STATISTICS WITH APPLIED PROBABILITY

Custom eBook for STA258

Nishan Mudalige
Nurlana Alili
Bryan Su

CANADA

Statistics with Applied Probability
Custom eBook for STA258

Nishan Mudalige
Department of Mathematical and Computational Sciences
University of Toronto Mississauga

Nurlana Alili
University of Toronto Mississauga

Bryan Xu
University of Toronto Mississauga

First edition: August 2025

Mudalige, M.; Alili, N.; Xu, B.
University of Guelph Bookstore Press

University of Toronto Mississauga,
Mississauga, Ontario, Canada

Contents

Chapter 0

Overview

Uncertainty is an inherent part of everyday life. We all face questions regarding uncertainty such as whether classes will go ahead as planned on any given day; will a flight leave on time; will a student pass a certain course? Uncertainties might also change depending on other factors, such as whether classes will still go ahead as planned when there is a snow warning in effect; if a flight is delayed can a person still manage to make their connection; will a student pass their course considering that the instructor is known to be a tough grader?

The ability to quantify uncertainty using rigorous mathematics is a powerful and useful tool. Calculating uncertainty on an intuitive level is something that is hard-wired in our DNA, such as the decision to fight or flight depending on a given set of circumstances. However we cannot always make such intuitive decisions based purely on hunches and gut feelings. Fortunes have been lost based on someone having a good feeling about something. If we have information available, we should make the best prediction possible using this information. For instance if we wanted to invest a lot of money in a company, we should use all available data such as past sales, market and industry trends, leadership ability of the CEO, forward looking statements etc. and with all this information we can then predict whether our investment will be profitable.

In order for companies to survive and remain competitive in todays environment it is essential to monitor industry trends and read markets properly. Companies that don't adapt and stick to an outdated business model tend to pay the price. At the other end of the spectrum, companies that understand the needs of the consumer, build their product around the consumer and keep evolving their product offerings based on consumer trends tend to perform well and remain competitive.

Statistics is the science of uncertainty and it is clearly a very useful subject for business. In this book you will be given an introduction to statistics and you will learn the framework as well as the language required at the introductory level. The material may be daunting at times, but the more you get familiar with the subject the more comfortable you will become with it. As business students, doing well in a statistics course will give you a competitive edge since the ability to interpret and perform quantitative analytics are skills that are highly desired by many employers.

Chapter 1

Descriptive Statistics and an Introduction to R

1.1 Introduction

Intuitively, statistics can be considered the science of uncertainty. Formally,

Definition 1.1 (Statistics).
Statistics is the science of collecting, classifying, summarizing, analyzing and interpreting data.

Population, Sample, Parameter
In statistics, researchers need to observe behavior, pattern, trends and other types of data to give a conclusion. To make the conclusion more persuasive, researchers require huge amount of data to support them, that's why study statistics need population.

Definition 1.2 (Population).
In statistics, a population is a set of similar observations which is of interest for some experimental questions. It can be a set of existing objects such as all people in Canada, or hypothetical group of existing objects such as the set of all possible hands in a game of poker.

However, data collection from population is a lot work. Usually, researchers select a finite number of observations to study.

Definition 1.3 (Sample).
It refers to a selection of a subset from population that researchers use it to estimate population characteristics.

Now, we have already chosen a sample, but how do we use it to estimate population characteristics? This is the point where parameter comes to play.

Definition 1.4 (Parameter Statistics).

A parameter is a quantity of statistical population which summerizes characteristics of the population. For example, mean, variance and standard deviation.

Descriptive and Inferential Statistics

Now, we have set everything we need. A population, a chosen sample in that population with its parameters. Next step is studying. There are two major types of analysis: descriptive and Inferential statistics. In this section, we are only going to give you a rough idea about what they are, more detailed materials will be introduced in later chapters.

Definition 1.5 (Descriptive Statistics).

It refers to the summation of all quantitive values that describe characteristics of the population. Usually, we use descriptive statistics to summerize characteristics of a data set.

Furthermore, we use inferential statistics to do statistical analysis.

Definition 1.6 (Inferential Statistics).

It refers to the process of using data analysis to indicate properties of a population. For example, testing hypothesis and confidence interval (both will be introduced in later chapters).

Qualitative and Quantitative Data

At this point, assume that we have finished all procedures such as obtaining parameters and analyzing properties. Now, another important thing is illustrating all the discovery.

Definition 1.7 (Qualitative Data).

This type of illustration refers to showing categorical data. For example, lecture notes from a course, open-question survey.

To illustrate numerical data, we use quantitative data.

Definition 1.8 (Quantitative Data).

Unless the previous type of illustration, quantitative data is represented numerically, including anything that can be counted, measured, or given a numerical value. For example, STA258 final mark score range from 100 different students who have taken this course.

1.2 Descriptive Statistics

Previously, we defined descriptive statistics. Now, let's introduce what exact they are.

Sample Mean, Variance and Standard Deviation

Sample mean (or sample average) is the average value of a sample which is selected from an interested population of an experiment. Usually, the sample mean is used to estimate population mean. In other words, we say that the sample mean is an estimator of population mean.

Definition 1.9 (Sample Mean).

Let $x_1, x_2, x_3, ..., x_n$ be a sample of data points. We define sample mean of the sample data points ($\bar{x}$) as the following:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Also, we define sample variance of the sample data points (s^2) as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Moreover, the standard deviation of the sample of data points (s) is:

$$s = \sqrt{s^2}, \quad \textit{for } s > 0.$$

Now, let's move to variance. It refers to the expected value of the squared deviation from the mean of a random variable in a population. Similarly, we do have sample variance as well, which is the expected value of the squared deviation from the mean of a random variable in a selected sample. At this point, we can still use sample variance to estimate population variance with adjustment, because the sample variance may differ significantly based on what data points are chosen from that population.

Definition 1.10 (Sample Variance).

Let $x_1, x_2, x_3, ..., x_n$ be a sample of data points, we define sample variance of the sample data points (s^2) as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \ \textit{where } \bar{x} \textit{ is the sample mean of the data points.}$$

Next is standard deviation. It is a measure of the amount of variation of the values of a variable about its mean. If standard deviation is relatively larger, then data points are widely spread out from the mean. Otherwise, data points stay close from the mean. Also, standard deviation is obtained by taking squared root from variance which is dependent on the choices of data points as well. To use sample standard deviation as an estimator to population standard deviation, we still need to adjust it.

Definition 1.11 (Sample Standard Deviation).
Let $x_1, x_2, x_3, ..., x_n$ be a sample of data points. The standard deviation of the sample of data points (s) is:

$$s = \sqrt{s^2}, \quad \textit{for } s > 0.$$

Median and Mode

The median and mode are two important measures of central tendency used in statistics to summarize and understand data. The median represents the middle value in a sorted dataset, giving a sense of the center that is not affected by extreme values or outliers. In contrast, the mode is the value that appears most frequently in a dataset, making it useful for identifying common or repeated observations.

Definition 1.12 (Median).
Let: $x_1, x_2, x_3, ..., x_n$ be a collection of data points which is arranged in ascending order from the smallest value to the largest value (or descending order from the largest value to the smallest value in that collection). The median of the given collection of data points is the middle value in that collection, which equally spreads the collection into two parts. Half of all the collection values are above the median value and the rest of the values in the collection is below the median value.

- *Case 1: when n is an odd number. (i.e. $1, 3, 11, 237, ...$). Then, the median M is defined as:*

$$M = \frac{n+1}{2}, \textit{ where } n \textit{ represents the } n^{th} \textit{ position.}$$

- *Case 2: when n is an even number (i.e. $2, 6, 100, 500, ...$). Then, the median M is: the average value of $\frac{n}{2}$'s and $\frac{n+2}{2}$'s position, where n represents the n^{th} position.*

Now, let's introduce mode.

Definition 1.13 (Mode).
It refers to a value that appears the most frequent than the appearance of all other values in a given dataset.

Percentile and Quartile

Percentiles and quartiles are statistical measures used to describe the distribution of data. A percentile indicates the value below which a given percentage of observations fall, helping to understand relative standing within a dataset. Quartiles, a specific type of percentile, divide the data into four equal parts (Q1, Q2/median, and Q3), providing insights into the spread and central tendency.

Definition 1.14 (Percentile and Quartile).

Let: $x_1, x_2, ..., x_n$ be a collection of data points in either ascending order. Percentile is denoted as: p^{th}, which indicates $p\%$ of observations are below to a such value. Quartiles, are special cases of percentile which equally spread the collection of data into four parts. Each part contains 25% of the entire collection. More specifically, we define quartiles as the following:

- *Q_1: the 25 percentile (or 25^{th}), which shows that 25% of the data points are below the value Q_1.*
- *Q_2: the 50 percentile (or 50^{th}), which shows that 50% of the data points are below the value Q_2.*
- *Q_3: the 75 percentile (or 75^{th}), which shows that 75% of the data points are below the value Q_3.*
- *Q_2 is qual to median.*

Moreover, we use $Q_3 - Q_1$ to calculate interquartile range (I.P.R), which shows the spread of the whole data set.

Skewness and Symmetry

The two terms 'skewness' and 'symmetry' are used to describe the shape of probability distribution. There are two types of skewness: left (or negative) skew and right (or positive) skew. In real life, a famous distribution highly used in hypothesis testing which is χ^2_n with n degrees of freedom, is right skewed probability distribution function. Another example regarding to symmetry is normal distribution such that its probability under its curve greater than μ is same as the probability below than μ. Now let's introduce the proper definition of skewness and symmetry.

Definition 1.15 (Skewness).

Skewness refers to such a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative or undefined.

Now, let's break down the main definition of skewness and symmetry:

Definition 1.16 (Left (or Negative) Skew).

By observing given probability distribution curve, if the left tail of the curve is longer than the right tail the mass of the distribution is concentrated on the right of the figure, then we say that probability distribution is left skew or negative skew. (See figure below)

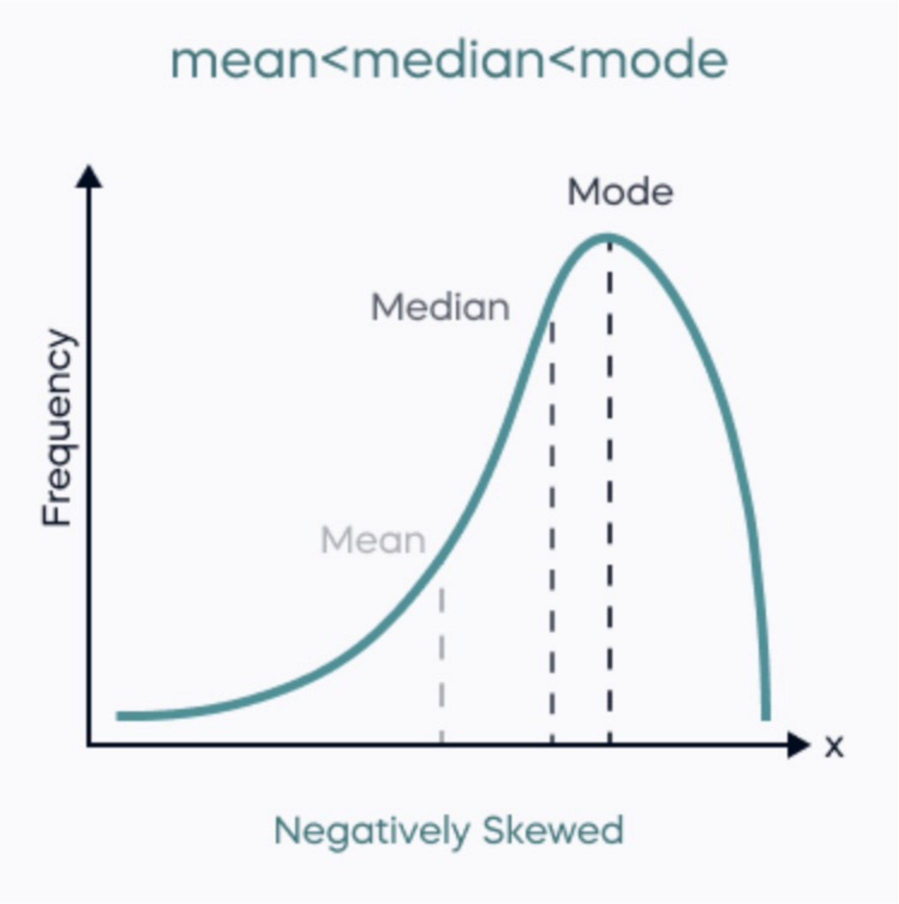


Figure 1.1: Visualization of left skew probability distribution

Definition 1.17 (Right (or Positive) Skew).

By observing given probability distribution curve, if the right tail of the curve is longer than the left tail the mass of the distribution is concentrated on the left of the figure, then we say that probability distribution is right skew or positive skew. (See figure below)

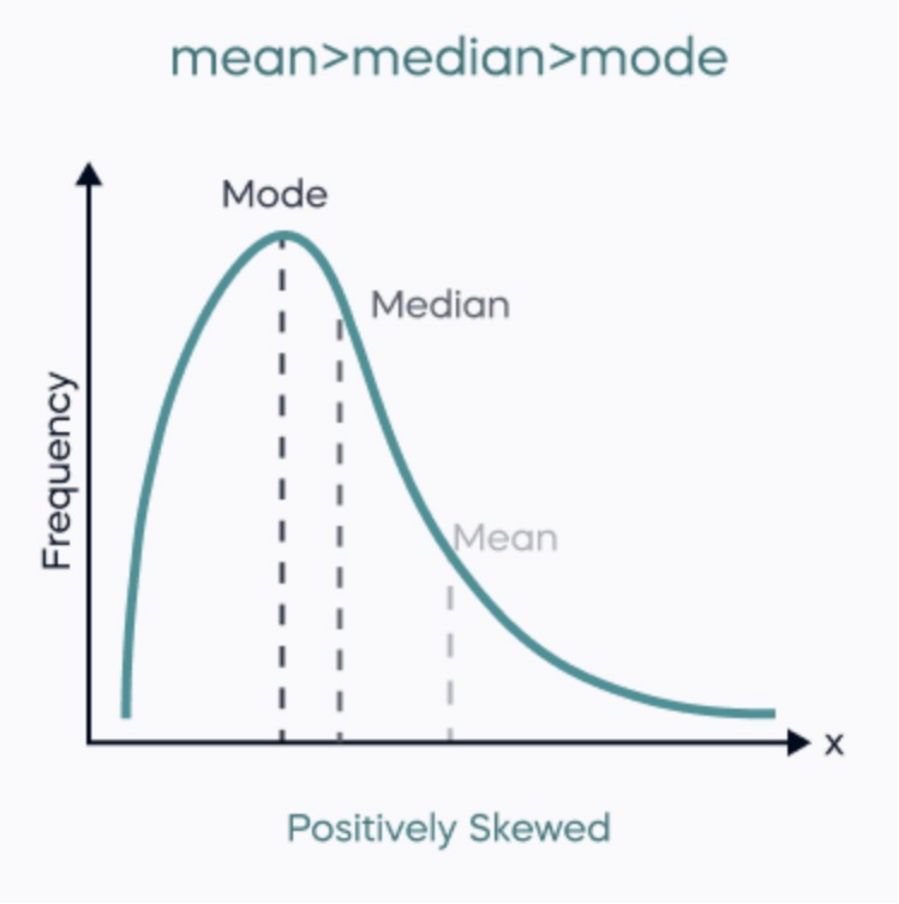


Figure 1.2: Visualization of right skew probability distribution

Symmetry is a special case of skewness when the value of skewness is 0.

Definition 1.18 (Symmetry)**.**
In statistics, symmetry s a probability distribution is reflected around a vertical line at some value of the random variable represented by the distribution. Probability under the curve below that value is equal to probability under the curve greater than that value. (see figure below)

Since symmetry is a special case, so that it has a unique property as the following:

Theorem 1.1 (Empirical Rule (or $68-95-99.7$ Rule))**.** *For any symmetric (bell-shaped) curve, let μ be its mean and σ be its standard deviation, the following probability set function is true:*

- 1.*:* $Pr(\mu - \sigma < X < \mu + \sigma) = 68.27\%$;
- 2.*:* $Pr(\mu - 2\sigma < X < \mu + 2\sigma) = 95.45\%$;
- 3.*:* $Pr(\mu - 3\sigma < X < \mu + 3\sigma) = 99.73\%$.

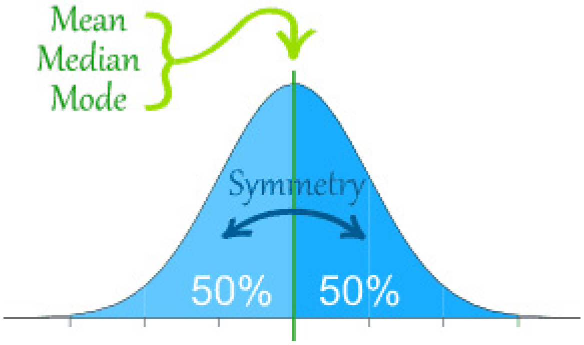


Figure 1.3: Visualization of symmetric probability distribution

Practice Example

Example 1.1.

[Calculating Sample Mean, Variance and Standard Deviation] Let: $x_1 = 1, x_2 = 3$ and $x_3 = 7$. Calculate the sample mean, sample variance and sample standard deviation for this collection of data points.

Solution (all results are kept in four digits):

By Definition 1.9, 1.10, 1.11, sample mean:

$$\bar{x} = \frac{1+3+7}{3} \approx 3.6667.$$

Then, we use sample mean to calculate sample variance:

$$s^2 = \frac{1}{3-1} \times [(1-3.6667)^2 + (3-3.6667)^2 + (7-3.6667)^2] \approx 9.3333.$$

Finally, we take the square root of sample variance to get sample deviation, and remember that $s > 0$:

$$s = \sqrt{s^2} \approx 3.0551.$$

Example 1.2.

[Median Calculation] Given two distinct collections of data points: $S_1 = \{2, 4, 6\}$ and $S_2 = \{1, 5, 16, 28\}$. Calculate the median of both two sets.

Solution:

For S_1, since $n = 3$ which is an odd number, so by $Definition$ 1.3, $M_{S_1} = 4$. For S_2, $n = 4$ in this case, so that we need to calculate the average of $\frac{n}{2}$ and $\frac{n+1}{2}$. Then,

$$M_{S_2} = \frac{5+16}{2} = 10.5.$$

Example 1.3.

Consider the data set $S = \{4, 25, 30, 30, 30, 32, 32, 35, 50, 50, 50, 55, 60, 74, 110\}$. Calculate its median and Q_1 (25^{th}).

Solution:

Simply counting the number of data points, $n = 15$, such that $M_S = \frac{15+1}{2} = 8$. Thus, the 8^{th} value in the set which is 35.

Since we know the median of this collection of data points, we just need to find the median of the lower half of this data, which is exactly going to be 25 percentile (25^{th}). In the lower half of the given collection (all values below the median), $n_{lower} = 7$. By $Definition$ 1.3, then median of the lower half (25^{th}) is going to be:

$$25^{th} = \frac{7+1}{2} = 4, \text{ the } 4^{th} \text{ position in the data set.}$$

Thus, Q_1 $(25^{th}) = 30$. To find Q_3 (75^{th}), apply the same strategy will guide you to find the correct answer, and we leave this as an exercise to you.

1.3 Graphical Techniques

In statistics, there are lots of types of graph to illustrate data, for example histograms and box-plots. This technique is used in the field of statistics for data visualization. Our objective is to both be able to identify some classical types of graph and interpret key statistical values (descriptive statistical values) from it.

1.3.1 Histograms

Introduction to Histograms

Histogram is a graphical representation of data that uses bars to display the frequency distribution of a dataset. Unlike bar graphs, which represent categorical data, histograms group numerical data into intervals (bins) and show how many values fall into each range. This makes histograms ideal for visualizing the shape, spread, and central tendency of continuous data, helping identify patterns such as symmetry, skewness, and outliers.

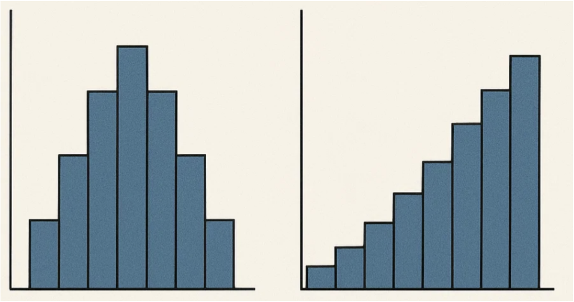


Figure 1.4: Visualization of histograms

Advantages and Disadvantages of Histograms

- Advantages of Histograms:
  1. Histograms are easily to used for visualise data (relatively). It allows us to get the idea of the "shape" of distribution (i.e. skewness which will be discussed late in this section).
  2. It is also flexible that people are able to modify bin widths.
- Disadvantages of Histograms:
  1. It is not suitable for small data sets.
  2. The values from histograms close to breaking points are likely similar, in fact they need to be classified into different bins (i.e. Student A and B scores 79 and 80 respectively in STA258, we consider a breaking point between 79 and 80. The two students have similar score, but student A is $B+$ and student B is $A-$ in GPA from).

Histograms with Skewness and Symmetry

A histogram visually represents the distribution of numerical data, making it a useful tool for assessing skewness and symmetry. It is quite straightforward to estimate the skewness of histograms by simply drawing a curve above bins on the histogram.

For a histogram to have a left (or negative) skew probability distribution:

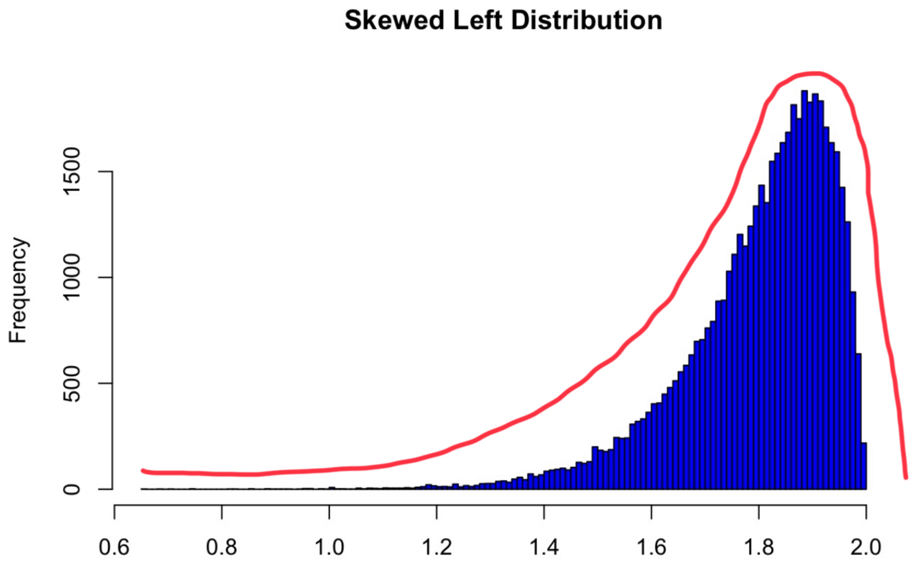


Figure 1.5: Visualization of a histogram has a left (or negative) skew probability distribution

For a histogram to have a right (or positive) skew probability distribution:

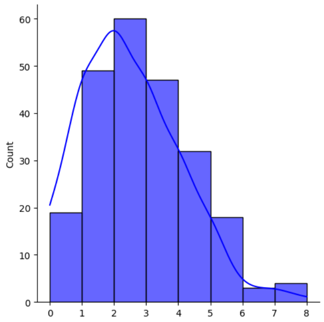


Figure 1.6: Visualization of a histogram has a right (or positive) skew probability distribution

For a histogram to have a symmetric probability distribution:

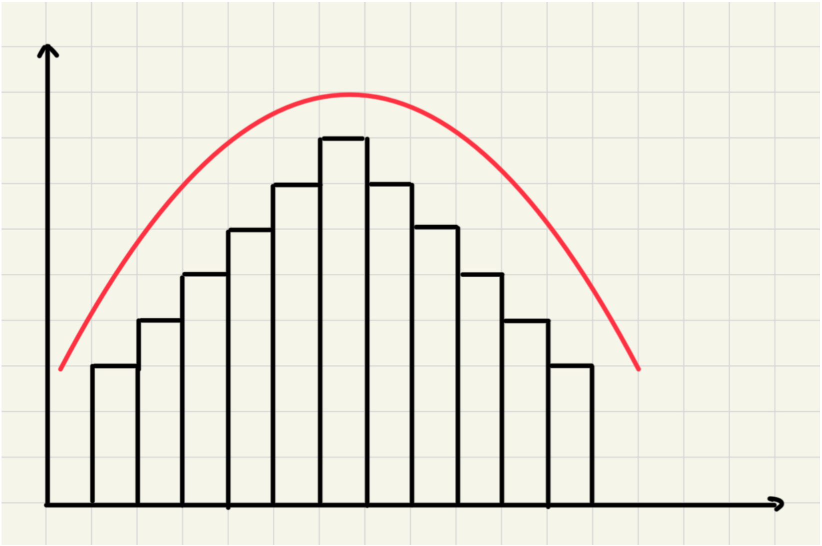


Figure 1.7: Visualization of a histogram has a symmetric probability distribution

1.3.2 Box-Plots

A boxplot (or box-and-whisker plot) is a standardized way to display data distribution based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. The box represents the interquartile range (IQR), while the whiskers show variability outside Q1 and Q3. Outliers are plotted as individual points. Boxplots efficiently compare distributions and highlight skewness, spread, and outliers. (See figure below)

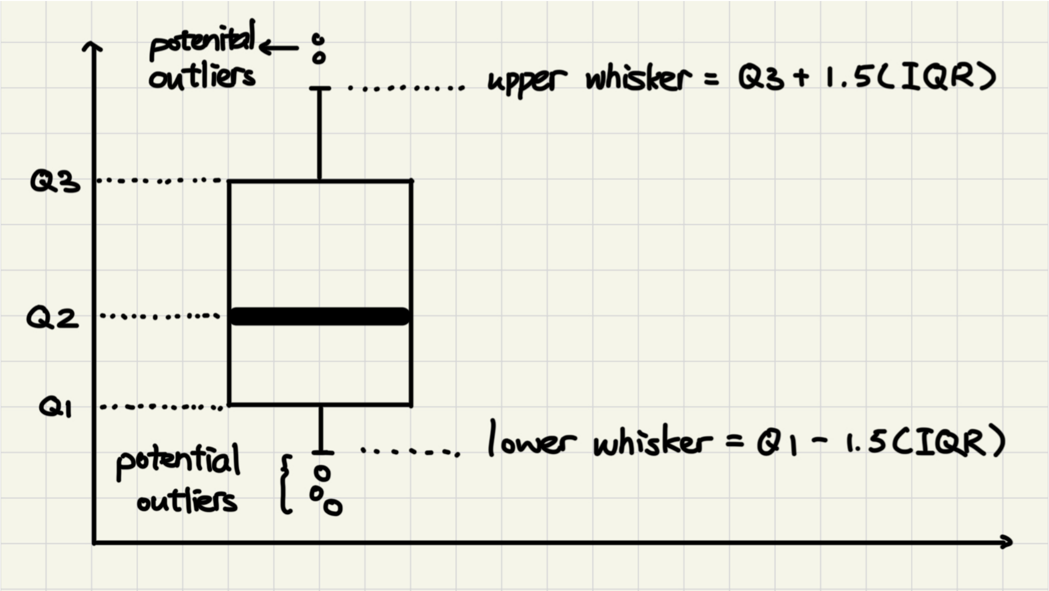


Figure 1.8: Visualization of a box-plot

Similar to histograms, we can still obtain information about skewness and symmetry, by observing the cut from the line of Q2.

If the median (Q2) cuts the box with upper area smaller than lower area, then we say that box-plot with left skew probability distribution. Or, if the median (Q2) cuts the box with upper area larger than lower area, then we say that box-plot with right skew probability distribution.

Otherwise, if the median (Q2) cuts the box with upper area equal to lower area, then we say that box-plot with symmetric probability distribution.

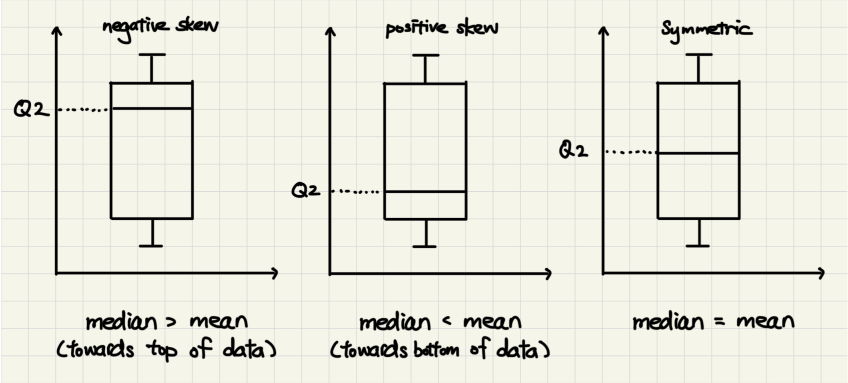


Figure 1.9: Visualization of a box-plot with skew and symmetric probability distribution

1.4 Introduction to R

R is used for data manipulation, statistics, and graphics. It is made of: operations (+,−, <) which is for calculations on vectors, arrays and matrices; a huge collection of functions; facilities for making unlimited types quality graphs; user contributed packages (sets of related functions); the ability to interface with procedures written in C, C+, or FORTRAN and to write additional primitives. R is also an open-source computing package which has seen a huge growth in popularity in the last few years (Please use this website: https://cran.r-project.org, to download R).

What is R-studio?

RStudio is a relatively new editor specially targeted at R. RStudio is cross-platform, free and open-source software (Please use: https://www.rstudio.com, to download Rstudio).

Make a Histogram Using R-studio

This is just a demonstration of how to start and use R-studio.
1. First of all, we need to know which dataset are we going to make into a histogram. In this case, as an example, we are going to use the waiting time in faithful in R-studio.
2. For any dataset, use the code: names(faithful) to get it. (inside the parentheses, type the names of variables you want in faithful dataset)
3. Then, we proceed with the code: hist(faithful$waiting) to get a basic plot.

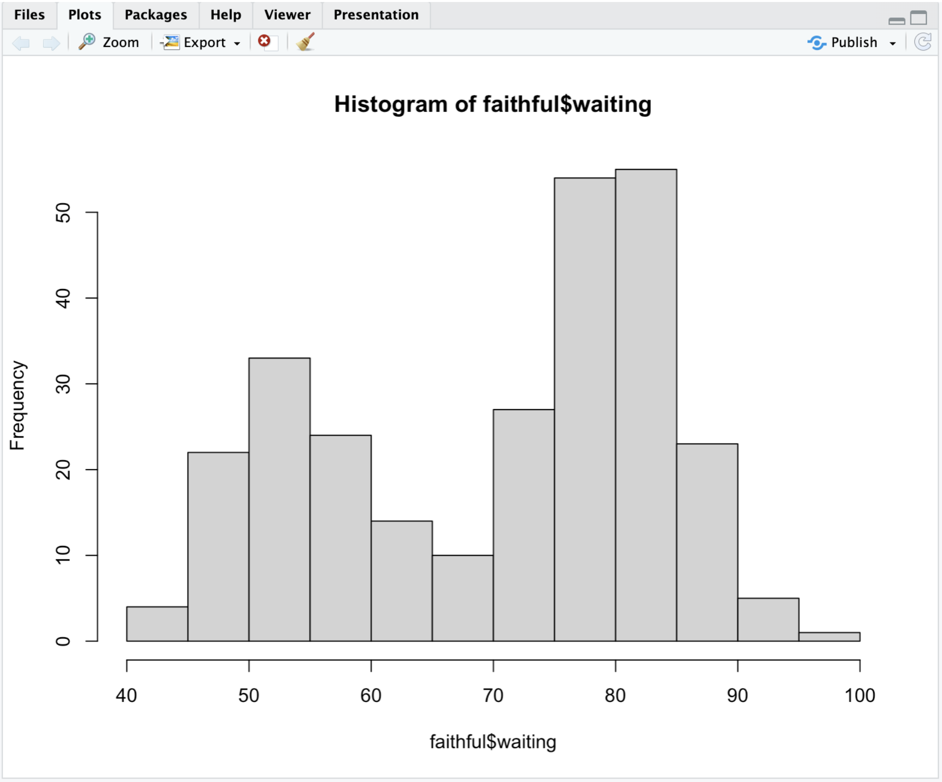
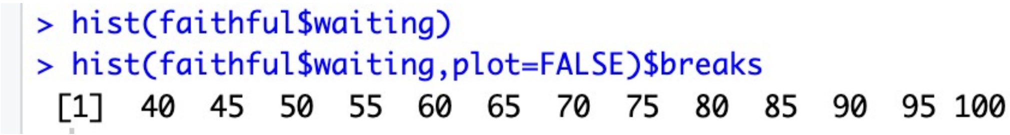


Figure 1.10: R-studio first three steps (by following the instructions, you should get this histogram)

4. Furthermore, we can also get more information. For example, by keep proceeding with the code: hist(faithful$waiting,plot=FALSE)$breaks, R-studio will show you all the breaking points between histogram cells.



```
> hist(faithful$waiting)
> hist(faithful$waiting,plot=FALSE)$breaks
 [1]  40  45  50  55  60  65  70  75  80  85  90  95 100
```

Figure 1.11: R-studio the forth step(by following the instructions, you should get this histogram)

Chapter 2

Sampling Distributions Related to a Normal Population

Previously, we have introduced lots of definitions and given you a rough idea about what really statistics it and what people do in statistics. Now, we are going to proceed statistical distributions.

2.1 Normal Distribution

In probability theory and statistics, normal distribution also called Gaussian distribution which is discovered by a famous German mathematician Johann Carl Friedrich Gauss in 1809. It is one of the most important distribution that used to approximate other types of probability distribution, such as binomial, hypergeometric, inverse (or negative) hypergeometric, negative binomial and Poisson distribution. Generally, it is denote as $N(\mu, \sigma^2)$ with probability density function as the following:

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

Formally, let's begin with its definition:

Definition 2.1 (Normal Distribution)**.**
Suppose a random variable $X \sim N(\mu, \sigma^2)$, then $E(X) = \mu$ and $Var(X) = \sigma^2$. And $-\infty < \mu < \infty, \sigma^2 > 0$. Moreover, X has probability density function as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \textit{ for } -\infty < x < \infty \textit{ (same as above).}$$

The only special case of normal distribution is standard normal distribution, such that a random variable $Y \sim N(\mu = E(Y) = 0, \sigma^2 = Var(Y) = 1)$, then Y has probability density function as:

$$f(y) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{y^2}{2}}.$$

2.2 Gamma and Chi-square Distribution

The Chi-square and Gamma distributions are two fundamental probability distributions widely used in statistical theory and applications. The Gamma distribution is a continuous distribution characterized by its shape and scale parameters, making it versatile for modeling waiting times and various positively skewed data. The Chi-square distribution, a special case of the Gamma distribution, arises naturally in the context of hypothesis testing and confidence interval estimation, especially in tests involving variance and categorical data.

Gamma Distribution

Definition 2.2 (Gamma Distribution)**.**
Suppose a random variable X is Gamma distributed with $\alpha > 0$ (shape parameter) and $\beta > 0$ (scale parameter) if and only if the probability density function of X is

$$f(x) = \frac{x^{\alpha-1}e^{\frac{-x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}, \ \textit{for } 0 < x < \infty.$$

Then, $E(X) = \alpha\beta$, $Var(X) = \alpha\beta^2$ and its moment generating function is $M_X(t) = \frac{1}{(1-\beta t)^{\alpha}}$, for $t < \frac{1}{\beta}$.

Now, let's introduce some properties of Gamma function:

- Gamma function (**not a distribution**):

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}\,dt, \ \text{for } x > 0.$$

- Properties
  - 1. $\Gamma(x) = x \cdot \Gamma(x-1)$;
  - 2. For all $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$;
  - 3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Chi-square Distribution

Definition 2.3 (Chi-square Distribution)**.**
A random variable X has a Chi-squared distribution with n degrees of freedom (χ^2_n) if and only if X is a random variable with a Gamma distribution with parameters $\alpha = \frac{n}{2}$ and $\beta = 2$. Then, the probability density function of X is given by

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{k}{2}-1}e^{\frac{-x}{2}}.$$

Moreover, $E(X) = n$, $Var(X) = 2n$ *and moment generating function of* X *is* $M_X(t) = (1-2t)^{\frac{-n}{2}}$, *for* $t < \frac{1}{2}$.

We claim that Chi-square distribution is a special case of Gamma distribution with $\alpha = \frac{n}{2}$ and $\beta = 2$. Now, let's prove it by using moment generating function.

The proof is quite straightforward as the following shows:

Proof. Suppose $X \sim Gamma(\alpha = \frac{n}{2},\, \beta = 2)$.
Then the following moment generating function holds for X:

$$M_X(t) = (1-2t)^{\frac{-n}{2}}, \text{ for } t < \frac{1}{2}.$$

Compare the moment generating function of X under Gamma distribution with Chi-square distribution, we can conclude that $X \sim \chi_n^2$. □

Obtaining Chi-square Distribution by Normal Distribution
Previously, we showed how to use Gamma distribution to get Chi-square distribution by moment generating function method. Now, let's do something interestingly, to use normal distribution to get Chi-square distribution. We will begin with a theorem, then prove it.

Theorem 2.1 ($Z^2 \sim \chi_1^2$)**.** *Suppose a random variable* Z *is standard normally distributed, such that* $Z \sim N(0,\,1)$. *Then,* Z^2 *is Chi-square distributed with* 1 *degree of freedom, so that* $Z^2 \sim \chi_1^2$.

The proof of Theorem 2.1 isn't that trivial to see. We still need moment generating function, but in a different way. Before we get into the proper proof, let's grab everything we need:

- 1. Recall STA256 about how to get moment generating function for a given continuous random variable that:

$$M_Z(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx.$$

- 2. We also need Gaussian integral:
  - (i)

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi};$$

  - (ii)

$$\int_{-\infty}^{\infty} e^{-kx^2}\, dx = \sqrt{\frac{\pi}{k}}, \text{ for } k > 0;$$

  - (iii)

$$\int_{-\infty}^{\infty} e^{kx^2}\, dx = \sqrt{\frac{\pi}{-k}}, \text{ for } k < 0.$$

Proof. Suppose that $Z \sim N(0,1)$, then

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-z^2}{2}}.$$

Next, computing M.G.F for Z^2:

$$M_{Z^2}(t) = E(e^{tz^2}) = \int_{-\infty}^{\infty} e^{tz^2} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-z^2}{2}}\, dz.$$

After rearranging:

$$M_{Z^2}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(\frac{1}{2}-t)z^2}\, dz.$$

Let $u^2 = (\frac{1}{2}-t)z^2$, then $u = z\sqrt{\frac{1}{2}-t}$, so that $du = \sqrt{\frac{1}{2}-t} \cdot dz$, and $dz = \frac{1}{\sqrt{\frac{1}{2}-t}} du$.

By substitution:

$$M_{Z^2}(t) = \frac{1}{\sqrt{2\pi}} \int_{u=-\infty}^{u=\infty} e^{-u^2} \cdot \frac{1}{\sqrt{\frac{1}{2}-t}} du$$

Rearranging and using 2(i) from above:

$$M_{Z^2}(t) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\frac{1}{2}-t}} \int_{u=-\infty}^{u=\infty} e^{-u^2} = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\frac{1}{2}-t}} \cdot \sqrt{\pi}.$$

After all simplification, we get:

$$M_{Z^2}(t) = (1-2t)^{-\frac{1}{2}}, \text{ which is the M.G.F of } \chi^2_{n=1}, \text{ as required.}$$

□

Now, we can do another proof by using Theorem 2.1.

Theorem 2.2 ($\sum_{i=1}^n Z_i^2 \sim \chi_n^2$)**.** *Suppose* $Z_1, Z_2, ..., Z_n \overset{i.i.d.}{\sim} N(0,1)$*, then the sum of* n *independent* Z^2 *is going to be Chi-square distributed with* n *degrees of freedom, as the following:*

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2.$$

We need Theorem 2.1 to prove this, but it going to be easier.

Proof. Assume that $Z_1, Z_2, ..., Z_n \overset{\text{i.i.d.}}{\sim} N(0,1)$, then we compute the M.G.F for the entire summation as well, and we let $\delta = \sum_{i=1}^n Z_i^2$, which as the following shows:

$$M_\delta(t) = E[e^{t\delta}] = E[e^{t(Z_1^2+\cdots+Z_n^2)}] = E[e^{tZ_1^2} \cdots e^{tZ_n^2}].$$

Since $Z_1, Z_2, ..., Z_n$ are independent and identically distributed, so that:

$$M_\delta(t) = E[e^{tZ_1^2}] \cdots E[e^{tZ_n^2}].$$

By Theorem 2.1, we have: $Z^2 \sim \chi_1^2$, then:

$$M_\delta(t) = E[e^{t\delta}] = E[e^{tZ_1^2} \cdots e^{tZ_n^2}] = [(1-2t)^{-\frac{1}{2}}] \cdots [(1-2t)^{-\frac{1}{2}}].$$

Since there are n individuals of Z^2, hence:

$$M_\delta(t) = (1-2t)^{-\frac{n}{2}}, \text{ which is the M.G.F of } \chi_n^2, \text{ as required.}$$

□

Here is the last theorem for Chi-square and normal distribution, but we won't show you the proof due to its complexity. For people who are interested in that, please see STA260 lecture notes or power point slide to figure out.

Theorem 2.3 **($\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$).** *Let n be sample size, s^2 be sample variance and σ^2 be population variance, then $\frac{(n-1)s^2}{\sigma^2}$ is Chi-square distributed with $n-1$ degrees of freedom. As the following:*

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2.$$

2.3 Student's t-Distribution and F-Distribution

The t-distribution and F-distribution are essential tools in inferential statistics, particularly in the context of hypothesis testing and variance analysis. The t-distribution, which resembles the normal distribution but with heavier tails, is primarily used when estimating population means in situations where the sample size is small and the population standard deviation is unknown. On the other hand, the F-distribution is used to compare variances between two populations and plays a central role in analysis of variance (ANOVA) and regression analysis.

Student's t-Distribution

Definition 2.4 (Student's t-Distribution)**.**
Suppose X is t-distributed with n degrees of freedom, then the probability density function of X is given by:

$$f_X(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{\frac{-n+1}{2}}.$$

Alternatively, define a new variable T is the following:

$$T = \frac{W}{\sqrt{\frac{V}{r}}}, \text{ for } W \sim N(0,1) \text{ and } V \sim \chi_r^2.$$

Or suppose $X_1, ..., X_n \overset{i.i.d.}{\sim} N(\mu,\ \sigma^2)$, *then* $\bar(X) \sim N(\mu,\ \frac{\sigma^2}{n})$. *Thus,*

$$T = \frac{\bar{x} - \mu}{\frac{s}{(\sqrt{n})}}.$$

Same as normal distribution, student's t-distribution is also symmetric. Also, as the degrees of freedom of t-distribution getting larger, the curve of student's t-distribution getting closer to standard normal distribution.

F-Distribution

Definition 2.5.
We define a new variable F as the following shows:

$$F = \frac{(\frac{W_1}{v_1})}{(\frac{W_2}{v_2})} \sim F_{v_1,\ v_2};\ \textit{for}\ W_1 \sim \chi^2_{v_1}\ \textit{and}\ W_2 \sim \chi^2_{v_2};\ \textit{also both}\ W_1\ \textit{and}\ W_2\ \textit{are independent.}$$

Alternatively, we select two samples (with same population variance) with size n *and* m, *and also sample variance* s_x *and* s_y *respectively. Then, F-distribution is:*

$$F = \frac{[\frac{(\frac{(n-1)}{\sigma^2})s_x^2}{n-1}]}{[\frac{(\frac{(m-1)}{\sigma^2})s_y^2}{m-1}]} \sim F_{n-1,\ m-1}.$$

Both student's t-distribution and F-distribution are highly used in inferential statistics, until confidence interval, testing hypothesis and ANOVA analysis, these two distributions will come to play a lot. At this point, just guarantee that you know how to obtain those distribution from random given information is sufficient.

Chapter 3

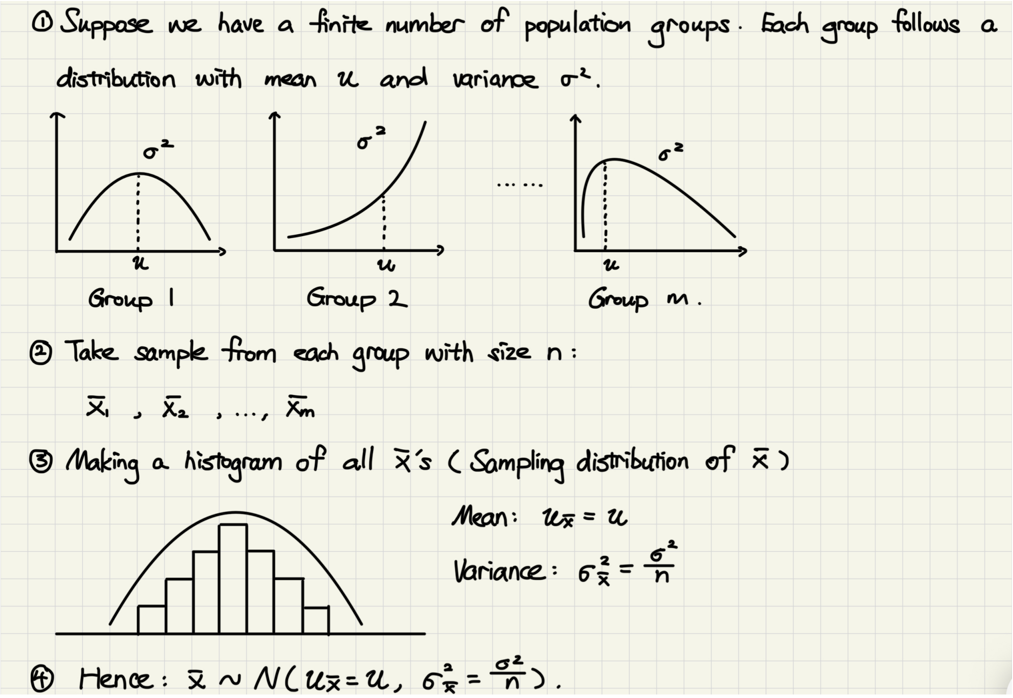
The Central Limit Theorem

The Central Limit Theorem (CLT) is one of the most important results in probability and statistics. It states that, given a sufficiently large sample size, the distribution of the sample mean of independent and identically distributed (i.i.d.) random variables approaches a normal distribution, regardless of the shape of the original distribution. Real-life Application of Central Limit Theorem in Financial Analysis. The CLT is often used by financial experts to examine stock market results.

Now, let's discuss Central Limit Theorem with more details. Suppose we have a finite number of populations and each population follows a distribution with population mean μ and population variance σ^2.. Then we take samples of same size n from each population, such that we have $\bar{x}_1, \bar{x}_2, ..., \bar{x}_m$ from population group 1 to m, respectively. Next, we make a histogram using the large collection of sample taken from each population group. Then, what we are doing right row is sampling distribution of $\bar{x}$. As a result, $\bar{x}$ follows a normal distribution with mean $\mu_{\bar{x}} = \mu$ and variance $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$, which is denoted as the following:

$$\bar{x} \sim N(\mu_{\bar{x}} = \mu,\ \sigma^2_{\bar{x}} = \frac{\sigma^2}{n}).$$

There is a figure shows the entire procedure from above paragraph.



① Suppose we have a finite number of population groups. Each group follows a distribution with mean u and variance σ^2.

σ^2
u
Group 1
σ^2
u
Group 2
......
σ^2
u
Group m.

② Take sample from each group with size n:

$\bar{X}_1$, $\bar{X}_2$, ..., $\bar{X}_m$

③ Making a histogram of all $\bar{X}$'s (Sampling distribution of $\bar{X}$)

Mean: $u_{\bar{x}} = u$

Variance: $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$

④ Hence: $\bar{x} \sim N(u_{\bar{x}} = u, \sigma^2_{\bar{x}} = \frac{\sigma^2}{n})$.

Figure 3.1: Procedure of the Central Limit Theorem

Now, let's begin with the proper definition of Central Limit Theorem.

Definition 3.1 (Central Limit Theorem).

Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then, we define the following:

$$U_n = \frac{\bar{X} - \mu}{(\frac{\sigma}{\sqrt{n}})} \sim N(\mu = 0, \sigma^2 = 1), \text{ where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then the distribution function of U_n converges to the standard Normal distribution function as $n \longrightarrow \infty$. That is,

$$\lim_{n\to\infty} P(U_n \leq u) = \int_{\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt; \text{ for all } u.$$

For this course in particular, we do not need to pay that much attention to the proving part of the definition above. However, we use Central Limit Theorem to approximate distributions. Here are the two important approximations:

- $$\bar{X}_n \approx N(\mu, \frac{\sigma^2}{n});$$

- $$T = \sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2).$$

A reminder that the distribution of U_n in definition 3.1 and the two approximation of distribution above are extremely important in this course, until later chapters you may see some materials that are similar.

Chapter 4

Normal Approximation to the Binomial Distribution

4.1 Introduction

Definition 4.1 (Statistic).

A statistic is a function of the observable random variables in a sample and known constants. Since statistics are functions of the random variables observed in a sample, they themselves are random variables. As such, all statistics have a corresponding probability distribution, which we refer to as their sampling distribution.

Review from STA256

Bernoulli Distribution:

A Bernoulli trial is a single experiment with two outcomes:

- Success: $X = 1$ with probability p
- Failure: $X = 0$ with probability $1 - p$

$X = x$	0	1
$P(X = x)$	$1 - p$	p

The probability mass function (PMF) is:

$$f(x) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}$$

Binomial Distribution:

A binomial distribution arises from n independent Bernoulli trials. Let:

$$X = \text{number of successes in } n \text{ trials}$$

Then:

$$X \sim \text{Binomial}(n, p)$$

where:

- Each trial results in either success (with probability p) or failure (with probability $1 - p$)
- $X \in \{0, 1, \ldots, n\}$

The PMF is:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Moment Generating Function (MGF):
The moment generating function (MGF) of a random variable X is defined as:

$$M_X(t) = \mathbb{E}[e^{tX}]$$

The MGF uniquely characterizes the distribution of X (if it exists in an open interval around 0), and it can be used to compute moments such as the mean and variance.

4.2 Bernoulli Distribution

Bernoulli random variable is a discrete random variable that has exactly two possible outcomes which are either a **success** or a **failure**. An experiment in which there are exactly 2 outcomes (which are success or failure) is called a **Bernoulli trial**.

When $x = 1$ we have a success and when $x = 0$ we have a failure. The term success and failure are relative to the problem being studied.

TIP: "success" need not be something positive

We chose to label a person who refuses to administer the worst shock a "success" and all others as "failures". However, we could just as easily have reversed these labels. The mathematical framework we will build does not depend on which outcome is labeled a success and which a failure, as long as we are consistent.

Consider the random experiment of rolling a die once. Define the random variable:

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th roll is a six,} \\ 0 & \text{otherwise} \end{cases}$$

Then $X_i \sim \text{Bernoulli}(p)$, where $p = P(\text{rolling a six})$.

Let $X \sim$ Bernoulli(p). The mass function of X is

$$P(X = x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

where p represents the probability of success.

Definition 4.2 (Mean and Variance of a Bernoulli Random Variable)**.**
Let $X \sim$ Bernoulli(p). The mean of X is

$$E(X) = \mu = p$$

and the variance of X is

$$Var(X) = \sigma^2 = p(1-p)$$

To support the earlier result, we now provide a derivation of the mean, variance, and standard deviation of a Bernoulli random variable.

Let X be a Bernoulli random variable with the probability of a success as p. Then

$$\begin{aligned} E[X] = \mu = \sum_{i=1}^{n} x_i \cdot P(X = x_i) \\ = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ = 0 \cdot (1-p) + 1 \cdot p \\ = p \end{aligned}$$

Similarly, the variance of X can be computed:

$$\begin{aligned} V(X) = \sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot P(X = x_i) \\ = (0-p)^2 \cdot P(X = 0) + (1-p)^2 \cdot P(X = 1) \\ = p^2(1-p) + (1-p)^2 p \\ = p(1-p) \end{aligned}$$

The standard deviation is

$$\begin{aligned} \sigma = \sqrt{\sigma^2} \\ = \sqrt{p(1-p)} \end{aligned}$$

4.3 Sampling Distribution of the Sum and MGF Derivation

Consider determining the sampling distribution of the sample total:

$$T_n = X_1 + X_2 + \cdots + X_n$$

Suppose $X_i \overset{iid}{\sim} \text{Bernoulli}(p)$. Then the moment-generating function of T_n is:

$$\begin{aligned}
M_{T_n}(t) &= \mathbb{E}[e^{tT_n}] \\
&= \mathbb{E}\left[e^{t(X_1+X_2+\cdots+X_n)}\right] \\
&= \mathbb{E}\left[e^{tX_1}e^{tX_2}\dots e^{tX_n}\right] \quad \text{(independence)} \\
&= \mathbb{E}[e^{tX_1}] \cdot \mathbb{E}[e^{tX_2}] \cdots \mathbb{E}[e^{tX_n}] \\
&= M_{X_1}(t) \cdot M_{X_2}(t) \cdots M_{X_n}(t) \\
&= \left[pe^t + (1-p)\right]^n
\end{aligned}$$

Since this is the MGF of a binomial random variable with parameters n and p, we conclude:

$$T_n \sim \text{Binomial}(n, p)$$

Example: Binomial Distribution from Die Rolls

We can think of rolling a die n times as an example of the binomial setting. Each roll gives either a six (a "success") or a number different from six (a "failure").
Knowing the outcome of one roll doesn't tell us anything about the others, so the n rolls are independent.
If we call a six a success, then:

- The probability of success on each trial is $p = P(\text{rolling a six}) = \frac{1}{6}$
- The probability of failure is $1 - p = \frac{5}{6}$

Let Y be the number of sixes rolled in n trials. Then $Y \sim \text{Binomial}(n, p)$, and the distribution of Y is called a **binomial distribution**.

4.4 Binomial Distribution

In section 4.2 we learnt about Bernoulli random variables in which we were interested in the outcome of just a single trial. A **binomial random variable** is a generalization of several independent Bernoulli trials. Instead of performing just a single Bernoulli trial and observing whether we have a success or not, we are now performing several Bernoulli trials and observing whether we have a certain number of successes and failures. The **binomial distribution** describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p.

Let $X \sim Bin(n,p)$. The probability of observing x successes in these n independent trials is given by

$$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$$

where

- n represents the number of trials,
- x represents the number of successes,
- p represents the probability of success on any given trial,

$$\binom{n}{x}=\frac{n!}{x!(n-x)!} \quad \text{is the binomial coefficient.}$$

Definition 4.3 (Mean and Variance of a Binomial Random Variable)**.**
Let $X \sim Bin(n,p)$. The mean of X is

$$E(X)=\mu=np$$

and the variance of X is

$$Var(X)=\sigma^2=np(1-p)$$

4.4.1 Visualizing the PMF of Binomial Distributions

R code:

```
## Pmf of Binomial with n=10 and p=1/6.

x ← seq(0, 10, by=1);
y ← dbinom(x, 10, 1/6);

plot(x, y, type= "p", col="blue", pch=19);
```

Probability Mass Functions (PMFs) for increasing n:

The following plots display the probability mass functions (PMFs) for a binomial distribution with $p=\frac{1}{6}$ and increasing values of n. As n increases, the binomial distribution begins to resemble a normal distribution.

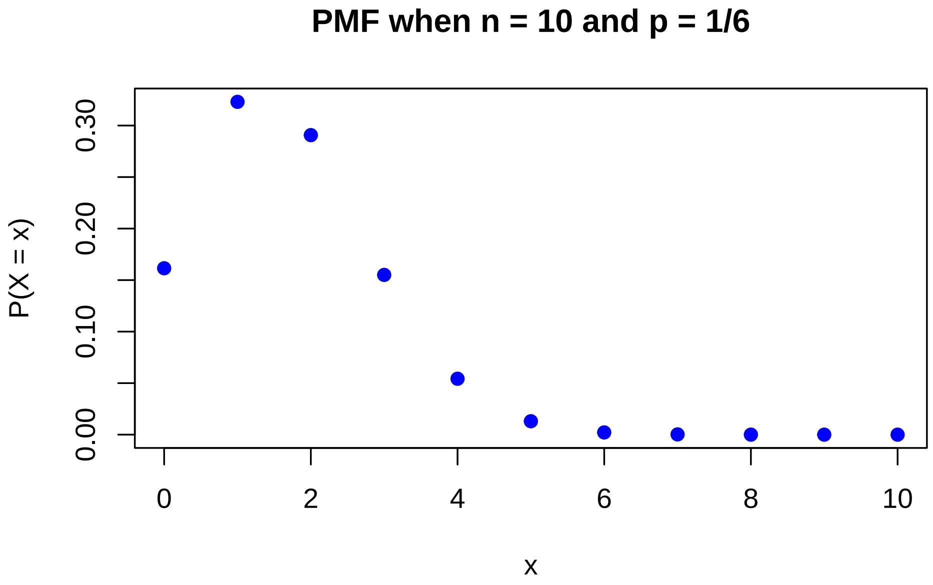


Figure 4.1: PMF of Binomial distribution with $n = 10$ and $p = \frac{1}{6}$.

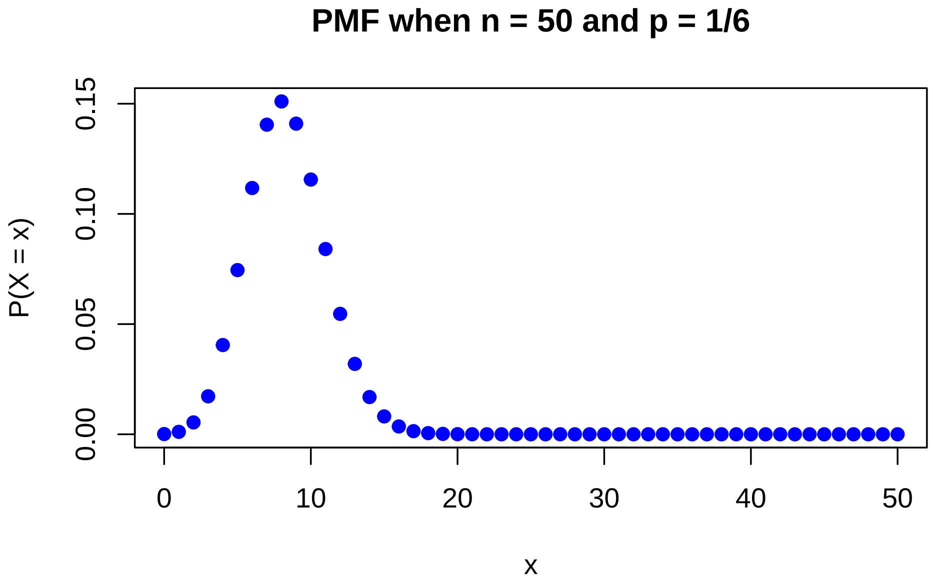


Figure 4.2: PMF of Binomial distribution with $n = 50$ and $p = \frac{1}{6}$.

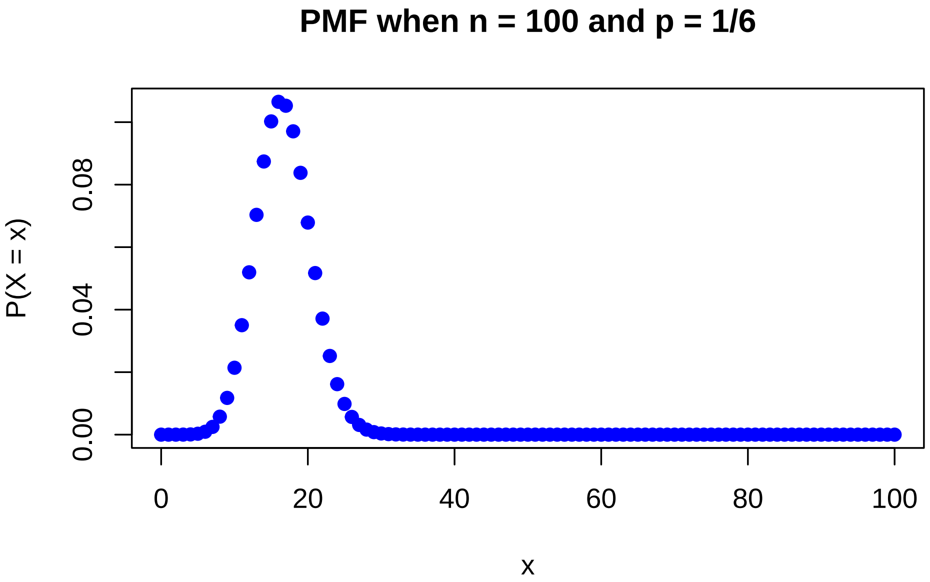


Figure 4.3: PMF of Binomial distribution with $n = 100$ and $p = \frac{1}{6}$.

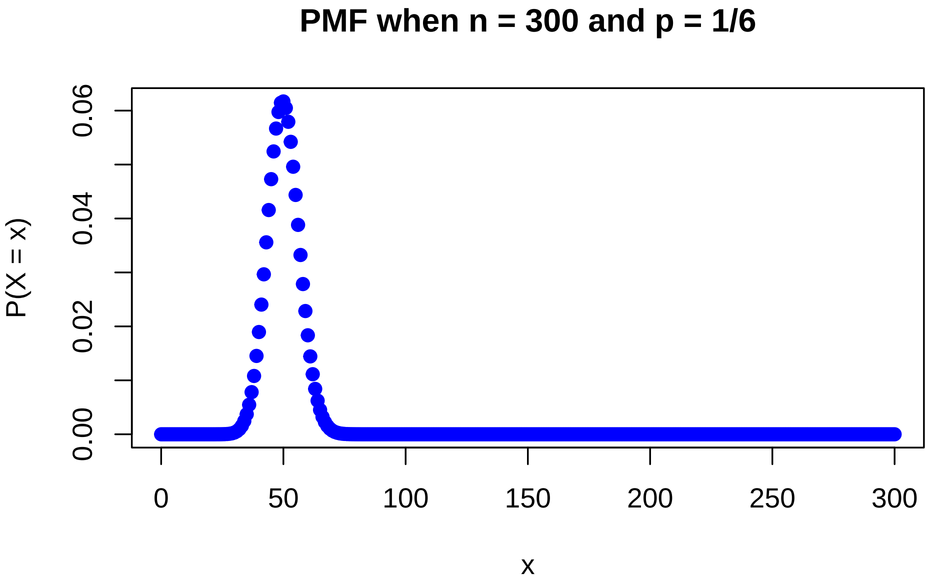


Figure 4.4: PMF of Binomial distribution with $n = 300$ and $p = \frac{1}{6}$.

4.5 Sampling Distribution of a Sample Proportion and the Normal Approximation

When studying categorical data, we are often interested not just in individual outcomes, but in the proportion of successes observed in a sample. Understanding how this proportion behaves across repeated samples is crucial for making inferences about a population. In this section, we explore the sampling distribution of a sample proportion and how it can be approximated by a normal distribution under certain conditions.

Draw a *Simple Random Sample (SRS)* of size n from a large population that contains proportion p of "successes". Let $\hat{p}$ be the ***sample proportion*** of successes:

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

Then:

- The **mean** of the sampling distribution of $\hat{p}$ is p.
- The **standard deviation** of the sampling distribution is $\sqrt{\frac{p(1-p)}{n}}$.

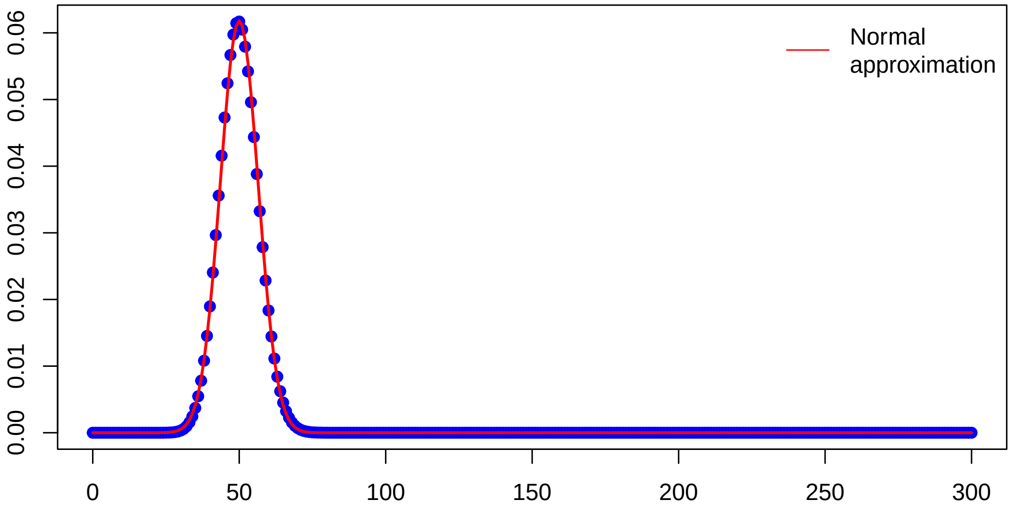


Figure 4.5: Binomial distribution with $n = 300$, $p = \frac{1}{6}$ and its Normal approximation.

According to the Central Limit Theorem (CLT), the sampling distribution of a sample proportion becomes approximately normal as the sample size increases.
That is:

$$\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

This approximation is most accurate when both $np \geq 10$ and $n(1-p) \geq 10$.
These are called the **success-failure conditions**.

Key Point: When the success-failure conditions are met, the normal approximation to the sampling distribution of $\hat{p}$ can be used for probability calculations.

Conditions for Using the Normal Approximation

Suppose $X \sim \text{Binomial}(n, p)$. Then:

$$\mu = np, \quad \sigma^2 = np(1-p)$$

Binomial probabilities can be approximated by the normal distribution:

$$X \approx \mathcal{N}(np,\, np(1-p))$$

This approximation is *useful for large n* and valid under the following conditions:

Standard Conditions

The binomial setting holds (i.e., independent trials, fixed n, same probability p) and

$$np \geq 10 \quad \text{and} \quad np(1-p) \geq 10$$

Alternatively, a more conservative criterion for using the normal approximation is:

$$n > 9 \cdot \left(\frac{\max(p,\, 1-p)}{\min(p,\, 1-p)} \right)$$

These ensure that the binomial distribution is sufficiently symmetric and smooth to approximate with the normal distribution.

We derive the sampling distribution of $\hat{p}$ using properties of the Bernoulli distribution.

Bernoulli Distribution (Binomial with $n = 1$)

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th roll is a six} \\ 0 & \text{otherwise} \end{cases}$$

$$\mu = \mathbb{E}(X_i) = p, \quad \sigma^2 = \text{Var}(X_i) = p(1-p)$$

Let $\hat{p}$ be our estimate of p. Note that $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$. Let $\hat{p} = \frac{\text{\# successes } (X)}{\text{sample size } (n)}$
Recall that for $X \sim \text{Binomial}(n, p)$:

$$X \dot{\sim} \mathcal{N}(np, np(1-p))$$

Let $\hat{p} = \frac{X}{n}$

Mean of $\hat{p}$:

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{X}{n}\right) = \frac{1}{n} \cdot \mathbb{E}(X) = \frac{1}{n} \cdot np = p$$

Variance of $\hat{p}$:

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\cdot \mathrm{Var}(X) = \frac{1}{n^2}\cdot np(1-p) = \frac{p(1-p)}{n}$$

By the Central Limit Theorem (CLT), for sufficiently large n:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Standardization of $\hat{p}$:

$$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

If n is large, then by the Central Limit Theorem:

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \Rightarrow \quad \hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Example 4.1.

[Normal Approximation for Proportions] In the last election, a state representative received 52% of the votes cast. One year after the election, the representative organized a survey that asked a random sample of 300 people whether they would vote for him in the next election. If we assume that his popularity has not changed, what is the probability that more than half the sample would vote for him?

Solution 1 (using Normal Approximation)

We want to determine the probability that the sample proportion is greater than 50%. In other words, we want to find $P(\hat{p} > 0.50)$.
We know that the sample proportion $\hat{p}$ is roughly Normally distributed with mean $p = 0.52$ and standard deviation

$$\sqrt{p(1-p)/n} = \sqrt{(0.52)(0.48)/300} = 0.0288.$$

Thus, we calculate

$$P(\hat{p} > 0.50) = P\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} > \frac{0.50-0.52}{0.0288}\right)$$

$$= P(Z > -0.69) = 1 - P(Z < -0.69) \quad \text{(Z is symmetric)}$$
$$= P(Z > -0.69) = 1 - P(Z > 0.69)$$
$$= 1 - 0.2451 = 0.7549.$$

If we assume that the level of support remains at 52%, the probability that more than half the sample of 300 people would vote for the representative is 0.7549.

R code (Normal approximation) Just type in the following:

```
1 - pnorm(0.50, mean = 0.52, sd = 0.0288)
## [1] 0.7562982
```

Recall that, `pnorm` will give you the area to the left of 0.50, for a Normal distribution with mean 0.52 and standard deviation 0.0288.

Solution 2 (using Binomial)

We want to determine the probability that the sample proportion is greater than 50%. In other words, we want to find $P(\hat{p} > 0.50)$. We know that $n = 300$ and $p = 0.52$.
Thus, we calculate

$$\begin{aligned}
P(\hat{p} > 0.50) &= P\left(\frac{\sum_{i=1}^{n} x_i}{n} > 0.50\right) \\
&= P\left(\sum_{i=1}^{300} x_i > 150\right) \\
&= 1 - P\left(\sum_{i=1}^{300} x_i \leq 150\right) \\
&\text{(it can be shown that } Y = \sum_{i=1}^{300} x_i \text{ has a Binomial distribution with} \\
&n = 300 \text{ and } p = 0.52) \\
&= 1 - F_Y(150)
\end{aligned}$$

R code (using Binomial distribution) Just type in the following:

```
1- pbinom(150, size = 300, prob = 0.52);
## [1] 0.7375949
```

Recall that, `pbinom` will give you the CDF at 150, for a Binomial distribution with $n = 300$ and $p = 0.52$.

Solution 3 (using continuity correction)

We have that $n = 300$ and $p = 0.52$. Thus, we calculate

$$\begin{aligned}
P(\hat{p} > 0.50) &= P\left(\frac{\sum_{i=1}^{n} x_i}{n} > 0.50\right) \\
&= P\left(\sum_{i=1}^{300} x_i > 150\right) \\
&= 1 - P\left(\sum_{i=1}^{300} x_i \leq 150\right)
\end{aligned}$$

(it can be shown that $Y = \sum_{i=1}^{300} x_i$ has a Binomial distribution with $n = 300$ and $p = 0.52$).

$$\begin{aligned}
&\approx 1 - P\left(\sum_{i=1}^{300} x_i \leq 150.5\right) \quad \text{(continuity correction)} \\
&= 1 - P\left(\frac{\sum_{i=1}^{300} x_i}{n} \leq \frac{150.5}{300}\right) \\
&= 1 - P(\hat{p} \leq 0.5017) \\
&= 1 - P\left(Z \leq -0.6354\right) \quad \text{(Why?)}
\end{aligned}$$

R code (Normal approximation with continuity correction) Just type in the following:

```
1 - pnorm(0.5017, mean = 0.52, sd = 0.0288)
## [1] 0.7374216
```

Recall that, `pnorm` will give you the area to the left of 0.5017, for a Normal distribution with mean 0.52 and standard deviation 0.0288.

4.6 Continuity Correction

The normal distribution is continuous, while the binomial distribution is discrete. When we approximate a binomial probability using the normal distribution, this mismatch can lead to inaccuracy—especially near the boundaries of discrete values. A continuity correction improves the approximation by adjusting for this difference. In this section, we explore how and why this correction is applied.

Suppose that Y has a Binomial distribution with $n = 20$ and $p = 0.4$. We will find the exact probabilities that $Y \leq y$ and compare these to the corresponding values found by using two Normal approximations. One of them, when X is Normally distributed with $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$. The other one, W, a shifted version of X.

For example,

$$P(Y \leq 8) = 0.5955987$$

As previously stated, we can think of Y as having approximately the same distribution as X.

$$P(Y \leq 8) \approx P(X \leq 8) = P\left[\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{8-8}{\sqrt{20(0.4)(0.6)}}\right] = P(Z \leq 0) = 0.5$$

$$P(Y \leq 8) \approx P(W \leq 8.5) = P\left[\frac{W - np}{\sqrt{np(1-p)}} \leq \frac{8.5 - 8}{\sqrt{20(0.4)(0.6)}}\right] = P(Z \leq 0.2282) = 0.5902615$$

Example 4.2.

Fifty-one percent of adults in the U. S. whose New Year's resolution was to exercise more achieved their resolution. You randomly select 65 adults in the U. S. whose resolution was to exercise more and ask each if he or she achieved that resolution. What is the probability that exactly forty of them respond yes?

We are given that $p = 0.51$, $n = 65$, and we want to find $P(X = 40)$ where $X \sim Binomial(n = 65, p = 0.51)$.

Use Normal Approximation We use normal approximation to the binomial. First, compute the mean and standard deviation:

$$\begin{aligned} \mu &= np = 65 \times 0.51 = 33.15 \\ \sigma^2 &= np(1-p) = 65 \times 0.51 \times 0.49 = 16.485 \\ \sigma &= \sqrt{16.485} \approx 4.06 \end{aligned}$$

We apply continuity correction:

$$P(X = 40) = P(39.5 \leq X \leq 40.5)$$

$$= P\left(\frac{39.5 - 33.15}{4.06} \leq Z \leq \frac{40.5 - 33.15}{4.06}\right) = P(1.56 \leq Z \leq 1.81)$$

From the standard normal table:

$$= P(Z \leq 1.81) - P(Z \leq 1.56) = 0.0594 - 0.0352 = 0.0242$$

So the approximate probability is:

$$P(X = 40) \approx 0.0242$$

Normal Approximation to Binomial

Let $X = \sum_{i=1}^{n} Y_i$ where $Y_1, Y_2, \ldots, Y_n$ are iid Bernoulli random variables. Note that $X = n\hat{p}$.

1. $n\hat{p}$ is approximately Normally distributed provided that $np \geq 10$ and $n(1-p) \geq 10$.

2. Another criterion is that the Normal approximation is adequate if

$$n > 9\left(\frac{\text{larger of } p \text{ and } q}{\text{smaller of } p \text{ and } q}\right)$$

3. The expected value: $E(\hat{p}) = np$.
4. The variance: $V(\hat{p}) = np(1-p) = npq$.

Chapter 5

Law of Large Numbers

5.1 Convergence in Probability

Definition 5.1 (Convergence in Probability).
The sequence of random variables $X_1, X_2, X_3, \ldots, X_n, \ldots$ is said to ***converge in probability*** *to the constant c, if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(|X_n - c| \leq \epsilon\right) = 1$$

or equivalently,

$$\lim_{n\to\infty} P\left(|X_n - c| > \epsilon\right) = 0$$

Notation: $X_n \xrightarrow{P} c$

This concept plays a key role in the Law of Large Numbers, where the sample mean of independent and identically distributed random variables converges in probability to the population mean as the sample size grows.

Definition 5.2 (Chebyshev's Inequality).
Let X be a random variable with finite mean μ and variance σ^2. Then, for any $k > 0$,

$$P\left(|X - \mu| \geq k\right) \leq \frac{\sigma^2}{k^2}$$

Using complements:

$$P\left(|X - \mu| < k\right) \geq 1 - \frac{\sigma^2}{k^2}$$

5.2 Weak Law of Large Numbers (WLLN)

Definition 5.3 (Weak Law of Large Numbers (WLLN)).
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having finite mean $E(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$. Then, for any $\epsilon > 0$,

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0 \quad \textit{as } n \to \infty$$

Notation: $\bar{X}_n \xrightarrow{P} \mu$

Proof of the Weak Law of Large Numbers (WLLN)

We aim to show that for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) = 0$$

where $\bar{X}_n$ is the sample mean of n independent and identically distributed (i.i.d.) random variables with

$$E(X_i) = \mu, \quad \text{and} \quad \mathrm{Var}(X_i) = \sigma^2.$$

Let

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

By the Central Limit Theorem (CLT), we know that

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Now, applying **Chebyshev's Inequality**, which states that for any random variable X with mean μ and variance σ^2,

$$P\left(|X - \mu| > k\right) \leq \frac{\sigma^2}{k^2} \quad \text{for } k > 0,$$

to $\bar{X}_n$, we set $k = \epsilon$, and obtain:

$$P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Taking the limit as $n \to \infty$, we have:

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0.$$

Since probabilities are always non-negative, we conclude:

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) = 0.$$

By the definition of convergence in probability,

$$\bar{X}_n \xrightarrow{P} \mu.$$

□

Example 5.1.
[Poisson Convergence via WLLN]
Let X_i, for $i = 1, 2, 3, \ldots$, be independent Poisson random variables with rate parameter $\lambda = 3$. Prove that:

$$\bar{X}_n \xrightarrow{P} 3$$

Properties of Poisson Distribution:

$$E(X_i) = \lambda, \quad \text{Var}(X_i) = \lambda$$

In this case, $\lambda = 3$, so:

$$E(X_i) = \text{Var}(X_i) = 3$$

Proof:
We know:

$$E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = 3, \quad \text{and} \quad \text{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{3}{n}$$

Applying Chebyshev's Inequality:

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - 3\right| \geq \epsilon\right) \leq \frac{3}{n\epsilon^2}$$

Taking the limit as $n \to \infty$:

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - 3\right| \geq \epsilon\right) \to 0$$

Conclusion:

$$\bar{X}_n \xrightarrow{P} 3$$

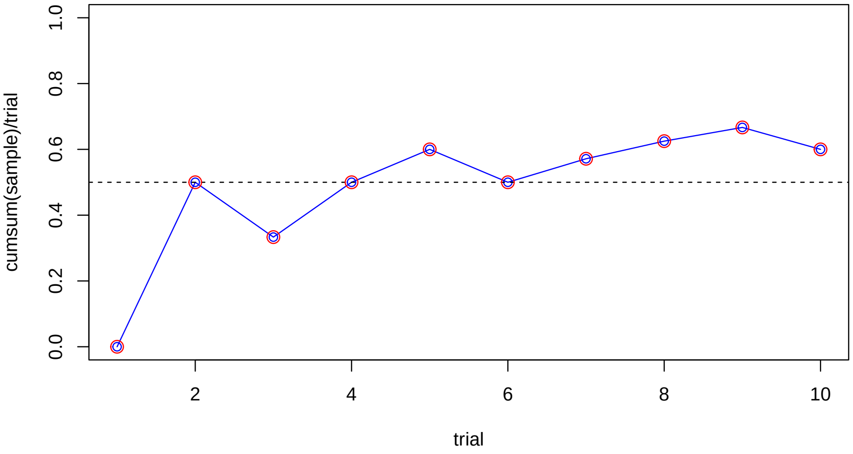


Figure 5.1: Simulation of running sample mean of Bernoulli($p = 0.5$) trials over time.

R Simulation Code (Single Sample Path)

```
n = 10
trial = seq(1, n, by = 1)
sample = rbinom(n, 1, 1/2)

plot(trial, cumsum(sample)/trial, type = "l", ylim = c(0,1), col = "blue")
points(trial, cumsum(sample)/trial, col = "red")
abline(h = 0.5, lty = 2, col = "black")
```

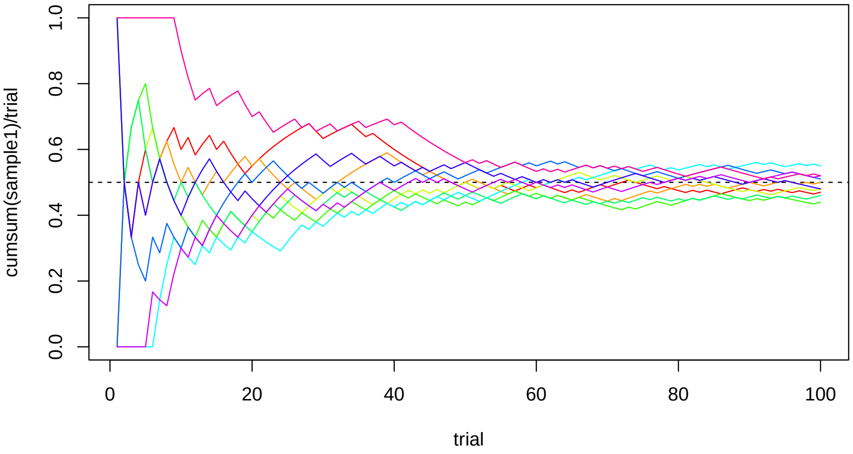


Figure 5.2: Simulation of 10 running sample means of Bernoulli($p = 0.5$) trials converging over 100 trials.

R Simulation Code (Multiple Sample Paths)

```
n = 100
trial = seq(1, 100, by = 1)

sample1 = rbinom(n, 1, 1/2)
sample2 = rbinom(n, 1, 1/2)
sample3 = rbinom(n, 1, 1/2)
sample4 = rbinom(n, 1, 1/2)
sample5 = rbinom(n, 1, 1/2)
sample6 = rbinom(n, 1, 1/2)
sample7 = rbinom(n, 1, 1/2)
sample8 = rbinom(n, 1, 1/2)

colors = rainbow(8)


plot(trial, cumsum(sample1)/trial, type = "l", col = colors[1], ylim = c(0,1))
lines(trial, cumsum(sample2)/trial, col = colors[2])
lines(trial, cumsum(sample3)/trial, col = colors[3])
lines(trial, cumsum(sample4)/trial, col = colors[4])
lines(trial, cumsum(sample5)/trial, col = colors[5])
lines(trial, cumsum(sample6)/trial, col = colors[6])
lines(trial, cumsum(sample7)/trial, col = colors[7])
lines(trial, cumsum(sample8)/trial, col = colors[8])
abline(h = 0.5, lty = 2, col = "black")
```

Empirical Probability Insight

The Law of Large Numbers gives us empirical probabilities. Consider tossing a fair coin. Define the random variable X as:

$$X = \begin{cases} 1 & \text{heads up} \\ 0 & \text{tails up} \end{cases}$$

Then as we sample more and more values of X, the sample mean $\bar{X}_n$ converges in probability to $P(\text{heads up})$, that is:

$$\bar{X}_n \xrightarrow{P} P(\text{heads up})$$

Chapter 6

One Sample Confidence Intervals on a Mean When the Population Variance is Known

6.1 Introduction

Statistical inference is concerned primarily with understanding the quality of parameter estimates. For example, a classic inferential question is, "How sure are we that the estimated mean, $\bar{x}$, is near the true population mean, μ?" While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics. We introduce these common themes by discussing inference about the population mean, μ, and set the stage for other parameters and scenarios. Some advanced considerations are discussed. Understanding this chapter will make the rest of this book, and indeed the rest of statistics, seem much more familiar.

Definition 6.1 (Key Terms).

Population: *A group of interest (typically large).*

Sample: *A subset of a population.*

Parameter (of population): *A numerical characteristic of a population. These are usually unknown in real-life settings.*

μ*: population mean*

σ^2*: population variance*

σ*: population standard deviation*

Note: Different from a parameter of a distribution.

Statistic (of sample): *A numerical characteristic of a sample, which is calculated and known (i.e., a function of the data).*

$\bar{x}$*: sample mean*

s^2*: sample variance*

s*: sample standard deviation*

Statistical Inference: *Use statistics (known) to make conclusions on parameters (un-*

known) and quantify the degree of certainty of statements made.

The sample mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is a number we use to estimate the population mean, μ. This is called a **point estimate**.

But, we know it's not equal to μ. Then, we'd rather estimate the population mean using an **interval estimate** that gives a *range of real numbers* that we hope contains the population mean, μ.

Example 6.1.

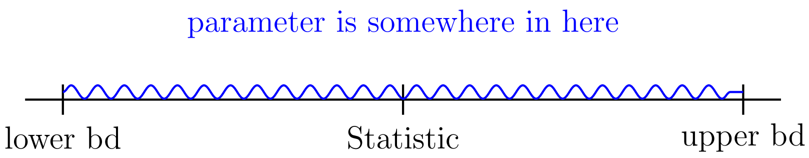
- $\bar{x}$ is a point estimate of μ
- s^2 is a point estimate of σ^2
- s is a point estimate of σ

(All calculated with data from a sample)

Due to the nature of randomness and calculating based on a subset, statistics are not guaranteed to be exactly equal to parameters.

Therefore, we create intervals around statistics which we believe capture the parameter.

Definition 6.2 (Confidence Interval)**.**
A confidence interval is a plausible range of values that captures a parameter with a quantified degree of confidence.



Suppose we are interested in the average mark for `STA258` for the current semester. We are 100% confident that the average mark is between 0 and 100; however, this is not useful information as we already know that the average mark must lie between 0 and 100. Using the marks of previous years, we can construct a 95% interval for the average mark. If it is determined that the average mark lies within 70% and 80%, this is much more meaningful as we can state with a high degree of certainty that the average mark is going to lie within a substantially narrow range.

In this course, all confidence intervals have the same basic skeleton:

$$\textit{estimator} \pm \underbrace{(\textit{value from reference distribution}) \times (\textit{standard error of estimate})}_{\textit{margin of error}}$$

The value from the reference distribution in the skeleton above will be either a value from the standard normal distribution or the Student t-distribution. The margin of error (MOE) can be considered as the distance around our estimator in which the true value of the parameter of interest will be found, with a specified level of confidence.

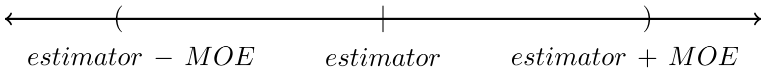


Figure 6.1: Visualization of a confidence interval on the real number line. The margin of error is abbreviated as MOE. The estimator is the centre of the interval. The confidence interval consists of all values between the estimator$-MOE$ and the estimator$+MOE$.

6.2 Interpretation

We use very specific language when we interpret a confidence interval.

Suppose we construct a C% confidence interval for some parameter such that C is between 0 and 100. In repeated sampling, we are C% confident that approximately C% of the intervals will capture the true value of the parameter.

By this we mean that if we constructed several C% confidence intervals using different samples (with or without replacing the units), then we should expect approximately C% of these intervals to capture the parameter of interest. For example suppose we construct 1000 95% confidence intervals for the population mean μ. We would expect approximately 95% of these 1000 intervals (i.e. 95% × 1000 = 950) to actually capture μ.

Note 6.1.

A more intuitive but equivalent interpretation is to state that we are $C\%$ confident that our target parameter is inside the interval constructed.

It is incorrect to state that there is a $C\%$ probability that the interval we constructed contains the parameter of interest. We assume that the value of a parameter is fixed. Therefore when we construct a confidence interval, the interval either contains the parameter or it does not.

6.3 Confidence Interval for μ (Known Variance)

When we know the population standard deviation σ, we can construct a confidence interval for μ in the following manner.

Confidence Interval 6.1 (Confidence Interval on μ when σ is Known)

A $(100-\alpha)\%$ confidence interval on μ when σ is known is

$$\bar{x} \ \pm \ z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

The $z_{\alpha/2}$ value is obtained from standard normal tables. The standard error is $\frac{\sigma}{\sqrt{n}}$ and the margin of error is $z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$.

Let $X_1, X_2, ..., X_n$ be iid $N(\mu, \sigma^2)$, where μ is unknown and σ is known.
We know that:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

and

$$P(-1.96 < Z < 1.96) = 0.95$$

Therefore:

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \Rightarrow P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Interpretation of Confidence Interval:

- This is a random interval $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$
- The interval is random since $\bar{X}$ is random due to sampling.
- The population mean μ is a fixed, but unknown, number.
- The probability μ is inside the random interval is 0.95 (success rate of the method).
- 95% of all samples give an interval that captures μ, and 5% do not.

Once we observe our sample:

- This is **not** a random interval $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$
- The probability μ is inside this interval is either 1 or 0

Confidence Interval Isn't Always Right:
Not all CIs contain the true value of the parameter. This can be illustrated by plotting many intervals simultaneously and observing.
—

R Output:

```
## Step 1. Generate random samples;
set.seed(2017)
m = 50;          # m = number of samples;
n = 25;          # n = number of obs in sample;
mu.i = 0;        # mu.i = mean of obs;
sigma.i = 5;     # sigma.i = std. dev. of obs;

mu.total = n * mu.i;              # mean of Total;
sigma.total = sqrt(n) * sigma.i;  # std. dev. of Total;
```

R Output:

```
## Step 2. Construct CIs;
xbar = rnorm(m, mu.total, sigma.total) / n;
SE = sigma.i / sqrt(n);

alpha = 0.10;
z.star = qnorm(1 - alpha / 2);
```

R Output:

```
## Step 3. Graph CIs;
matplot(rbind(xbar - z.star * SE, xbar + z.star * SE),
        rbind(1:m, 1:m),
        type = ”l”, lty = 1,
        xlab = ”´”, ylab = ”´”);
abline(v = 0, lty = 2);
```

Confidence Interval for the Mean of a Normal Population

Draw an SRS (Simple Random Sample) of size n from a Normal population having unknown mean μ and **known** standard deviation σ. A level C confidence interval for μ is:

$$\bar{x} \pm z_* \cdot \frac{\sigma}{\sqrt{n}}$$

The critical value z_* is illustrated in a Figure below and depends on C.

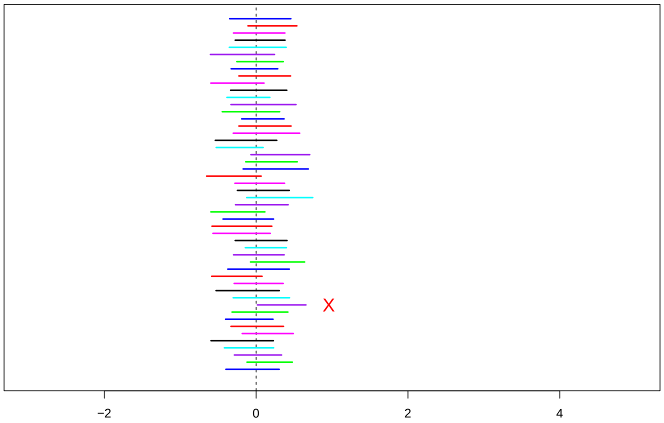


Figure 6.2: Simulated 95% confidence intervals for the population mean. Red “X” marks indicate intervals that do not contain the true mean ($\mu = 0$).

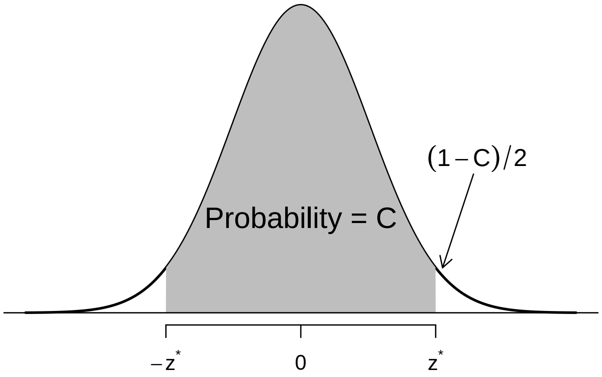


Figure 6.3: The central area under the standard normal curve with confidence level C.

Large Sample CI for μ (Normal data)

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Valid if:

- n large
- random sample from a Normal distribution
- independent observations

Some definitions:

- $1-\alpha$ is the confidence coefficient
- $100(1-\alpha)\%$ is the confidence level

One Sample CI on the Population Mean μ

- When population standard deviation σ is **known**
- Formula: $\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
- Margin of error comes from standard normal and standard error

How to find $z_{\alpha/2}$?
Example: Find $z_{\alpha/2}$ for a 95% CI on μ:

$$1-\alpha = 0.95, \quad \alpha = 0.05, \quad \alpha/2 = 0.025$$

$$z_{\alpha/2} = 1.96 \quad \text{(from table or R: \texttt{qnorm(0.975)})}$$

Table of Common z-values

Confidence coefficient	Confidence level	z
0.90	90%	1.645
0.95	95%	1.96
0.99	99%	2.576

Example 6.2.
Playbill magazine reported that the mean annual household income of its readers is \$119,155. Assume this estimate is based on a sample of 80 households, and that the population standard deviation is known to be $\sigma = 30{,}000$.

- $\bar{x} = 119{,}155$
- $n = 80$
- $\sigma = 30{,}000$

Tasks:

(a) Develop a 90% confidence interval estimate of the population mean.

(b) Develop a 95% confidence interval estimate of the population mean.

(c) Develop a 99% confidence interval estimate of the population mean.

90% CI Calculation

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 119{,}155 \pm 1.645 \cdot \frac{30{,}000}{\sqrt{80}}$$

$$= 119{,}155 \pm 5{,}500.73$$
$$= (113{,}654.27,\ 124{,}655.73)$$

95% CI Calculation

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 119{,}155 \pm 1.96 \cdot \frac{30{,}000}{\sqrt{80}}$$
$$= 119{,}155 \pm 6{,}574.04$$
$$= (112{,}580.96,\ 125{,}729.04)$$

99% CI Calculation

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 119{,}155 \pm 2.576 \cdot \frac{30{,}000}{\sqrt{80}}$$
$$= 119{,}155 \pm 8{,}620.04$$
$$= (110{,}534.96,\ 127{,}775.04)$$

Interpretation
We are 99% confident the mean household income of magazine readers is between $110,534.96 and $127,775.04.

Example 6.3.

Scenario:

The number of cars sold annually by used car salespeople is known to be **normally distributed**, with a population standard deviation of $\sigma = 15$. A random sample of $n = 15$ salespeople was taken, and the number of cars each sold is recorded below. Construct a **95% confidence interval** for the population mean number of cars sold, and provide an interpretation.

Raw data:

79	43	58	66	101
63	79	33	58	71
60	101	74	55	88

The sample mean is:

$$\bar{x} = \frac{79 + 43 + \cdots + 55 + 88}{15} = 68.6$$

R function:

```
simple.z.test = function(x, sigma, conf.level = 0.95) {
  n = length(x);
  xbar = mean(x);
  alpha = 1 - conf.level;
  zstar = qnorm(1 - alpha/2);
  SE = sigma / sqrt(n);
  xbar + c(-zstar * SE, zstar * SE);
}
```

R output:

```
# Step 1. Entering data;
cars = c(79, 43, 58, 66, 101, 63, 79,
         33, 58, 71, 60, 101, 74, 55, 88)

# Step 2. Finding CI;
simple.z.test(cars, 15)

## [1] 61.00909 76.19091
```

Interpretation: We estimate that the mean number of cars sold annually by all used car salespeople lies between 61 and 76, approximately. This type of estimate is correct 95% of the time.

Cases Where Valid

- Large samples where population is **normal**.
- Large samples where population is **not normal** (By CLT).
- Small samples where population is **normal**.

Note: A sample is considered large if $n \geq 30$.

Example 6.4.

Suppose a student measuring the boiling temperature of a certain liquid observes the readings (in degrees Celsius) 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2 on 6 different samples of the liquid. He calculates the sample mean to be 101.82. If he knows that the distribution of boiling points is Normal, with standard deviation 1.2 degrees, what is the confidence interval for the population mean at a 95% confidence level?

A confidence interval uses sample data to estimate an unknown population parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct.

The **interval** often has the form
estimate ± margin of error

The **confidence level** is the success rate of the method that produces the interval. A level C **confidence interval for the mean** μ of a Normal population with **known** standard deviation σ, based on an SRS of size n, is given by

$$\bar{x} \pm z^{\star} \frac{\sigma}{\sqrt{n}}$$

The **critical value** $z^{\star}$ is chosen so that the standard Normal curve has area C between $-z^{\star}$ and $z^{\star}$.

Other things being equal, the **margin of error** of a confidence interval gets smaller as

- the confidence level C decreases,
- the population standard deviation σ decreases, and
- the sample size n increases.

6.4 APPENDIX

Interval estimators are commonly called **confidence intervals**. The upper and lower endpoints of a confidence interval are called the **upper** and **lower confidence limits**, respectively. The probability that a (random) confidence interval will enclose θ (a fixed quantity) is called the **confidence coefficient**.

Suppose that $\hat{\theta}_L$ and $\hat{\theta}_U$ are the (random) lower and upper confidence limits, respectively, for a parameter θ. Then, if

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha,$$

the probability $(1 - \alpha)$ is the **confidence coefficient**.

Pivotal quantities

One very useful method for finding confidence intervals is called the **pivotal method**. This method depends on finding a pivotal quantity that possesses two characteristics:

- It is a function of the sample measurements and the unknown parameter θ, where θ is the **only** unknown quantity.
- Its probability distribution does not depend on the parameter θ.

Index