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 Companies that don't adapt and stick to an outdated business model tend to pay the price.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

In this section, we will briefly introduce R, a useful tool for statistical computing and data visualisation, discuss quantitative data with its properties, basic statistical value calculation and box plot.

1.1 Overview

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

Definition 1.1 (Statistics)**.** *Statistics is the science of collecting, classifying, summarising, analysing and interpreting data.*

1.2 Descriptive statistics

Descriptive statistics refers to the entire progress of summarising numerical and categorical data, then analysing those.

1.2.1 Sample Mean, Sample variance and Sample Standard Deviation

Sample mean is the average value of a sample of numbers taken from a population. Sample variance is a measure of dispersion in that sample which shows how far a set of numbers is spread out from sample mean. Sample standard deviation is a measure of the amount of variation of the values of a variable about its sample mean.

Definition 1.2. *Let $x_1, x_2, x_3, ..., x_n$ be a sample of data points. We define sample mean of the sample data*

points ($\bar{x}$) with n observations 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:

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

Let's proceed with an example, to see how those values are calculated.

Example 1.1.
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 1.2, 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 $s > 0$:

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

1.2.2 Median and Mode

Median indicates the information about the central value of a given collection of data points. Mode refers to a value which appears most frequently in a dataset.

Median:

Definition 1.3.
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:*
$$M = \frac{n+1}{2}, \text{ where } n \text{ represents the } n^{th} \text{ position.}$$
- *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.*

Example 1.2.
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,

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

Mode

Definition 1.4.
The mode of a dataset is a value which appears most frequently in it

Example 1.3.
Consider a dataset: $\{1, 2, 2, 3, 3, 3, 4, 4, 4, 4\}$. Find the mode of this dataset.
Solution:
Since '4' appears 4 time in the set, which is the most frequent. Thus, the mode of this given dataset is: 4.

1.2.3 Percentile and Quartile

In statistics, percentiles and quartiles are tools used to describe the relative standing of data points within a dataset. They help us understand how individual values compare to the rest of the data.

Definition 1.5.
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, 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.

Example 1.4.
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.

Five Number Summery

The five number summery are: minimum, $Q1$, $Q2$, $Q3$ and maximum.

1.2.4 Skewness and Symmetry

In probability theory and statistics, the word 'skewness' is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. There are two types of skewness: left skewed (or negative skew) and right skewed (or positive skew).

Left Skewed (Negative Skew)

For a probability distribution of a real-valued random variable about its mean to be left skewed or negative skewed, we observe the tail of its graph, where the tail locates on the left. (See figure 1.1 below)

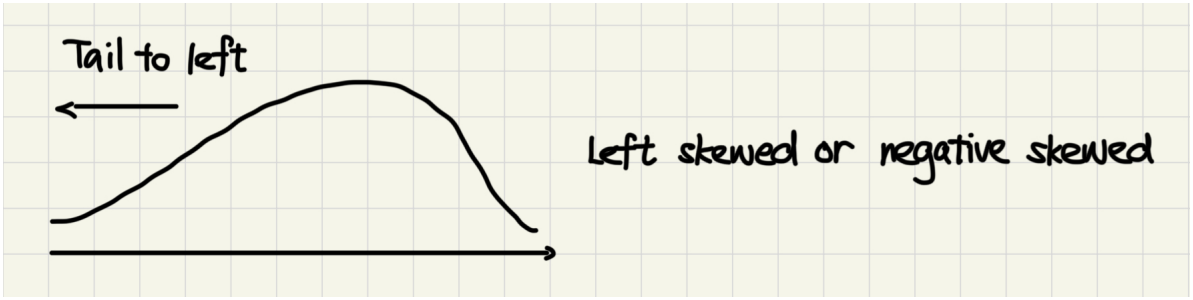

Figure 1.1: Visualization of a left skewed distribution

Right Skewed (Positive Skew)

Similarly, if a probability distribution of a real-valued random variable about its mean has graph where its tail locates on the right, then we say that distribution is right skewed or positive skewed. (See figure 1.2 below)

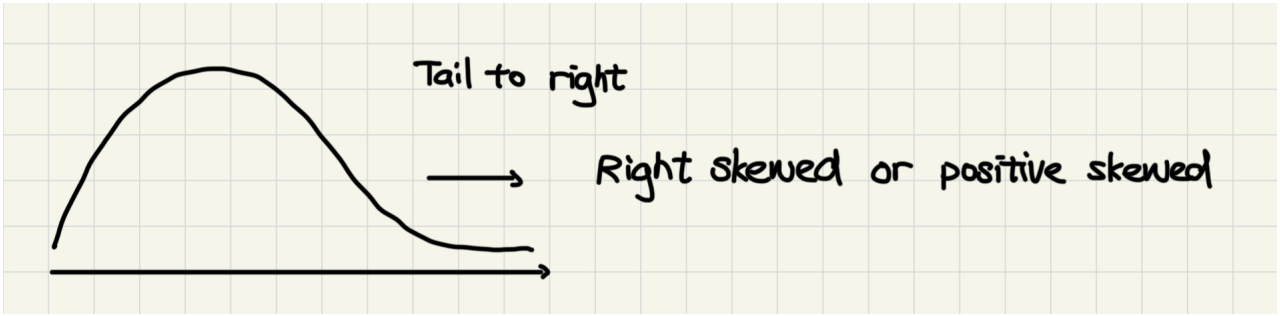

Figure 1.2: Visualization of a right skewed distribution

Two classic example of right skewed probability distribution is χ^2-distribution and F-distribution.

Symmetry

In statistics, a symmetric probability distribution is reflected around a vertical line at a certain value. That means the probability on the left side of that line at a certain point is equal to the probability on the right. (See figure 1.3 below)

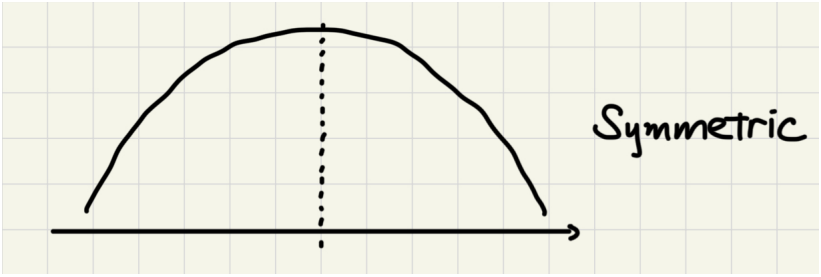

Figure 1.3: Visualization of a symmetric distribution

Some common examples of symmetric probability distribution in real world are: normal distribution and student's t-distribution.

The Empirical Rule

Definition 1.6 (The 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.: $Pr(\mu - 3\sigma < X < \mu + 3\sigma) = 99.73\%$.

1.3 Graphical Techniques

In this section, we are going to introduce some common statistical graphs. Then, we will obtain information and do interpretation from the graphs.

1.3.1 Histograms

A histogram is constructed by placing the variables of interest on the horizontal axis and the frequency, relative frequency, or percent frequency on the vertical axis, which is a type of quantitative or numerical data for statistical analysis. For example: STA258 final mark from 100 different students at UTM. The following figure is an example of histogram:

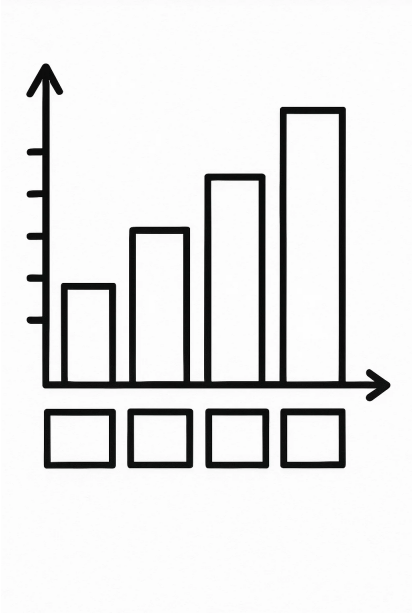

Figure 1.4: Visualization of a histogram: its horizontal axis lists the bins of a data, its vertical axis represents the frequency of the occurance

Advantages and Disadvantages of Histograms

- Advantages:
  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:
  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 Relate to Skewness and Symmetry

We can determine skewness and symmetry by observing and drawing a curve above columns on histograms. We use this method to approximate the shape of probability distribution baed on histograms. As the following figures show:

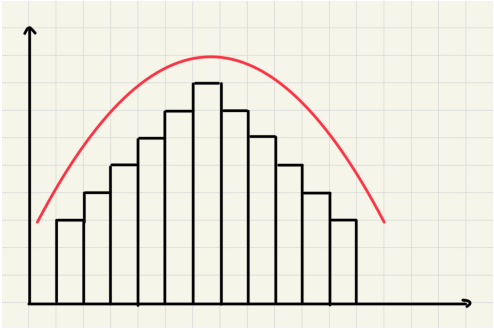

Figure 1.5: Visualisation of a histogram with approximate symmetric probability distribution

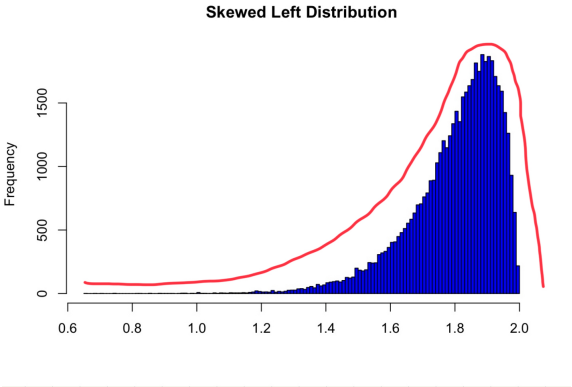

Figure 1.6: Visualisation of a histogram with approximate left skew probability distribution

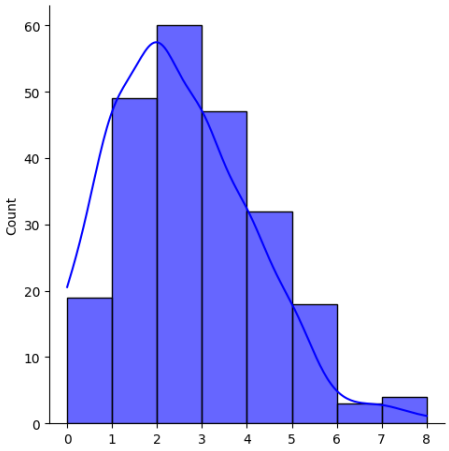

Figure 1.7: Visualisation of a histogram with approximate right skew probability distribution

1.3.2 Box Plot

In descriptive statistics, we use box plot to demonstrating graphically the locality, spread and skewness groups of numerical data through their quartiles. Let's use a figure to demonstrate a box plot.

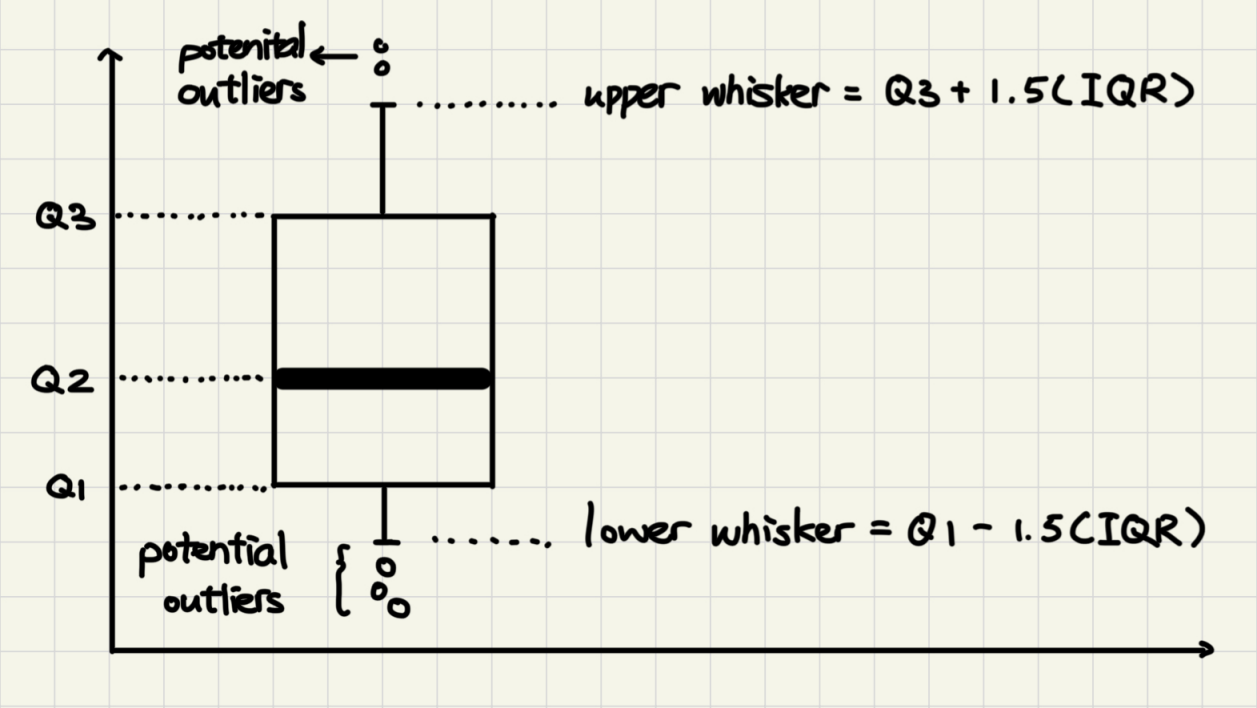

Figure 1.8: Visualisation of a box plot

In this course, we should be able to obtain some key values to help us to continue our statistical analysis. These values are: minimum, $Q1$, $Q2$ (median), $Q3$, maximum and outliers.

Box Plot and Skewness of Probability Distribution

We can obtain skewness from box plot as well. From box plot instead of drawing a curve, we observe how median ($Q2$) separates the 'box' in box plot. In words, if median is greater than mean in a box plot, then it has a negative skewed probability distribution. Otherwise, if median is less than mean in a box plot, then it has a positive skewed probability distribution. However, when median is equal to mean in a box plot, then it has a symmetric probability distribution. (See figure below)

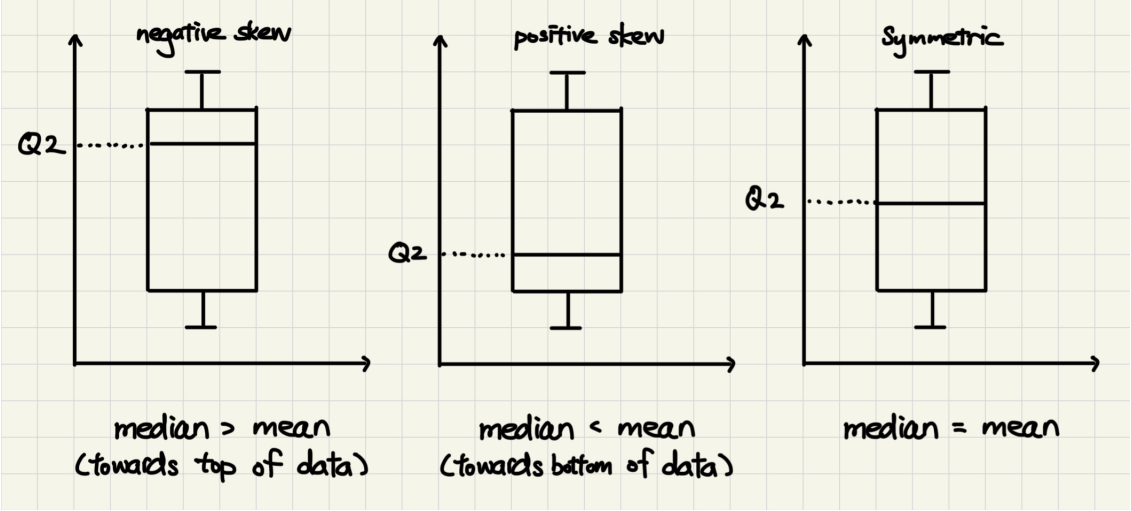

Figure 1.9: Visualisation of how to obtain skewness and symmetry from a box plot

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).

Chapter 2

Sampling Distributions Related to a Normal Population

In Chapter 1, we introduced some basic statistical values, now we are going to introduce some distributions.

2.1 Normal Distribution

Normal distribution or Gaussian distribution in probability theory and statistics, is a type of continuous probability distribution. It is discovered by a German mathematician Carl Friedrich Gauss in 1809, and denoted as $N(\mu, \sigma^2)$. Generally, its probability density function is the following:

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

Normal distribution is important in statistics and is widely used in the natural and social sciences to represent real-valued random variables whose distributions are unknown. Based on that, we have central limit theorem (C.L.T, which we will discuss it in the next chapter) that helps mathematicians and statisticians to solve real world problems.

Definition 2.1.
Let: $x_1, x_2, x_3, ..., x_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ is normally distributed with mean } \mu_{\bar{x}} = \mu \text{ and variance } \sigma^2_{\bar{x}} = \frac{\sigma^2}{n}.$$

We write as:

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

Then, for a singe variable x_i, for $i \in \{1, 2, ..., n\}$, it follows:

$$z = \frac{x_i - \mu}{\sigma}, \text{ and } z \sim N(\mu = 0, \sigma^2 = 1) \text{ which } z \text{ is standard normal distribution.}$$

Next, $\bar{x}$ (sample mean: average on multiple observations) follows:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{(\frac{\sigma}{\sqrt{n}})},\ z \text{ is standard normal distribution as well.}$$

By using $Definition 2.1$, we are able to solve some probability questions regarding to normal distribution. Next, let's proceed with a classic example.

Example 2.1.

Consider marks on a standardised test are **normally distributed** with $\mu = 75$ and $\sigma = 15$. What is the probability the **class average**, for a class of 30, is greater than 76?

Solution: We are asked to find the probability of class average, which is $\bar{x}$. Then we proceed the transformation of sample mean: $P(\bar{x} > 76)$.

Then:

$$P(\bar{x} > 76) = P(\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{76 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}) = P(z > \frac{76 - 75}{(\frac{15}{\sqrt{30}})}) = P(z > 0.37)$$

Using the standard normal distribution table, the final answer is: $P(z > 0.37) = 0.3557$.

2.2 Chi-squared (χ^2) and Gamma Distribution:

2.2.1 Chi-squared distribution

Index