Mr Ong Statistics Textbook

統計學课本

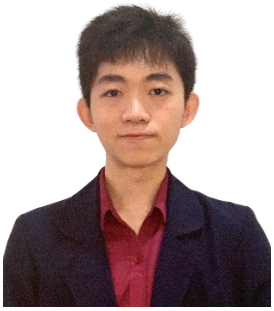
2022 Edition

Last update on August 1, 2022

An intuitive guide on data description, combinatorics, probability, probability distribution, and hypothesis testing
(Mastering Statistics in **46 pages**!)

Ong Zhi Zhung

About the Author



Ong Zhi Zhung（王子聪）

- Member of MENSA Malaysia, **IQ 156**
- **Champion** of National Statistics Competition 2021
- **Silver Medalist** of International Youth Mathematics Challenge 2021
- **Gold Medalist** of International Mathematical Olympiad National Selection Test Malaysia (Full Marks) 2020
- Three-Time **Gold Medalist** of Kangaroo Mathematics Competition 2017, 2019, 2020
- Graduate of **ASASIpintar Program**, Pusat GENIUS@Pintar Negara
- Owner of **"Mr Ong Talk 子聪学长说" YouTube Channel**

Contents

Preface

The initial motive of writing this textbook is to prepare students to sit for the National Statistics Competition (NSC). However, it is important to note that having the knowledge of understanding and using statistics and statistical procedures has now become important skills in almost every profession and academic discipline – an obvious example, you will be doing research projects in your final year of bachelor's degree, as well as master's and PhD.

There are five main topics covered in the textbook, namely data description, combinatorics, probability, probability distribution, and hypothesis testing. Intuitive and daily-life examples are used in this book to give the reader a better and easier understanding of the use of statistics in real-life. Mastering the basic statistical concepts and techniques allow you to "analyse the present, predict the future" prompting you to apply your statistical knowledge in real-life situation.

This book comes together with "Mr Ong Statistics Video Course", where video-clip lectures have been pre-recorded by Mr Ong, explaining all the contents of the statistics course, with the reference of the book. It is highly recommended to watch the video lectures and read this book at the same time to maximise your effectiveness of learning.

Before you start reading this book, PLEASE WATCH THIS VIDEO!!!

Click here to watch!

Chapter 1 Data Description

1.1 Measures of Central Tendency

Central tendency（集中趋势）: Mean, Mode, Median

Example 1:

The Malaysia team, with a group of five, has participated in the first round of the "Super Brain" Challenge. The team that scores 75 points or above, on average, will promote to the 2nd round. Table shows the score of each contestant of team Malaysia. Does Malaysia eventually get into the 2nd round?

Contestant	A	B	C	D	E
Score	72	73	76	76	78

1.1.1 Mean

Mean（均值） = Average = Arithmetic Average
Definition: The mean of a data set is the sum of the observations divided by the number of observations.

Formula:

$$\bar{X} = \frac{\sum x_i}{n}$$

Example:
Find the mean score of Malaysia team.

Contestant	A	B	C	D	E
Score	72	73	76	76	78

Solution:

1.1.2 Median

Median （中位数）= Middle Value
Definition: The median of a data set is the number that divides the bottom 50% of the data from the top 50%.

How to find median?

Arrange the data in increasing order.

- If the number of observations is odd, then the median is the observation exactly in the middle of the ordered list.
- If the number of observations is even, then the median is the mean of the two middle observations in the ordered list.

In both cases, if we let n denote the number of observations, then the median is at position **(n+1)/2** in the ordered list.

Example:
Find the median of the data.

Contestant	A	B	C	D	E
Score	72	73	76	76	78

Solution:

1.1.3 Mode

Definition: The mode （众数） of a data set is its most frequently occurring value.

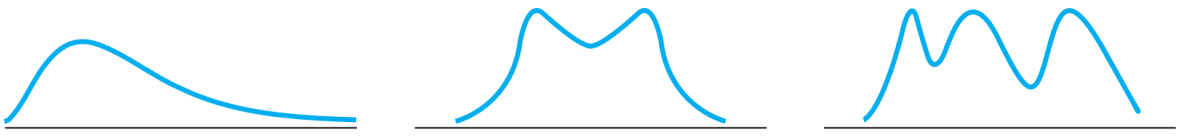
Example:
Find the median of the data.

Contestant	A	B	C	D	E
Score	72	73	76	76	78

Solution:

Modality

A distribution is unimodal if it has one peak, bimodal if it has two peaks, and multimodal if it has three or more peaks.



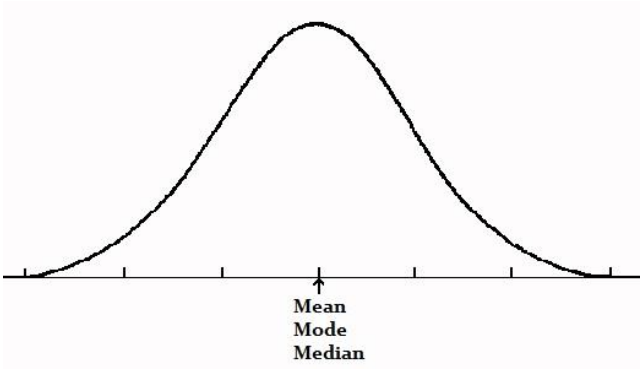
(a) Unimodal (b) Bimodal (c) Multimodal

Example:

Unimodal: 72, **73, 73, 73, 73**, 74, 76, 76, 78, 78 (1 mode)
Bimodal: 72, **73, 73, 73**, 74, **76, 76, 76**, 78, 78 (2 modes)
Multimodal: 72, **73, 73, 73**, **76, 76, 76**, **78, 78, 78** (3 or more modes)

1.1.4 Comparison of the Mean, Median, and Mode

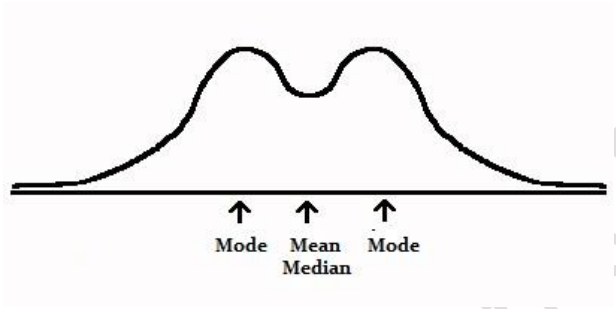
Mean, Mode and Median in a **Symmetric Distribution** （对称分布）



Requirement of Symmetric Distribution:

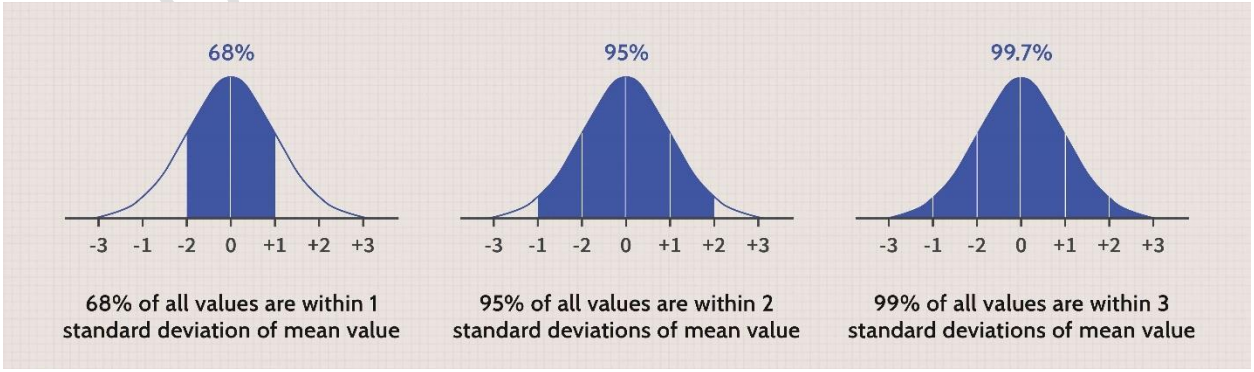
Mean, Median, Mode are equal

An exception is the bimodal distribution. The mean and median are still in the center, but there are two modes: one on each peak.



Normal Distribution （正态分布）

- One example of the symmetric distribution
- Symmetric bell shape
- Mean and median are equal; both located at the center of the distribution
  ≈68% of the data falls within 1 standard deviation of the mean
  ≈95% of the data falls within 2 standard deviations of the mean
  ≈99.7% of the data falls within 3 standard deviations of the mean



Selecting an Appropriate Measure of Center

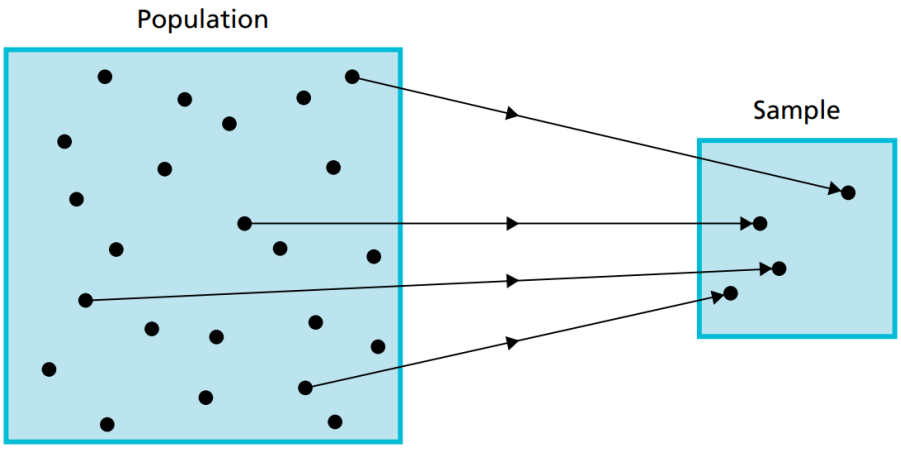
- Mode is not usually selected.
  The mean is sensitive to extreme (very large or very small) observations, whereas the median is not.

Consequently, when the choice for the measure of center is between the mean and the median, the median is usually preferred for data sets that have extreme observations.

1.1.5 Centers of the Population and Sample

Population and Sample Data

Population data: The values of a variable for the entire population.
Sample data: The values of a variable for a sample of the population.



Example:

Population （总体）	**Sample** （样本）
All countries of the world	Countries with published data available on birth rates since 2000
Undergraduate students in Malaysia	300 undergraduate students from UM, UKM, USM who volunteer for your psychology research study

Collecting data from a sample

Samples are easier to collect data from because they are practical, cost-effective, convenient, and manageable. With statistical analysis, you can use sample data to make estimates or test hypotheses about population data.

Formula:

Sample Mean

$$\bar{x} = \frac{\sum x_i}{n}$$

Population Mean

$$\mu = \frac{\sum x_i}{N}$$

* Only differ in symbol!

1.2 Measures of Variation

Measures of variation （差异量数）= measures of spread = measures of difference

Example:
Two data sets have the same mean, median, mode.

Data 1: Team Malaysia in Super Brain Challenge

Contestant	A	B	C	D	E
Score	72	73	76	76	78

Data 2: Team Singapore in Super Brain Challenge

Contestant	A	B	C	D	E
Score	67	72	76	76	84

1.2.1 Range

The range （极差） of a data set is given by the formula

$$\text{Range} = \text{Max} - \text{Min}$$

where Max and Min denote the maximum and minimum observations, respectively.

Example:
What is the range of Data 1? How about Data 2?

Solution:

1.2.2 Variance

Formula:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Why is the formula so?

Example 1: Team Malaysia in Super Brain Challenge

Contestant	A	B	C	D	E
Score	72	73	76	76	78

Solution:

First, find the mean.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{72 + 73 + 76 + 76 + 78}{5} = 75$$

Deviation from mean:						Sum, Σ
Score, x	72	73	76	76	78	
Deviation from mean, $x - \bar{x}$						

What is the total deviation? $\Sigma(x - \bar{x})$

Square the deviations! → 负负得正, sum no longer be zero!

Remember, we get this "measure of variation" by squaring the values! Is that what we want?

1.2.3 Standard Deviation

Formula:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Solution:

Recall from Example 1 (Team Malaysia),

$s^2 = 6$

$s = \sqrt{s^2} =$

So, why we need to know the standard deviation of a data set?

→ to measure how wide a set of values spread out.

A low standard deviation indicates that the values tend to be close to the mean of the set.
A high standard deviation indicates that the values are spread out over a wider range.

Example 2: Team Singapore in Super Brain Challenge

Contestant	A	B	C	D	E
Score	67	72	76	76	84

Find the variance （方差） and standard deviation （标准差） of the data.

Solution:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{67 + 72 + 76 + 76 + 84}{5} = 75$$

Deviation from mean:						Sum, Σ
Score, x	67	72	76	76	84	
Deviation from mean, $x - \bar{x}$						

Computing Formula for Sample Standard Deviation:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}$$

1.2.4 Variations of the Population and Sample

Formula for Population Standard Deviation:

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

Computing Formula:

$$\sigma = \sqrt{\frac{\sum x_i^2}{N} - \mu^2}$$

Population Variance = σ^2

Now, let's reveal the answer of "**Why using (n-1) as the denominator in sample variance or sample standard deviation**"!

To put it simply (n−1) is a smaller number than (n). When you divide by a smaller number you get a larger number. Therefore, when you divide by (n−1) the sample variance will work out to be a larger number.

Let's think about what a larger vs smaller sample variance means. If the sample variance is larger, than there is a greater chance that it captures the true population variance. That is why when you divide by (n−1) we call that an unbiased sample estimate. Whereas dividing by (n) is called a biased sample estimate.

Because we are trying to reveal information about a population by calculating the variance from a sample set, we probably do not want to underestimate the variance. Basically, by just dividing by (n) we are underestimating the true population variance, that is why it is called a biased estimate.