Module 4
Data Summary of
One Numerical Variable

Module Learning Outcomes

- Calculate manually all numerical statistics and differentiate their roles.
- Use Excel to calculate all numerical statistics.
- Provide a description of key parameters and key statistics.
- Determine if any observed values are outliers.
- Find the five-number summary and draw boxplot manually.
- Use Excel to make boxplot and histogram.
- Choose the most appropriate graph to summarize one numerical variable and provide a description of the graph.
- Perform Data Summary for one variable by using only one graph and one statistic (or one set of statistics).
- Calculate weighted average either manually or using Excel.

Summarizing a Single Numerical Variable Using Statistics

- There are three measures of statistics for one numerical variable: (1) measures of the **centre**, and (2) measures of the **location**, and (3) the measures of the **variability** (or spread, dispersion, variation).

4.1 Measures of the Centre

- The two key statistics measuring the centre are the mean (or average) and median.
- The **sample mean** or **sample average** (denoted as $\bar{X}$; pronounced as "X bar") can be calculated by dividing the sum of all observed values (or responses, in a more general term) by the sample size (n), i.e.

$$\bar{X} = \frac{1}{n}\sum X$$

- The **sample median** is defined as the "middle-most" value of an ordered list (of the observed values). An ordered list is simply a list of all observed values arranged in ascending order (from smallest to largest).
- When n is too big, we could first find the location of the median (i) using $i = \frac{n+1}{2}$ and the median can be found at location i.
- For example, the "middle-most" value of the ordered list (1, 3, 4, 7, 8) is at location 3 $\left(i = \frac{5+1}{2} = 3\right)$. Hence, the median has a value of "4".
- In a case where the location index is not a whole number, we would use the midpoint (or average) of the two adjacent (or nearest) values on the ordered list.
- For example, the "middle-most" value of the ordered list (1, 3, 4, 6, 7, 8) is at location 3.5 $\left(i = \frac{6+1}{2} = 3.5\right)$. Hence, the median is 5 (the midpoint between the 3rd and 4th positions).
- In this course, **mode** is not classified under the measure of the centre, as it does not really measure the "centre". For example, 14 students got A, 12 students got B and 4 students got C in a course. The mode is A, which is not the "centre".
- MLO: Calculate manually all numerical statistics and differentiate their roles.
- MLO: Use Excel to calculate all numerical statistics.

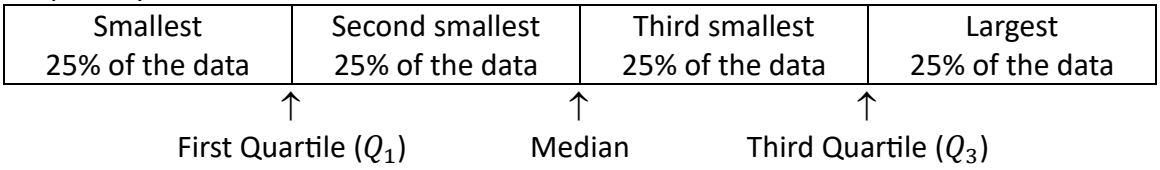
4.2 Measures of the Location

- Besides the centre, the second measure is about **locations**, and there are basically two pairs to consider: 1) the minimum and the maximum, and 2) the first quartile and the third quartile.
- The first pair is straight forward. The **minimum** represents the smallest value in the data set (or the left-hand most value in the ordered list) and the **maximum** represents the largest value (or the right-hand most value in the ordered list).

- **Percentiles** are rarely used but it does pop up from time to time.
- Imagine that the ordered list is divided into 100 (from *cent*) equal parts.

- There will be 99 percentiles (1st percentile, 2nd percentile, …, 99th percentile) because the first and the last ones are called the minimum and maximum, respectively.
- Note: Please ask Michael to draw the diagram about percentiles to better visualize it.

- Similarly, to understand **quartiles**, imagine that an ordered list is divided into 4 (from *quar-*) equal parts.
- How many quartiles should we have here?
- Note that the second quartile has the same definition as median. Hence, median is used over the second quartile.
- Therefore, we only have two "real" quartiles in practice – the **first quartile** (Q_1) and the **third quartile** (Q_3).
- To find the quartiles, all observed values are arranged in ascending order (i.e. ordered list).
- The **first quartile** is the "median" of all the observed values strictly to the left of the overall median.
- The **third quartile** is the "median" of all the observed values strictly to the right of the overall median.
- Graphically, their use can be illustrated here.



- Note: There are different formula being used in different textbook and statistical software application (like MS Excel), and their values may be slightly different. I guess the point here is not to be too concerned about the "exact" value (because it's based on how things are defined), but rather focus on the bigger picture in understanding the concept and how to apply it when needed.
- MLO: Calculate manually all numerical statistics and differentiate their roles.
- MLO: Use Excel to calculate all numerical statistics.

4.3 Measures of the Variability

- The final measure is **variability** (or sometimes called **spread**) and it has three distinct values: 1) range, 2) interquartile range, and 3) standard deviation.
- The **range**, in statistics, is defined as the difference between the maximum and the minimum.
- Note that it always has a non-negative (zero or positive) value. Do you know why?

- The **interquartile range** (or IQR) is defined as the difference between the third quartile and the first quartile, i.e. $IQR = Q_3 - Q_1$.
- Again, it always carries a non-negative value. Do you know why?

- Loosely speaking, **standard deviation** (or simply denoted as small letter s) is defined as average of the n **deviations**, where deviation is the difference between an observed value and the sample mean.
- It tells you about <u>what the typical difference each observed value is from the sample mean</u>.
- By and large, this is *the* measure of variability you absolutely need to master to do well in EDA or statistics in general.
- A very close "cousin" of standard deviation is the **variance**. As its name implies, it measures the variations of all observed values.
- The standard deviation is simply the "square root" of variance. Here are their formulas.

Variance	Standard Deviation
$s^2 = \frac{1}{n-1}\sum(X-\bar{X})^2$	$s = \sqrt{\frac{1}{n-1}\sum(X-\bar{X})^2}$

- To calculate standard deviation, we take the approach "from inside to outside":
  0) It is preferred, but not necessary, to put all observed values in an ordered list.
  1) Find the sample mean ($\bar{X}$).
  2) Take the difference between each value (X) and the sample mean ($\bar{X}$).
  3) Square the differences from (2).
  4) Sum all square values from (3).
  5) Divide the sum from (4) by $n-1$.
  6) Take the square root ($\sqrt{\ }$) of (5).
- Note: Being able to calculate its value is nice, but don't kill yourself by spending too much time on this. I personally would recommend students understand its use more than the calculation.
- There are a couple of facts related to standard deviation that you need to know.
  1) The sum of difference (between observed values and mean) is always zero, i.e. $\sum(X-\bar{X}) = 0$.
  2) Standard deviation (like any other statistics) has the same the units as the variable of interest.
- MLO: Calculate manually all numerical statistics and differentiate their roles.
- MLO: Use Excel to calculate all numerical statistics.

4.4 Parameters vs. Statistics

- A **parameter** is defined as a number that describes a variable (or multiple variables later) of the **population**.
- Likewise, a **statistic** is defined as a number that describes (or summarizes) a variable (or multiple variables later) of the **sample**.
- A very simple mnemonic (or an easy way to memorize things) is P vs. S (**P**arameter from **P**opulation and **S**tatistic from **S**ample).
- Values of parameters are always <u>unknown</u> to us, but we should be able to describe it in words.

- The description of parameter has three main parts: (1) a single value, (2) of a variable, (3) from the population.
- Values of statistics vary from sample to sample and we can describe it in words too.
- The description of statistic also has three main parts: (1) a single value, (2) of the variable, (3) from the sample.
- One basic use of (the discipline of) statistics is to use statistics to estimate unknown parameters.
- MLO: Provide a description of key parameters and key statistics.

4.5 Detection of Outliers

- One of the application of quartiles or inter-quartile range is to determine if any observed values are outliers in the data set or not.
- An **outlier** is loosely defined as an observed value that is significantly different from the rest of the values.
- How "significant"? We have a rule for that.
- An observed value is defined as an **outlier** if it is either strictly bigger than the upper limit (UL) or strictly smaller than the lower limit (LL), where LL and UL are defined as:

$$LL = Q_1 - 1.5 \times IQR \text{ and } UL = Q_3 + 1.5 \times IQR$$

- Do ask Michael to draw the diagram of this rule for your understanding.
- MLO: Determine if any observed values are outliers.

4.6 Five-Number Summary

- The **five-number summary** refers to a collection of five important statistics.

Minimum Q_1 Median Q_3 Maximum

- It has two main purposes: one is to make boxplot, and another is to summarize a data set (more on this in Data Summary subsection later).
- MLO: Find the five-number summary.

4.7 Summarizing a Single Numerical Variable Using Graphs

- There are two "numerical graphs": 1) histogram, and 2) boxplot.
- **Box and whisker plot** or simply **boxplot** is a rare breed with both statistics and graphs in one.
- To draw a horizontal boxplot:
  1) Draw the x-axis, label it with variable of interest, and put down a proper scale of the axis.
  2) Draw two short vertical lines for Q1 and Q3 at their corresponding points, above the x-axis.
  3) Connect the two lines to form a box.
  4) Draw another vertical line for median inside the box in (3).

5) Draw the right-hand whisker from Q3 to the maximum.
6) Draw the left-hand whisker from Q1 to the minimum.

- Note: Excel only has vertical boxplot. But when you are asked to draw one in a test, please make sure you do horizontal boxplot. Do you know why?
- MLO: Draw a boxplot manually.
- MLO: Use Excel to make boxplot.

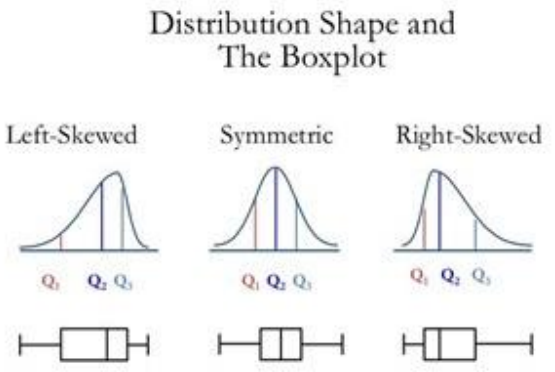
- Not many of us would have seen a boxplot before. In contrast, **histogram** is the most commonly used numerical graphs.
- In this course, we will only focus on histograms with *vertical bars*. In other words, the variable of interest is on the x-axis and frequency or percentage on the y-axis.
- However, drawing a histogram is not the focus here. In fact, students in my class will never be asked to histograms manually in the course, but make sure you are able to use Excel to produce one.
- Besides, students are expected to know about the description of the graphs or share with others what story the graph tells us.
- Histograms are great to show the shape (and everything we are looking for in numerical graphs). However, the shape seems "distorted" when the sample size is small.
- *As a rule of thumb: Histograms are typically the first choice when summarizing one numerical data because it shows the shape readily. But when sample size is less than 30, boxplots are used over histograms. Also, boxplots are preferred when we want to compare data from multiple samples.*
- MLO: Use Excel to make histogram.

4.8 Description of Numerical Graphs

- There are four key areas when describing numerical graphs: 1) **Centre**, 2) **Spread**, 3) **Shape** (will be discussed in the following subsection), and 4) **Outliers**.
- It is rather common that subjectivity is involved when discussing centre, spread and outliers. As long as the key meaning of them is captured, you will be fine in general.
- In other words, when providing the description, we do so by only looking at the graphs (not peeking at the statistics).
- MLO: Choose the most appropriate graph to summarize one numerical variable and provide a description of the graph.

4.9 Shape of Numerical Graphs

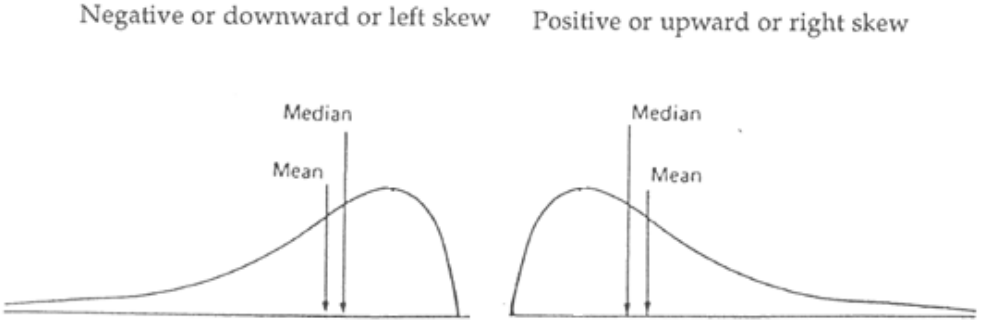
- In this course, we have four shapes to know: 1) **Normal shape** (or bell shape), 2) **symmetric** (but not necessarily Normal), 3) **right-skewed** (or **skewed to the right** or **positively skewed**), and 4) **left-skewed** (or **skewed to the left** or **negatively skewed**).
- Here is an image from the internet that shows the shape as well as how they look like in boxplots.



- MLO: Choose the most appropriate graph to summarize one numerical variable and provide a description of the graph.

Relative Location of Mean and Median

- For data that have symmetrical shape, the values of mean and median are similar (or the same).
- With skewed data, however, the mean is always being pulled to the side of the outliers. Do you know why?
- More specifically, mean would be larger than median in right-skewed data, whereas mean would be smaller than median in left-skewed data.



- Therefore, median will be used for skewed data to measure the centre.

4.10 Measure Sensitivity

- Note that we typically have 2-3 statistics under the same measure.
- The million-dollar question is which statistic shall we use. It comes down to what data do we have.
- Some statistics are drastically affected by the existence of outliers. We typically say that the statistics are "sensitive to outliers".
- In contrast, some statistics are not affected by outliers so much (and some of them are quite robust too). We typically say that those statistics are "not sensitive to outliers".

- The two lists are shown below.

	Non-Sensitive Measures	Sensitive Measures
Measures of Centre	Median	Mean
Measures of Location	First Quartile, Third Quartile	Minimum, Maximum
Measures of Variability	Interquartile Range	Range, Standard Deviation

- Statistics that are sensitive to outliers are not generally used to summarize such data set (with outliers and/or shows skewness).
- MLO: Perform Data Summary for one variable by using only one graph and one statistic (or one set of statistics).

4.11 Data Summary for One Numerical Variable

- **Graph**: Choose between histogram and boxplot.
- Criteria: Mainly based on sample size
- **Statistic(s)**: Choose between the "mean & standard deviation" combo and the *five-number summary*.
- Criteria: Use both the graph and outlier method to determine the skewness or presence of outliers in the data set. If the data set is skewed and/or contains outliers, use the non-sensitive measures in the *five-number summary*. Otherwise, it is safe to use the sensitive measures in the mean and SD combo.
- MLO: Perform Data Summary for one variable by using only one graph and one statistic (or one set of statistics).

4.12 Weighted Average Calculation

- Although this is not a formal part of the course, it is very useful to be able to calculate the **weighted average** and understand its applications.
- It is particularly important for you to calculate what score (in percentage) you need in the final exam in order to get certain grades.
- There are a few practice questions in the Problem Sets.
- So, hopefully you don't need to ask me "what percentage do I need to get in the final to pass the course" because you can do it yourself now! 😉
- MLO: Calculate weighted average either manually or using Excel.