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 Companies that don't adapt and stick to an outdated business model tend to pay the price.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

Introduction to R and Assessing Normality

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 Basic

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.*

1.2 What Is 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 RStudio?
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).

1.3 Quantitative Data

A common graphical representation of quantitative data is a histogram. This graphical summary can be prepared for data previously summarised in either a frequency, relative frequency, or percent frequency distribution. 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.

- Bar Charts v.s. Histograms:
  1. Bar chats are for qualitative or categorical data (i.e. Favourite food data for 100 different students at UTM).
  2. Histograms are for quantitative or numerical data (i.e. STA258 final mark from 100 different students at UTM).

- 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).

- Skewness and Empirical Rule (or $68-95-99.7$ Rule):
  There are three types of skewness that are right (or positive) skewed (i.e. χ^2 distribution), left (negative) skewed and symmetric. For any symmetric (bell-shaped) curve (i.e. normal distribution and $t-$ distribution), it follows the Empirical Rule as the following defined:

Definition 1.2 (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.: $Pr(\mu-3\sigma<X<\mu+3\sigma)=99.73\%$.

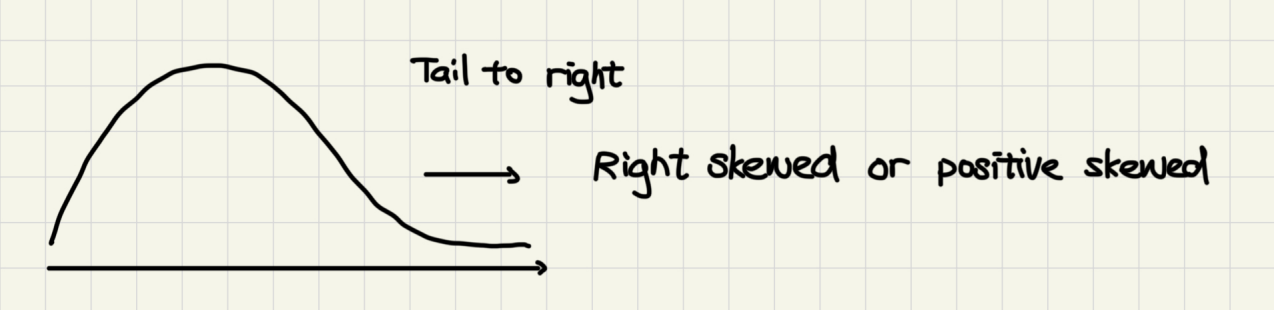

Figure 1.1: Visualization of a right skewed distribution

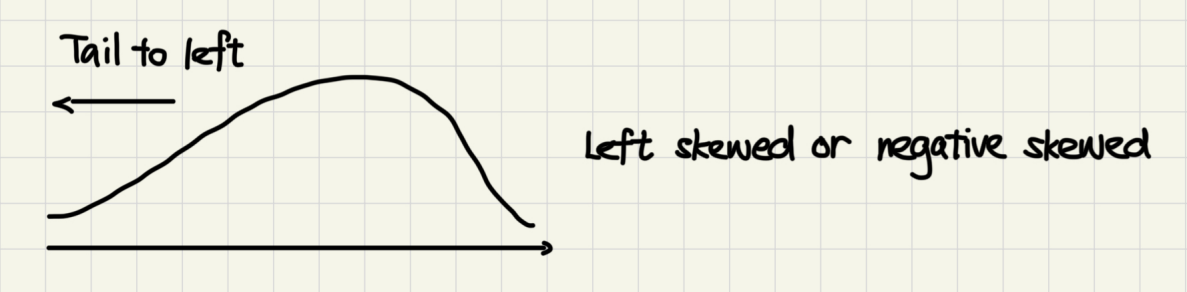

Figure 1.2: Visualization of a left skewed distribution

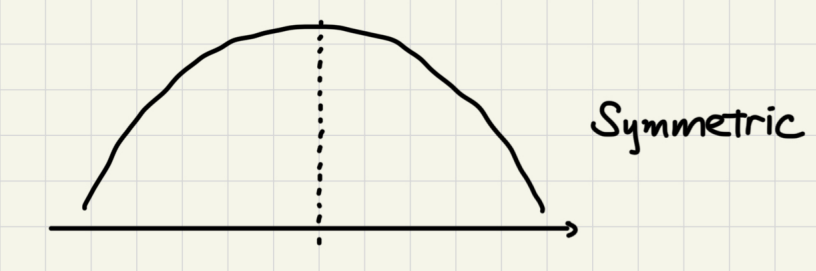

Figure 1.3: Visualization of a symmetric distribution

1.4 Basic Statistical Value Calculation:

To begin with this section, we start from three main measures in quantitative statistics are the mean, variance and standard deviation. These measures form the basis of any statistical analysis.

1.4.1 Sample Mean, Sample variance and Sample Standard Deviation

Definition 1.3.
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:

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

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 $s > 0$:

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

1.4.2 Median, Percentile and Quartile:

Now, we move to median and percentile. Median indicates the information about the central value of a given collection of data points; percentile: p^{th} percentile which is a value that indicates $p\%$ of observations are below it.

Median:

Definition 1.4.
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,

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

Percentile and Quartile:

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, 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.

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.

1.5 Box Plot:

It is a relatively simple visualisation of data used to determine skewness and potential outliers. As the following figure shows:

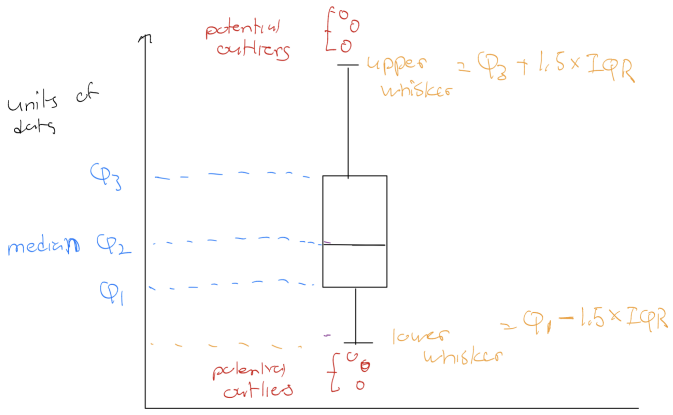

Figure 1.4: Visualization of a box plot

The most important thing from box plot interpretation in this course is obtaining key information. For example: four quartiles, location of potential outliers. Then we approximate those values to continue our statistical analysis.

Outliers:
In statistics, an outlier from a set of data points is the value that significantly differs the other observations from that data set.

Example 1.4.
Consider the following collection of data points:

$$S = \{2, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 200\}.$$

Find the outliers in the collection of data points.

Solution: Simply finding the values that have huge difference with the other data points, such that the outliers of S are: $2, 100$ and 200.

Five number summery:
In statistics, the five number summery are: minimum value, 25^{th} (Q_1), 50^{th} (Q_2), 75^{th} (Q_3) and maximum from a given collection of data points.

Chapter 2

Sampling Distribution

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.

Index