# Data Management

**MDM4U**

Qinghao Hu

October 27, 2025

# Contents

# Chapter 3

# One Variable Statistics

## 3.1 Variables and Data

### 3.1.1 Definitions

**Categorical** variables represent data that are generally grouped into categories, and are also known as qualitative variables

**Ordinary** variables are categorical variables whose data has a natural order but the difference between values cannot be determined or is not meaningful

**Nominal** variables describe names, labels, or categories that have no natural order

**Quantitative** variables describe data values that are numerical, and are also known as numerical variables

**Continuous** variables are numerical variables which can assume an infinite number of values in a given interval

**Descrete** variables are numerical variables that only take on a finite number of possible values in a given interval

**Primary** data are data that are collected by the statisticians who are analyzing the data, from first-hand sources such as surveys or experiments

**Secondary** data are data that the statisticians who are analyzing the data did not participate in the first hand data collection process (ie Surveys or experiments)

**Microdata** contains records for each individual surveyed

**Aggregate** or summary data are data that are combined or summarized in such a way that the individual microdata can no longer be determined.

Data gathered from a **cross** sectional study considers individuals from different groups at the same time

Data gathered from a **longitudinal** study considers how the characteristics of a specific sample changes over time

An **index** is a continous variable such that it is an arbitrarity defined number that provides a measure of scale. It is used to related the values of a variable to a base level

The **consumer** price index, CPI, provides a broad picture of the cost of living in Canada by comparing the cost of a wide variety of consumer goods, such as food, clothing, fuel, heating cost, transportation, shelter, and recreation

Health officials use the **body** mass index to determine whether a person is overweight. The BMI is calculated by dividing a person's mass in kilograms by the square of their height in meters

## 3.2 One Variable Graphs

### 3.2.1 Some Definitions

A **Frequency** bar (column) graph is a visual display of data in which quantities are represented by bars of equal width, typically used with categorical or discrete data

A **CIRCLE** graph or **PIE** chart contains a circle divided into sectors whose areas are proportional to the categories represented. It is used to show how each category is compared to the whole

A **PICTOGRAPH** is a graph that uses pictures or symbols to represent categorical quantities. It's advantage is being visually appealing, hence it is the most often used graphical format. However, it may be difficult to present exact values when using the format, depending on the data given

A **STEN** and **LEAF** plot can be created easily to see the distribution of a set of numerical data. However, its appearance is not as scientific as a histogram.

A **HISTOGRAM** is used to represent numerical data or data organized using intervals. The bars of a histogram are attached and each bar is placed between two intervals endpoints. The area of each bar is proportional to the frequency of data in the interval. Typically, 5  15 intervals/bins of equal length are used and every piece of data must fall into exactly one bin. The width of each bin is the **bin width**

Values of a continuous varialbe can be grouped into intervals in the form of $(a, b]$ such that this interval includes all values from $a$ to $b$, including $a$ but excluding $b$.

A bin width of **5** units witht he first bin being $[10, 15)$ is reasonable if a set of continuous data has 26 values, a minimum of 12

*Example* 1.

A group of students' heights, in centimetres, are shown. [35, 40)
Draw a stem and leaf graph for the data.
Use the first two digits of the numbers as the stems.

| | | | | | | |
|---|---|---|---|---|---|---|
| 150 | 150 | 154 | 161 | 162 | 163 | 165 |
| 174 | 175 | 175 | 176 | 179 | 180 | 182 |

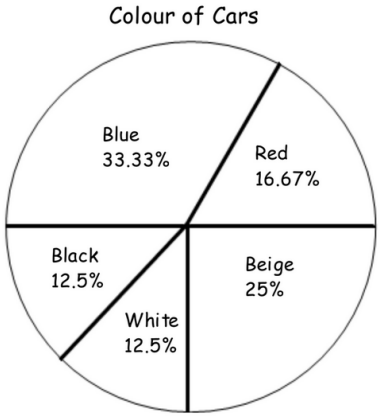
| Stem | Leaf |
|---|---|
| 15 | 0, 0, 4 |
| 16 | 1, 2, 3, 5 |
| 17 | 4, 5, 5, 6, 9 |
| 18 | 0, 2 |

This is an example of **STEM and LEAF graph**

*Example* 2.

3. Consider the given pie chart then complete the table.

| Colour | Frequency | Relative Frequency | Degrees |
|---|---|---|---|
| Beige | 30 | 1/4 | 90° |
| Black | 15 | 1/8 | 45° |
| Blue | 40 | 1/3 | 120° |
| Red | 20 | 1/6 | 60° |
| White | 15 | 1/8 | 45° |
| Total | 120 | 1 | 360° |




**This is an Example of PIE Graphs**

4. In a study of travel time to school, 29 students were surveyed.

   a) Use the grid on the left to create a (frequency) histogram for the data. Label fully. A title is not required.

   b) Relative frequency shows the frequency of each interval with respect to the total number of data values (i.e. the proportion of data contained in each interval is calculated). Complete the given table.

| Travel Time to School (min) | Frequency | Relative Frequency | Cumulative Frequency |
|---|---|---|---|
| [0, 5) | 2 | $\frac{2}{29}$ | 2 |
| [5, 10) | 5 | $\frac{5}{29}$ | 7 |
| [10, 15) | 10 | $\frac{10}{29}$ | 17 |
| [15, 20) | 7 | $\frac{7}{29}$ | 24 |
| [20, 25) | 3 | $\frac{3}{29}$ | 27 |
| [25, 30) | 2 | $\frac{2}{29}$ | 29 |

   c) Use the grid on the right to create a relative frequency histogram for the data. Label fully. A title is not required.

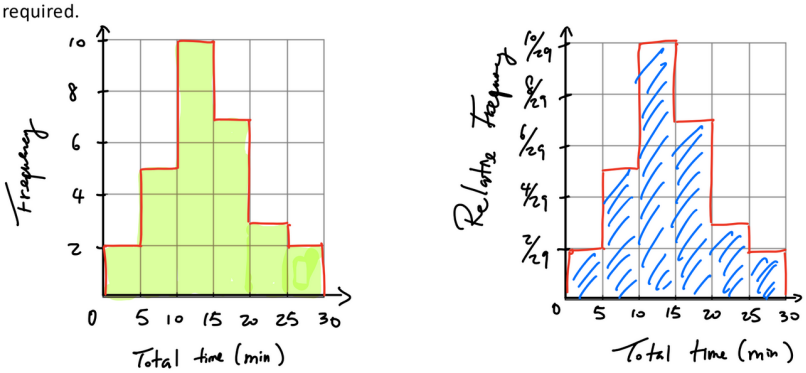



Figure 3.1: Example

Like a histogram, a **frequency polygon** gives an idea of the shape of the data distribution. It helps show the changes in frequency from one interval to the next. If the midpoint of each interval is used as an estimate for the all the values in the interval, then a frequency polygon is a line graph joining the mindpoints of the top of adjacent bars of a histogram.

Check the note

## 3.3 Central Tendency

### 3.3.1 Definitions of Central Tendency

**CENTRAL**

Measure of **CENTRAL** tendency are used to determine the averages of a set of data

**Mean**

The **Mean** of a set of numerical data is equal to the sum of the values of a variable divided by the number of values. That is:

$$\text{population mean} = \mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\text{sample mean} = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

If a set of numerical data is listed from least to greatest, then:

1. the median is the middle number (or the mean of the Two middle numbers), and,

2. the values can be ranked such that the minimum has rank=1 and the maximum has rank = N or n

The **Mode** is the most frequently occurred value in the data

The **MODAL** interval is the interval that contains the msot number of values

**OUTLIERS** are values that are significantly distant from the majority of the data

**How to choose form**

1. If the data is **not** numeric, use the Mode
2. if the data contain outliers and/or mean is skewed, use the median
3. otherwise use the mean

**Weighted**

the Weighted mean reflects the relative importance of each value of the data set. They could be calculated by two formulas:

Population:

$$\mu_w = \frac{w_1x_1 + w_2x_2 + \cdots + w_Nx_N}{w_1 + w_2 + \cdots + w_N} = \frac{\sum_{i=1}^{N} w_ix_i}{\sum_{i=1}^{N} w_i}$$

Sample:

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \cdots + w_Nx_N}{w_1 + w_2 + \cdots + w_N} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}$$

## 3.4 Standard Deviation

**Spread**

Measures of **spread** are frequently calculated when analyzing numerical data. These measures are values that quantify how consistent or spread out of a set of data is. A measure of spread is often used to determine which of serval sets of data is more consistent

The **spread** of a set of data is **zero** if all the values in the set are identical

Just as there are several measures of central tendency, there are **several** different measures of spread. Variance and standard deviation are two of these measures.

The **Deviation** of a piece of data is the difference between the value of that piece of data and the mean of the set

**Deviation of Detum**

Population:

$$x - \mu$$

Sample:

$$x - \bar{x}$$

**Definition 3.4.1.** *Variance: The mean of the squares of the deviations. Larger variance = spread of the data is larger*

Population Variance:

$$\sigma^s = \frac{\sum(x-\mu)^2}{N}$$

Sample variance:

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

Sample variance's denominator being $n-1$ as opposed to $n$ because in a sample the deviations tend to be underestimated.

#### Standard deviation

**Definition 3.4.2.** *Standard Deviation approximates the **TYPICAL** distance from the mean to each datum in the set*

Population standard deviation:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Sample standard deviation:

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

*Remark.* Always calculate the sample variance/standard deviation if it is not clear whether the data is from a census or from a sample

#### Z-score

A datum's z-score is the number of standard deviations that datum is above or below the mean of the data set. Hence:

$$z = \frac{x-\mu}{\sigma}$$
$$z = \frac{x-\bar{x}}{s}$$

## 3.5 Quartiles

### 3.5.1 Definitions

**Definition 3.5.1.** *Interquartiles range is the measures of spread*

#### Range

Range is the difference between the largest and smallest values and it does not give any information about the spread of the other data values in the set

#### Quartiles

Quartiles divide a set of ordered data into 4 equal groups, similar to a median that divides the data into 2 equal groups. The three dividers are the first quartile $Q_1$, the second quartile $Q_2$ and the third quartile $Q_3$

$Q_2$ can be seen as the median of a data set. $Q_1$ and $Q_3$ can be defined as the **Median** of the lower and upper halves of the set, respectively, with an understanding of that, if the ordered set has an odd number of values then the middle number is not part of the lower or upper half of the data set

**InterQuartile**

It is the range of the middle half of the data. It can be calculated by this formula:

$$IQR = Q_3 - Q_1$$

A larger $IQE$ indicates a greater spread of the central half of the data. The **semi-interquartile** range is the $IQE$ divide by 2

**Box-and-whisker plots**

This graph illustrate the spread of the data around the **MEDIAN**. The box shows three quartile values, a left and a right whisker that "lead" to the minimum and the maximum values of the data set

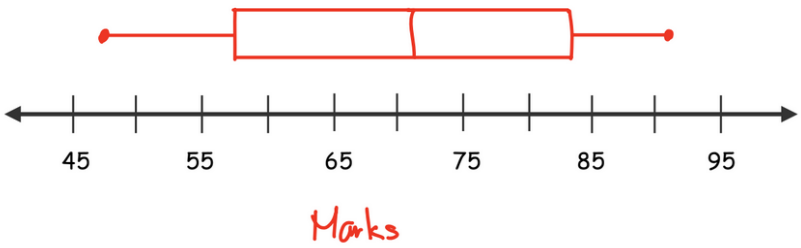



Figure 3.2: An example of box-and-whisker plot

**Outlier**

A **Modified** box plot may be used if the data contains outliers. Any value of $x$ that is at least 1.5 times the box length from the box are considered outliers. These outliers must be plot as separate points instead of including them as part of the whiskers

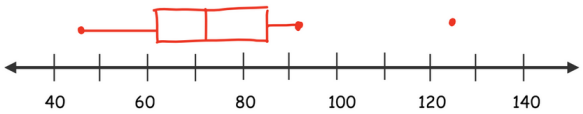



Figure 3.3: In this graph, 125 is an example of outlier

### 3.5.2 Percentiles

**Percentiles** divide a set of data into 100 equal intervals. A common definition for percentiles states that "A value corresponds to the $k^{th}$ percentile if $k$ % of the data are less than or equal to the value"

For a data set of n:

The Rank of the percentile p is:

$$R = \frac{p * n}{100}$$

Else linear interpolate the values with rank $R$ and $R+1$ to determine the value that corresponds to percentile $p$

The percentile that corresponds to a specific datum is:

$$p = \frac{100L}{n}$$

$L$ is the number of data values less than or equal to that specific datum

**Remainder**

$Q_1$ and the $25^{th}$ percentile corresponds to the same location of the data set, two different methods are being used to find the two statistics, for the purpose of convenience. Hence, they may not be equal to each other even though they should.

## 3.6 Spread Grouped Data

### 3.6.1 For weighted data

If $x_i$ represents the $i^{th}$ distinct value and $w_i$ represents the weighting of the value $x_i$ them

Average:

$$M \text{ or } x_i = \frac{\sum(w_i * x_i)}{N} \tag{3.1}$$

Population deviation:

$$\sigma = \sqrt{\frac{\sum w_i * (x_i - M)}{N}} \tag{3.2}$$

For sample deviation:

$$s = \sqrt{\frac{\sum w_i * (x_i - \bar{x})}{n-1}} \tag{3.3}$$

## 3.7 Collect data

### 3.7.1 Anonymous and not be anonymous

**Definition 3.7.1.** *(Anonymous): More likely to get honest data (hence more reliable data)*

**Definition 3.7.2.** *(Not Anonymous): Show credibility of Survey (answers)*

### 3.7.2 Survey Questions

**Open**

Respondents answer in their own words

**Closed**

**Definition 3.7.3.** *(checklist): Choose as many as apply (With a rating scale should be evenly distributed)*

**Definition 3.7.4.** *(Ranking): Give a rank based on the choices*

### 3.7.3 Questions should be avoided

**Definition 3.7.5.** *(Double-Barrelled): A question that akss more than one topic. (Ex. Do you like Math and Science?)*

**Definition 3.7.6.** *(Leading Question): Encourages a particular answer, onften becauses of this question is phrased or presented*

**Definition 3.7.7.** *(Loaded Question): A question contains assumptions, where answering it means the respondent accepts what the questioner is assuming*

### 3.7.4 Experimental vs Observational study

**Observational study**

- Study about how one factor affect another factor, without make any attempt to intervene

**Experimental study**

- Try to determine the cause and effect relationship between two variables by changing the value or characteristics of one variable to see what effect it has on the other variable
- Randomly place participants into the experimental group and the control group, with each group having a similar demographic make-up
- One experimental(Treatment) group of an experimental study receive the specific Treatment
- The other one do not receive the specific treatment being measured

## 3.8 sampling

**Definition 3.8.1.** *(Population): referss to the entire group that is being studied. Also called descriptive statistics*

**Definition 3.8.2.** *(Sample): Is a portion of the population. Summary values calculated from the sample data are called statistics. The term inferential statistics describes the process of generalizing about the population based on sample data. Therefore, it is important to have a "representative sample" when performing a statistical study*

### 3.8.1 Some type of sampling

**Definition 3.8.3.** *(Simple Random Sampling): every member of the population has an equal chance of being chosen for the study.*

**Definition 3.8.4.** *(SYSTEMIC Random Sampling): Individuals are selected at regular intervals, starting with a randomly chosen position.*

**Definition 3.8.5.** *(Stratified Random Sampling): Population is divied into groups/strata such that all members in each stratum share common characteristics but are different from members in other strata, then the same proportion of members from each stratum are randomly chosen*

**Definition 3.8.6.** *(Cluster random sampling): The population is divied into groups/clusters such that each cluster is a representative of the whole population, then every member of a random sample cluster are are surveyed*

**Definition 3.8.7.** *(Multi-stage random sampling): more than one loevel of random sampling techniques are applied*

**Definition 3.8.8.** *(Voluntary-response random sampling): members of the population are invited to participate in the survey and anyone who choose to participate is the sample*

**Definition 3.8.9.** *(Voluntary-response random sampling): members of the population are invited to participate in the survey and anyone who choose to participate is the sample*

**Definition 3.8.10.** *(Convenience random sampling): the sample is selected because it is easily accesible.*

## 3.9 Bias

**Definition 3.9.1.** *(Sampling bias): When the sample is not representative of the population*

**Definition 3.9.2.** *(Measurement bias): the data measuring tools are poorly designed*

**Definition 3.9.3.** *(Response bias): when participants in a survey deliberately give false or misleading answers*

**Definition 3.9.4.** *(Non-response bias): a form of sampling bias, particular groups are under-represented because they choose not to participate*