UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga
(Final project paper)

Interval zaupanja za populacijsko povprečje, ko porazdelitev preučevane spremenljivke ni normalna

(The effects of nonnormal distribution on confidence interval
around population mean)

Ime in priimek: Slađana Babić
Študijski program: Matematika
Mentor: doc. dr. Rok Blagus

Koper, junij 2015

Ključna dokumentacijska informacija

Ime in PRIIMEK: Slađana BABIĆ

Naslov zaključne naloge: Interval zaupanja za populacijsko povprečje, ko porazdelitev preučevane spremenljivke ni normalna

Kraj: Koper

Leto: 2015

Število listov: 39 Število slik: 6 Število tabel: 20

Število referenc: 11

Mentor: doc. dr. Rok Blagus

Ključne besede: interval zaupanja, pokritje intervala zaupanja, širina intervala zaupanja, bootstrapping, simulacije.

Math. Subj. Class. (2010): 62F25

Izvleček:
Cilj naloge je preučiti pokritje in širino intervala zaupanja za populacijsko povprečje. Izpeljali bomo cenilke za interval zaupanja za primer, ko je preučevana spremenljivka porazdeljena normalno ter za primer, ko spremenljivka ni porazdeljena normalno. Poleg klasičnih cenilk za oceno populacijskega povprečja bomo uporabljali tudi novejše metode samovzorčenja (ang. bootstrap), ki temeljijo izključno na podlagi vzorcev. Preučevali bomo vpliv velikosti vzorca in oblike porazdelitve.
Za preučevanje lastnosti cenilk pri majhnih vzorcih bomo uporabili simulacije, kjer nas bo zanimalo, katera izmed cenilk ima pri določeni velikosti vzorca in obliki populacijske porazdelitve najboljše pokritje. V kolikor bo imelo več cenilk isto pokritje, nas bo zanimala tudi širina dobljenih intervalov zaupanja.

Key words documentation

Name and SURNAME: Slađana BABIĆ

Title of final project paper: The effects of nonnormal distribution on confidence interval around population mean

Place: Koper

Year: 2015

Number of pages: 39 Number of figures: 6 Number of tables: 20

Number of references: 11

Mentor: Assist. Prof. Rok Blagus, PhD

Keywords: confidence interval, coverage probability, width of the confidence interval, bootstrapping, simulations.

Math. Subj. Class. (2010): 62F25

Abstract:
The aim of the final project paper is to first derive confidence interval when the variable that is being studied follows normal distribution and in case it does not. In order to construct confidence interval, we will rely on Central Limit Theorem and on the Slutsky Theorem. After that, we will examine the coverage probability and the width of the confidence intervals for different sample sizes and different probability distributions, using simulations. We will explain what are simulations, when we used them, and why they are good. Also, we will briefly present some basic properties of the distributions we have used for simulations. We will explain what is bootstrapping and how does it work. At the end will be given results obtained using different methods for constructing confidence interval.

Acknowledgement

I would like to express my deep gratitude to my mentor, assist. prof. Rok Blagus, for him guidance, useful help and advice for my final project paper.

I would also like to thank Faculty of Mathematics, Natural Sciences and Information Technologies for their support through scholarship.

Finally, I wish to thank my family, especially my mother for support and encouragement throughout my study.

Contents

List of Tables

List of Figures

List of Abbreviations

i.e. that is
e.g. for example
i.i.d. independent and identically distributed
CI confidence interval
BCa bias corrected and accelerated
etc. and so on

1 Introduction

In statistics, estimation refers to the process by which one makes inferences about the population, based on information obtained from a sample. The sample statistic is calculated from the sample data and the population parameter is estimated from this sample statistics. There are two types of estimates: point estimates and interval estimates. The point estimate is usually different from the population parameter, because of the sampling error. Because of that, it is better to give an interval estimate, which is a range of values used to estimate the parameter. Confidence interval is the most commonly used interval estimate to make inferences about the population parameters from the sample data.

Imagine you are trying to find out how many days of vacation Slovenians have taken in the past year. You could ask every Slovenian about his or her vacation schedule to get the answer, but this would be expensive and time consuming. To save time and money, you would probably survey a smaller group of Slovenians. However, your finding may be different from the actual value if you had surveyed the whole population. That is, it would be an estimate. Each time you repeat the survey, you would likely get slightly different results. Commonly, when researchers present this type of estimate, they will put a confidence interval around it. The confidence interval is a range of values, in which the actual value is likely to fall. It represents the accuracy or precision of an estimate.

Here, we are interested in the confidence interval around the population mean. First of all, we will explain why do we need confidence interval, and how do we interpret it. Then, we will continue with presenting two ways of the derivation of the confidence interval. First, when the population variance is known and the variable that is being studied follows normal distribution, and the second one, when the population variance is unknown and the variable that is being studied follows t-distribution. After that, in the third and the fourth chapter, to explain better the derivation of the confidence interval we will present two important theorems; Central Limit Theorem and Slutsky Theorem. In chapter five, we will present a powerful statistical technique, bootstrapping. Using it, we can deal with samples in cases we can not assume a normal distribution or the t-distribution. Moreover, our aim is to examine the coverage probability of the confidence interval and its width by using different distributions and

simulations. So for the end, there will be results obtained by using simulations.

2 Confidence interval

Interval estimators and derivation of the confidence interval will be the main topic in this chapter. The fundamental idea of statistics is to analyze a sample of data, and to make quantitative inferences about the population, from which the data were sampled. Confidence intervals are the most straightforward way to do this.

Example 2.1. Say we are interested in the mean weight of 18-year-old girls living in the Europe. Since it would have been impractical to weigh all the 18-year-old girls in the Europe, we take a sample of for example 10 girls with weights of 51, 55, 49, 57, 62, 47, 51, 53, 59, and 56 kg, and find that the mean weight is 54 kilograms. The sample mean of 54 kg is a point estimate of the population mean.

Example 2.2. Let us look once again at the example from the beginning. Say we have asked 10 Slovenians about their vacation, and we have got the following results: 10, 12, 6, 15, 5, 7, 14, 8, 14 and 9 days. The sample mean in this case is 10 days.

If we just give estimate alone, that does not reflect a measure of the sampling error of the obtained value; we do not have a good sense of how far this sample mean may be from the population mean. Because of that we need confidence intervals, since they provide more information than the point estimates.

Definition 2.3. An **interval estimate** of a real-valued parameter θ, is any pair of functions $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$, of a sample that satisfies $L(x) \leq U(x)$ for all $x = (x_1, \ldots, x_n)$ from a sample space. If X=x is observed, the inference $L(x) \leq \theta \leq U(x)$ is made. The random interval $[L(X), U(X)]$ is called an interval estimator.

We will denote by $[L(X), U(X)]$ an interval estimator of θ and by $[L(x), U(x)]$ the realized value of the interval, based on random sample $X = (X_1, \ldots, X_n)$. Typically, confidence intervals are expressed as a two-sided range. We call this interval a 'two sided', because it is bounded by both lower and upper confidence limits. In some circumstances, it can make more sense to express the confidence interval in only one direction, to either the lower or upper confidence limit. For example, if $L(x) = -\infty$ then we have the one-side interval $(-\infty, U(x)]$. In other situations, it can make sense to express a one-sided confidence limit as a lower limit only, so we take $U(x) = \infty$ and we have $[L(x), \infty)$.

Definition 2.4. For the interval estimator $[L(X), U(X)]$ of a parameter θ, the **coverage probability** of $[L(X), U(X)]$ is the probability that the random interval $[L(X), U(X)]$ covers the parameter θ. It is denoted by $P_\theta(\theta \in [L(X), U(X)])$.

Remark 2.5. The interval is random quantity, not the parameter.

Remark 2.6. The probability statements refer to X not θ.

Definition 2.7. For an interval estimator $[L(X), U(X)]$ of a parameter θ, the **confidence coefficient** of $[L(X), U(X)]$ is the infimum of the coverage probabilities.

Interval estimators, together with confidence coefficient are known as confidence intervals. In general, when we are not sure in the exact form of our set, we will speak about confidence set with the confidence coefficient $1 - \alpha$. Usually, we are looking for 95% and 99% confidence intervals. The meaning of a confidence interval is frequently misinterpreted.
For the given data of the weight of the girls, 95% confidence interval is $[51, 57]$. What does that mean? If repeated samples were taken and the 95% confidence interval computed for the each sample, 95% of the intervals would contain the population mean. So, confidence intervals provide more information than the point estimates.

2.1 Confidence interval around population mean

This chapter is principally concerned with the confidence interval around the population mean. We will assume that the population is of size N and that each member of population $X_1,\ldots,X_N$, is determined with a numerical value. These numerical values will be denoted by $x_1,x_2,\ldots,x_N$. The variable x_i may be a numerical value, such as age or height. The population mean or average is defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

We will also need to consider the population variance, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$.

A useful identity can be obtained by transforming the expression:

$$\begin{aligned}
\sigma^2 &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(X_i^2 - 2x_i\mu + \mu^2) \\
&= \tfrac{1}{N}\left(\textstyle\sum_{i=1}^{N} X_i^2 - 2\mu\sum_{i=1}^{N} X_i + N\mu^2\right) \\
&= \tfrac{1}{N}\left(\textstyle\sum_{i=1}^{N} X_i^2 - 2N\mu^2 + N\mu^2\right) \\
&= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} X_i^2 - \mu^2
\end{aligned}$$

In order to calculate the confidence interval, first of all we have to select a sample from our population.

Simple random sampling is a basic type of sampling, with which every object has the same probability of being chosen. We will consider two cases, when it is done with and without replacement. Since we take a sample randomly, the sample mean will also be random. Next step is to calculate a sample mean and a standard deviation.
As an estimator of the population mean we will consider sample mean:

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

$\bar{X}$ is a random variable whose distribution is called sampling distribution. The sampling distribution depends on the X_i.

Lemma 2.8. *If we denote the values assumed by the population members by* $x_1, x_2, \ldots, x_N$, *and we assume that all members of the population have distinct values, then* X_i *is a discrete random variable with probability mass function*

$$P(X_i = x_j) = \frac{1}{N}.$$

Also holds,

$$E(X_i) = \mu$$

$$Var(X_i) = \sigma^2$$

Proof. From probability theory, we know that probability mass function of a discrete random variable X_i, in this case will be exactly equal to $P(X_i = x_j) = \frac{1}{N}$.
The expected value of the random variable X_i is:

$$E(X_i) = \sum_{j=1}^{N} x_j P(X_i = x_j) = \frac{1}{N}\sum_{j=1}^{N} x_j = \mu$$

To show the last equation we will use the definition of variance:

$$\begin{aligned} Var(X_i) &= E(X_i^2) - [E(X_i)]^2 \\ &= \tfrac{1}{N}\textstyle\sum_{j=1}^{N} x_j^2 - \mu^2 \\ &= \sigma^2 \end{aligned}$$

□

Theorem 2.9. *With simple random sampling the expected value of a sample mean,* $E(\bar{X})$ *is* μ.

Proof.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} n\mu = \mu$$

. □

In statistics, the bias of an estimator is the difference between the estimator's expected value and the true value of the parameter being estimated. An estimator with zero bias is called unbiased estimator, in all other cases is biased.
Obviously, the sample mean defined as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, is unbiased estimator of the population mean.
In order to find a standard deviation we have to find a variance, since a standard deviation is the square root of the variance. The variance of a random variable is a measure of its variability, and the covariance of two random variables is a measure of their joint variability, or their degree of association.

Definition 2.10. If X and Y are jointly distributed random variables, with expectations $E(X)$ and $E(Y)$ respectively, the covariance of X and Y is

$$Cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

provided that the expectations exists.

This expression can be simplified by expanding the product and using the linearity of the expectation.

$$\begin{aligned} Cov(X,Y) &= E\left[XY - XE(Y) - YE(X) + E(X)E(Y)\right] \\ &= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

In the case when X and Y are independent, $E(XY) = E(X)E(Y)$ so, $Cov(X,Y) = 0$. What we can say about $Cov(X_i, X_j)$ when $i \neq j$?
First, we will look at the case when the sampling was done with replacement. That means that the population element can be selected more than one time. Then, X_i are independent, and for $i \neq j$ the $Cov(X_i, X_j) = 0$, while $Cov(X_i, X_i) = E(X_i^2) - E(X_i)^2 = Var(X_i) = \sigma^2$.
Using the property of the variance for a linear combination of random variables, that $Var(\sum_{i=1}^{n} b_i X_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j Cov(X_i, X_j)$, we have:

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j)$$

Now we can find $Var(\bar{X})$.

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

The standard deviation of $\bar{X}$ is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

The other case is when the sampling is done without replacement, when a population

element can be selected only one time. This causes dependency among X_i. In order to find variance, first we need to find $Cov(X_i, X_j)$ for $i \neq j$. First of all, $Cov(X_i, X_j)$ are the same for all $i \neq j$, since they have the same distribution.
$Cov(X_i, \sum_{j=1}^{N} X_j) = 0$, because $\sum_{j=1}^{N} X_j$ is a constant. From this, we have

$$Cov(X_i, X_i) + \sum_{j=1, j\neq i}^{N} Cov(X_i, X_j) = 0$$

$$Cov(X_i, X_i) + (N-1)Cov(X_i, X_j) = 0$$

This implies, $Cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$ for $i \neq j$. We see the covariance depends on the population size. If the population is very large, the covariance is very close to zero. Using once again the property of the variance for a linear combination of random variables we get:

$$\begin{aligned}
Var(\bar{X}) &= Var\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i\right) \\
&= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \\
&= \tfrac{1}{n^2}\left(\textstyle\sum_{i=1}^{n} Cov(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i} Cov(X_i, X_j)\right) \\
&= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} Var(X_i) + \tfrac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} Cov(X_i, X_j) \\
&= \tfrac{\sigma^2}{n} - \tfrac{1}{n^2} n(n-1)\tfrac{\sigma^2}{N-1} \\
&= \tfrac{\sigma^2}{n}\left(1 - \tfrac{n-1}{N-1}\right) \\
&= \tfrac{\sigma^2}{n}\left(\tfrac{N-n}{N-1}\right)
\end{aligned}$$

Since the variance is defined and calculated as the average squared deviation from the population mean, intuitively the estimator of the population variance will be defined as the average squared deviation from the sample mean. Uncorrected sample variance defined as $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, is biased estimator of the population variance.

$$\begin{aligned}
E(\hat{\sigma}^2) &= E\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= E\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2\right] \\
&= E\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}((X_i - \mu) - (\bar{X} - \mu))^2\right] \\
&= E\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2)\right] \\
&= E\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\tfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu) + (\bar{X} - \mu)^2\right] \\
&= E\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2\right] = \sigma^2 - E\left[(\bar{X} - \mu)^2\right] < \sigma^2
\end{aligned}$$

For unbiased estimator of sample variance we suggest $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. In statistics, this is known as Bessel's correction for the sample variance [8].

Let us show that holds $E(s^2) = \sigma^2$.

$$\begin{aligned} E(s^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\ &= \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right] \\ &= \frac{1}{n-1}\left[n\sigma^2 + n\mu - n(\frac{\sigma^2}{n} + \mu^2)\right] \\ &= \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right] \\ &= \frac{1}{n-1}\sigma^2(n-1) = \sigma^2 \end{aligned}$$

Remark 2.11. $Var(X_i) = E(X_i^2) - E(X_i)^2$

$$E(X_i^2) = Var(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

$$E(\bar{X}) = Var(\bar{X}) + E(\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

Remark 2.12.

$$\begin{aligned} \sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i}^{n} X_i + n\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2\bar{X}n\bar{X} + n\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \end{aligned}$$

To sum up:

- sample mean is defined as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- If the variance of the population is known, we have $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$.
- If it is not known, we will use unbiased estimator $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

First of all, we will derive the confidence interval for the population mean when the standard deviation of the population is known, and the variable that is being studied follows the normal distribution.
In practice, the population standard deviation is rarely known. However, learning how to compute a confidence interval when the standard deviation is known, is an excellent introduction to how to compute a confidence interval when the standard deviation has to be estimated. To obtain this confidence interval, we need to know the sampling distribution of the estimator. Once we know the distribution, we can talk about the confidence interval. We said before that our assumptions will be that the variable is normally distributed. The normal distribution is easy to use, since it does not bring with it too much complexity.

Theorem 2.13. *[9] If X and Y are independent random variables that are normally distributed, then their sum is also normally distributed, i.e. if $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ then $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.*

As the consequence of this theorem, we obtain the following. If X_1, X_2,...,X_n are independent, normally distributed random variables, with mean μ and the standard deviation σ, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is normally distributed with the sample mean μ and the standard deviation $\frac{\sigma}{\sqrt{n}}$.

If the original population is normally distributed, we will use the above theorem. In the case the random variables do not have normal distribution, we will use the Central Limit Theorem. More about the Central Limit Theorem will be in the next chapter. To sum up, we have $X_i \sim N(\mu, \sigma)$, $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. But how is distributed $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$?

$$E\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{1}{\frac{\sigma}{\sqrt{n}}}E(\bar{X}-\mu) = \frac{1}{\frac{\sigma}{\sqrt{n}}}(\mu-\mu) = 0$$

$$Var\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{1}{\frac{\sigma^2}{n}}Var(\bar{X}-\mu) = \frac{1}{\frac{\sigma^2}{n}}\frac{\sigma^2}{n} = 1$$

The special case for which $\mu = 0$ and $\sigma = 1$ is called standard normal distribution, which is denoted as $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$. If we are interested in the probability that a standard normal variable Z will fall between two values, for example -z in z, we can denote that as $P(-z < Z < z)$.

We will denote by z the value from the standard normal distribution, for the selected confidence level (e.g. for a 95% confidence level z=1.96). So, when we are looking for a $(1-\alpha)$ confidence interval we will have:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha$$

$$P(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}) = 1-\alpha$$

$$P(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(-\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

From the last expression, we obtain the formula for confidence interval for the population mean. If the standard deviation is known, it will be $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. The lower limit is obviously $\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, and the upper limit is $\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

In practice, we often do not know the value of the population standard deviation. In that case, we should use the t-distribution, rather than the normal distribution. First we have to do, is to estimate standard deviation from the sample data. Since we

use s as an estimator of the population variance, intuitively to estimate σ, we will use $s_{\bar{X}} = \frac{s}{\sqrt{n}}$. First, let us look how is distributed $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$.

$$\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{s^2}{\sigma^2}}\sqrt{\frac{n-1}{n-1}}}$$

We already know that $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Next we will do, is to look at the distribution of $\frac{s^2(n-1)}{\sigma^2}$.

$$\begin{aligned} s^2(n-1) &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(X_i-\bar{X})^2(n-1) \\ &= \textstyle\sum_{i=1}^{n}(X_i-\bar{X})^2 \\ &= \textstyle\sum_{i=1}^{n}(X_i-\mu+\mu-\bar{X})^2 \\ &= \textstyle\sum_{i=1}^{n}(X_i-\mu)^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2 \end{aligned}$$

This implies the following:

$$\begin{aligned} \tfrac{s^2(n-1)}{\sigma^2} &= \tfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \\ &= \textstyle\sum_{i=1}^{n}\left(\left(\frac{X_i-\mu}{\sigma}\right)^2-\left(\frac{\bar{X}-\mu}{\sigma}\right)^2\right) \\ &= \textstyle\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2-n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 \\ &= \textstyle\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2-\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \end{aligned}$$

Observe, $Z_i = \frac{X_i-\mu}{\sigma} \sim N(0,1)$ and $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Before we continue, a few remarks regarding to relationship between the normal distribution and χ^2-distribution.
If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$.
If $X_1, \ldots, X_n$ are independent, standard normal random variables, with mean 0 and variance 1, then the sum of their squares has the χ^2-distribution with n degrees of freedom.

$$X_1^2 + \cdots + X_n^2 \sim \chi^2_n$$

If $X_1 \sim \chi^2_n$ and $X_2 \sim \chi^2_m$, and they are independent, then $X_1 + X_2 \sim \chi^2_{n+m}$
Using the above definitions, we have the following:

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$$

and

$$\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \sim \chi^2_1.$$

And finally,

$$\frac{s^2(n-1)}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2-\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \sim \chi^2_{n-1}.$$

At the very beginning we said that we will use t-distribution.
Student's t-distribution with ν degrees of freedom can be defined as the distribution of the random variable T with

$$T = \frac{Z}{\sqrt{V/\nu}}$$

where

- Z has a standard normal distribution;
- V has a χ^2-distribution with ν degrees of freedom;
- Z and V are independent.

We are interested in distribution of $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$. The first thing we did, was the transformation of that expression into $\frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{s^2}{\sigma^2}}\sqrt{\frac{n-1}{n-1}}}$. If we look again carefully in our case, we will observe the following:

- $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
- $V = \frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$ with n-1 degrees of freedom.

The only thing we still have to check is the independence of Z and V.
If we look at the normally distributed variables $X_1, \ldots, X_n$, vector $(X_1, \ldots, X_n)$ is jointly normally distributed, i.e. it is so distributed that every linear combination $a_1X_1 + \cdots + a_nX_n$ has a 1-dimensional normal distribution.

Theorem 2.14. *[10] If $X_1, \ldots, X_n$ are jointly normally distributed, uncorrelated and $Cov(X_i, X_j) = 0$ for all $i \neq j$, then the X_i are independent.*

Using the above theorem, we have to check that $cov\left(\bar{X}, \begin{bmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix}\right) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$.

But it is enough to check that $Cov(\bar{X}, X_i - \bar{X}) = 0$.

$$Cov(\bar{X}, X_i - \bar{X}) = Cov(\bar{X}, X_i) - Cov(\bar{X}, \bar{X}) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} X_i, X_i\right) - Var(\bar{X})$$

$$= \frac{1}{n}Cov(\sum_{i=1}^{n} X_i, X_i) - \frac{\sigma^2}{n} = \frac{1}{n}\left[Cov(X_i, X_i) + \sum_{j\neq i} Cov(X_j, X_i)\right] - \frac{\sigma^2}{n}$$

$$= \frac{1}{n}Var(X_i) - \frac{\sigma^2}{n} = \frac{1}{n}\sigma^2 - \frac{\sigma^2}{n} = 0$$

It follows that $\frac{Z}{\sqrt{\frac{V}{n-1}}} \sim t_{n-1}$.
Let us denote $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$. Then:

$$P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$$

$$P(-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \leq t_{\alpha/2}) = 1 - \alpha$$

$$P(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}) = 1 - \alpha$$

Then, the formula for a confidence interval for μ when σ is unknown will be:

$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

The values of t are larger than the values of z, so confidence intervals when σ is estimated are generally wider than confidence intervals when σ is known.
Constructing confidence intervals with the t-distribution is the same as using the normal distribution, except it replaces the z-score with a t-score.
Recall the above formula for calculating the confidence interval for a mean. Notice again, in our calculations we used the sample standard deviation s, instead of the true population standard deviation σ. This estimation of σ introduces extra error, and this extra error can be pretty big when sample size is not enough large. Because s is a poor estimator of σ with a small sample size, we will not assume that the sample distribution is normal. Instead, we will use the t-distribution, which is designed to give us a better interval estimate of the mean when we have a small sample size.

For the end of this part just a few remarks regarding the formula of the confidence interval. We said $1 - \alpha$ is a confidence coefficient. So, α is the value we choose at the beginning, and the most commonly used confidence levels are 95%, 99% or sometimes 90%. To find the critical value, or $z_{\alpha/2}$ we use tables for a standard normal distribution, where the values of the cumulative distribution function of the normal distribution are given. Or, when we are using t-distribution, the critical value $t_{\alpha/2,df}$ is obtained from tables for a t-distribution.

3 Central Limit Theorem

Up to this point, we started from the assumption that the variable follows the normal distribution. In this chapter we will discuss one of the fundamental theorems of probability - the Central Limit Theorem, since CLT enables us to use the approximate formula for the CI based on standard normal distribution, even when the variable that is being studied does not follow the normal distribution.

Theorem 3.1. *Let $X_1, \ldots, X_n$ be a sequence of independent random variables having mean μ and variance σ^2. Let each X_i have the distribution function $P(X_i \leq x) = F(x)$ and the moment generating function $M(t) = E(e^{tX_i})$. Let $S_n = \sum_{i=1}^{n} X_i$. Then*

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$$

for $-\infty < x < \infty$.

Before we give a proof, we need a few facts about moment generating functions. Recall, the moment generating function of a random variable X is $M_X(t) = E(e^{tX})$. One of the properties of the moment generating functions is, if the moment generating function exists in an open interval containing zero, then $M^{(r)}(0) = E(X^r)$.

Proposition 3.2. *If X and Y are independent random variables with mgfs M_X and M_Y, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.*

Since the proof is quite simple, even though it isn't the main topic we will give it.

Proof.

$$M_{X+Y}(t) = E(e^{tX+tY}) = E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t)$$

□

Proposition 3.3. *If X is random variable with mgf M_X, and $Y = a + bX$, then $M_Y(t) = e^{at}M_X(bt)$.*

Proof.

$$M_Y(t) = E(e^{tY}) = E(e^{at+btX}) = E(e^{at}e^{btX}) = e^{at}E(e^{btX}) = e^{at}M_X(bt)$$

□

To prove the CLT, we will also need the following theorem, and we will skip the proof of it.

Theorem 3.4. *Let F_n be a sequence of a cumulative distribution functions, or just distribution functions, with the correspond moment generating functions M_n. Let F be a distribution function with the moment generating function M. If $M_n(t) \to M(t)$, for all t in an open interval containing zero, then $F_n(x) \to F(x)$ for all x at which F is continuous.*

So now we can give the proof of the Central Limit Theorem.

Proof. It suffices to do the proof in the case $\mu = 0$. In the case $\mu \neq 0$, let $Y_i = X_i - \mu$ for each i. Let $T_n = Y_1 + \cdots + Y_n$. Then we have

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \lim_{n\to\infty} P\left(\frac{T_n}{\sigma\sqrt{n}} \leq x\right).$$

Obviously, it is enough to prove the theorem for $\mu = 0$.
Let us denote $Z_n = \frac{S_n}{\sigma\sqrt{n}}$. Using the above theorem, we see it is enough to show that the mgf of a standardized sum of n independent, identically distributed random variables approaches the mgf of a standard normal random variable as $n \to \infty$. So, we will show that the mgf of Z_n tends to the mgf of the standard normal distribution. Since S_n is a sum of independent random variables, using the first proposition, we have $M_{S_n}(t) = [M(t)]^n$, and by second proposition, we have $M_{Z_n}(t) = \left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$.
We will look at the limit of $\log[M_{Z_n}(t)]$. First, $\log[M_{Z_n}(t)] = \log\left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n = n\log\left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]$. We will denote $\frac{1}{\sqrt{n}}$ by x. Then we have,

$$L = \lim_{x\to 0} \frac{\log\left[M(\frac{tx}{\sigma})\right]}{x^2}.$$

Since $M(0) = 1$, to calculate the limit we will use l'Hospital's rule.

$$L = \lim_{x\to 0} \frac{\frac{M'(\frac{tx}{\sigma})\frac{t}{\sigma}}{M(\frac{tx}{\sigma})}}{2x} = \frac{t}{2\sigma}\lim_{x\to 0}\frac{M'(\frac{tx}{\sigma})}{xM(\frac{tx}{\sigma})} = \frac{t^2}{2\sigma^2}\lim_{x\to 0}\frac{M''(\frac{tx}{\sigma})}{M(\frac{tx}{\sigma}) + xM'(\frac{tx}{\sigma})\frac{t}{\sigma}}$$

$$= \frac{t^2}{2\sigma^2}\frac{M''(0)}{M(0) + 0M'(0)\frac{t}{\sigma}} = \frac{t^2}{2\sigma^2}\frac{M''(0)}{M(0)}$$

Using property that $M^{(r)}(0) = E(X^r)$, we get $M(0) = E(1) = 1$, $M'(0) = E(X)$ and $M''(0) = E(X^2) = Var(X) + E(X)^2 = \sigma^2$. So, we have $L = \frac{t^2}{2}$, which is exactly the logarithm of the moment generating function of the standard normal distribution [5].
□

But how the Central Limit Theorem is connected with a finding confidence interval for the population mean?

The central limit theorem states, if you have a population with mean μ and standard deviation σ, and take sufficiently large random sample from the population, then the distribution of the sample mean will be approximately normal. This will be true regardless of whether the distribution in the population is normal or not, provided that the sample size is sufficiently large. But what do we do if we want to calculate confidence interval for a sample of insufficiently large size?

We use the t-distribution, but only if we feel it is appropriate to assume that the population distribution itself is normal, or close to normal. By CLT, $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is distributed normally, even if the population distribution is not normal. But in practice σ is rarely known. Because of that, we look at the $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$, and we are interested in the distribution of it in the case when the population distribution is not normal. In that case we rely on the results of both the CLT, and Slutsky theorem. More about Slutsky theorem will be in the next chapter.

4 Slutsky Theorem

Looking for a confidence interval when we did not know the value of the population standard deviation, what we have done first was to find the distribution of $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$.

Using Slutsky's theorem we will show that $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \xrightarrow{d} Z$, where Z is random variable with standard normal distribution. So, first of all Slutsky's theorem.

Theorem 4.1. *[6] Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be random variables. Suppose that X_n converges in distribution to random variable X, i.e. $(X_n \xrightarrow{d} X)$, and Y_n converges in probability to a constant c, i.e. $(Y_n \xrightarrow{p} c)$, then:*

$$X_n + Y_n \xrightarrow{d} X + c$$

$$Y_n X_n \xrightarrow{d} cX$$

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c} \quad if \quad c \neq 0$$

Remark 4.2. A sequence $X_1, X_2, \ldots$ of random variables is said to converge in distribution to a random variable X, if $\lim_{n\to\infty} F_n(x) = F(x)$ for every $x \in R$ at which F is continuous. Here, F_n and F are distributions functions of random variables X_n and X respectively.

A sequence $X_1, X_2, \ldots$ of random variables is said to converge in probability towards the random variable X, if for all $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0.$$

If $X_n \xrightarrow{d} X$ and $P(X = c) = 1$, where c is constant, then $X_n \xrightarrow{p} c$. Convergence in probability implies the convergence in distribution, so we also have a definition of convergence in probability towards a constant.

As we said before, we will show that $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \xrightarrow{d} Z$, where Z is random variable with standard normal distribution. Before we do that, we need one theorem.

Theorem 4.3. *If $X_1, \ldots, X_n$ are i.i.d. with $E(X_i) = \mu$, the weak law of large numbers states that the sample average, $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$, converges in probability towards the expected value, when $n \to \infty$:*

$$\bar{X} \xrightarrow{p} \mu.$$

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2.
The CLT tell us that $Z_n = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is approximately $N(0,1)$. But we rarely know σ. We have seen before that we can estimate it by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.
We will show that if we replace σ by s, then for $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ will still hold that approximately is $N(0,1)$.
Denote $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.
First, we will show that $S^2 \xrightarrow{p} \sigma^2$, where $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2$.

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

Let us define $Y_i = X_i^2$. Then, by the weak law of large numbers we have:

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 = \frac{1}{n}\sum_{i=1}^{n}Y_i \xrightarrow{p} E(Y_i) = E(X_i^2) = Var(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

Again, by the same law we have $\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{p} E(X_i) = \mu$.
Since $f(x) = x^2$ is continuous, we will have $\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2 \xrightarrow{p} \mu^2$, because continuous functions are limit-preserving. So, $S^2 \xrightarrow{p} (\sigma^2 + \mu^2) - \mu^2 = \sigma^2$.
But we want to see what will happen if we replace σ by s.

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

Since, $S^2 \xrightarrow{p} \sigma^2$ and $\frac{n}{n-1} \to 1$, we will have $s^2 \xrightarrow{p} \sigma^2$.
Once again, we can use that continuous functions are limit-preserving, so $s \xrightarrow{p} \sigma$.
Then we have, $\frac{S}{\sigma} \xrightarrow{p} 1$ and using that continuous functions are limit-preserving we obtain $\frac{\sigma}{s} \xrightarrow{p} 1$.
Finally,

$$T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{X}-\mu)}{s}\frac{\sigma}{\sigma} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\frac{\sigma}{s}$$

Let us denote $Z_n = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ and $V = \frac{\sigma}{s}$. By Slutsky's theorem, $T = Z_n V \xrightarrow{d} Z \cdot 1 = Z$, since $Z_n \xrightarrow{d} Z$ and $V \xrightarrow{p} 1$.

If we do not know the distribution of the data we are working with, or do not feel comfortable making assumptions of normality, we rely on CLT and the Slutsky theorem, since we can not use the t-distribution without assumption of normality. Therefore, by the Central Limit and the Slutsky theorem we can use the asymptotic properties of the statistic $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$, to form confidence intervals based on the standard normal distribution, without making any assumptions about the distribution of the sample data, and using s^2 to estimate σ^2. An important note to remember, it is often the

case that people say as n becomes large the normal distribution approximates the t-distribution, but in fact, as shown above, as n becomes large the formulation above (T) actually approximates the normal distribution (again based on the CLT and the Slutsky theorems).

5 Bootstrapping

Bootstrapping is a computer-based method which can answer questions that are complicated for traditional statistical analysis. Most commonly, bootstrap is used to estimate the variance of the estimators that can not be evaluated theoretically. It is a powerful statistical technique, which works quite well, even with samples of a small size and when we do not know anything about the distribution of our data.

Up to now, when we wanted to determine the confidence interval, we had to assume the distribution of the population, and in some cases we also had to know the standard deviation.

But bootstrapping method does not require anything other than the sample, and assumes that each sample is identically and independently distributed. Basic idea of bootstrapping is that inference about population from the sample data, can be modeled by resampling the sample data and preforming inference on resample. Bootstrap samples are obtained by randomly sampling with replacement, to obtain samples with the same size as the original sample.

So, sample from the population becomes 'population' and resample is a 'sample'. As the population is unknown, the quality of the inference from the sample is also unknown, we can not be sure about the sampling error. But using a bootstrap method, the population is in fact the sample and that is known, so the quality of the inference from the resample data is measurable. With the following numerical example we will demonstrate how the process works.

Example 5.1. Assume that our sample is 1,2,3,3,10. Our goal will be a 90% confidence interval about the mean of the sample. We begin with a sample from a population that we know nothing about. Next we do is to form bootstrap samples. Each bootstrap sample will have the same size as a original sample. In our case, that is five. Bootstrap samples may be different from the original sample and from each other, since we are randomly selecting and replacing each value. We will take 20 bootstrap samples:
2,1,10,3,2; 3,10,10,2,3; 1,3,1,3,3; 3,1,1,3,10; 3,3,1,3,2;
3,10,10,10,3; 2,3,3,2,1; 2,3,1,10,3; 1,10,2,10,10; 3,3,3,3,3;
3,3,3,3,1; 1,2,3,3,2; 3,3,10,10,2; 3,2,1,3,3; 3,1,10,1,10;
3,2,3,1,1; 3,3,3,2,3; 10,3,1,3,3; 3,2,1,10,2; 10,2,2,1,1.
Now we calculate the means of each of our bootstrap samples. These means, arranged

in ascending order are: 2, 2.2, 2.2, 2.2, 2.4, 2.4, 2.6, 2.8, 3, 3.2, 3.6, 3.6, 3.6, 3.8, 4, 11, 5.6, 5.6, 6.6, 7.2. From this bootstrap sample means, we can obtain a confidence interval. For our example above we have a confidence interval $[2.2, 6.6]$. The CI is than obtained by calculating the 5th and the 95th percentile of the obtained distribution.

Next we will do, is to present conditions that must be satisfied in order to bootstrapping procedure gives a reliable results.
Suppose we have a random sample $X_1, \ldots, X_n$ with values $x_1, \ldots, x_n$. Its empirical distribution function is defined as $\hat{F}(x) = \frac{\#\{x_j \leq x\}}{n}$.

Remark 5.2. $\#A$ means the number of times the event A occurs.

Using this empirical distribution function we want to estimate some properties of some quantity, say T. So we want to estimate a distribution function $G_{F,n}(t) = P(T \leq t)$. Here, the term $\{F, n\}$ indicates that we take a sample of size n from the F. The bootstrap estimate of the last expression will be $G_{\hat{F},n}(t) = P(T \leq t)$, and similarly, we take a sample of size n from the $\hat{F}$.

Since we have a sample, next step is to take resamples from it. Say we take B resamples. It is important to emphasize that, we are not using resamples to obtain some information about the population. We are using them to learn something about the distribution of the sample statistic. So, we try to approximate the sampling distribution of some statistic by resampling the sample, and calculating the statistic on the resamples. In the end, the distribution of the wanted parameter T is approximated through the empirical distribution of the B estimates for T, since we have taken B resamples.

In order to obtain reliable results, or in other words, in order to $G_{\hat{F},n}$ approaches $G_{F,n}$ as $n \to \infty$, three conditions must hold.

Suppose that $\mathcal{N}$ is the neighbourhood for F, in a suitable space of distributions. If we want to $\hat{F}$ walls into $\mathcal{N}$ with probability 1, then the following conditions must hold: [2]

1. For any $A \in \mathcal{N}, G_{A,n}$ must converge weakly to a limit $G_{A,\infty}$.
2. This convergence must be uniform on $\mathcal{N}$.
3. The function mapping A to $G_{A,\infty}$ must be continuous.

The first condition tells us that there is a limit for $G_{F,n}$. As n increases, $\hat{F}$ changes, and the second and third conditions are needed to ensure that $G_{\hat{F},n}$ approaches $G_{F,\infty}$ along every possibly sequences $\hat{F}s$.

Remark 5.3. Weak convergence of $G_{A,n}$ to $G_{A,\infty}$ means that for all integrable functions h,

$$\int h(u)dG_{A,n}(u) \to \int h(u)dG_{A,\infty}(u)$$

as $n \to \infty$.

Under this conditions the bootstrap is reliable, meaning that for any t and ϵ, $P(G_{\hat{F},n}(t) - G_{F,n}(t) > \epsilon) \to 0$ as $n \to \infty$.
There are several methods for constructing confidence intervals from the bootstrap distribution of a real parameter: basic, percentile, studentized, bias-corrected and BCa method. Later on, for simulations we will use percentile and BCa method, because of that we will briefly explain how do they work.

Percentile method uses B statistics, computed from the bootstrap samples. We arranged them in an ascending order and if we are looking for the $100(1-\alpha)$ confidence interval we take 50α and $100 - 50\alpha$ percentiles as limits of the interval.

Important issues for the bootstrap, and inference in general, are skewness and bias since bias estimates can have high variability.

The computation of the BCa confidence interval is a bit more complicated. It proceeds in three steps. First, we take a B resamples. Next we have to do is to calculate a bias correction value. The bias correction coefficient adjusts for the skewness in the bootstrap sampling distribution. If the bootstrap sampling distribution is perfectly symmetric, then the bias correction will be zero. At the end we calculate acceleration value. The acceleration coefficient adjusts for nonconstant variances, within the resampled data sets [4]. The formulas for calculating this parameters are quite complicated, and not so intuitive. More about that is given in Efron and Tibshirani [3].

6 Probability distributions

In this section, we will present some basic and most commonly used probability distributions that we will use later for simulations. A function describing the possible values of a random variable, and their associated probabilities is known as a probability distribution. We know that random variables can be discrete, that is, taking any of a specified finite or countable list of values, with a probability mass function, or continuous, taking any numerical value in an interval or collection of intervals, via a probability density function.

6.1 The Normal Distribution

The most familiar continuous distribution is the normal distribution. The normal distribution has two parameters, usually denoted by μ and σ^2, which are its mean and variance. The probability density function of the normal distribution is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

for $-\infty < x < \infty$.
One of the reasons why the normal distribution is one of the most important is Central Limit Theorem, which shows that normal distribution can be used to approximate a large variety of distributions in large samples. Some of the properties of the normal distribution are:

- the mean, median(the middle number in a set of data when it is ranked from lowest to highest) and mode(the number that occurs most frequently in a data set) are equal.

- it is symmetrical. This means that if the distribution is cut in half, each side would be the mirror of the other

- the total area under the curve is equal to one. The total area, however, is not shown. This is because the tails extend to infinity.

- the area under the curve can be determined. If the standard deviation is known, one can determine the percentage of data under sections of the curve.

- around 68% of the area of a normal distribution is within one standard deviation of the mean.
- Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

The figure below shows how the percentages of area under the normal curve are distributed in terms of standard deviation units from the mean.

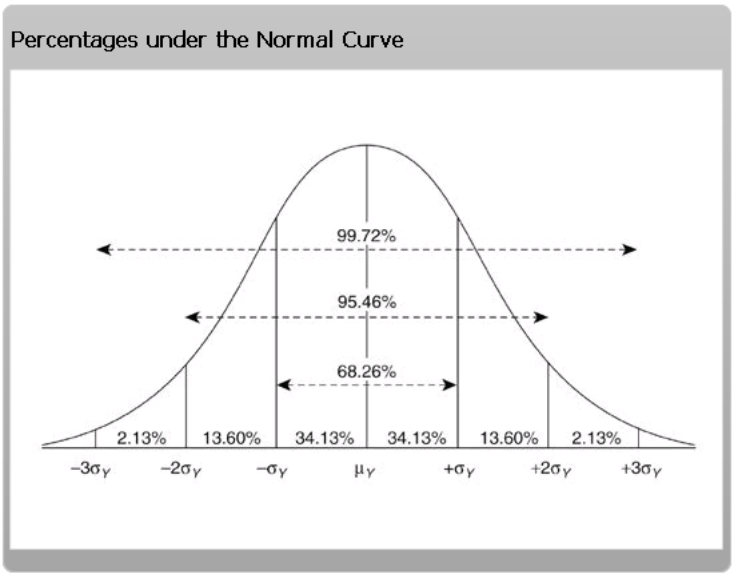

Figure 1: The percentages of area under the normal curve(figure is from [7])

The normal distribution is very important distribution, since it is based on theory, rather than on real data. Many things in life never match this model perfectly, but approximately they have the normal distribution. Sometimes, we say that normal distribution is actually a family of normal distributions, since each of them is characterized by its mean and a standard deviation.

The n-th moment of the probability distribution of the variable X, if exists, is defined as $\mu_n = E(X^n)$. The zeroth moment is the total probability, the first moment is the mean, the second moment is the variance, and because they are used so frequently we will give them for each distribution individually. As we said, the normal distribution is actually determined by them. For example, below are shown normal distributions with $\mu = 0, \sigma = \frac{1}{2}$ (solid curve), $\mu = 1, \sigma = 1$ (dotted curve) and $\mu = 0, \sigma = 2$ (dashed curve).

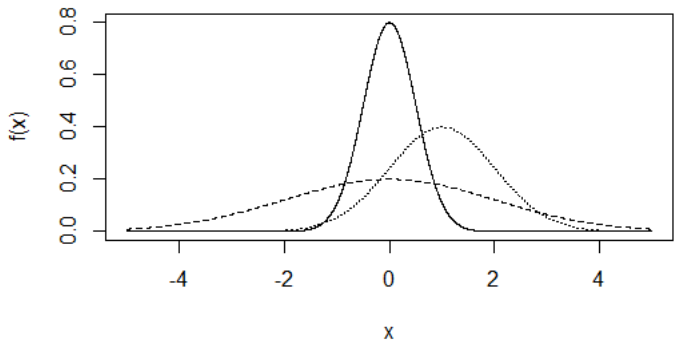

Figure 2: Probability density function of normal distribution

6.2 The Gamma Distribution

The gamma distribution is another widely used distribution. A continuous random variable X is said to have a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$, shown as $X \sim Gamma(\alpha, \lambda)$, if its probability density function is given by

$$f(x) = \frac{\lambda^{\alpha} x^{\alpha-1} e^{-x\lambda}}{\Gamma(\alpha)}$$

where $x > 0$, and $\Gamma(\alpha)$ is a gamma function.
Some properties of the gamma probability density function are:

- if $0 < \alpha < 1$, f is decreasing with $f(x) \to \infty$ as $x \to 0$.
- if $\alpha = 1$, f is decreasing with $f(0) = 1$.
- if $\alpha > 1$, f increases and then decreases, with mode at $(\alpha - 1)\lambda$.
- if $0 < \alpha \leq 1$, f is concave upward.
- if $1 < \alpha \leq 2$, f is concave downward and then upward, with inflection point at $\lambda(\alpha - 1 + \sqrt{\alpha - 1})$.
- if $\alpha > 2$, f is concave upward, then downward, then upward again, with inflection points at $\lambda(\alpha - 1 \pm \sqrt{\alpha - 1})$.
- $E[X] = \lambda\alpha$
- $var(X) = \lambda^2\alpha$.

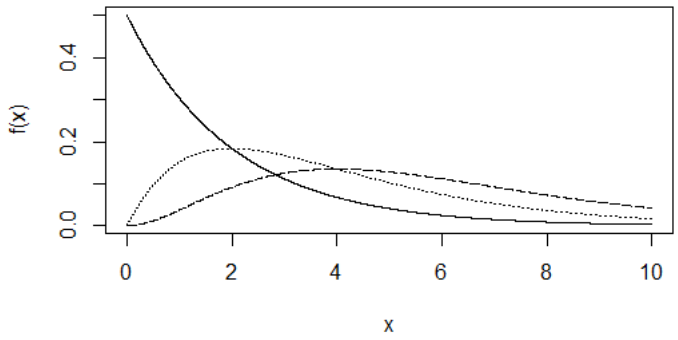

Figure 3: Probability density function of gamma distribution

Above are shown probability density function of the gamma distribution for different values of α, for $\alpha = 1$ and $\lambda = 1/2$ (solid curve), for $\alpha = 2$ and $\lambda = 1/2$ (dotted curve) and for $\alpha = 3$ and $\lambda = 1/2$ (dashed curve).

In case when $\alpha = 1$ we have a exponential distribution.

6.3 The Exponential Distribution

The probability density function of an exponential distribution is

$$f(x) = \begin{cases} \lambda e^{-x\lambda} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{6.1}$$

Some properties of it are:

- f is decreasing on $[0, \infty)$.
- f is concave upward on $[0, \infty)$.
- $f(x) \to 0$ as $x \to \infty$.
- $E[X] = \frac{1}{\lambda}$
- $var(X) = \frac{1}{\lambda^2}$.

Below are shown probability density functions of the exponential distribution for different values of λ. For $\lambda = 1/2$ (solid curve), for $\lambda = 1$ (dotted curve) and for $\lambda = 2$ (dashed curve).

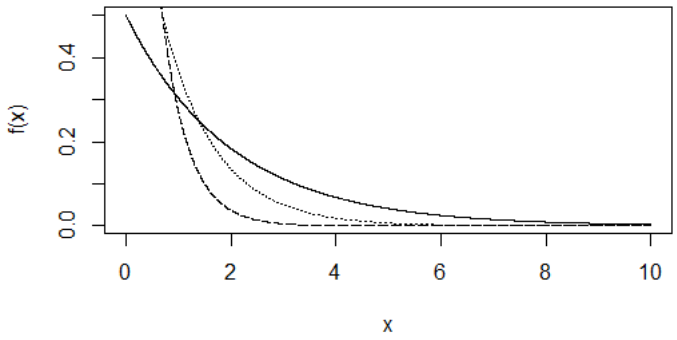

Figure 4: Probability density function of exponential distribution

6.4 The Uniform Distribution

The uniform distribution is the simplest continuous random variable you can imagine. A continuous random variable X is said to have a uniform distribution over the interval $[a, b]$, shown as $X \sim Uniform(a, b)$, if its probability density function is given by:

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x < a \quad or \quad x > b \end{cases} \tag{6.2}$$

Expected value of a random variable X that is uniformly distributed is $E[X] = \frac{1}{2}(a+b)$, and the variance is $var(X) = \frac{1}{12}(b-a)^2$. Below is shown a probability density function of the uniform distribution over the interval $[1, 3]$.

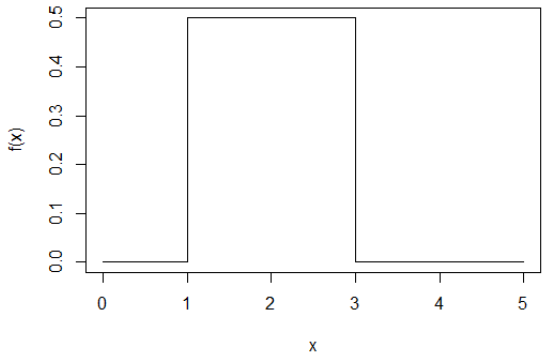

Figure 5: Probability density function of uniform distribution

6.5 The Pareto type I Distribution

The Pareto Distribution was first proposed as a model for the distribution of incomes. The probability density function is $f(x) = \frac{ab^a}{x^{a+1}}$, for $a > 0$ and $b \leq x < \infty$. The parameter b is a lower bound on the possible values that a Pareto distributed random variable can take on. A well known properties of it are:

- $E[X] = b\frac{a}{a-1}$ if $a > 1$.
- $Var(X) = b^2\frac{a}{(a-1)^2(a-2)}$ if $a > 2$.

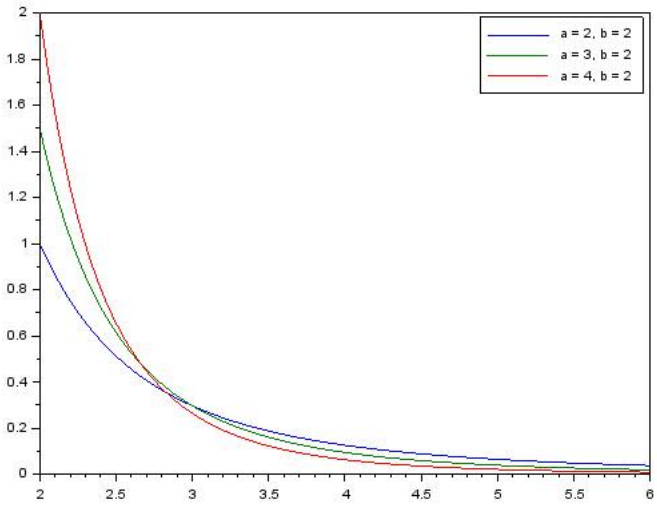

Figure 6: Probability density function of Pareto distribution

Above are shown probability density functions of the Pareto distribution for different values of a.

7 Simulations

Up to now, for the construction of the confidence interval around population mean, we have relied on the CLT or the Slutsky's theorem in case $n \to \infty$. We would also like to know what happens in case we have small n. Also, we are interested in the influence of the distribution. What happens if the distribution is symmetric, and what if it is asymmetric. Using simulations, we provide answers to these questions.

Simulation is a modeling of a random events, by using random numbers to specify random event outcomes, in order to closely match real-world outcomes. It is a numerical technique for performing experiments on the computer. Properties of the statistical simulations must be determined in a such way that method gives a reliable results, but exact derivations of properties are rarely possible.

There are many reasons for using simulations in statistics. For example, some situations are difficult to analyze, time-consuming and very often expensive. Using simulations, we approximate real-world results, and at the same time we save our time, money and we need less effort. Simulation is useful only if simulated outcomes closely match real-world outcomes. If we want to produce a useful simulation, first of all we describe what are possible outcomes. Next we do, is to assign to each outcome one or more random numbers. Also we have to choose a source of random numbers. For example, that can be random number generator, since in case we have sample of large size, this is not time consuming. Then we choose a random number, and based on it we have simulated outcome. We select the numbers and states the simulated results until we get a stable pattern. In the end, we just have to analyze the simulated outcomes.

Here we will use simulations to check the coverage probability of the confidence interval we have found and its width. We will check what happens if the population has normal, exponential, gamma, uniform and Pareto distribution. Also, we will change sample sizes and observe how that affects on the coverage and on the width. We will observe for $n = 10$, 25, 50 and 100. Narrow width and high confidence level are desirable, and because of that we are looking for how large n we can get coverage close to 0.95.

We assumed that true mean in the population is 10, and code for simulations is made in such way, to return also expected value of the sample mean. Furthermore, code also provides the standard error of the sample mean. But the most important

results are the average width of the confidence intervals, and the coverage. We will calculate all of that, using four methods, namely: derivation of the confidence interval from normal distribution, from t-distribution, using percentile method and using BCa method. Code was written in R programming language [11].

7.1 Results of simulations

In case the population is distributed normally and we have a sample of size $n = 10$ the expected value of the sample mean was 9.999085 and standard error was 0.316792. Regarding the average width and the coverage probability we obtain the following:

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	1.205833	1.391752	1.142680	1.162475
coverage probability	0.91774	0.94920	0.90214	0.90012

Table 1: Normal distribution and $n = 10$

For $n = 25$ the expected value is 10.0004 and the standard error is 0.199707.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	0.7764539	0.8176284	0.7629257	0.7661205
coverage probability	0.93788	0.95006	0.93200	0.93138

Table 2: Normal distribution and $n = 25$

In case for $n = 50$ the expected value was 9.999718 and the standard error was 0.1408864.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	0.5514976	0.5654573	0.5477670	0.5484882
coverage probability	0.94562	0.95146	0.94254	0.94298

Table 3: Normal distribution and $n = 50$

For $n = 100$ the expected value was 10.00033 and the standard error was 0.1003401.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	0.3909936	0.3958318	0.3904553	0.3906429
coverage probability	0.94668	0.94924	0.94610	0.94494

Table 4: Normal distribution and $n = 100$

Regardless of the sample size, the estimate of the population mean was unbiased and

as expected, the standard error decreased when the sample size was larger. If we look at the above tables, we see the coverage probability is the best in case when we used the t-distribution for constructing the confidence interval, since for any size of a sample, the coverage probability was almost 0.95. We obtained the narrowest confidence interval in case we used the bootstrapping, but coverage probability for $n = 10$ was too much liberal, around 0.90. Later, for $n = 50$ and more, the coverage was almost 0.95. So, the bootstrapping works in case of a large sample sizes. But, as n became bigger the results were more or less the same in all cases. We can see that in table for $n = 100$.

Then we have observed what happens if the distribution is exponential. We will give results for sample sizes as in case for normal distribution.

In case for $n = 10$ the expected value was 10.00971 and the standard error was 3.17514.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	11.46271	13.23006	10.73971	11.71961
coverage probability	0.86934	0.89980	0.86290	0.87670

Table 5: Exponential distribution and $n = 10$

For $n = 25$ the expected value was 9.982807 and the standard error was 1.98841.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	7.556746	7.957471	7.389153	7.817267
coverage probability	0.91278	0.92412	0.91276	0.92158

Table 6: Exponential distribution and $n = 25$

For $n = 50$ the expected value was 10.00011 and the standard error was 1.412062.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	5.439803	5.577497	5.390721	5.576672
coverage probability	0.92948	0.93534	0.93028	0.93462

Table 7: Exponential distribution and $n = 50$

For $n = 100$ the expected value was 10.00424 and the standard error was 1.002702.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	3.884842	3.932914	3.874950	3.948652
coverage probability	0.94014	0.94320	0.94080	0.94192

Table 8: Exponential distribution and $n = 100$

Again, regardless of the sample size, the estimate of the population mean was unbiased and the standard error decreased when the sample size was larger.
When the sample size was small ($n = 10$), the coverage probabilities were too small, especially when using a normal distribution and BS percentile method, where the coverage probability was only at around 0.86. The best, although still much too liberal confidence interval, was obtained when using the t-distribution, where the average width of the confidence interval was the largest.
When the sample size increased, the coverage probabilities of all methods improved substantially, and were very close to 0.95 when there were 100 samples.
Also we observed what happens if we have gamma distribution.
For $n = 10$ we obtained the expected value equal to 9.99489 and standard error was 2.232085.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	8.299800	9.579488	7.820109	8.257750
coverage probability	0.89320	0.92384	0.88302	0.88920

Table 9: Gamma distribution and $n = 10$

For $n = 25$ the expected value was 9.994553 and the standard error was 1.420917.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	5.408255	5.695049	5.299634	5.472949
coverage probability	0.92360	0.93496	0.92164	0.92430

Table 10: Gamma distribution and $n = 25$

For $n = 50$ the expected value was 10.00054 and the standard error was 0.9965776.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	3.875380	3.973475	3.844670	3.916925
coverage probability	0.93766	0.94308	0.93670	0.93910

Table 11: Gamma distribution and $n = 50$

For $n = 100$ the expected value was 10.0011 and the standard error was 0.7061736.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	2.756238	2.790345	2.750383	2.778415
coverage probability	0.94434	0.94716	0.94474	0.94520

Table 12: Gamma distribution and $n = 100$

In this situation we noticed that for small sample size ($n = 10$), the coverage probability was around 0.88, which is much smaller than the nominal level of 0.95. Only for the t-distribution was 0.92, but the average width of the confidence interval obtained by that method was the largest. The percentile interval, and the adjusted percentile interval seem to be more narrow then others. We can observe that, as sample size increased from $n = 10$ to $n = 100$, the coverage probabilities tended to improve for all methods. So, for the largest sample the coverage probability and the average width of the confidence intervals were more or less similar in all cases.
Also, we can observe that higher confidence levels have wider intervals, but an increase in sample size leads to a decreased interval width.
In case we had the uniform distribution, the results were the following.
For $n = 10$ the expected value was 10.00116 and the standard error was 1.827826.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	7.052209	8.139540	6.677899	6.750578
coverage probability	0.91502	0.94498	0.91130	0.93224

Table 13: Uniform distribution and $n = 10$

For $n = 25$ the expected value was 10.01008 and the standard error was 1.153798.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	4.511717	4.750968	4.428073	4.436661
coverage probability	0.93770	0.94900	0.93778	0.94768

Table 14: Uniform distribution and $n = 25$

For $n = 50$ the expected value was 10.00324 and the standard error was 0.8182043.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	3.194766	3.275633	3.170603	3.172679
coverage probability	0.94316	0.94888	0.94256	0.94796

Table 15: Uniform distribution and $n = 50$

For $n = 100$ the expected value was 9.998707 and the standard error was 0.5792864.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	2.261493	2.289477	2.257330	2.258044
coverage probability	0.94632	0.94956	0.94622	0.94834

Table 16: Uniform distribution and $n = 100$

Again, for $n = 100$ the results were almost the same for all methods. For smaller sample sizes, the width was again the best using the bootstrapping methods, and the coverage probability was almost 0.95 for all sample sizes in case we used the t-distribution. As we see, the results are similar to those we obtained in case the population was normally distributed. That is because both distributions are symmetric.

The results for Pareto distribution were not so good. For $n = 10$ the expected value was 10.18909 and the standard error was 87.79223.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	17.40290	20.08614	14.67850	21.17690
coverage probability	0.54676	0.58420	0.55722	0.62804

Table 17: Pareto distribution and $n = 10$

For $n = 25$ the expected value was 10.0214 and the standard error was 31.29265.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	13.80296	14.53491	12.18321	17.69689
coverage probability	0.60288	0.61626	0.62516	0.70348

Table 18: Pareto distribution and $n = 25$

For $n = 50$ the expected value was 9.866963 and the standard error was 12.8919.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	11.18041	11.46341	10.13730	14.42519
coverage probability	0.63318	0.64402	0.66070	0.73614

Table 19: Pareto distribution and $n = 50$

For $n = 100$ the expected value was 10.08289 and the standard error was 20.79693.

	Normal distribution	t-distribution	BS Percentile	BS BCa
average width	10.187482	10.313543	9.256738	13.236690
coverage probability	0.66162	0.66530	0.69058	0.76100

Table 20: Pareto distribution and $n = 100$

In this case regardless of the sample size and method we used, the standard error was enormous. For small sample size ($n = 10$), the best result of the coverage probability was only 0.62, which is much less than 0.95. As the sample size increased, the coverage

probability was not so better. For $n = 100$, we obtained the best result using BCa method, and it was just 0.76.

Regarding to average width, the confidence intervals were the narrowest in case we used percentile method. But comparing to the previous results, for example when population was normally or gamma distributed, we can not say that we are satisfied with the results.

The reason of this, is because Pareto distribution is not an exponential family and also it is not symmetric. Although the results were not good, the bootstrapping methods appeared to be the best in this case. Bootstrap methods do not require just large sample, but also a lot of bootstrap replications in order to be useful. We had 1000 replications.

In case the population was normally distributed we obtained the best results regardless of sample sizes. But when we observed Pareto distribution, the results were not even close to good. We can conclude that distribution and sample size have a big influence on the confidence interval.

8 Conclusion

We were interested in the confidence interval around the population mean. First of all, we defined and explained why do we need them. We presented two ways of derivation CI; in case population variance was known, and when it was unknown. Also, we stated two important theorems, the CLT and the Slutsky theorem, and explained how we rely on them for construction CI. One of the most advanced statistical techniques, bootstrapping, was also presented. For the end, we used simulations to check coverage probability and width of the confidence intervals, in cases we had different population distributions and different sample sizes.

For small sample sizes we obtained liberal confidence intervals, meaning that the coverage probability was much less than 0.95. As we increased the sample size to $n = 100$, confidence intervals become conservative, or in other words, the coverage probability was almost 0.95. This was not the case only with Pareto distribution where we obtained wide confidence intervals with low confidence level.

We concluded that distribution and sample size have a big influence on the confidence interval. In all cases, the narrowest CI we obtained using a percentile method. In most cases, coverage probability was near 0.95 when we used the t-distribution for constructing CI, except when population had Pareto distribution. In that case, as the best method for constructing CI proved to be BCa method.

Regarding to sample sizes, when n was 100, almost in all the cases, except when population had Pareto distribution, we obtained more or less the same results. Looking at the results for Pareto distribution, we see that it is recommended to use the bootstrapping methods, in case we do not have distribution with so nice properties, for example when it is not symmetrical, when mean and median are not the same, when it is not an exponential family, etc.

More about CI using bootstrap method can be found in Efron and Tibshirani [3].

9 Povzetek naloge v slovenskem jeziku

Cilj naloge je preučiti pokritje in širino intervala zaupanja če izhajamo iz normalne porazdelitve in ko preučevana spremenljivka ni normalno porazdeljena. Ker so točkovne ocene za populacijski parameter pogosto nezanesljive, ponavadi iščemo interval zaupanja (IZ), recimo $[L, D]$ v katerem bo z neko stopnjo zaupanja ležal iskani parameter. Interval zaupanja za populacijski parameter θ s stopnjo zaupanja $(1-\alpha)$, na podlagi vzorca $X_1, \ldots, X_n$ definiramo kot par statistik $[L, D]$, $L = L(X_1, \ldots, X_N)$ in $D = D(X_1, \ldots, X_N)$ tako, da velja $P(L \leq \theta \leq D) \geq (1-\alpha)$. Za intervalsko oceno $[L, D]$ parametra θ, pokritje intervala $[L, D]$ je verjetnost da $[L, D]$ vsebuje parameter θ. V nalogi smo preučevali interval zaupanja za populacijsko poprečje. Na začetku smo pokazali da, ko izhajamo iz normalne porazdelitve, oziroma, ko je populacijski odklon σ znan, potem IZ izračunamo po formuli $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Ampak, v praksi skoraj nikoli ne poznamo populacijskega standardnega odklona in zaradi tega nas zanima kako lahko v tem primeru izračunamo IZ. Najprej smo pokazali, da je cenilka za populacijski odklon $s = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ in da ima izraz $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ t-porazdelitev, z $n-1$ stopinjami prostosti. Se pravi, ko ne poznamo σ, uporabljamo t-porazdelitev, in v tem primeru izračuamo IZ kot $\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$.

Ogledali smo si tudi in enega izmed najbolj pomembnih izrekov v verjetnosti - centralni limitni izrek. Centralni limitni izrek nam pove, da če imamo populacijo s povprečjem μ in standardnim odklonom σ, in če vzamemo dovolj velik vzorec, potem bo porazdelitev vzorčnega povprečja približno normalna. Pomembna stvar je, da to velja, ne glede na to ali je populacija porazdeljena normalno ali pa ne.

Ko ne poznamo populacijske porazdelitve in ne moremo predpostaviti, da je le-ta porazdeljena normalno, ne moremo uporabiti t-porazdelitve za konstrukcijo IZ. Izraz, $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$, zapišemo kot, $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{X}-\mu)}{s}\frac{\sigma}{\sigma} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\frac{\sigma}{s}$. Označimo $Z_n = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ in $V = \frac{\sigma}{s}$. Centralni limitni izrek nam pove da Z_n gre proti Z, kjer je $Z \sim N(0,1)$. Pokazali smo tudi, da $V \xrightarrow{p} 1$. Z uporabo izreka Slutsky sledi, da T gre proti Z. Oziroma, lahko konstruiramo IZ z uporabo standardne normalne porazdelitve, brez predpostavke o porazdelitvi populacije in z uporabo s namesto σ.

Ena izmed novejših statističnih metod, ki se uporablja za računanje intervala zaupanja

je samovzorčenje (ang. bootstrapping). Na kratko smo razložili, kako deluje in zakaj ga lahko uporabljamo. Pomembna lastnost samovzorčenja je, da statistični izračuni temeljijo samo na podlagi različnih vzorcev. Pri samovzorčenju na vzorec iz populacije gledamo kot na populacijo in potem vzorčimo iz tega vzorca, od tu ime samovzorčenje. 'Nova populacija' je znana, ker je to v resnici vzorec in zaradi tega lahko opazimo napako, ki smo jo naredili pri sklepanju na podlagi ponovnih vzorcev iz začetnega vzorca, kateri predstavlja populacijo.

Za konstrukcijo intervala zaupanja smo uporabljali centralni limitni izrek in izrek Slutskya za velike vzorce. Kaj se dogaja z malimi vzorcimi, smo preverili z uporabo simulacij. Se pravi, z uporabo simulacij smo preverili pokritje in širino intervala zaupanja v primeru, ko je populacija imela normalno, eksponentno, gama, enakomerno in Pareto porazdelitev. Spreminjali smo velikost vzorca in pogledali, kaj se dogaja za $n = 10$, 25, 50 in 100. Želeli smo, da bi imele različne metode za izračun IZ pokritje čim bližje 0.95 in čim ožji interval zaupanja. Predpostavili smo, da je populacijsko povprečje 10 in koda za simulacijo pa je bila zasnovana tako, da je vsakič vrnila tudi vzorčno povprečje in standardno napako. Ampak, najbolj pomembni rezultati simulacije so povprečna širina in pokritje intervala zaupanja. Za izračune IZ smo uporabili štiri metode, konstrukcija intrevala zaupanja z uporabo normalne porazdelitve, z uporabo t-porazdelitve, z uporabo metode percentilov in BCa metode.

Na koncu se je izkazalo, da ko smo vzorec povečali, so bili rezultati za vse štiri metode zelo podobni. V primeru, ko je bila populacija normalno porazdeljena, je bilo pokritje najbolj blizu 0.95, ko smo uporabljali t-porazdelitev, ne glede na velikost vzorca. Širina intervala je bila najmanjša ko smo uporabljali bootstrapping metode, ampak pokritje je bilo v tem primeru dobro zgolj pri največji velikosti vzorca. Ko smo izhajali iz eksponentne porazdelitve, smo imeli najožji interval z uporabo metode percentilov, ampak pokritje ni bilo dobro. Ko smo povečali velikost vzorca, je bilo pokritje blizu 0.95. V primeru, ko je populacija imela gama porazdelitev, smo z metodami samovzorčenja zopet dobili najožji interval in spet je bilo pokritje dobro samo za velike vzorce. Pokritje je bilo najboljše z uporabo t-porazdelitve. Tudi v primeru, ko smo izhajali iz enakomerne porazdelitve, je bilo pokritje za vse velikosti vzorca najboljše z uporabo t-porazdelitve. Rezultati pokritja in širine intervala zaupanja za $n = 100$ so bili skoraj isti za vse štiri metode. Ko smo izhajali iz Pareto porazdelitve, rezultati niso bili tak dobri. Napaka je bila velika, širina intervala je tudi bila velika in pokritje zelo slabo. V tem primeru so še najbolje delovale metode samovzorčenja. Z uporabo metode percentilov smo imeli najožji interval in z uporabo BCa metode smo imeli najboljše pokritje vendar pa je bilo pokritje še vedno precej manjše od željenih 0.95. Opazili smo torej, da ima oblika porazdelitve lahko velik vpliv na izračunane IZ, še posebej, ko je velikost vzorca majhna.

10 Bibliography and Sources

[1] G. Casella and R.L. Berger, *Statistical Inference*. Duxbury-Thomson Learning, Second Edition 2002. *(Not cited.)*

[2] A.C. Davison and D.V. Hinkley, *Bootstrap methods and their application*. Cambridge University Press, First Edition 1997. *(Not cited.)*

[3] B. Efron and R.J. Tibshirani, *An introduction to the bootstrap*. Chapman & Hall, First Edition 1993. *(Not cited.)*

[4] K. Kelley, The effects of nonnormal distribution on confidence intervals around the standardized mean difference: bootstrap and parametric confidence intervals. *Educational and Psychological Measurement* 65 (2005) 51–69. *(Not cited.)*

[5] J.A. Rice, *Mathematical Statistics and Data Analysis*, Thomson Brooks/Cole, Third Edition, 2007. *(Not cited.)*

[6] J. Shao, *Mathematical Statistics*, Springer, Second Edition, 2003. *(Not cited.)*

[7] *Properties of the Normal Distribution*, Boston University. `https://learn.bu.edu/bbcswebdav/pid-826908-dt-content-rid-2073693_1/courses/13sprgmetcj702_ol/week03/metcj702_W03S01T02_normal.html`. (Viewed on: 6/6/2015.) *(Not cited.)*

[8] *Bessel's correction*, `http://en.wikipedia.org/wiki/Bessel%27s_correction`. (Viewed on: 11/6/2015.) *(Not cited.)*

[9] *Sum of normal random variables*, `http://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables`. (Viewed on: 11/6/2015.) *(Not cited.)*

[10] *Jointly Gaussian and uncorrelated variables are independent*, Department of Mathematics UIUC. `http://www.math.uiuc.edu/∼r-ash/Stat/StatLec21-25.pdf`. (Viewed on: 12/6/2015.) *(Not cited.)*

[11] *Programming language R*, `http://www.r-project.org/`. (Viewed on: 30/6/2015.) *(Not cited.)*