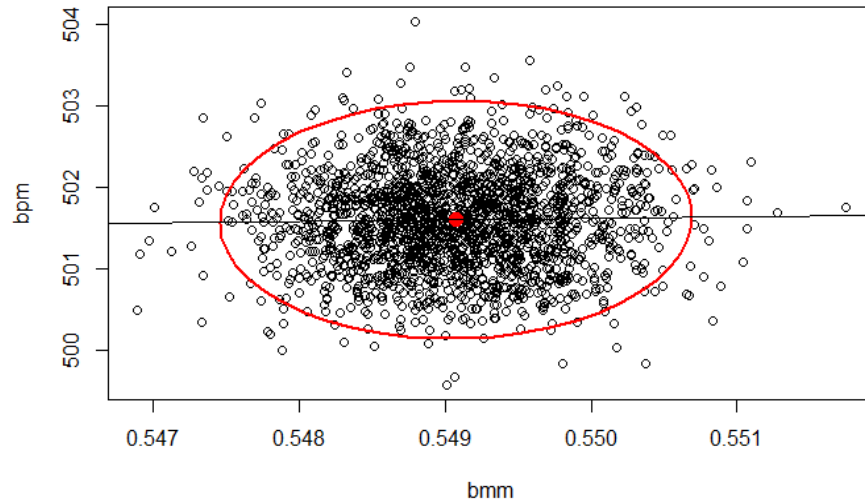
3.7.2 Area of Confidence Region

The mean of parameters \hat{p} and \hat{m} of each bootstrap samples are used to find the area of confidence region. The areas of confidence region of each method (independent bootstrap samples, dependent bootstrap samples with different k) are evidences for effectiveness of each method. To determine the area of confidence region, we need three steps.

1. find the center of confidence region

2. calculate the maximum and minimum distance on the confidence region to center
3. area of confidence region equals to π times the maximum and minimum distance on the confidence region to center.

Figure 3.1: Figure shows the confidence region of mean of \hat{p} and \hat{m} from 2000 independent bootstrap sample follows $\text{Bin}(500, 0.55)$



The area $\hat{A} = \pi ab = 0.1289$ where a is the maximum distance on the confidence region to center, and b is the minimum distance on the confidence region to center.

Chapter 4

Statistical Simulation

4.1 Introduction

The simulation study generated $n = 1000$ random samples from $Bin(m, p)$. Then, 8 different re-sampling methods are independent bootstrap and dependent bootstrap when $k = 5, 10, 25, 50, 100, 250, 500$. Each bootstrap samples has same sample size $m = 1000$ and total $B = 2000$ bootstrap samples in one experiment. For one experiment, 16,000 samples are generated, and in total of 16,000,000 random variables. Moreover, 10 values for population parameter p , and 4 values for population parameter m , so there are 40 experiments. In this research, 40 experiments need to generate 640,000 samples and 640,000,000 random variables.

4.2 Simulation

In this section, I will fix the value for population parameters p and m . Let p equal to 0.05 to 0.90 by 0.05 and m equal to 100; 250; 500; 750. This gives combinations of $10 \times 4 = 40$ of 2 parameters.

I will set up tables of the mean and median of coverage probabilities of different values of p and m for independent bootstrap samples and dependent bootstrap samples with $k = 5, 10, 25, 50, 100, 250, 500$.

4.3 Coverage Probability of 95% Confidence Region

The goal of this section is to find out that whether the coverage probability increases by the two methods of bootstrap re-sampling. In addition, at what value of p that the coverage probability will be closest to confidence level $1 - \alpha$. There are 8 tables of mean and median of coverage probability by different bootstrap methods.

For the independent bootstrap method, when $m = 100$, probability $p = [0.2, 0.25]$ have the best outcome of $\delta \approx (1 - \alpha)$. Similarly, that $m = 250, p = [0.35, 0.4]; m = 500, p = [0.55, 0.6]; m = 750, p = [0.65, 0.7]$. The similar result can be applied for dependent bootstrap method with different value of k , since different values of k slightly change coverage probabilities.

Table 4.1: Coverage probabilities of 95% confidence regions of Independent

Bootstrap samples

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Mean	Median	Mean	Median	Mean	Median	Mean	Median
0.05	0.5708	0.5710	0.392	0.392	0.2847	0.2840	0.2344	0.2340
0.1	0.7532	0.7530	0.5603	0.5610	0.4071	0.4070	0.3468	0.3470
0.15	0.8662	0.8660	0.6801	0.6800	0.5194	0.5200	0.4334	0.4340
0.2	0.9314	0.9310	0.746	0.746	0.6122	0.6120	0.5087	0.5090
0.25	0.9659	0.9660	0.8199	0.8200	0.6688	0.6690	0.5616	0.5620
0.3	0.9847	0.9850	0.8199	0.820	0.6688	0.6690	0.5616	0.5620
0.35	0.9938	0.9940	0.9265	0.9270	0.7808	0.7810	0.7001	0.700
0.4	0.9983	0.9980	0.9512	0.9510	0.818	0.818	0.7493	0.7490
0.45	0.9989	0.9990	0.9687	0.9690	0.8676	0.8680	0.7798	0.7800
0.5	0.9998	1	0.9853	0.9850	0.9137	0.9140	0.8244	0.8240
0.55	0.9999	1	0.9913	0.9910	0.9320	0.9321	0.8657	0.8660
0.6	1	1	0.9976	0.998	0.9559	0.9560	0.912	0.912
0.65	1	1	0.9991	0.9990	0.9785	0.9790	0.9469	0.9470
0.7	1	1	0.9994	1	0.9923	0.9920	0.9663	0.9670
0.75	1	1	1	1	0.9973	0.998	0.9833	0.9840
0.8	1	1	1	1	0.9984	0.9990	0.9954	0.9960
0.85	1	1	1	1	0.9999	1	0.9982	0.9980
0.9	1	1	1	1	1	1	0.9998	1

Table 4.2: Coverage probabilities of 95% confidence regions of dependent Bootstrap

samples $k = 5$

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.5700	0.5703	0.3920	0.3919	0.2850	0.2846	0.2340	0.2344
0.1	0.7530	0.7528	0.5600	0.5602	0.4080	0.4078	0.3470	0.3474
0.15	0.8660	0.8659	0.6795	0.6795	0.5190	0.5189	0.4330	0.4331
0.2	0.9310	0.9312	0.7460	0.7459	0.6130	0.6127	0.5090	0.5086
0.25	0.9660	0.9658	0.8200	0.8199	0.6690	0.6682	0.5620	0.5627
0.3	0.9850	0.9848	0.8790	0.8792	0.7340	0.7341	0.6400	0.6396
0.35	0.9940	0.9939	0.9270	0.9265	0.7810	0.7811	0.7000	0.6991
0.4	0.9990	0.9983	0.9510	0.9508	0.8180	0.8182	0.7500	0.7495
0.45	0.9990	0.9989	0.9690	0.9686	0.868	0.868	0.7800	0.7797
0.5	1.0000	0.9997	0.9850	0.9851	0.9140	0.9138	0.8240	0.8244
0.55	1	1	0.9910	0.9912	0.932	0.932	0.8660	0.8654
0.6	1	1	0.9980	0.9976	0.9560	0.9561	0.912	0.912
0.65	1	1	0.9990	0.9991	0.9560	0.9561	0.912	0.912
0.7	1	1	1	0.9994	0.9920	0.9923	0.9670	0.9665
0.75	1	1	1	1	0.9970	0.9973	0.9830	0.9833
0.8	1	1	1	1	0.9990	0.9984	0.9960	0.9954
0.85	1	1	1	1	1	0.9998	0.9980	0.9982
0.9	1	1	1	1	1	1	1	0.9998

Table 4.3: Coverage probabilities of 95% confidence regions of dependent Bootstrap

samples $k = 10$

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.5700	0.5705	0.3920	0.3916	0.2850	0.2845	0.2340	0.2344
0.1	0.7530	0.7527	0.5600	0.5603	0.4080	0.4073	0.347	0.347
0.15	0.8650	0.8655	0.6800	0.6801	0.5200	0.5192	0.4340	0.4335
0.2	0.9310	0.9311	0.7450	0.7451	0.7450	0.7451	0.5080	0.5075
0.25	0.9660	0.9657	0.8190	0.8194	0.6690	0.6688	0.5620	0.5619
0.3	0.9850	0.9847	0.8800	0.8799	0.7340	0.7338	0.6400	0.6403
0.35	0.9940	0.9938	0.9270	0.9265	0.781	0.781	0.6990	0.6991
0.4	0.9980	0.9983	0.9510	0.9507	0.8185	0.8185	0.7490	0.7494
0.45	0.9990	0.9989	0.9990	0.9989	0.9990	0.9989	0.780	0.780
0.5	1	0.9997	0.985	0.985	0.9140	0.9136	0.8245	0.8246
0.55	1	0.999	0.9910	0.9911	0.932	0.932	0.8660	0.8656
0.6	1	1	0.9980	0.9976	0.932	0.932	0.8660	0.8656
0.65	1	1	0.9990	0.9991	0.9790	0.9785	0.9470	0.9465
0.7	1	1	1	0.9994	0.9920	0.9923	0.9670	0.9663
0.75	1	1	1	1	0.9970	0.9973	0.9830	0.9832
0.8	1	1	1	1	0.9990	0.9984	0.9950	0.9953
0.85	1	1	1	1	1	0.9998	0.9980	0.9982
0.9	1	1	1	1	1	1	1	0.9998

Table 4.4: Coverage probabilities of 95% confidence regions of dependent Bootstrap

samples $k = 25$

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.5700	0.5705	0.3920	0.3918	0.2840	0.2840	0.2340	0.2343
0.1	0.7530	0.7526	0.7530	0.7526	0.4070	0.4073	0.3470	0.3468
0.15	0.8650	0.8653	0.6810	0.6803	0.5190	0.5192	0.4340	0.4336
0.2	0.9310	0.9307	0.746	0.746	0.6120	0.6122	0.5080	0.5083
0.25	0.9660	0.9657	0.8200	0.8194	0.6690	0.6685	0.5630	0.5628
0.3	0.9850	0.9845	0.8790	0.8789	0.734	0.734	0.6400	0.6402
0.35	0.9940	0.9939	0.9270	0.9267	0.7810	0.7809	0.7000	0.6995
0.4	0.9980	0.9983	0.9510	0.9509	0.8180	0.8185	0.7490	0.7494
0.45	0.9990	0.9989	0.9990	0.9989	0.8680	0.8679	0.7790	0.7791
0.5	1	0.9998	0.9850	0.9853	0.9140	0.9138	0.8240	0.8239
0.55	1	0.9999	0.9910	0.9911	0.932	0.932	0.8660	0.8656
0.6	1	1	0.9980	0.9976	0.956	0.956	0.9120	0.9121
0.65	1	1	0.9990	0.9991	0.9790	0.9785	0.9470	0.9465
0.7	1	1	1	0.9994	0.9930	0.9924	0.9660	0.9663
0.75	1	1	1	1	0.9970	0.9973	0.9830	0.9833
0.8	1	1	1	1	0.9990	0.9984	0.9960	0.9954
0.85	1	1	1	1	1	0.9998	0.9980	0.9982
0.9	1	1	1	1	1	1	1	0.9998

Table 4.5: Coverage probabilities of 95% confidence regions of dependent Bootstrap

samples $k = 50$

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.5700	0.5705	0.3920	0.3917	0.2840	0.2843	0.2340	0.2344
0.1	0.7530	0.7526	0.560	0.560	0.4070	0.4076	0.3470	0.3469
0.15	0.8650	0.8655	0.6790	0.6797	0.5180	0.5184	0.4330	0.4328
0.2	0.931	0.931	0.7460	0.7458	0.6130	0.6132	0.5080	0.5084
0.25	0.9660	0.9657	0.8200	0.8197	0.669	0.669	0.5630	0.5629
0.3	0.9850	0.9847	0.8790	0.8792	0.7340	0.7341	0.640	0.640
0.35	0.9940	0.9939	0.9260	0.9259	0.7810	0.7809	0.7000	0.6997
0.4	0.9980	0.9983	0.9510	0.9508	0.8185	0.8186	0.7500	0.7494
0.45	0.9990	0.9989	0.9680	0.9685	0.8680	0.8678	0.7800	0.7798
0.5	1	0.9998	0.9850	0.9851	0.9130	0.9136	0.8250	0.8247
0.55	1	0.9999	0.9910	0.9911	0.9320	0.9321	0.8660	0.8659
0.6	1	1	0.9980	0.9976	0.9560	0.9559	0.9120	0.9121
0.65	1	1	0.9990	0.9991	0.9790	0.9783	0.9460	0.9464
0.7	1	1	1	0.9994	0.9920	0.9922	0.9660	0.9663
0.75	1	1	1	1	0.9970	0.9973	0.9840	0.9834
0.8	1	1	1	1	0.9990	0.9984	0.9960	0.9954
0.85	1	1	1	1	1	0.9998	0.9980	0.9982
0.9	1	1	1	1	1	1	1	0.9998

Table 4.6: Coverage probabilities of 95% confidence regions of dependent Bootstrap

samples $k = 100$

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.5710	0.5709	0.392	0.392	0.285	0.285	0.2340	0.2347
0.1	0.7530	0.7534	0.5610	0.5605	0.4070	0.4074	0.3470	0.3473
0.15	0.8660	0.8656	0.6800	0.6803	0.5195	0.5192	0.4330	0.4334
0.2	0.9310	0.9309	0.7460	0.7458	0.6130	0.6127	0.5090	0.5084
0.25	0.9660	0.9658	0.820	0.820	0.6700	0.6692	0.563	0.563
0.3	0.9850	0.9847	0.8800	0.8795	0.7350	0.7346	0.6390	0.6398
0.35	0.9940	0.9939	0.9260	0.9262	0.7810	0.7812	0.7000	0.6993
0.4	0.9980	0.9983	0.9510	0.9511	0.8190	0.8186	0.7490	0.7495
0.45	0.9990	0.9989	0.9690	0.9689	0.8670	0.8674	0.7800	0.7798
0.5	1	0.9997	0.9850	0.9852	0.9140	0.9136	0.8240	0.8244
0.55	1	0.9999	0.9910	0.9911	0.9320	0.9321	0.8660	0.8659
0.6	1	1	0.9981	0.9976	0.9562	0.9559	0.912	0.9121
0.65	1	1	0.9990	0.9991	0.9790	0.9779	0.9461	0.9464
0.7	1	1	1	0.999	0.992	0.9925	0.9663	0.9659
0.75	1	1	1	1	0.9968	0.9977	0.9840	0.9835
0.8	1	1	1	1	0.9991	0.9986	0.9960	0.9958
0.85	1	1	1	1	1	0.9998	0.9980	0.9982
0.9	1	1	1	1	1	1	1	0.9998

Table 4.7: Coverage probabilities of 95% confidence regions of dependent Bootstrap

samples $k = 250$

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.571	0.570	0.3910	0.3915	0.2840	0.2842	0.2340	0.2339
0.1	0.7530	0.7528	0.5600	0.5596	0.4080	0.4076	0.3470	0.3471
0.15	0.8660	0.8656	0.6810	0.6803	0.5190	0.5186	0.433	0.433
0.2	0.9310	0.9309	0.7450	0.7451	0.6120	0.6124	0.5080	0.5084
0.25	0.9660	0.9657	0.8200	0.8197	0.669	0.669	0.5630	0.5629
0.3	0.9850	0.9847	0.8790	0.8792	0.7340	0.7341	0.640	0.640
0.35	0.9940	0.9939	0.9260	0.9259	0.7810	0.7809	0.7000	0.6998
0.4	0.9980	0.9983	0.9512	0.9508	0.8184	0.8186	0.7500	0.7494
0.45	0.9990	0.9989	0.9680	0.9685	0.8680	0.8678	0.7800	0.7799
0.5	1	0.9998	0.9850	0.9851	0.9132	0.9136	0.8251	0.8247
0.55	1	0.9999	0.9910	0.9911	0.9320	0.9323	0.8660	0.8659
0.6	1	1	0.9980	0.9974	0.9559	0.9559	0.9120	0.9121
0.65	1	1	0.9990	0.9991	0.9790	0.9783	0.9460	0.9464
0.7	1	1	1	0.9994	0.9920	0.9923	0.9660	0.9663
0.75	1	1	1	1	0.9970	0.9973	0.9840	0.9834
0.8	1	1	1	1	0.9990	0.9985	0.9960	0.9954
0.85	1	1	1	1	1	0.9998	0.9980	0.9982
0.9	1	1	1	1	1	1	1	0.9998

Table 4.8: Coverage probabilities of 95% confidence regions of dependent Bootstrap

samples $k = 500$

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.571	0.571	0.3920	0.3922	0.2850	0.2847	0.2350	0.2346
0.1	0.7530	0.7536	0.5610	0.5607	0.4080	0.4075	0.3475	0.3472
0.15	0.8660	0.8661	0.6800	0.6804	0.5190	0.5188	0.4330	0.4337
0.2	0.9310	0.9312	0.7460	0.7463	0.6120	0.6123	0.5080	0.5084
0.25	0.9660	0.9658	0.8200	0.8198	0.6690	0.6684	0.5620	0.5623
0.3	0.9850	0.9847	0.8790	0.8794	0.7340	0.7336	0.640	0.640
0.35	0.9940	0.9939	0.9260	0.9263	0.7810	0.7811	0.7000	0.6992
0.4	0.9980	0.9983	0.9510	0.9509	0.8190	0.8185	0.750	0.750
0.45	0.999	0.999	0.969	0.969	0.8680	0.8675	0.7800	0.7798
0.5	1	0.9997	0.9850	0.9853	0.9140	0.9135	0.8250	0.8244
0.55	1	0.9999	0.9910	0.9912	0.9325	0.9322	0.8660	0.8658
0.6	1	1	0.9980	0.9976	0.956	0.956	0.9120	0.9121
0.65	1	1	0.9990	0.9991	0.9790	0.9785	0.9460	0.9464
0.7	1	1	1	0.9994	0.9920	0.9923	0.9660	0.9663
0.75	1	1	1	1	0.9970	0.9973	0.9830	0.9831
0.8	1	1	1	1	0.9990	0.9984	0.9960	0.9954
0.85	1	1	1	1	1	0.9998	0.9980	0.9982
0.9	1	1	1	1	1	1	1	0.9998

4.4 Coverage Probability of 90% Confidence Region

I also set tables (table 4.9 and 4.10) of 90% confidence region. Because there is no strong evidence to show that the coverage probabilities are changed by dependent bootstrap method with different value of k , I just select $k = 5, 50$ for table 4.10 and table 4.11. An independent bootstrap method of coverage Probability of 90%

confidence region is shown in table 4.9.

Table 4.9: Coverage probabilities of 90% confidence regions of independent
Bootstrap samples

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.5190	0.5196	0.392	0.392	0.2840	0.2847	0.2340	0.2344
0.1	0.7020	0.7015	0.5045	0.5045	0.3620	0.3615	0.3070	0.3067
0.15	0.8190	0.8194	0.618	0.618	0.4650	0.4645	0.3850	0.3849
0.2	0.8960	0.8962	0.686	0.686	0.5520	0.5525	0.4540	0.4543
0.25	0.9420	0.9417	0.7620	0.7622	0.6060	0.6056	0.5030	0.5032
0.3	0.9690	0.9693	0.8290	0.8285	0.6710	0.6707	0.5785	0.5785
0.35	0.9850	0.9849	0.8860	0.8855	0.7190	0.7191	0.6370	0.6368
0.4	0.9940	0.9935	0.9160	0.9155	0.7580	0.7585	0.6870	0.6873
0.45	0.9970	0.9966	0.9430	0.9427	0.8160	0.8159	0.7180	0.7179
0.5	0.9990	0.9989	0.9680	0.9684	0.8700	0.8695	0.7660	0.7659
0.55	1	0.9996	0.9790	0.9792	0.8940	0.8936	0.8110	0.8110
0.6	1	1	0.9920	0.9917	0.9230	0.9228	0.8680	0.8672
0.65	1	1	0.9970	0.9964	0.9570	0.9566	0.9100	0.9101
0.7	1	1	0.9980	0.9982	0.9800	0.9803	0.9380	0.9382
0.75	1	1	1	0.9997	0.9910	0.9913	0.9640	0.9634
0.8	1	1	1	1	0.9960	0.9954	0.9870	0.9866
0.85	1	1	1	1	0.9990	0.9993	0.9950	0.9945
0.9	1	1	1	1	1	1	1	0.9994

Table 4.10: Coverage probabilities of 90% confidence regions of dependent

Bootstrap samples $k = 5$

Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.5190	0.5194	0.3490	0.3489	0.2510	0.2514	0.2070	0.2069
0.1	0.7010	0.7012	0.5040	0.5044	0.3620	0.3619	0.3070	0.3073
0.15	0.8190	0.8192	0.6180	0.6176	0.4640	0.4641	0.3850	0.3848
0.2	0.8960	0.8959	0.6860	0.6862	0.5530	0.5532	0.4540	0.4539
0.25	0.9420	0.9415	0.7630	0.7624	0.6060	0.6053	0.5040	0.5042
0.3	0.9700	0.9694	0.8280	0.8278	0.6700	0.6711	0.5780	0.5781
0.35	0.9850	0.9851	0.8850	0.8855	0.7200	0.7197	0.6360	0.6357
0.4	0.9940	0.9936	0.9150	0.9153	0.7590	0.7586	0.6870	0.6873
0.45	0.9970	0.9966	0.9430	0.9427	0.8170	0.8165	0.7180	0.7179
0.5	0.9990	0.9989	0.9680	0.9683	0.8700	0.8699	0.7660	0.7658
0.55	1	0.9996	0.9790	0.9792	0.894	0.894	0.8110	0.8109
0.6	1	1	0.9920	0.9916	0.9230	0.9229	0.8670	0.8673
0.65	1	1	0.9970	0.9964	0.9570	0.9564	0.9090	0.9094
0.7	1	1	0.9980	0.9982	0.9800	0.9803	0.9380	0.9384
0.75	1	1	1	0.9997	0.9910	0.9913	0.9640	0.9635
0.8	1	1	1	1	0.9960	0.9955	0.9870	0.9866
0.85	1	1	1	1	0.9990	0.9993	0.9950	0.9944
0.9	1	1	1	1	1	1	1	0.9994

Table 4.11: Coverage probabilities of 90% confidence regions of dependent

Bootstrap samples $k = 50$

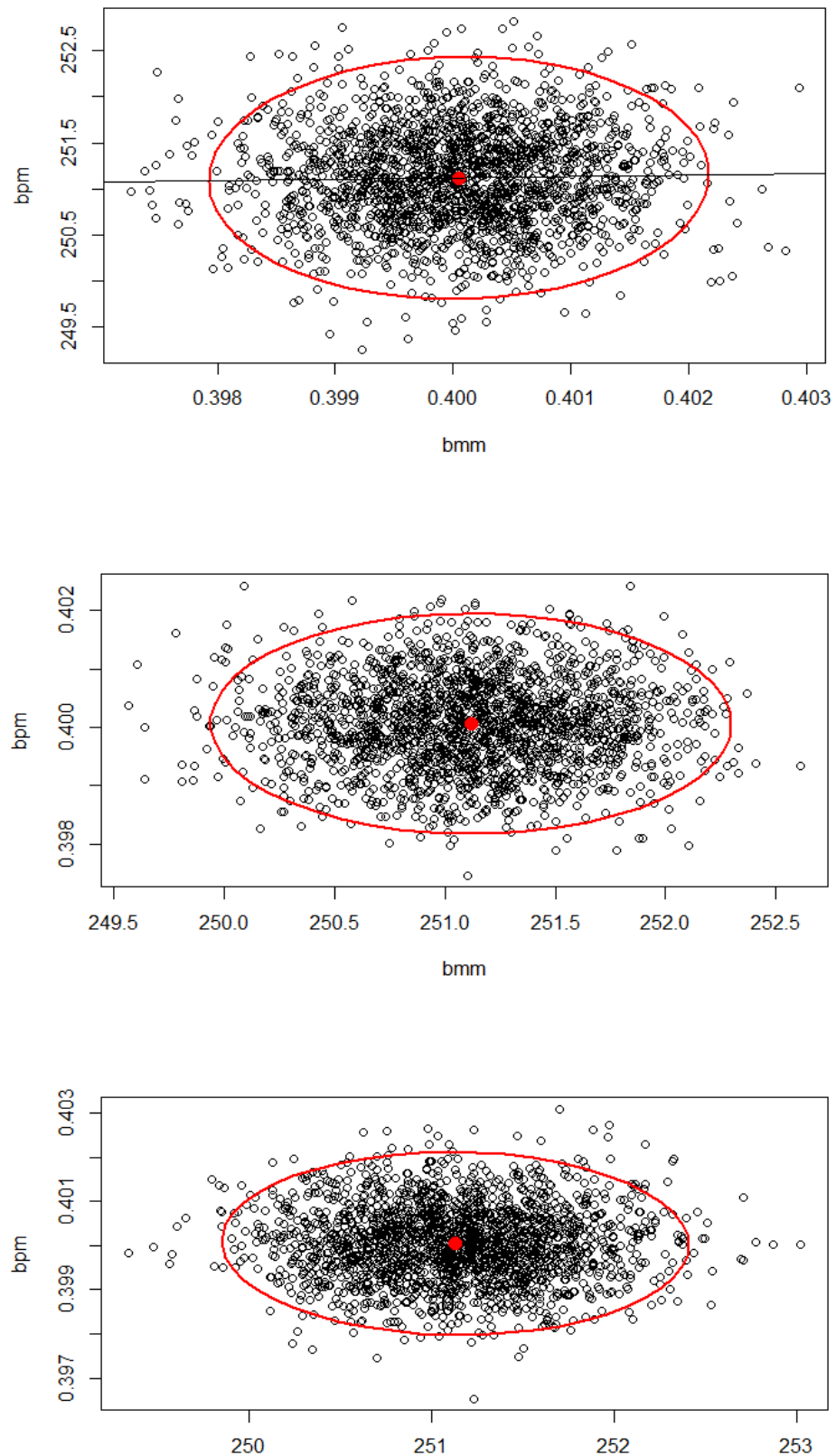
Probability\Sucess trial	m=100		m=250		m=500		m=750	
p	Median	Mean	Median	Mean	Median	Mean	Median	Mean
0.05	0.5190	0.5193	0.3480	0.3487	0.2510	0.2513	0.2070	0.2069
0.1	0.7010	0.7008	0.5040	0.5040	0.362	0.362	0.3070	0.3069
0.15	0.8190	0.8191	0.6170	0.6177	0.4640	0.4637	0.3840	0.3845
0.2	0.8960	0.8958	0.6860	0.6857	0.5540	0.5537	0.4540	0.4539
0.25	0.9410	0.9414	0.762	0.762	0.6060	0.6059	0.5040	0.5046
0.3	0.9690	0.9694	0.8280	0.8276	0.6710	0.6708	0.5780	0.5785
0.35	0.9850	0.9851	0.8850	0.8851	0.7190	0.7192	0.6360	0.6365
0.4	0.9940	0.9936	0.9150	0.9152	0.7590	0.7589	0.6880	0.6876
0.45	0.9970	0.9966	0.9430	0.9426	0.8160	0.8162	0.7170	0.7177
0.5	0.9990	0.9989	0.9680	0.9682	0.8690	0.8695	0.7660	0.7661
0.55	1	0.9996	0.979	0.979	0.8940	0.8936	0.8110	0.8111
0.6	1	1	0.9920	0.9916	0.9230	0.9231	0.8670	0.8674
0.65	1	1	0.9970	0.9964	0.9570	0.9564	0.9100	0.9099
0.7	1	1	0.9980	0.9982	0.9810	0.9803	0.9380	0.9382
0.75	1	1	1	0.9997	0.9910	0.9912	0.9640	0.9635
0.8	1	1	1	1	0.9960	0.9955	0.9870	0.9867
0.85	1	1	1	1	0.9990	0.9993	0.9950	0.9945
0.9	1	1	1	1	1	1	1	0.9994

From the table 4.9, we can see that for independent bootstrap method, when $m = 100$, probability $p = [0.2, 0.25]$ have the best outcome of $\delta \approx (1 - \alpha)$. Similarly, that $m = 250, p = [0.35, 0.4]$; $m = 500, p = [0.55, 0.6]$; $m = 750, p = [0.6, 0.65]$. From table 4.10 and 4.11, the similar results can be applied for dependent bootstrap method with different value of k , since different values of k slightly change coverage probabilities.

4.5 Areas of Confidence Region

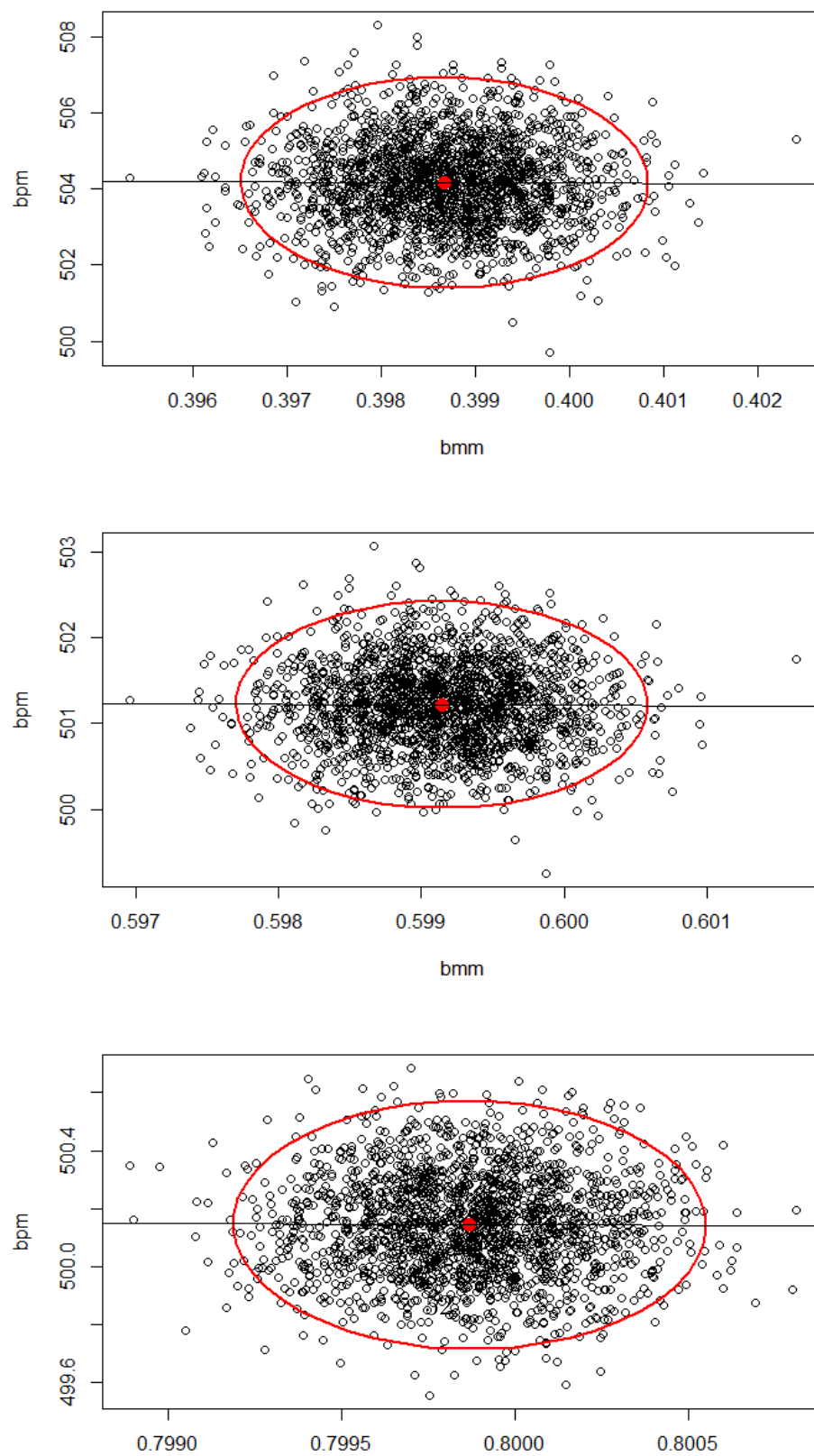
The areas of confidence regions of $Bin(p = 0.4, m = 250)$ are different by independent bootstrap and dependent bootstrap with $k = 5, 50$. The area of independent bootstrap method confidence region is $\hat{A} = 0.01224842$. The area for dependent bootstrap method $k = 5$ confidence region is $\hat{A} = 0.009778989$. The area for dependent bootstrap method $k = 50$ confidence region is $\hat{A} = 0.01155577$. The Figure 4.1 shows the areas of confidence regions of three methods. We can conclude that the independent bootstrap confidence region has the largest area than the dependent bootstrap confidence regions. Moreover, as k increases, the area of dependent bootstrap confidence regions are closer to the independent bootstrap confidence region.

Figure 4.1: The Areas 95% Confidence regions of $\text{Bin}(p = 0.4, m = 250)$ for independent bootstrap and dependent bootstrap with $k = 5, 50$



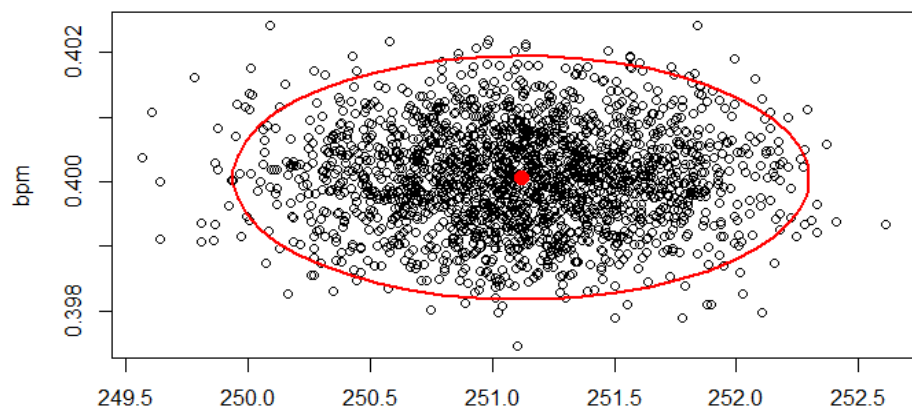
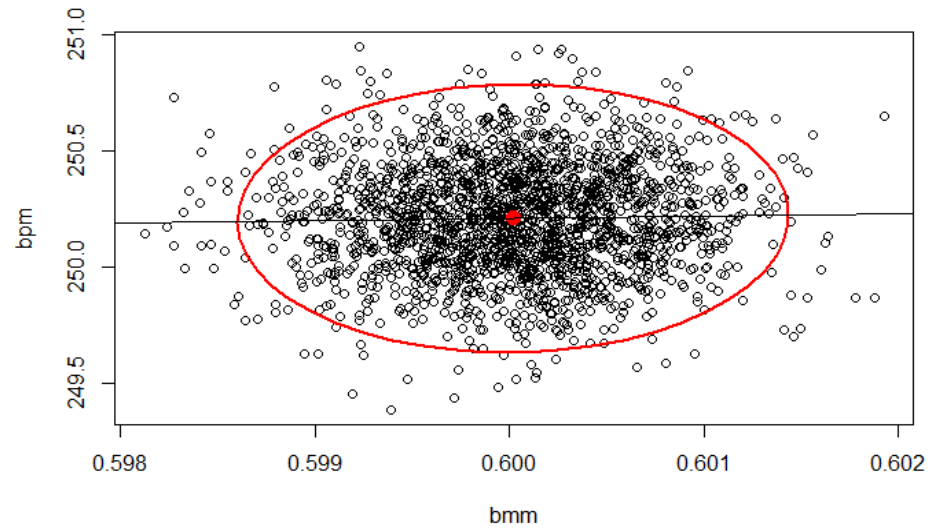
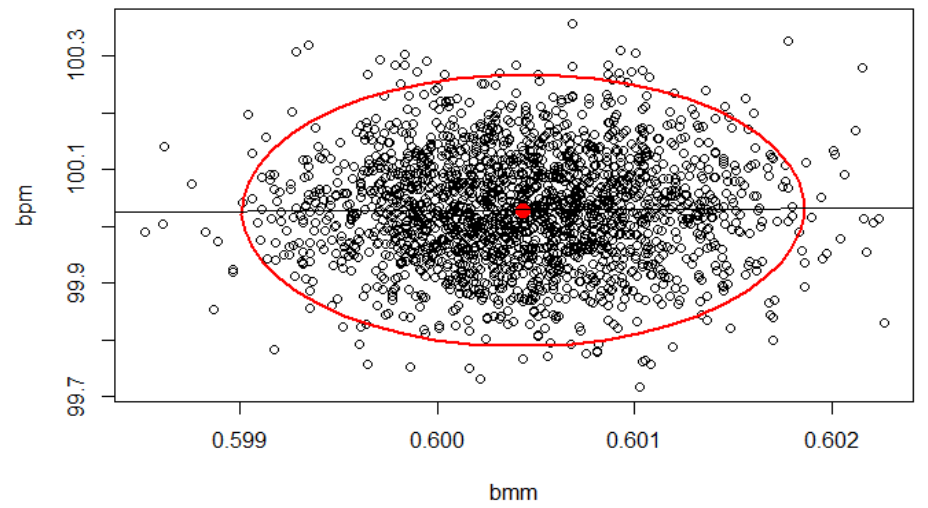
For the same bootstrap method, the value of population parameter p and m can influence the areas of confidence regions. The figure 4.2 shows independent bootstrap confidence regions of fixed m and $p = \{0.4, 0.6, 0.8\}$. The area of confidence region of $Bin(p = 0.4, m = 500)$ is $\hat{A} = 0.04209973$. The area of confidence region of $Bin(p = 0.6, m = 500)$ is $\hat{A} = 0.009022365$. The area of confidence region of $Bin(p = 0.8, m = 500)$ is $\hat{A} = 0.001286966$. Therefore, as p increases, and m is fixed, the area of independent bootstrap confidence region is decreasing.

Figure 4.2: The Areas 95% Confidence regions of $Bin(p = \{0.4, 0.6, 0.8\}, m = 500)$
for independent bootstrap method



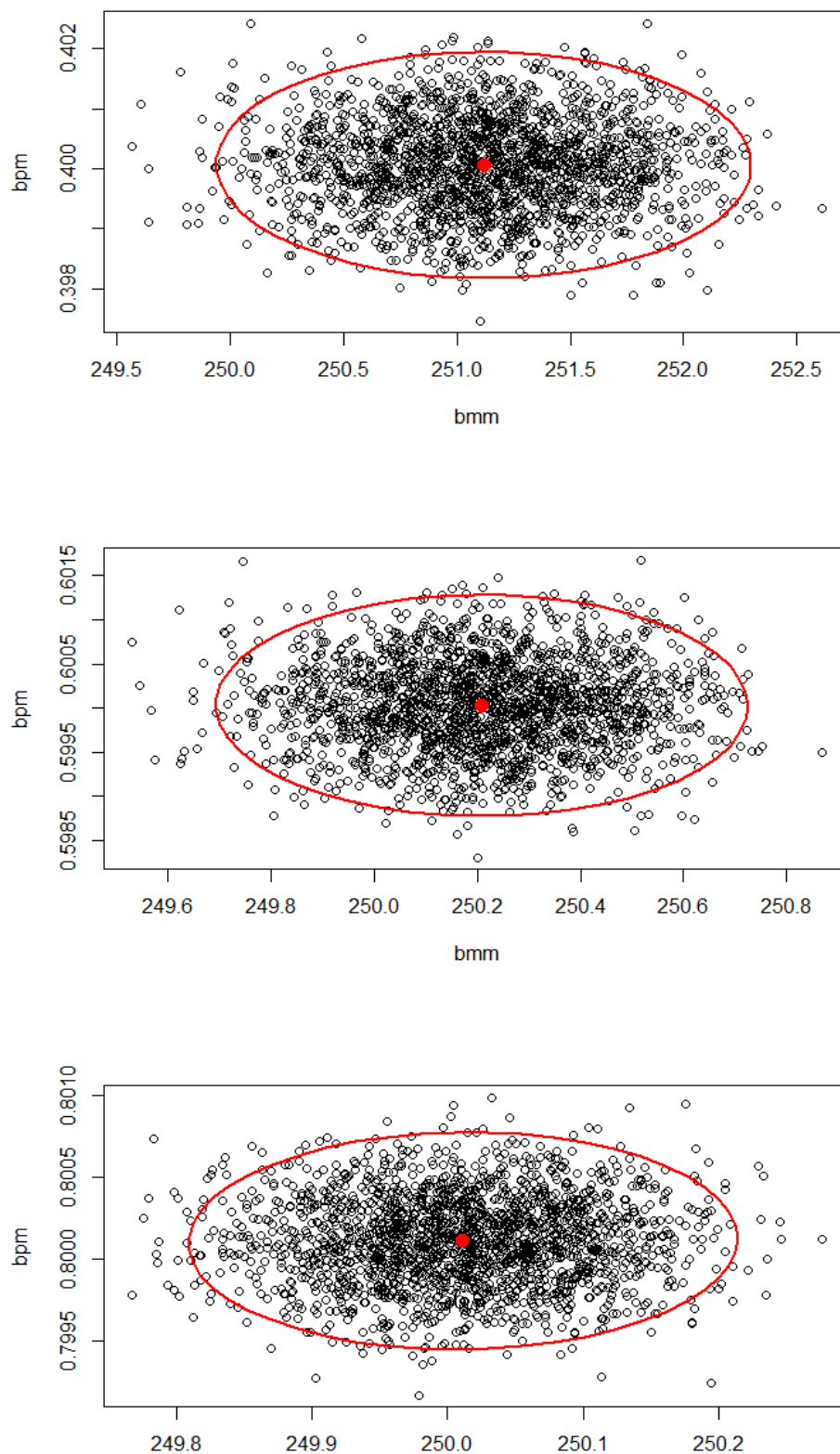
The figure 4.3 shows that independent bootstrap confidence regions of fixed $p = 0.6$ and $m = \{100, 250, 500\}$. The area of confidence region of $Bin(p = 0.6, m = 100)$ is $\hat{A} = 0.001112208$. The area of confidence region of $Bin(p = 0.6, m = 250)$ is $\hat{A} = 0.003049787$. The area of confidence region of $Bin(p = 0.6, m = 500)$ is $\hat{A} = 0.009022365$. Therefore, as m increases, and p is fixed, the area of independent bootstrap confidence region is increasing.

Figure 4.3: The Areas 95% Confidence regions of $Bin(p = 0.6, m = \{100, 250, 500\})$
for independent bootstrap method



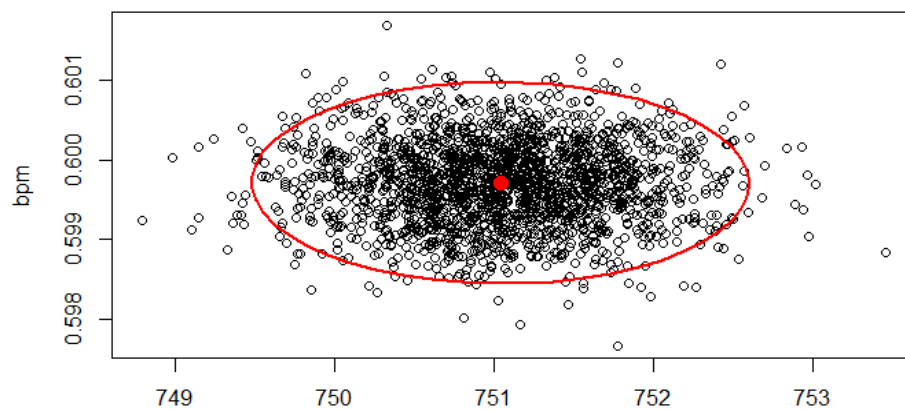
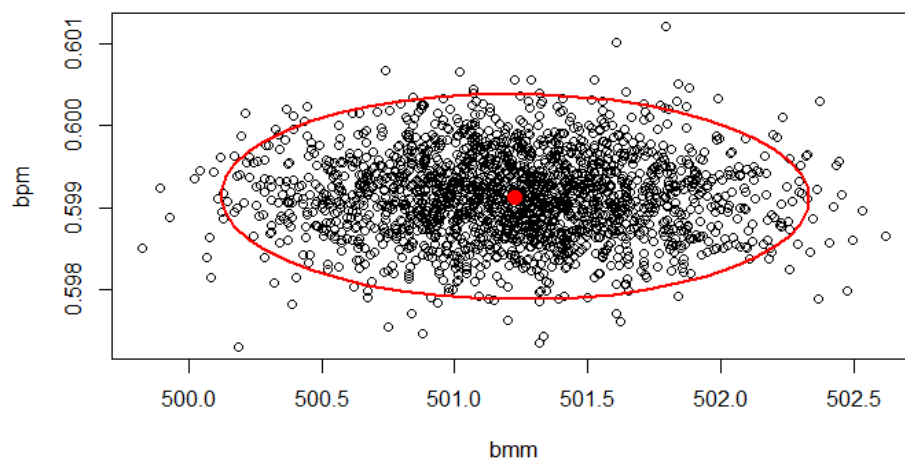
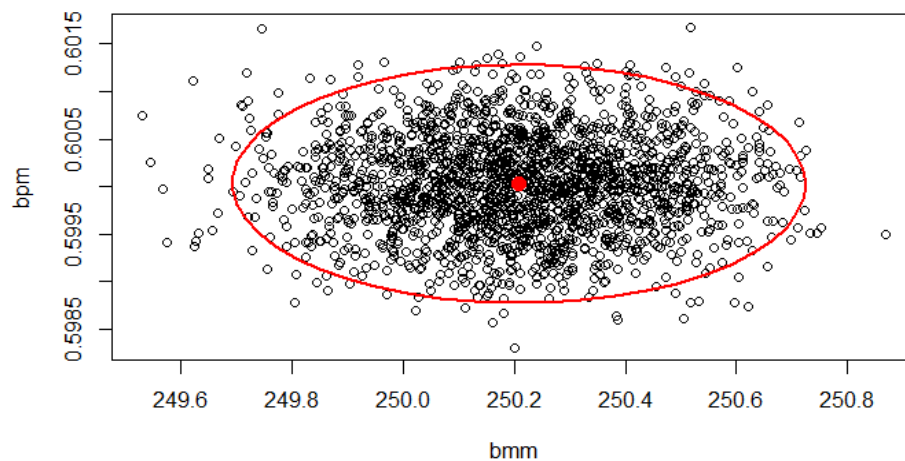
The figure 4.4 shows dependent bootstrap with $k = 5$ confidence regions of fixed $m = 250$ and $p = \{0.4, 0.6, 0.8\}$. The area of confidence region of $Bin(p = 0.4, m = 250)$ is $\hat{A} = 0.009778989$. The area of confidence region of $Bin(p = 0.6, m = 250)$ is $\hat{A} = 0.002416341$. The area of confidence region of $Bin(p = 0.8, m = 250)$ is $\hat{A} = 0.0004680606$. Therefore, as p increases, and m is fixed, the area of dependent bootstrap confidence region is decreasing.

Figure 4.4: The Areas 95% Confidence regions of $Bin(p = 0.6, m = \{100, 250, 500\})$
for dependent bootstrap method



The figure 4.5 shows that dependent bootstrap with $k = 5$ confidence regions of fixed $p = 0.6$ and $m = \{250, 500, 750\}$. The area of confidence region of $Bin(p = 0.6, m = 250)$ is $\hat{A} = 0.002416341$. The area of confidence region of $Bin(p = 0.6, m = 500)$ is $\hat{A} = 0.007431883$. The area of confidence region of $Bin(p = 0.6, m = 750)$ is $\hat{A} = 0.0135053$. Therefore, as m increases, and p is fixed, the area of dependent bootstrap confidence region is increasing.

Figure 4.5: The Areas 95% Confidence regions of $Bin(p = 0.6, m = \{100, 250, 500\})$
for dependent bootstrap method



Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this section, the results from chapter 4 will be concluded. Firstly, this research is designed to recognize the performance of bootstrap method. After generating the original samples of $Bin(p, m)$, I calculate the coverage probabilities of confidence regions without resampling. The coverage probabilities of original sample confidence regions is smaller than the coverage probabilities of bootstrap confidence regions. For example, the mean of coverage probabilities of 95% confidence regions $Bin(p = 0.5, m = 500)$ is 0.786, but mean of coverage probabilities of independent bootstrap 95% confidence regions $Bin(p = 0.5, m = 500)$ is 0.9137. Similarly, the mean of coverage probabilities of dependent bootstrap($k = 5$) 95% confidence regions $Bin(p = 0.5, m = 500)$ is 0.9138. Thus, the bootstrap re-sampling method increases

the coverage probability δ . In other words, the bootstrap re-sampling method positively influences the coverage probability of confidence region.

Next, the difference between independent and dependent bootstrap is not evident. For the fixed value of parameters p and m , independent and dependent bootstrap confidence regions have cognate coverage probabilities. Furthermore, different k of dependent bootstrap confidence regions have closer coverage probabilities.

The areas of independent and dependent bootstrap confidence regions are slightly different. For the same value of parameters p and m , the area of independent bootstrap confidence region is larger than dependent bootstrap confidence region. However, Both independent and dependent bootstrap confidence region follow:

1. as p increases, and m is fixed, the area of dependent bootstrap confidence region is decreasing.
2. as m increases, and p is fixed, the area of independent bootstrap confidence region is increasing.

5.2 Future Work

1. In this research, the confidence region of binomial distribution was investigated. Other distributions have two or more dimensions will be my next goal, such as weibull distribution (2-dimension) or generalized gamma distribution (3-dimension).

2. In chapter 4, the results are based on simulated outcomes. I will find some real data follows binomial distribution, and test whether these conform the conclusion of this research.

3. Different methods of estimating the parameters compare with method of moment to show which method gives the better estimators, in other words, the estimators obtain lower mean square error.

Bibliography

- [1] Ando,T. Totally Positive Matrices. *Linear Algebra Appl.* **90** (1987), 165–219.
- [2] Casella,G.;Berger,R.L. (2002). "Statistical Inference". 2nd Ed.
- [3] Cramér, H. Mathematical Methods of Statistics, 2016, Princeton university press
- [4] Dekking, F.M. (Frederik Michel), 1946- (2005). A modern introduction to probability and statistics : understanding why and how. Springer.
- [5] DiCiccio TJ, Efron B. Bootstrap confidence intervals (with Discussion). *Science* 11: 189–228, 1996.
- [6] Efron, B. (1979). "Bootstrap methods: Another look at the jackknife". The Annals of Statistics. 7 (1): 1–26.
- [7] Efron, B.; Tibshirani, R. (1993). An Introduction to the Bootstrap. software Archived, 2012.

- [8] Goldberger, Arthur S. (1964). *Econometric Theory*. New York: Wiley. pp. 117–120.
- [9] Haldane, B.S. The fitting of Binomial distributions. *Ann. Eugenics*. 1941. V.11. P.179–181.
- [10] Lehmann, E. L.; Casella, George (1998). *Theory of Point Estimation* (2nd ed.). New York: Springer.
- [11] Lehmann, E. L. *Elements of Large – Sample Theory*, 2004, Springer Science & Business Media.
- [12] Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. (1992). "Least Squares as a Maximum Likelihood Estimator". *Numerical Recipes in FORTRAN: The Art of Scientific Computing* (2nd ed.). Cambridge: Cambridge University Press. pp. 651–655. ISBN 0-521-43064-X.
- [13] Rossi, R. J. *Mathematical Statistics : An Introduction to Likelihood Based Inference*. p. 227, 2018.
- [14] Saad,S. Asymptotic Analysis of method of moments estimators of parameters p and m for the binomial distribution.(2019).
- [15] Smith, W.D. and Taylor, R.L. (2001). Dependent bootstrap confidence intervals. *Selected Proceeding of the Symposium on Inference for Stochastic Processes*

- (Athens, GA, 2000). IMS Lecture Notes – Monograph Series, 37, 2001b, 91-107.
- [16] Towards data science <https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60>
- [17] Varian, H.(2005). "Bootstrap Tutorial". *Mathematica Journal*, 9, 768–775.
- [18] Yates, Daniel S.; David S. Moore; Daren S. Starnes (2008). *The Practice of Statistics*, 3rd Ed. Freeman. ISBN 978-0-7167-7309-2.

Appendix R code

```
# install packages

library(survival)

library(Matrix)

library(MatrixModels)

library(matrixStats)

library(ellipse)

library(car)

library(jocre)

rm(list=ls())


#Independent Bootstrap 95% confidence region of Bin(0.55,500)

#generating 1000 Bin samples

set.seed(2)

times=1000
```

```

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,500,0.55)

}


#Method of moment of estimators P and M

m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

#coverage probability of confidence region without bootstrapping

p=.55;m=500;a=2*(1-p)*(m-1)

E=c(m,-p,-p,((p^2)*(p+a))/(m*a))

inv_E=1/(p*(1-p))*matrix(E,nrow = 2)

chi=(1/(p*(1-p)))*((m*(p_hat-p)^2)-(2*p*(m_hat-m)*(p_hat-p))+(((p^2)*(p+a))/(m*a))

count(chi<=qchisq(.95,df=2))


#B=2000 independent bootstrap samples

nboot=2000;n=times

#data=data.frame(p_hat,m_hat)

```

```

tmpdata = sample(p_hat,n*nboot, replace=TRUE)

pbs= matrix(tmpdata, nrow=nboot, ncol=n)

tmpdata = sample(m_hat,n*nboot, replace=TRUE)

mbs= matrix(tmpdata, nrow=nboot, ncol=n)

#coverage probability of confidence region with independent bootstrapping

chi=(1/(p*(1-p)))*((m*(pbs-p)^2)-(2*p*(mbs-m)*(pbs-p))+((p^2)*(p+a))/(m*a))*(mbs-

mt=(rowCounts(chi<=qchisq(.95,df=2)))/n

summary(mt)

plot(mt)


bmm=rowMeans(pbs);bpm=rowMeans(mbs)

df=data.frame(bmm,bpm)

plot(df)

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

abline(lm(bpm~bmm))

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

```

```

plot(df)

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

abline(lm(bpm~bmm))

#area of mean of confidence region

require(car)

dataEllipse(df$bmm, df$bpm, levels=0.5)


me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area

#Independent Bootstrap 95% confidence region of Bin(0.6,250)

#generating 1000 Bin samples

set.seed(2)

```

```

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,250,0.6)

}


#Method of moment of estimators P and M

m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

#coverage probability of confidence region without bootstrapping

p=.6;m=250;a=2*(1-p)*(m-1)

E=c(m,-p,-p,((p^2)*(p+a))/(m*a))

inv_E=1/(p*(1-p))*matrix(E,nrow = 2)

chi=(1/(p*(1-p)))*((m*(p_hat-p)^2)-(2*p*(m_hat-m)*(p_hat-p))+(((p^2)*(p+a))/(m*a)))

count(chi<=qchisq(.95,df=2))


#B=2000 independent bootstrap samples

nboot=2000;n=times

```



```

#data=data.frame(p_hat,m_hat)

tmpdata = sample(p_hat,n*nboot, replace=TRUE)

pbs= matrix(tmpdata, nrow=nboot, ncol=n)

tmpdata = sample(m_hat,n*nboot, replace=TRUE)

mbs= matrix(tmpdata, nrow=nboot, ncol=n)

#coverage probability of confidence region with independent bootstrapping

chi=(1/(p*(1-p)))*((m*(pbs-p)^2)-(2*p*(mbs-m)*(pbs-p))+((p^2)*(p+a))/(m*a))*(mbs-
mt=(rowCounts(chi<=qchisq(.95,df=2)))/n

summary(mt)

plot(mt)


bmm=rowMeans(pbs);bpm=rowMeans(mbs)

df=data.frame(bmm,bpm)

plot(df)

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

abline(lm(bpm~bmm))

cset(df, method="boot.kern",alpha = 0.05)

```

```

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

abline(lm(bpm~bmm))


#area of mean of confidence region

require(car)

dataEllipse(df$bmm, df$bpm, levels=0.5)


me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area


#Independent Bootstrap 95% confidence region of Bin(0.6,100)

#generating 1000 Bin samples

```

```

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,100,0.6)

}


#Method of moment of estimators P and M

m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

#coverage probability of confidence region without bootstrapping

p=.6;m=100;a=2*(1-p)*(m-1)

E=c(m,-p,-p,((p^2)*(p+a))/(m*a))

inv_E=1/(p*(1-p))*matrix(E,nrow = 2)

chi=(1/(p*(1-p)))*((m*(p_hat-p)^2)-(2*p*(m_hat-m)*(p_hat-p))+(((p^2)*(p+a))/(m*a))

count(chi<=qchisq(.95,df=2))


#B=2000 independent bootstrap samples

```

```

nboot=2000;n=times

#data=data.frame(p_hat,m_hat)

tmpdata = sample(p_hat,n*nboot, replace=TRUE)

pbs= matrix(tmpdata, nrow=nboot, ncol=n)

tmpdata = sample(m_hat,n*nboot, replace=TRUE)

mbs= matrix(tmpdata, nrow=nboot, ncol=n)

#coverage probability of confidence region with independent bootstrapping

chi=(1/(p*(1-p)))*((m*(pbs-p)^2)-(2*p*(mbs-m)*(pbs-p))+((p^2)*(p+a))/(m*a))*(mbs-

mt=(rowCounts(chi<=qchisq(.95,df=2)))/n

summary(mt)

plot(mt)


bmm=rowMeans(pbs);bpm=rowMeans(mbs)

df=data.frame(bmm,bpm)


plot(df)

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

abline(lm(bpm~bmm))

```

```

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

abline(lm(bpm~bmm))


#area of mean of confidence region

require(car)

dataEllipse(df$bmm, df$bpm, levels=0.5)


me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area


#Independent Bootstrap 95% confidence region of Bin(0.4,250)

```

```

#generating 1000 Bin samples

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,250,0.4)

}


#Method of moment of estimators P and M

m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

#coverage probability of confidence region without bootstrapping

p=.4;m=250;a=2*(1-p)*(m-1)

E=c(m,-p,-p,((p^2)*(p+a))/(m*a))

inv_E=1/(p*(1-p))*matrix(E,nrow = 2)

chi=(1/(p*(1-p)))*((m*(p_hat-p)^2)-(2*p*(m_hat-m)*(p_hat-p))+(((p^2)*(p+a))/(m*a))

count(chi<=qchisq(.95,df=2))

```

```

#B=2000 independent bootstrap samples

nboot=2000;n=times

#data=data.frame(p_hat,m_hat)

tmpdata = sample(p_hat,n*nboot, replace=TRUE)

pbs= matrix(tmpdata, nrow=nboot, ncol=n)

tmpdata = sample(m_hat,n*nboot, replace=TRUE)

mbs= matrix(tmpdata, nrow=nboot, ncol=n)

#coverage probability of confidence region with independent bootstrapping

chi=(1/(p*(1-p)))*((m*(pbs-p)^2)-(2*p*(mbs-m)*(pbs-p))+((p^2)*(p+a))/(m*a))*(mbs-

mt=(rowCounts(chi<=qchisq(.95,df=2)))/n

summary(mt)

plot(mt)


bmm=rowMeans(pbs);bpm=rowMeans(mbs)

df=data.frame(bmm,bpm)

plot(df)

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

```

```

abline(lm(bpm~bmm))

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

abline(lm(bpm~bmm))


#area of mean of confidence region

require(car)

dataEllipse(df$bmm, df$bpm, levels=0.5)


me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area

```



```

#Independent Bootstrap 95% confidence region of Bin(0.4,100)

#generating 1000 Bin samples

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,100,0.4)

}


#Method of moment of estimators P and M

m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

#coverage probability of confidence region without bootstrapping

p=.4;m=100;a=2*(1-p)*(m-1)

E=c(m,-p,-p,((p^2)*(p+a))/(m*a))

inv_E=1/(p*(1-p))*matrix(E,nrow = 2)

chi=(1/(p*(1-p)))*((m*(p_hat-p)^2)-(2*p*(m_hat-m)*(p_hat-p))+(((p^2)*(p+a))/(m*a)))

count(chi<=qchisq(.95,df=2))

```

```

#B=2000 independent bootstrap samples

nboot=2000;n=times

#data=data.frame(p_hat,m_hat)

tmpdata = sample(p_hat,n*nboot, replace=TRUE)

pbs= matrix(tmpdata, nrow=nboot, ncol=n)

tmpdata = sample(m_hat,n*nboot, replace=TRUE)

mbs= matrix(tmpdata, nrow=nboot, ncol=n)

#coverage probability of confidence region with independent bootstrapping

chi=(1/(p*(1-p)))*((m*(pbs-p)^2)-(2*p*(mbs-m)*(pbs-p))+((p^2)*(p+a))/(m*a))*(mbs-

mt=(rowCounts(chi<=qchisq(.95,df=2)))/n

summary(mt)

plot(mt)


bmm=rowMeans(pbs);bpm=rowMeans(mbs)

df=data.frame(bmm,bpm)

plot(df)

```

```

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

abline(lm(bpm~bmm))

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

with(df, dataEllipse(bmm, bpm, level = 0.95, add = TRUE))

abline(lm(bpm~bmm))


#area of mean of confidence region

require(car)

dataEllipse(df$bmm, df$bpm, levels=0.5)


me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z))-me))^2))

area=pi*min(dist2center)*max(dist2center)

area

```

```

#Dependent Bootstrap with k=500 95% confidence region of Bin(0.5,500)

rm(list=ls())

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){
  x[i,]=rbinom(n=1000,500,0.5)
}


m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=500

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

#tpdata = sample(ordp,n, replace=FALSE)

```

```

samp <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

}

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samm[i,] <- sample(ordm, size = n, replace = F)

}


p=.5;m=500;a=2*(1-p)*(m-1)


chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s

mt=(rowCounts(chi<=qchisq(.9,df=2)))/n

summary(mt)


dbpm=rowMeans(samp)

dbmm=rowMeans(samm)

```

```

df=data.frame(dbmm,dbpm)

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))

abline(lm(dbpm~dbmm))

require(car)

dataEllipse(df$dbmm, df$dbpm, levels=0.5)


me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area


#Dependent Bootstrap with k=250 95% confidence region of Bin(0.5,500)

rm(list=ls())

```

```

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,500,0.5)

}


m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=250

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

#tpdata = sample(ordp,n, replace=FALSE)


samp <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

```

```

}

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samm[i,] <- sample(ordm, size = n, replace = F)

}


p=.5;m=500;a=2*(1-p)*(m-1)


chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s

mt=(rowCounts(chi<=qchisq(.9,df=2)))/n

summary(mt)


dbpm=rowMeans(samp)

dbmm=rowMeans(samm)

df=data.frame(dbmm,dbpm)

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

```



```

plot(df)

with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))

abline(lm(dbpm~dbmm))

require(car)

dataEllipse(df$dbmm, df$dbpm, levels=0.5)


me = apply(df, 2, mean)
v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area


#Dependent Bootstrap with k=100 95% confidence region of Bin(0.5,500)

rm(list=ls())

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

```

```

for (i in 1:times){

  x[i,]=rbinom(n=1000,500,0.5)

}

m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=500

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

#tpdata = sample(ordp,n, replace=FALSE)

samp <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

}

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

```

```

for(i in 1:2000){

  samm[i,] <- sample(ordm, size = n, replace = F)

}

p=.5;m=500;a=2*(1-p)*(m-1)

chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s

mt=(rowCounts(chi<=qchisq(.9,df=2)))/n

summary(mt)


dbpm=rowMeans(samp)

dbmm=rowMeans(samm)

df=data.frame(dbmm,dbpm)

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))

abline(lm(dbpm~dbmm))

```

```

require(car)

dataEllipse(df$dbmm, df$dbpm, levels=0.5)


me = apply(df, 2, mean)
v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))
z = ellipse(me, v, rad, segments=1001)
dist2center = sqrt(rowSums((t(t(z)-me))^2))
area=pi*min(dist2center)*max(dist2center)

area


#Dependent Bootstrap with k=50 95% confidence region of Bin(0.5,500)

rm(list=ls())

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,500,0.5)

}

```

```

m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=50

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

#tpdata = sample(ordp,n, replace=FALSE)


samp <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

}

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samm[i,] <- sample(ordm, size = n, replace = F)

}

```

```
p=.5;m=500;a=2*(1-p)*(m-1)
```

```
chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s
```

```
mt=(rowCounts(chi<=qchisq(.9,df=2)))/n
```

```
summary(mt)
```

```
dbpm=rowMeans(samp)
```

```
dbmm=rowMeans(samm)
```

```
df=data.frame(dbmm,dbpm)
```

```
cset(df, method="boot.kern",alpha = 0.05)
```

```
plot(cset(df, method="boot.kern",alpha = 0.05))
```

```
plot(df)
```

```
with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))
```

```
abline(lm(dbpm~dbmm))
```

```
require(car)
```

```
dataEllipse(df$dbmm, df$dbpm, levels=0.5)
```

```

me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area

#Dependent Bootstrap with k=25 95% confidence region of Bin(0.5,500)

rm(list=ls())

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,500,0.5)

}

m=rowMeans(x)

v=rowVars(x)

```

```

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=25

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

#tpdata = sample(ordp,n, replace=FALSE)


samp <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

}

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samm[i,] <- sample(ordm, size = n, replace = F)

}


p=.5;m=500;a=2*(1-p)*(m-1)

```



```

chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s
mt=(rowCounts(chi<=qchisq(.9,df=2)))/n
summary(mt)

```

```

dbpm=rowMeans(samp)
dbmm=rowMeans(samm)
df=data.frame(dbmm,dbpm)
cset(df, method="boot.kern",alpha = 0.05)
plot(cset(df, method="boot.kern",alpha = 0.05))
plot(df)
with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))
abline(lm(dbpm~dbmm))
require(car)
dataEllipse(df$dbmm, df$dbpm, levels=0.5)

```

```

me = apply(df, 2, mean)
v =var(df)

```

```

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area

#Dependent Bootstrap with k=10 95% confidence region of Bin(0.5,500)

rm(list=ls())

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,500,0.5)

}

m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=10

```

```

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

#tpdata = sample(ordp,n, replace=FALSE)


samp <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

}

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samm[i,] <- sample(ordm, size = n, replace = F)

}


p=.5;m=500;a=2*(1-p)*(m-1)


chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s

mt=(rowCounts(chi<=qchisq(.9,df=2)))/n

```

```
summary(mt)
```

```
dbpm=rowMeans(samp)
```

```
dbmm=rowMeans(samm)
```

```
df=data.frame(dbmm,dbpm)
```

```
cset(df, method="boot.kern",alpha = 0.05)
```

```
plot(cset(df, method="boot.kern",alpha = 0.05))
```

```
plot(df)
```

```
with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))
```

```
abline(lm(dbpm~dbmm))
```

```
require(car)
```

```
dataEllipse(df$dbmm, df$dbpm, levels=0.5)
```

```
me = apply(df, 2, mean)
```

```
v =var(df)
```

```
rad = sqrt(2*qf(0.95, 2, nrow(df)-1))
```

```
z = ellipse(me, v, rad, segments=1001)
```

```
dist2center = sqrt(rowSums((t(t(z)-me))^2))
```

```

area=pi*min(dist2center)*max(dist2center)

area

#Dependent Bootstrap with k=5 95% confidence region of Bin(0.5,500)

rm(list=ls())

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,500,0.5)

}


m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=5

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

```

```

#tpdata = sample(ordp,n, replace=FALSE)

samp <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

}

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samm[i,] <- sample(ordm, size = n, replace = F)

}


p=.5;m=500;a=2*(1-p)*(m-1)


chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s

mt=(rowCounts(chi<=qchisq(.9,df=2)))/n

summary(mt)

```

```

dbpm=rowMeans(samp)

dbmm=rowMeans(samm)

df=data.frame(dbmm,dbpm)

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))

abline(lm(dbpm~dbmm))

require(car)

dataEllipse(df$dbmm, df$dbpm, levels=0.5)


me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z))-me))^2))

area=pi*min(dist2center)*max(dist2center)

area

```

```

#Dependent Bootstrap with k=5 95% confidence region of Bin(0.6,500)

rm(list=ls())

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,500,0.6)

}


m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=5

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

#tpdata = sample(ordp,n, replace=FALSE)


samp <- matrix(NA, ncol = 1000, nrow = 2000)

```



```

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

}

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samm[i,] <- sample(ordm, size = n, replace = F)

}


p=.6;m=500;a=2*(1-p)*(m-1)


chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s

mt=(rowCounts(chi<=qchisq(.9,df=2)))/n

summary(mt)


dbpm=rowMeans(samp)

dbmm=rowMeans(samm)

df=data.frame(dbmm,dbpm)

```

```

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))

abline(lm(dbpm~dbmm))

require(car)

dataEllipse(df$dbmm, df$dbpm, levels=0.5)


me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area


#Dependent Bootstrap with k=5 95% confidence region of Bin(0.6,250)

rm(list=ls())

set.seed(2)

```

```

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

  x[i,]=rbinom(n=1000,250,0.6)

}


m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=5

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

#tpdata = sample(ordp,n, replace=FALSE)


samp <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

}

```

```

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samm[i,] <- sample(ordm, size = n, replace = F)

}

p=.6;m=250;a=2*(1-p)*(m-1)

chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s

mt=(rowCounts(chi<=qchisq(.9,df=2)))/n

summary(mt)


dbpm=rowMeans(samp)

dbmm=rowMeans(samm)

df=data.frame(dbmm,dbpm)

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

```

```

with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))

abline(lm(dbpm~dbmm))

require(car)

dataEllipse(df$dbmm, df$dbpm, levels=0.5)


me = apply(df, 2, mean)
v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z)-me))^2))

area=pi*min(dist2center)*max(dist2center)

area


#Dependent Bootstrap with k=5 95% confidence region of Bin(0.6,100)

rm(list=ls())

set.seed(2)

times=1000

x=matrix(nrow=times, ncol=1000)

for (i in 1:times){

```

```

    x[i,]=rbinom(n=1000,100,0.6)
  }

m=rowMeans(x)

v=rowVars(x)

p_hat=(m-v)/m

m_hat=(m^2)/(m-v)

k=5

nboot=2000;n=times

ordp=c(rep(p_hat,k))

ordm=c(rep(m_hat,k))

#tpdata = sample(ordp,n, replace=FALSE)


samp <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

  samp[i,] <- sample(ordp, size = n, replace = F)

}

#tmdata = sample(ordm,n, replace=FALSE)

samm <- matrix(NA, ncol = 1000, nrow = 2000)

for(i in 1:2000){

```

```

    samm[i,] <- sample(ordm, size = n, replace = F)
  }

p=.6;m=100;a=2*(1-p)*(m-1)

chi=(1/(p*(1-p)))*((m*(samp-p)^2)-(2*p*(samm-m)*(samp-p))+(((p^2)*(p+a))/(m*a))*(s
mt=(rowCounts(chi<=qchisq(.9,df=2)))/n

summary(mt)


dbpm=rowMeans(samp)

dbmm=rowMeans(samm)

df=data.frame(dbmm,dbpm)

cset(df, method="boot.kern",alpha = 0.05)

plot(cset(df, method="boot.kern",alpha = 0.05))

plot(df)

with(df, dataEllipse(dbmm, dbpm, level = 0.95, add = TRUE))

abline(lm(dbpm~dbmm))

require(car)

```

```

dataEllipse(df$dbmm, df$dbpm, levels=0.5)

me = apply(df, 2, mean)

v =var(df)

rad = sqrt(2*qf(0.95, 2, nrow(df)-1))

z = ellipse(me, v, rad, segments=1001)

dist2center = sqrt(rowSums((t(t(z))-me))^2))

area=pi*min(dist2center)*max(dist2center)

area

```