Descriptive Statistics

STAT 1F92 ❈ Component 01 ❈ September 06 – September 19, 2023[1]

Table of Content

❈ September 06, 2023 ❈

1 Descriptive Statistics

The objective is to cover what the idea of statistics is, organizing data and differentiating between qualitative and quantitative variables. We will also go over organizing quantitative data as well as finding the mean, standard deviation, outliers and boxplots.

1.1 Introduction to Statistics

To objective is to go over a broad definition of what statistics is, understand the difference between statistic and parameter, qualitative versus quantitative variables as well as discrete and continuous variables. Let us give a rigorous[2] definition of what statistics is.

[1] Complied at: Friday October 6th, 2023 1:21pm -04:00

[2] Rigorous means ‘specifically well-defined’, a word that will be used regularly

Definition 1.1: The study of statistics

Statistics is the study of collecting, organizing, summarizing, and analysing information which yield to some conclusions that is associated with a degree of confidence.

Statistics is the approach to mathematically answer a question with certain degree of confidence. For instance, how can we find the average height of human beings? How about finding the number of people currently alive? How about finding the weather for the next week? All of these questions are based off statistics and degree of confidence. While it is unfeasible to measure each single person to calculate the total average or count each single person in a country, we can *approximate* the target based on data that we collect. We need to defined our scope of interest (*i.e.*, who are we interested in).

Definition 1.2: Defining the study scope

- The entire group that will be studied is referred to as the *population*
- An *individual* is a person or object that is a member of the population
  - It is what we are 'studying'
  - It consists of characteristics with some values (age, weight, height, colour, etc.)
- A *sample* is a subset of the population

In other words, population is a collection of individuals (people or objects) and a sample is a subset of the population. Note that the characteristics don't have to be mathematical/numeric. For instance, the notion of 'colour' isn't defined to perform mathematical operations on. What does red + blue or magenta ÷ green give? Nonsense. It doesn't generate meaningful result.

Definition 1.3: Statistic vs. parameter

A *statistic* is a numerical summary of a sample. A *parameter* is a numerical summary of a population.

To remember which is associated with which, **s**tatistic and **s**ample both start with 's', whereas **p**arameter and **p**opulation both start with 'p'.

Example 1.1: Differentiating between a statistic and parameter

- The statement *75.4% of Brock students have smartphones* describes a parameter as it summaries a numerical value of the population
- The statement *a sample of 100 Brock students concludes that 70% use their smartphones daily* is a statistic because it describes a sample of the population (100 students is a subset of all students)

Data that is gathered will be in the form of numbers, tables and collection of input/entries. For example, there are 250 students in our course which has five assignments, two midterms and two exams. Each student is associated with nine entries, which means the course will have $250 \cdot 9 = 2250$ entries in total. Human beings are not meant to process thousands of values. Instead, we resort to visualization, also referred to as descriptive statistics.

Definition 1.4: Descriptive statistics

The notion of organizing and summarizing data through numerical summaries, tables and graphs.

For example, we can graph the data and visualize the average, minimum and maximum of each entry (so nine graphs are created in total). There are different plots we can use including scatter plots, line graph, bar graph, etc. which will be discussed in this course.

Suppose we want to find the average height of Brock students. Since it is unfeasible to measure the height of *every* student, we decide to take a sample of, say 100 students, measure their height and find the average. With some certainty, we can somewhat conclude that the average of *all* Brock student is close to the average we found from the sample[3]. This is referred to as *inferential statistics.*

Definition 1.5: Inferential statistics

Inferential statistics derive some result(s) from a sample and applies it to the population with an associated certainty.

1.2 Quantitative and Quantitative Variables

In general, there are two types of data variables: qualitative (categorical) variables and quantitative variables.

Definition 1.6: Types of variables

- **Qualitative/categorical:** variables classify an individual based on attribute or characteristic that doesn't allow for mathematical operations (*i.e.*, cannot apply addition, multiplication, etc. on them)
  - Colour is an example (as mentioned above) and gender is another example (what does male × female give? Undefined.)
  - Another example is postal code. What does L2S 3A1 ÷ 7 give? How about L2S 3A1 + L2S 3A1? Again, undefined
- **Quantitative:** variables that represent numerical values of individuals (quantitative is derived from quantity, which means amount)
  - The values allow for applying mathematical operations on them
  - For example, age, weight, height, temperature, time are all valid examples

The values of a qualitative/categorical variable are fixed in the sense they are what they are. Quantitative variables on the other hand can be divided into two categories: discrete variable and continuous variable.

[3] Realistically, we need to mathematically define what 'close' means or what is the degree of certainty (5%, 50%, 99.9%, etc.?)

Definition 1.7: Discrete vs. continuous variable

- **Discrete variable:** a quantitative variable that has finite number of possible values such as $0, 1, 2, \ldots, n$, where n is some integer (*i.e.*, whole number). There must be 'gaps' between the values
  - Examples are persons, cars, soccer balls, etc.
  - What does $\frac{1}{2}$ of a person or car represent? Undefined, again
- **Continuous variable:** a quantitative variable that has an infinite number of possible values and gaps don't exist. Examples include:
  - Time. We can have 2.563 seconds or 0.3841 of an hour or 9 days, etc.
  - Distance. We can have 34.942201 metres or 50.239 inches or 70 kilometres, etc.

❄ September 12, 2023 ❄

Example 1.2: Differentiating between a discrete and continuous variable

- Scenarios
  - The number of smartphones you have owned
  - The distance you can run within 10 seconds
  - The number of students arriving to Brock from 8:00AM to 10:00AM
  - The time it will take you to walk 500m
- Solutions
  - The number of smartphones needs to be a discrete (whole number) because a phone is a single unit, it doesn't make sense to have 0.5 of a phone
  - The distance you can run within 10 seconds is continuous because you can be as precise as possible. Maybe you run 90m or 90.1m or 90.001m, etc.
  - The number of students arriving is, surprisingly, discrete because we are counting the number of students. It happened that they were recorded based off a timeslot, but the number students is an integer
  - The time it will take you to walk 500m is continuous as you could finish the task in 40 seconds, 40.53 seconds, 40.3451 seconds, etc.

Definition 1.8: What is data

Each variable is associated with a value, that value is referred to as data.

We can have a variable that is qualitative or quantitative (discrete or continuous) and the value we assign to that variable is referred to as data. For example, consider having an individual as a car, which contains a qualitative variable colour. That colour variable could be red, green, blue, magenta, orange, etc. These colour are referred to as data, which means:

- Qualitative data are values corresponding to a qualitative variable (*e.g.*, colour)
- Discrete data are values corresponding to a discrete variable (*e.g.*, people, whole numbers)
- Continuous data are values corresponding to a continuous variable (*e.g.*, time, whole and

decimal numbers)

Example 1.3: Identifying data and types of variables

Given the following table, identify and characterize the types of variables (qualitative or quantitative discrete/continuous) and data:

Table 1.1: General summary of some countries

Country	Continent	Life Expectancy (in years)	Population
Australia	Oceania	83.28	26,461,166
Colombia	South America	76.10	218,689,757
Canada	North America	83.99	38,516,736
Ireland	Europe	81.87	5,323,991
Kuwait	Asia	79.35	3,103,580
Moldova	Europe	72.72	3,250,532
Philippines	Asia	70.48	116,434,200
Seychelles	Africa	76.36	97,617
United States	North America	80.75	339,665,118

Source: CIA The World Factbook (2023)

The individuals are the countries we are studying as they have some characteristics. These countries are Australia, Colombia, Canada, Ireland, Kuwait, Moldova, Philippines, Seychelles and United States. The variables are the characteristics, continent (qualitative/categorical), life expectancy (quantitative continuous) and population (quantitative discrete). The data are the values associated with the variables. For example, the values Oceania, Europe, Asia, 83.99, 76.36, 3,103,580, 97,617, etc. are all data.

There are times where we need to understand the data correctly before concluding whether the data is discrete or continuous. For example, suppose the population data is given as the column in Table 1.2. Based on the data in the column, it seems that we are dealing with decimal data, hence, we might want to conclude that the population variable is a continuous variable. Of course, this *isn't* the case as the column mentions the data to be in millions (*i.e.*, multiple each data value by one million to get the expanded data).

1.3 Level of Measurement of a Variable

While we have variables that could either be qualitative, quantitative (discrete or continuous), we can go a level further and assign a level of measurement to each variable.

Population in Millions
26.461166
218.689757
38.516736
5.323991
3.10358
3.250532
116.4342
0.097617

Table 1.2: The population column in millions

Definition 1.9: Level of measurement

- **Nominal:** the values of the variable name, label, or category where order *doesn't* matter and cannot perform mathematical operations
  - The variable 'colour' is an example where the values don't have an order (blue is not better than magenta and red isn't worse than green)
  - We cannot perform mathematical operations for reasons mentioned earlier
- **Ordinal:** same as nominal properties but order matters
  - The variable 'letter grade' (A, B, C, D and F) is an example where the values are labels but do have an order
  - Grade A is higher than B and D is lower than C
  - Again, we cannot perform mathematical operations, what does A − D give?
- **Interval:** same as ordinal properties but adding/subtracting values have meaning and a value of zero doesn't mean the absence of the quantity
  - The variable 'heat', which is measured by temperature (in Celsius to Fahrenheit), is an example where the values can be added (*e.g.*, $4° + 5° = 9°$) or subtracted (*e.g.*, $12° - 20° = -8°$)
  - A temperature of 0° doesn't mean the absence of heat as negative degrees means less heat
- **Ratio:** same as interval properties but also multiplying/dividing values have meaning, the value of zero means the absence of the quantity and ratios of the values of the variable have meaning
  - The variable 'height' is an example where the values can be multiplied (*e.g.*, $200\text{cm} \cdot 4\text{cm} = 800\text{cm}$) or divided (*e.g.*, $9\text{cm} \div 5\text{cm} = 1.8\text{cm}$) and a height of 0cm means no height available (absence of height)
  - We can also have a building A to be three times taller than building B (*e.g.*, suppose building B is 20 metres tall, then building A is 60 metres tall)

	Nominal	Ordinal	Interval	Ratio
Categorizes variables	✓	✓	✓	✓
Order matters		✓	✓	✓
Adding/subtracting			✓	✓
Zero isn't absence of quantity			✓	
Multiplying/dividing				✓
Zero is absence of quantity				✓
Ratios defined				✓

1.4 Simple Random Sampling

Suppose we want to conduct a study where people are involved. How can we select our individuals? An approach is to select the individuals randomly.

Definition 1.10: Random sampling

The notion of selecting a sample from the population by relying on chance or luck

In other words, each individual in the population has equal probability of getting selected as a part of the sample group. The objective is to gather a sample that represent the population through chance/luck rather than simplicity[4] or favouring individuals. In case the process wasn't completely random, then the conclusion will contain biases.

Definition 1.11: Simple random sample

When the probability of any sample of size n obtained from a population of size N is equally likely to occur, we refer to that sample as a *simple random sample*

In other words, from a population of size N, choose sample A of size n, sample B of size n, sample C of size n, etc., we realize that the probability of forming sample A is equal to forming sample B, which is equal to forming sample C, etc.

Example 1.4: Selecting simple random sampling

Suppose we want to randomly choose three people ($n = 3$) to participate in a clinical trial from the following population:

List all possible samples of $n = 3$. In case an individual is selected, then they cannot be selected again in the same sample. Give a conclusion regarding the likelihood

[4] It is difficult to obtain random individuals as it might require resources or more time to achieve

containing Amy, Emma and Bob. The solution is:

Amy Bob Carly	Amy Bob David	Amy Bob Emma	Amy Bob Felix
Amy Carly David	Amy Carly Emma	Amy Carly Felix	Amy David Emma
Amy David Felix	Amy Emma Felix	Bob Carly David	Bob Carly Emma
Bob Carly Felix	Bob David Emma	Bob David Felix	Bob Emma Felix
Carly David Emma	Carly David Felix	Carly Emma Felix	David Emma Felix

There are 20 samples generated. Note that each name is found ten times. There are $4 \cdot 5 = 20$ samples and each cell has three values, which means $20 \cdot 3 = 60$ names in total. Since we are randomly and fairly generating the samples of six people, each individual is recorded ten times. As for the sample that contains Amy, Bob and Emma, it is generated only once and highlighted. Since there are 20 samples in total and it was generated only once, there is 1 in 20 (or $\frac{1}{20}$ or 5%) chance of occurring. This is the same for any sample found. In this example, the order of the sample doesn't matter. For instance, Amy, Emma and Bob is the same as Amy, Bob and Emma.

1.5 Different Sampling Methods

Let us discuss the different ways to choose samples, such as stratified sampling, systematic sampling and cluster sampling and convenience sampling.

Definition 1.12: Stratified sampling

Achieved by dividing the population into non-overlapping groups, called *strata*[a] based on some variable (*e.g.*, gender), and then selecting a simple random sample from each of the strata.

[a] Stratum is singular and strata is plural.

Example 1.5: Measuring IQ with stratified sampling

Suppose we wanted to measure the 'IQ' scores of students at Brock University to see if there is a pattern between students' IQ score and their major. We can divide our strata based on majors. Let us assume we only have the following majors with equal number of students (say 100 students): Mathematics, English, Biology, History and Chemistry. Then, we can use simple random sampling and select, say ten students, from each major to get 50 students in total and study those 50 students.

Definition 1.13: Systematic sampling

Achieved by selecting every kth individual (*e.g.*, every 5th individual) from the population.

Example 1.6: Measuring height with systematic sampling

Suppose we wanted to measure the 'height' of students in this class. Let us say we measure the height of every 4th student who enters the room.

Definition 1.14: Cluster sampling

Achieved by dividing the population into sections (or clusters) based on natural/already existing characteristic, then, randomly select some of those clusters by choosing *all* the individuals from those selected clusters.

Example 1.7: Measuring age with cluster sampling

Suppose we wanted to measure the 'age' of people living in Ontario. Since Ontario is divided based on cities, we can define a cluster to be a city. Now, we are able to study all individuals in a city. For instance, let us select St. Catharines, Niagara Falls and Toronto[a] as the clusters and ignore the rest of cities. Now, we only measure the age of people who live in either St. Catharines, Niagara Falls and/or Toronto.

[a] We need to select more cities for actual studies but ignored in this example

Definition 1.15: Convenience sampling

Achieved when individuals are easily obtained and not based on randomness.

Typical convenience sampling example could be the individuals themselves deciding to participate in some survey. Here is a visual representation of the different sampling methods discussed above:

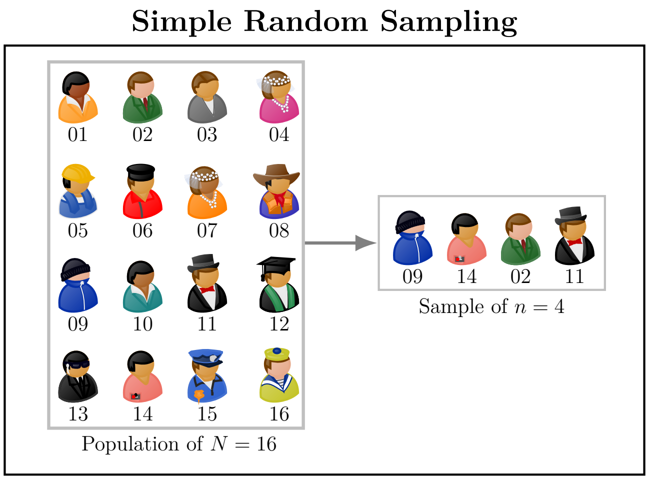


Figure 1.1: Simple random sampling where the sample is randomly selected.

Stratified Sampling

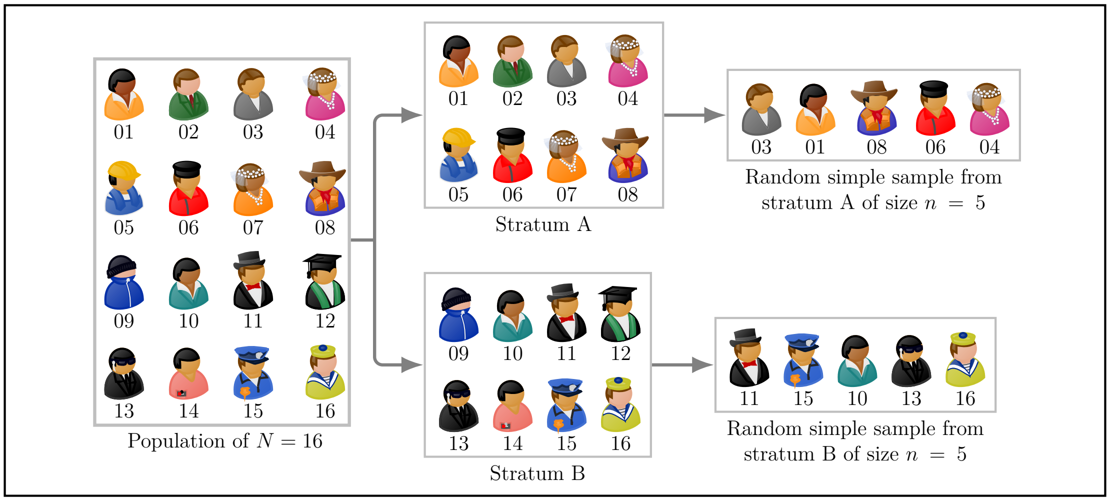


Figure 1.2: The two strata are based of some characteristic. Suppose stratum A includes parents whereas stratum B includes non-parent individuals. Note that our sample is the total individuals selected (*i.e.*, $n = 10$ which contains: 03, 01, 08, 06, 04, 11, 15, 10, 13 and 16)

Systematic Sampling

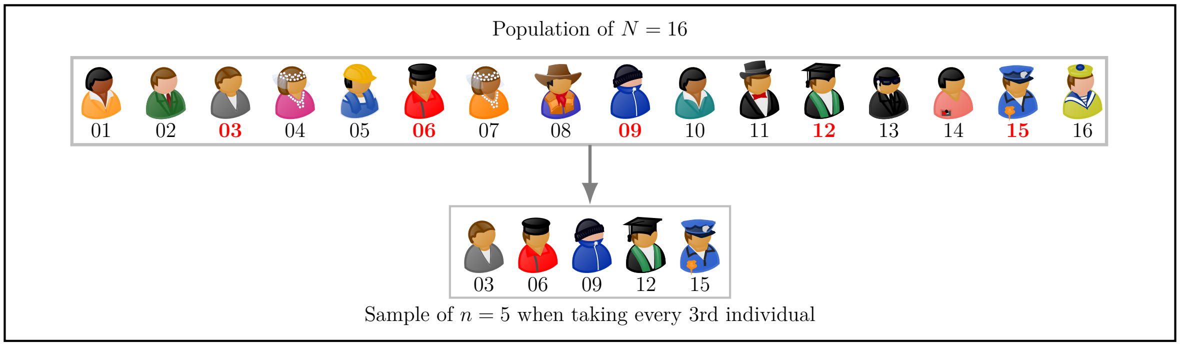


Figure 1.3: Systematic sampling where every third person has been selected in the sample

Cluster Sampling

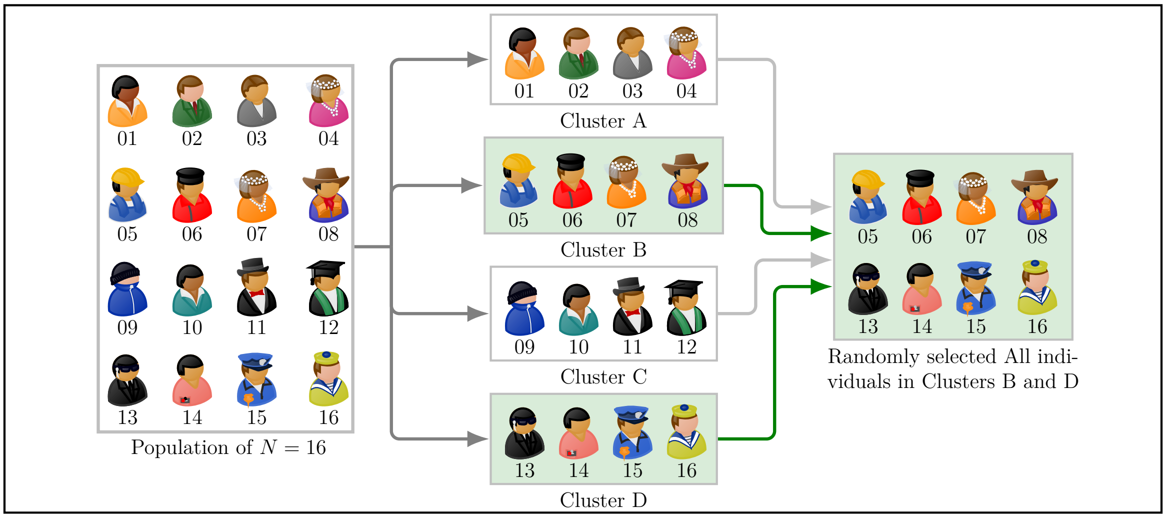


Figure 1.4: Cluster sampling where the population is divided into some clusters based on some natural characteristic (where they live, occupation, age, etc.) and then randomly choosing the *entire* cluster to study. Clusters B and D were randomly chosen.

1.6 Encountering Biases

When selecting individuals, there is a chance for encountering bias.

Definition 1.16: Bias

Bias is encountered when the results of the sample (statistic) don't represent the result of the population (parameter).

Moreover, there are two different sources of bias when choosing a sample. They are sampling bias and response bias.

Definition 1.17: Sampling biases

- **Sampling bias:** describes the approach used to obtain the sample's individuals favours a part of the population over the rest
  - Any convenience sample has sampling bias because the individuals are not randomly chosen
  - An example is asking the population about sensitive topics, such as what is your salary if we want to find the average salary, or what is your age if we wanted to find the average age
- **Response bias:** encountered when the answers on a survey do not reflect the true response of the individual, which could be met by
  - **Interviewer error:** the interviewer isn't as skillful as they should (*e.g.*, when the interviewer fears for being to personal or not realizing the significance of the response)

- **Misrepresented answers:** the responses received are to realistic because people might over/underestimate their abilities (*e.g.*, asking people how many push-ups they can do in a minute and then asking them to do the push-ups, how accurate are they?)
- **Wording of questions:** could lead to response bias in the survey. We need to stay away from yes/no questions and focus on open-ended but balanced questions. Consider the variants of the same question:
  - i. Are you against increasing taxes?
  - ii. Do you favour or oppose the increase of taxes?

  The second version is balanced and doesn't ask for a yes/no question
- **Data-entry error:** It is easy to introduce typos and mistakes when entering the data collected on paper into a computer

1.7 Organizing Quantitative Data

We will discuss how we can visually represent discrete and continuous variables graphically using a histogram.

1.7.1 Visualizing Discrete Data

Suppose we have data of some discrete variable in a table, how can we visualize the data? One simple approach is by constructing frequency (number of times something occurred) and relative frequency (probability of occurring) distributions. We create classes of the distinct values of the discrete variable. Suppose we asked 50 students about the number of hours they study per week and the data recorded as:

10	00	06	02	01	06	01	08	08	08
02	08	04	13	02	08	00	12	05	08
14	10	13	14	11	14	04	02	02	01
12	10	05	09	12	14	13	08	13	05
09	06	01	13	04	11	09	10	03	04

Table 1.3: The number of hours spent studying per week of 50 random individuals

Let us construct the table to summarize the data, all found in Table 1.4. The frequency is found by counting the number of times we see a value in the table and relative frequency is dividing that number found by the number of people.

We have a nice organized table summarizing the data. How can we visualize it‽ We can use a type of a diagram called histogram, as shown in Figure 1.5.

Number of hours (classes)	**Frequency**	**Relative frequency**
00	2	$\frac{2}{50} = 0.04$
01	4	$\frac{4}{50} = 0.08$
02	5	$\frac{5}{50} = 0.10$
03	1	$\frac{1}{50} = 0.02$
04	4	$\frac{4}{50} = 0.08$
05	3	$\frac{3}{50} = 0.06$
06	3	$\frac{3}{50} = 0.06$
07	0	$\frac{0}{50} = 0.00$
08	7	$\frac{7}{50} = 0.14$
09	3	$\frac{3}{50} = 0.06$
10	4	$\frac{4}{50} = 0.08$
11	2	$\frac{2}{50} = 0.04$
12	3	$\frac{3}{50} = 0.06$
13	5	$\frac{5}{50} = 0.10$
14	4	$\frac{4}{50} = 0.08$
The total frequency is 50 (which is $n = 50$) and total relative frequency is 1 (or 100%)		

Table 1.4: Finding the frequency and relative frequency of the number of hours spent studying per week

Definition 1.18: Histogram

A *histogram* is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and the rectangles touch each other.

Instead of plotting the frequency, we can plot the relative frequency. The plot will be the same but we have to change the y-axis labels to show the relative frequency data. Also, compare both Figure 1.6a and Figure 1.6b.

September 14, 2023

1.7.2 Visualizing Continuous Data

Suppose we have data of some continuous variable in a table, how can we visualize the data? Similar to discrete data, we can construct a frequency (number of times something occurred) and relative frequency (probability of occurring) distributions. However, note that

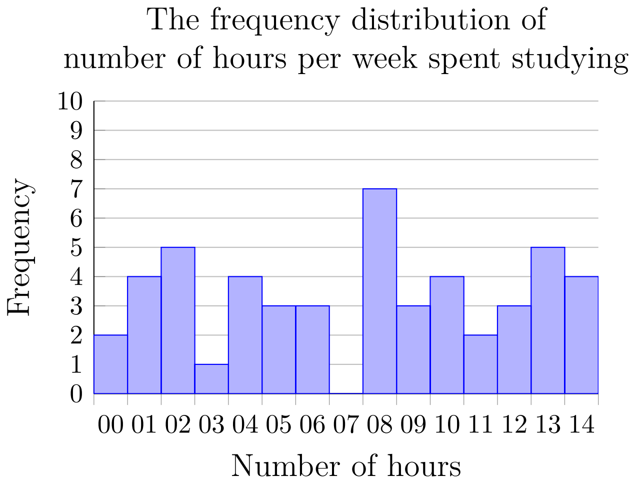
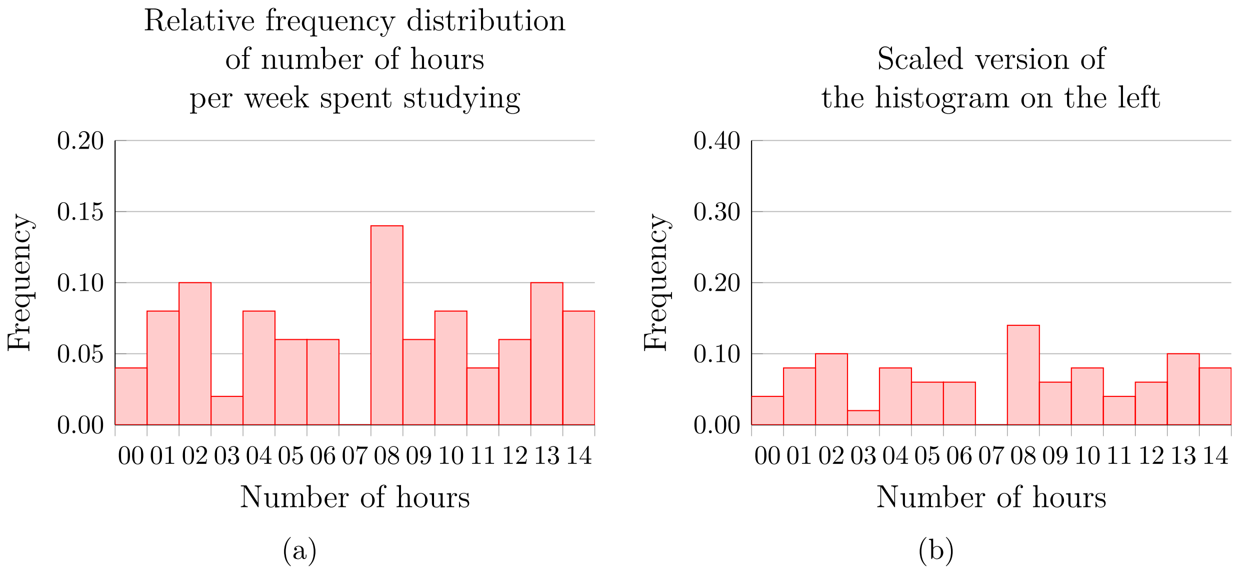


Figure 1.5: A frequency histogram of the data regarding the number of hours spent studying



(a)

(b)

Figure 1.6: A relative frequency histogram of the data regarding the number of hours spent studying where the y-axis is scaled, which makes the data look smaller. In reality, both histograms (Figure 1.6a and Figure 1.6b) represent the *exact* same data.

continuous means a range not single unit. To understand that, suppose we want to know the number of numbers between 0 and 1 (including both endpoints). Then, discretely (*i.e.*, whole numbers only), we would have a count of just two (the value '0' and value '1'). Continuously, we would have infinitely many possibilities. For instance, we can have 0.5, 0.55, 0.555, etc. How can we count the values? Instead of counting *values*, we construct *ranges* and count the number of values that fall within a range. A simple example is counting how many students received A+ $(90 - 100)$, A $(80 - 89.9)$, B $(70 - 79.9)$, C $(60 - 69.9)$ and D $(50 - 59.9)$ in a course. We created four ranges (also referred to as 'classes' or 'bins') and we can count the number of students within each class. For example, suppose we timed 50 people to run 100 metres and the data recorded in Table 1.5.

30.17	21.29	22.20	19.99	26.97	31.47	31.65	28.24	19.86	23.16
11.83	16.89	19.02	33.45	39.79	20.35	25.95	24.15	26.50	31.57
36.80	13.40	29.43	13.61	34.68	20.32	17.97	30.28	35.71	22.39
32.63	24.67	24.31	35.13	28.93	33.21	11.81	20.15	31.73	29.92
15.74	14.09	17.74	14.16	12.36	18.55	28.30	18.49	11.79	27.79

Table 1.5: The times of the 50 random individuals completing the run with the slowest and fastest values highlighted.

Let us construct the table to summarize the data. We will define our classes as the following ranges:

$$10.00 - 14.99$$
$$15.00 - 19.99$$
$$20.00 - 24.99$$
$$25.00 - 29.99$$
$$30.00 - 34.99$$
$$35.00 - 39.99$$

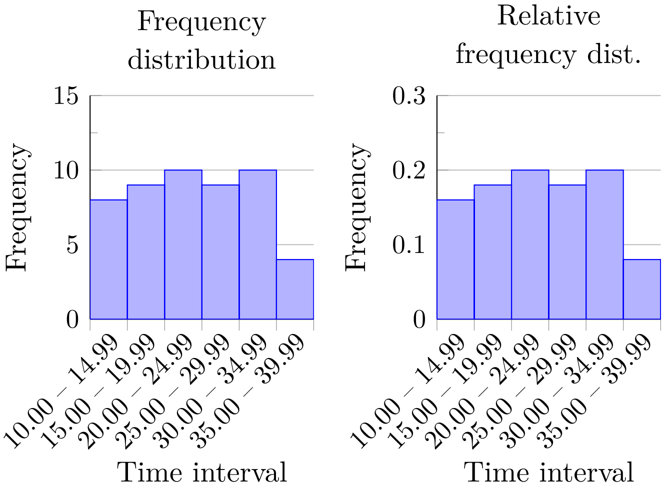
Why the size classes above? That what was chosen. We could have simply chosen five classes or ten classes. There is a suggestion on how to define this which will be discussed later. Now, the number of times we have encountered a time between 10.00 and 14.99 is eight. We will do the same for the other classes to generate the table. It is easier to first sort the values first and then create the table, but this step is not required.

11.79	11.81	11.83	12.36	13.40	13.61	14.09	14.16	15.74	16.89
17.74	17.97	18.49	18.55	19.02	19.86	19.99	20.15	20.32	20.35
21.29	22.20	22.39	23.16	24.15	24.31	24.67	25.95	26.50	26.97
27.79	28.24	28.30	28.93	29.43	29.92	30.17	30.28	31.47	31.57
31.65	31.73	32.63	33.21	33.45	34.68	35.13	35.71	36.80	39.79

Table 1.6: The sorted times of the 50 random individuals completing the run with the slowest and fastest values highlighted.

Time interval (classes in seconds)	**Frequency**	**Relative frequency**
10.00 – 14.99	8	$\frac{8}{50} = 0.16$
15.00 – 19.99	9	$\frac{9}{50} = 0.18$
20.00 – 24.99	10	$\frac{10}{50} = 0.20$
25.00 – 29.99	9	$\frac{9}{50} = 0.18$
30.00 – 34.99	10	$\frac{10}{50} = 0.20$
35.00 – 39.99	4	$\frac{4}{50} = 0.08$
The total frequency is 50 (which is $n = 50$) and total relative frequency is 1 (or 100%)		

Table 1.7: Finding the frequency and relative frequency of the times of the runs



(a) (b)

Figure 1.7: The frequency and relative frequency of the data found in Table 1.5.

Plotting the histograms of continuous data is the same plotting discrete data. Ensure to place the class labels in the x-axis names. We had our data to be divided by an increment of five seconds (*i.e.*, the class width is five). We can use the following formula for dividing the data:

$$\text{Class width} \approx \frac{\text{largest data value} - \text{smallest data value}}{\text{number of classes}}. \tag{1}$$

The symbol ≈ means approximately. In our case, we know the smallest value is 11.79 and the largest is 39.79. Suppose we want the number of classes to be six, then the class width is:

$$\begin{aligned}
\text{Class width} &\approx \frac{\text{largest data value} - \text{smallest data value}}{\text{number of classes}} \\
\text{Class width} &\approx \frac{39.79 - 11.79}{6} \\
\text{Class width} &\approx 4.666666666666667 \\
\text{Class width} &= 5 \text{ (round up)}
\end{aligned} \tag{2}$$

We should round up/down because it is easy to work with whole numbers. Rounding up/down may result in fewer/more classes than were originally intended. A more formal notation to use when dealing with classes (or ranges in general) is to use brackets (*i.e.*, $[\ldots]$) for inclusivity or parentheses (*i.e.*, $(\ldots)$) for exclusivity. Here are the possible combinations:

$$\begin{aligned}
[0,1] &\rightarrow \text{The range is anything between 0 and 1 including both 0 and 1} \\
(0,1) &\rightarrow \text{The range is anything between 0 and 1 but not including 0 nor 1} \\
[0,1) &\rightarrow \text{The range is anything between 0 and 1 including 0 but not 1} \\
(0,1] &\rightarrow \text{The range is anything between 0 and 1 not including 0 but including 1}
\end{aligned} \tag{3}$$

Here are some examples using discrete values:

$$\begin{aligned}[0,5] &= 0,1,2,3,4,5\\ (0,5) &= 1,2,3,4\\ [0,5) &= 0,1,2,3,4\\ (0,5] &= 1,2,3,4,5\end{aligned} \tag{4}$$

So, we can rewrite our classes as:

$$\begin{aligned}10.00 - 14.99 &\rightarrow [10,15)\\ 15.00 - 19.99 &\rightarrow [15,20)\\ 20.00 - 24.99 &\rightarrow [20,25)\\ 25.00 - 29.99 &\rightarrow [25,30)\\ 30.00 - 34.99 &\rightarrow [30,35)\\ 35.00 - 39.99 &\rightarrow [35,40)\end{aligned} \tag{5}$$

Our histogram should have the following labels when using interval notations:

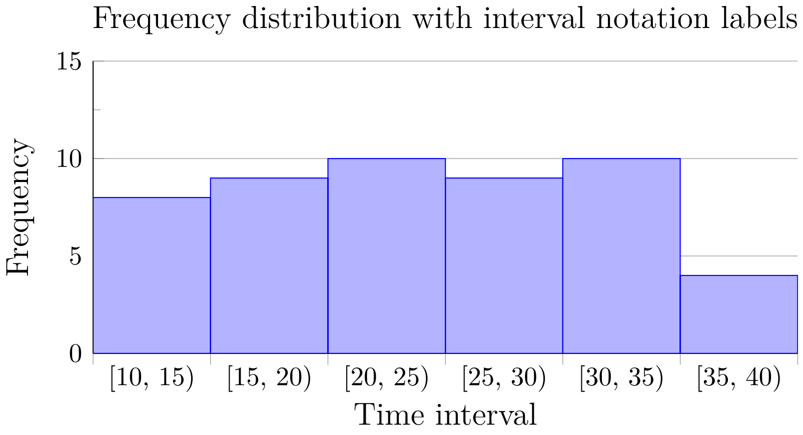
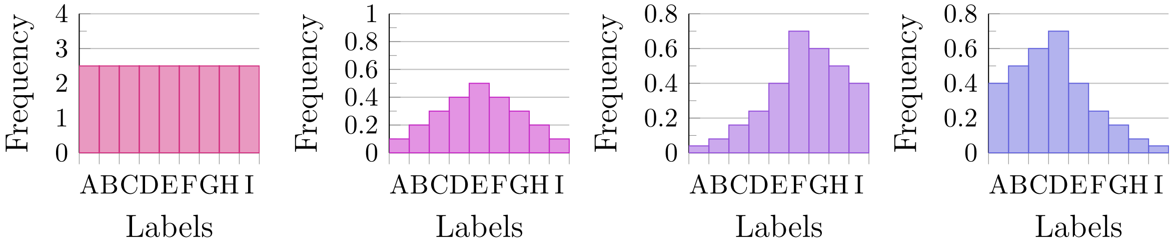


Table 1.8: A proper frequency distribution of continuous data that describes each class using the interval notation

A distribution could be described as one of the four main shapes: *uniform* distribution (equal), *bell-shaped* distribution (looks like a bell), *skewed right* (low on the right side) and *skewed left* (low on the left side).

1.8 Mean, Median and Mode

We will learn how to find the mean (average), median and mode. Let us start with the mean. There are two different means we can find with each having its own formula: the mean of a sample (*i.e.*, a statistic) and the mean of a population (*i.e.*, parameter). Let us recall that lowercase n refers to the size of a sample and uppercase N refers to the size of a population.



(a) Uniform (symmetric)

(b) Bell-shaped (symmetric)

(c) Skewed left

(d) Skewed right

Figure 1.8: The four typical distribution

Definition 1.19: The mean $\overline{x}$ of a sample

The mean of a sample, denoted as $\overline{x}$ (pronounced as 'x bar'), is the symbol used to find the mean (or average) of a sample size n using the following formula:

$$\overline{x} = \frac{x_{\boxed{1}} + x_{\boxed{2}} + x_{\boxed{3}} + \cdots + x_{\boxed{n}}}{n} = \frac{\sum_{i=1}^{n} x_{\boxed{i}}}{n}, \tag{6}$$

What does $\sum_{i=1}^{n}$, pronounced 'sigma', describe? It is a fancy notation to denote that i starts from 1 and increments by one each time until reaching n. In other words, i will be $1, 2, 3, 4, \ldots, n$. Now, Σ is the symbol that denotes the sum. Hence, if we want to find the sum of 1 + 2 + 3, we could use $\sum_{i=1}^{3}(\boxed{i})$. What this means is that we start at $i = 1$, plug that value in the boxed formula (just i in this case) to get 1. Do the same thing for $i = 2$ and $i = 3$ to get the sum of $1 + 2 + 3 = 6$. Another example is $\sum_{i=1}^{4}(\boxed{2 \cdot i})$. We have i to start at 1 and go to 4, each time we plug the current value of i into the boxed formula $2 \cdot i$ and then find the sum of all four values. It will be $2 \cdot 1 + 2 \cdot 2 + 2 \cdot 3 + 2 \cdot 4 = 20$.

Example 1.8: Finding the mean $\overline{x}$ of a sample

Find the mean of the following height data (in centimetres) from a sample of five people:

$$\begin{array}{cccccc c} 194 & 173 & 160 & 188 & 176 & \leftarrow & \text{data} \\ \boxed{1} & \boxed{2} & \boxed{3} & \boxed{4} & \boxed{5} & \leftarrow & \text{position} \end{array}$$

We need to add all of the values and then divide by five:

$$\begin{aligned}
\overline{x} &= \frac{\sum\limits_{i=1}^{n} x_{\boxed{i}}}{n} \\
\overline{x} &= \frac{x_{\boxed{1}} + x_{\boxed{2}} + x_{\boxed{3}} + \cdots + x_{\boxed{n}}}{n} \\
\overline{x} &= \frac{194 + 173 + 160 + 188 + 176}{5} \\
\overline{x} &= 178.2\text{cm}
\end{aligned} \tag{7}$$

Definition 1.20: The mean μ of a population

The mean of a population, denoted as μ (pronounced as 'meu'), is the symbol used to find the mean (or average) of a population using the following formula:

$$\mu = \frac{x_{\boxed{1}} + x_{\boxed{2}} + x_{\boxed{3}} + \cdots + x_{\boxed{N}}}{N} = \frac{\sum\limits_{i=1}^{N} x_{\boxed{i}}}{N}, \tag{8}$$

where N is the sample size.

Example 1.9: Finding the mean μ of a population

Find the mean of the following height data (in centimetres) from a population of ten people:

$$\begin{array}{cccccccccccc}
176 & 177 & 174 & 160 & 187 & 167 & 183 & 172 & 160 & 178 & \leftarrow & \text{data} \\
\boxed{1} & \boxed{2} & \boxed{3} & \boxed{4} & \boxed{5} & \boxed{6} & \boxed{7} & \boxed{8} & \boxed{9} & \boxed{10} & \leftarrow & \text{position}
\end{array}$$

We need to add all of the values and then divide by ten:

$$\begin{aligned}
\mu &= \frac{\sum\limits_{i=1}^{N} x_i}{N} \\
\mu &= \frac{x_{\boxed{1}} + x_{\boxed{2}} + x_{\boxed{3}} + \cdots + x_{\boxed{N}}}{N} \\
\mu &= \frac{176 + 177 + 174 + 160 + 187 + 167 + 183 + 172 + 160 + 178}{10} \\
\mu &= 173.4\text{cm}
\end{aligned} \tag{9}$$

❈ September 19, 2023 ❈

Definition 1.21: What is median M

The *median*, denoted by M, of data is the middle value when the data is *sorted*.

To find the median, we need to first sort our data (lowest to left and highest to right) and then find the middle number. In case we have odd number of data, then the middle number will be M. In case we have an even number of data, then we need to find the two middle numbers, add them together and then divide by two. In general, when there are odd number of data, the median is just the $\left(\frac{n+1}{2}\right)^{th}$ position, in case there are even number of data, then we add the values in the $\frac{n}{2}^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ positions and divide by two.

Example 1.10: Finding the median M with an odd number of data (easy)

Find the median given the following height data (in centimetres) from a population (or sample) of nine people:

176	177	174	160	187	167	183	172	160	←	data
1	2	3	4	5	6	7	8	9	←	position

The data given is raw (non-sorted), let us sort it first:

160	160	167	172	174	176	177	183	187	←	data
1	2	3	4	5	6	7	8	9	←	position

Hence, the median is 174cm.

Example 1.11: Finding the median M with an even number of data (a little complex)

Find the median given the following height data (in centimetres) from a population (or sample) of eight people:

176	177	174	160	167	183	172	160	←	data
1	2	3	4	5	6	7	8	←	position

The data given is raw (non-sorted), let us sort it first:

160	160	167	172	174	176	177	183	←	data
1	2	3	4	5	6	7	8	←	position

We see that the values in position 4 and 5 are the middles ones. Let us add them and then divide by two to get $\frac{172+174}{2} = 173\text{cm}$.

Definition 1.22: Mode

The most value repeated in the data. It could be none (if the elements are all unique), one or multiple values.

Example 1.12: Mode examples

Given the following data, what is the mode?

160 160 167 172 174 176 177 183

The mode is 160, because it is the most common value and was found twice.

Given the following data, what is the mode?

$$160 \quad 160 \quad 167 \quad 172 \quad 174 \quad 177 \quad 177 \quad 183$$

The mode is 160 and 177, because they were both found twice.
Given the following data, what is the mode?

$$160 \quad 161 \quad 167 \quad 172 \quad 174 \quad 176 \quad 177 \quad 183$$

The mode doesn't exist as all values are unique. It is *not* zero, it doesn't exist or it is none.

1.9 Weighted Mean

The weighted mean find the average but associate some 'scaling' to each data value. The best use case is calculating your course grade. Suppose your class consists of one 10% assignment and one 90% exam. You receive 99% on the assignment[5] and 50% on the exam. What is your average? We cannot just add 99 with 50 and divide by two as the exam weighs more than the assignment. Instead, we need to scale the two data values accordingly. The assignment value will be multiply the assignment grade by 0.10 (which is 10%) and the exam mark by 0.90 (which is the 90%) and all divided by the sum $0.10 + 0.90$:

$$\text{total} = \frac{\left(0.10 \cdot \dfrac{99}{100}\right) + \left(0.90 \cdot \dfrac{50}{100}\right)}{0.10 + 0.90} = 0.549 = 54.9\%. \tag{10}$$

Definition 1.23: Weighted mean formula

The weighted mean formula, $\overline{x}_w$, finds the weighted average where data values don't necessarily have the same significance to one another. Each data value is associated with a weight and that weight is multiplied by the data value. We then divide by the sum of the weights:

$$\overline{x}_w = \frac{\sum_{i=1}^{n}(w_i \cdot x_i)}{\sum_{i=1}^{n}(w_i)} = \frac{(w_1 \cdot x_1) + (w_2 \cdot x_2) + (w_3 \cdot x_3) + \cdots + (w_n \cdot x_n)}{w_1 + w_2 + w_3 + \cdots + w_n} \tag{11}$$

Example 1.13: Weighted mean example

Suppose a student received the following grades in some course:

[5] Recall that $x\%$ is the same as $\dfrac{x}{100}$

Entry	Weight	Grade
Assignment 1	$5\% = 0.05$	$\frac{8}{10} = 0.8$
Assignment 2	$5\% = 0.05$	$\frac{10}{10} = 1$
Assignment 3	$5\% = 0.05$	$\frac{7}{10} = 0.7$
Midterm 1	$10\% = 0.10$	$\frac{37}{40} = 0.925$
Exam 1	$20\% = 0.20$	$\frac{48}{50} = 0.96$

To find their average in the course, we need to multiply the grade by the weight and sum all of the entries. For assignment 1, it will be $0.8 \cdot 0.05$, or equivalently, $\frac{8}{10} \cdot 0.05$. We will do the same for all of the other entries and then divide by the sum of the weights (which happens to be 0.45):

$$\begin{aligned}
\overline{x}_w = \frac{\sum_{i=1}^{n}(w_i \cdot x_i)}{\sum_{i=1}^{n}(w_i)} &= \frac{(w_1 \cdot x_1) + (w_2 \cdot x_2) + (w_3 \cdot x_3) + \cdots + (w_n \cdot x_n)}{w_1 + w_2 + w_3 + \cdots + w_n} \\
&= \frac{(0.05 \cdot 0.8) + (0.05 \cdot 1) + (0.05 \cdot 0.7) + (0.10 \cdot 0.925) + (0.20 \cdot 0.96)}{0.05 + 0.05 + 0.05 + 0.10 + 0.20} \\
&= 0.91 \text{ or } 91\%.
\end{aligned} \tag{12}$$

The conclusion is that the student has completed 45% of the course and received 91% on their submitted work. In case they have completed the *entire* course, we will see the denominator adds up to one. In examples other than grades and marks, the sum of the weights doesn't have to add up to one.

1.10 Standard Deviation and Variance

The standard deviation is a way to measure the deviation from the mean. A *small* standard deviation indicates the data are placed densely around the mean. A *high* standard deviation indicates the data to be more spread out. For instance, have a look at the following data where each of the three samples have a mean of $\overline{x} = 60$:

	x_1	x_2	x_3	x_4	x_5	x_6	
Sample A:	60	60	60	60	60	60	$\overline{x} = 60$
Sample B:	360	0	0	0	0	0	$\overline{x} = 60$
Sample C:	100	32	89	122	5	12	$\overline{x} = 60$

While the samples have a mean of 60, we see they are quite different from one another.

How can we measure this difference? We use the standard deviation formula. It will tell us how much each data value is away from the mean. For instance, Sample A has a standard deviation of zero because each point *is* the mean, hence, there is no spread relative to the mean. Sample B has a significant deviation. All the five zeros are 60 units away from the mean and the 360 data value is 300 units away. Sample B has a standard deviation of around 147. Lastly, Sample C has a standard deviation of around 50.

Definition 1.24: Sample standard deviation formula

The sample standard deviation formula, denoted by s, is given as

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1}}, \tag{13}$$

where x_i is the current data value, $\overline{x}$ is the sample mean and n is the sample size.

Example 1.14: Finding the sample standard deviation

Find the standard deviation of the following sample data:

$$\begin{array}{ccccccc} 12 & 71 & 33 & 64 & 99 & \leftarrow & \text{data} \\ 1 & 2 & 3 & 4 & 5 & \leftarrow & \text{position} \end{array}$$

We know that $n = 5$, let us find the sample mean and then the sample standard deviation.

$$\overline{x} = \frac{12 + 71 + 33 + 64 + 99}{5} = \frac{279}{5} = 55.8 \tag{14}$$

To find the standard deviation, get the first value in the data, subtract it from $\overline{x}$ and then square the value. Do that for all of the values, then divide by $n - 1$ and then take the square root:

$$\begin{aligned} s &= \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + (x_4 - \overline{x})^2 + (x_5 - \overline{x})^2}{n - 1}} \\ &= \sqrt{\frac{(12 - \overline{x})^2 + (71 - \overline{x})^2 + (33 - \overline{x})^2 + (64 - \overline{x})^2 + (99 - \overline{x})^2}{n - 1}} \\ &= \sqrt{\frac{(12 - 55.8)^2 + (71 - 55.8)^2 + (33 - 55.8)^2 + (64 - 55.8)^2 + (99 - 55.8)^2}{5 - 1}} \\ &= \sqrt{\frac{(-43.8)^2 + (15.2)^2 + (-22.8)^2 + (8.2)^2 + (43.2)^2}{4}} \\ &= \sqrt{\frac{1918.44 + 231.04 + 519.84 + 67.24 + 1866.24}{4}} \\ &= \sqrt{\frac{4602.8}{4}} \\ &= \sqrt{1150.7} \\ &= 33.92197 \end{aligned} \tag{15}$$

Hence, the sample standard deviation is 33.92197 of the sample provided.

Definition 1.25: Population standard deviation formula

The population standard deviation formula, denoted by σ (pronounced as 'sigma'), is given as

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}, \tag{16}$$

where x_i is the current data value, μ is the population mean and N is the population size.

Example 1.15: Finding the population standard deviation

Find the standard deviation of the following population data:

$$\begin{array}{cccccccc} 40 & 94 & 39 & 88 & 58 & 5 & \leftarrow & \text{data} \\ 1 & 2 & 3 & 4 & 5 & 6 & \leftarrow & \text{position} \end{array}$$

We know that $N = 6$, let us find the population mean and then the population standard deviation.

$$\mu = \frac{40 + 94 + 39 + 88 + 58 + 5}{6} = \frac{324}{6} = 54 \tag{17}$$

To find the standard deviation, get the first value in the data, subtract it from μ and then square the value. Do that for all of the values, then divide by N and then take the square root:

$$\begin{aligned} \mu &= \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + (x_4 - \mu)^2 + (x_5 - \mu)^2 + (x_6 - \mu)^2}{N}} \\ &= \sqrt{\frac{(40 - \mu)^2 + (94 - \mu)^2 + (39 - \mu)^2 + (88 - \mu)^2 + (58 - \mu)^2 + (5 - \mu)^2}{N}} \\ &= \sqrt{\frac{(40 - 54)^2 + (94 - 54)^2 + (39 - 54)^2 + (88 - 54)^2 + (58 - 54)^2 + (5 - 54)^2}{6}} \\ &= \sqrt{\frac{(-14)^2 + (40)^2 + (-15)^2 + (34)^2 + (4)^2 + (-49)^2}{6}} \\ &= \sqrt{\frac{196 + 1600 + 225 + 1156 + 16 + 2401}{6}} \\ &= \sqrt{\frac{5594}{6}} \\ &= \sqrt{932.33333} \\ &= 30.53413 \end{aligned} \tag{18}$$

Hence, the population standard deviation is 30.53413 of the population data provided.

Now that we have learned about standard deviation, variance is very similar to the calculation above but without the square root.

Definition 1.26: Sample variance formula

The sample variance formula, denoted by s^2, is given as

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1}, \tag{19}$$

where x_i is the current data value, $\overline{x}$ is the sample mean and n is the sample size.

Definition 1.27: Population variance formula

The population standard deviation formula, denoted by σ^2 (pronounced as 'sigma squared'), is given as

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N} = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}, \tag{20}$$

where x_i is the current data value, μ is the population mean and N is population size.

Example 1.16: Variance example

The sample variance of the data given in Example 1.14 is 1150.7. This is because we said that the sample variance is s^2, in other words, get the sample standard deviation (which is s), and then square it. The sample standard deviation is $s = 33.92197$, squaring it gives us $(33.92197)^2 = 1150.7$.

The population variance of the data given in Example 1.15 is 932.33333. This is because we said that the population variance is σ^2, in other words, get the population standard deviation (which is σ), and then square it. The population standard deviation is $\sigma = 30.53413$, squaring it gives us $(30.53413)^2 = 932.33333$.

In general, the metric variance is not used as much as it changes the units we are dealing with. For example, suppose we are measure lengths of road in kilometres (km). Then calculating the variance will give us the units squared, *i.e.*, km^2. In other words, we originally deal with road/lines and variance changes this so that we start dealing with squares instead of lines. Standard deviation on the other hand deals with the same roads we started with.

Example 1.17: Finding standard deviation and variance using Excel

Suppose we have the following excel data (SD is short for standard deviation):

	A	B	C	D	E
1	Data	Sample SD	Population SD	Sample Variance	Population Variance
2	40				
3	94				
4	39				
5	88				
6	58				
7	5				
8					

Our data is found in `A2:A7` (this is the range where we start at `A2` and go down to `A7`).

- To find the sample SD, in cell `B2`, write `=STDEV.S(A2:A7)`, the equal sign is the beginning of a formula, otherwise, it will write text. The `STDEV` is short for standard deviation and `.S` means finding the sample of the data from cells `A2` to `A7`.
- In `C2`, we can find the population SD by writing `=STDEV.P(A2:A7)`. Similar to the previous command but now using `.P`, which stands for population.
- To find the sample variance, in cell `2D`, write `=VAR.S(A2:A7)`. The `VAR` is short for variance and `.S` means finding the sample of the data from cells `A2` to `A7`.
- In `E2`, we can find the population variance by writing `=VAR.P(A2:A7)`. Similar to the previous command but now using `.P`, which stands for population.
- Note that we chose cells `B2`, `C2`, `D2` and `E2` to write the formulas because we had the titles above them. We are able to choose any cells. Also, this is an example to show how to use Excel functions, not practical exercise as we don't usually consider data to be sample *and* population.

Writing the above formulas in the appropriate cells gives the following output:

	A	B	C	D	E
1	Data	Sample SD	Population SD	Sample Variance	Population Variance
2	40	33.44846783	30.53413391	1118.8	932.3333333
3	94				
4	39				
5	88				
6	58				
7	5				
8					

1.11 Understanding What Standard Deviation Represents

Suppose the class mean is $\mu = 70$ and the standard deviation is $\sigma = 10$. How does a standard deviation of, say 5 or 20, differ from 10? Assuming that the data have a distribution that is bell shaped, which is also symmetric (*i.e.*, most of the data is close to the mean and not as much to the far left/right of the mean), then roughly 68% of the data lies between $\mu \pm 1 \cdot \sigma$. In our example, since the mean is 70 and the standard deviation is 10, this means that 68%

of the students have grades between 60 and 80 (remember the formula is $\mu \pm 1 \cdot \sigma = 70 \pm 10$). In case we want to find how much of our data lie between $\mu \pm 2 \cdot \sigma$, then roughly 95% of the data lies between ± two standard deviations. In our example, since the mean is 70, then 95% of the students will have a grade between 50 and 90. Lastly, we can extend the pattern to ± three standard deviations (*i.e.*, $\mu \pm 3 \cdot \sigma$). This will give us 99.7% of the data to be within that range. In our example, 99.7% of the students will have a grade between 40 and 100. We can go further to 4 standard deviation or even one million standard deviations but three is enough for practical situations. A visualization can be seen in Figure 1.9 and Figure 1.10 showing a typical normal distribution and how the division of standard deviation work.

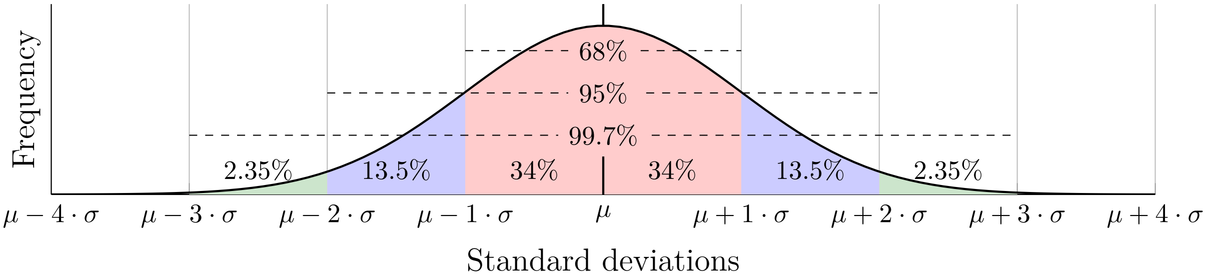


Figure 1.9: A typical normal distribution curve where it is symmetric about the mean

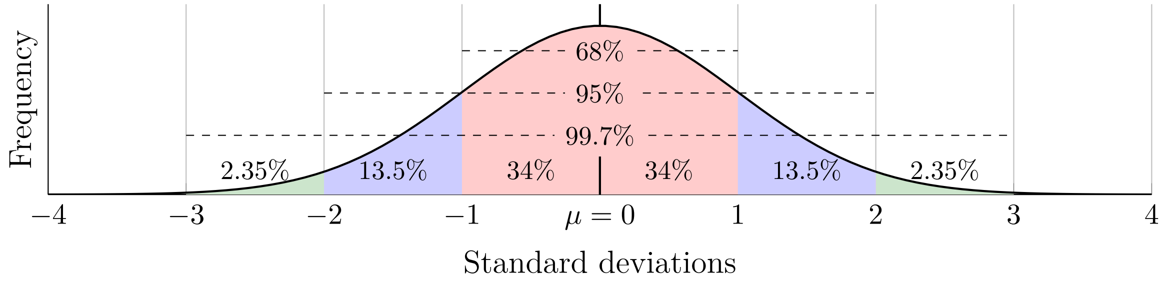


Figure 1.10: The same curve found in Figure 1.9 with easier labels to read because of the assumptions that the mean is 0 and standard deviation is 1

The data given in our example can be represented as the normal distribution curve shown in Figure 1.11.

Definition 1.28: The empirical rule

Given a distribution is roughly bell shaped, then:

- Approximately 68% of the data will lie within one standard deviation of the mean
- Approximately 95% of the data will lie within two standard deviation of the mean
- Approximately 99.7% of the data will lie within three standard deviation of the mean
- The same logic applies when we are dealing with sample data (change the population mean μ to the sample mean $\overline{x}$ and the population standard deviation σ with the sample standard deviation s).

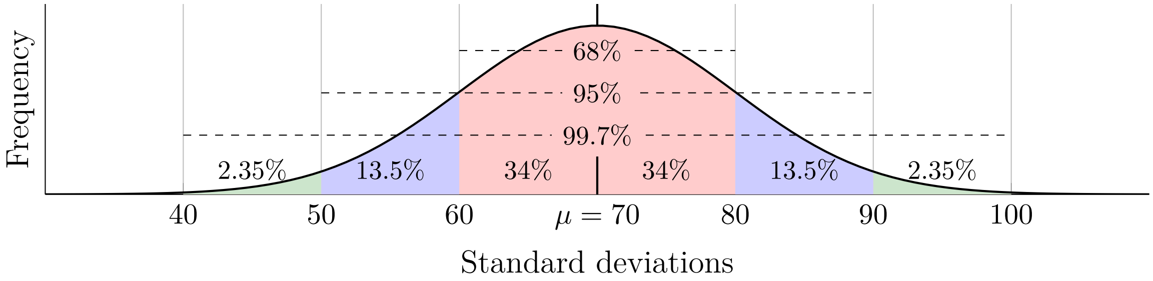


Figure 1.11: The data of the example explained where the mean of the course is 70 percent and standard deviation is 10

1.12 The z-score

Suppose that you receive a grade of 16 out of 33, what does that equate to? Perhaps a fail, D, C, B or A? It is difficult for us to judge. It would be easier for us to change the problem so that our range is not between 0 and 33, but between 0 and 100. Then, our grade of 16 will lie somewhere between 0 and 100 and now, we are able to understand the grade more easily. Let us go back to the discussion given in Section 1.11. We said that the mean of the class grades is 70 and the standard deviation is 10. Our important values are 40, 50, 60, 70, 80, 90 and 100. While it is correct, this is not the most optimal approach to describe our data. We want to have our mean to be zero and standard deviation to be 1, similar to what was shown in Figure 1.10. The simple manipulation of changing our range from 40 to 100 into -3 to 3 is referred to as the z-score. This is given by two similar formulas depending on whether we are dealing with sample or population data.

Definition 1.29: z-score formulas

The z-score, which has a mean of zero and standard deviation of one, represents the distance that a data value is from the mean in terms of the number of standard deviations.

Population z-score Sample z-score

$$z = \frac{x-\mu}{\sigma} \qquad\qquad z = \frac{x-\overline{x}}{s},$$

where x is the data value. A positive z-score means the data value is *larger* than the mean. A negative z-score means the data value is *smaller* than the mean. A z-score of zero means the data value is the mean. For instance, a z-score of 2.21 means the data value is 2.21 standard deviations above the mean. A z-score of -0.98 means the data value is 0.98 standard deviations below the mean. Note that the z-score is unitless.

Example 1.18: z-score example

Suppose we have a sample data to have $\overline{x} = 823.35$ and $s = 110.78$. What are the z-scores of the following data values:

$$1023, \quad 810.53, \quad 38, \quad 2013.41$$

Since this is sample data, we need to use the sample formula:

$$z = \frac{x - \overline{x}}{s}$$

Let us find the z-scores:

$$z = \frac{1023 - 823.35}{110.78} \quad , \quad z = \frac{810.53 - 823.35}{110.78} \quad , \quad z = \frac{38 - 823.35}{110.78} \quad , \quad z = \frac{2013.41 - 823.35}{110.78}$$

$$z = \frac{199.65}{110.78} \quad , \quad z = \frac{-12.82}{110.78} \quad , \quad z = \frac{-785.35}{110.78} \quad , \quad z = \frac{1190.06}{110.78}$$

$$z = 1.80222 \quad , \quad z = -0.11572 \quad , \quad z = -7.08928 \quad , \quad z = 10.74255$$

Hence, the data value 1023 is 1.80222 standard deviations from the mean, the data value 810.53 is -0.11572 standard deviations from the mean, the data value 38 is -7.08928 standard deviations from the mean and the data value 2013.41 is 10.74255 standard deviations from the mean.

1.13 Quartiles

Quartiles is a way to divide the data points into four parts (Q_1, Q_2, Q_3 and Q_4) and used to examine the properties of each part. Sort the values, find the median, the same metric from Definition 1.21, which is also Q_2 here. Now, we have data starting from the left to the median and from median to the right. The metric Q_1 is the midpoint from the left to the median, and The metric Q_3 is the midpoint from the median to the right. The metric Q_4 is the largest value. In summary, suppose the data is sorted from smallest (left) to largest (right):

- The median M is also Q_2, which is the middle value of the data
- Q_1 is the middle between the smallest value and the median
- Q_3 is the middle between the median value and the largest value
- Q_4 is the largest value

There are two different main approaches to calculate the quartiles, either we include the median value in the calculations of Q_1 and Q_3, or we don't. In this course, you have the choice to include it or not as long as you mention it. In Excel, there are two functions where one function (`=QUARTILE(...)`) includes the median in the calculation and the other (`=QUARTILE.EXC(...)`) doesn't. In the next examples, we will see how to find Q_1 and Q_3 when the median is included and excluded for odd values. For even values, let us just use a single easy method. In Excel, the function `=QUARTILE.EXC(...)` uses the excluded version[6].

1.13.1 Example with Odd Number of Data Points

Find the quartiles of theses values:

$$3 \quad 6 \quad 7 \quad 1 \quad 10 \quad 1 \quad 5 \quad 8 \quad 2 \quad 4 \quad 7$$

First, we need to sort them:

[6] When trying to create a boxplot in Excel, the function `=QUARTILE.EXC(...)` is internally used

$$1 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 7 \quad 8 \quad 10$$

The division and steps are shown in Figure 1.12. The conclusion is:

$$\begin{array}{lcccccccccccc} \text{Excluding the median:} & Q_1 & = & 2 & Q_2 & = & 5 & Q_3 & = & 7 & Q_4 & = & 10 \\ \text{Including the median:} & Q_1 & = & 2.5 & Q_2 & = & 5 & Q_3 & = & 7 & Q_4 & = & 10 \end{array}$$

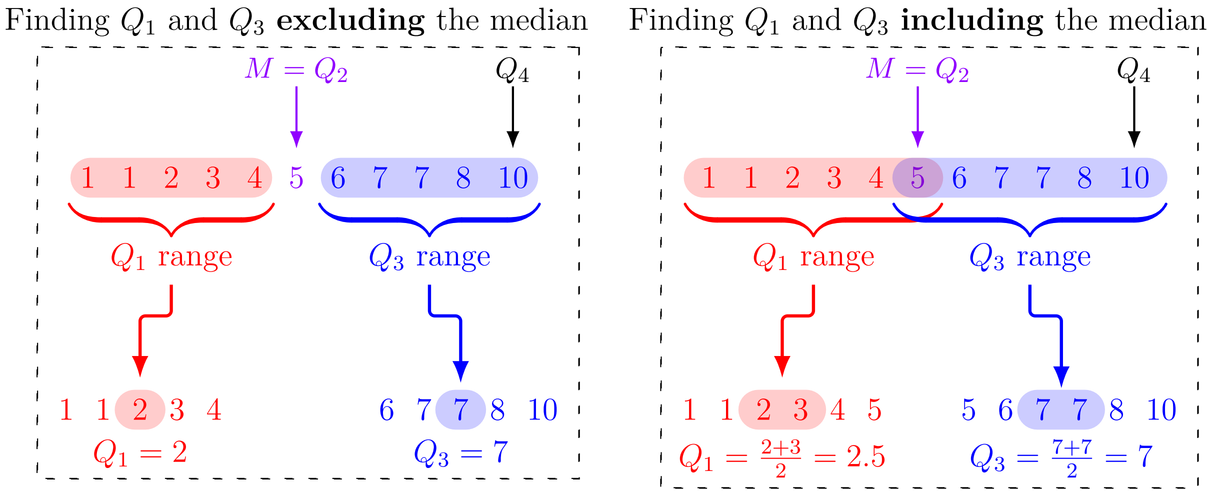


Figure 1.12: Dealing with odd number of elements and dividing the ranges of Q_1 and Q_3 when the median is excluded (left) or included (right)

1.13.2 Example with Even Number of Data Points

Find the quartiles of theses values:

$$3 \quad 6 \quad 7 \quad 7 \quad 1 \quad 10 \quad 1 \quad 5 \quad 8 \quad 2 \quad 4 \quad 7$$

First, we need to sort them:

$$1 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 7 \quad 7 \quad 8 \quad 10$$

The division and steps are shown in Figure 1.13. Basically, find the midpoint of the two middle values to find the midpoint and always exclude the median as we can divide the two ranges in half nicely. Note that Excel's `QUARTILE.EXC` function calculates this a *little* differently when there are even number of elements[7]. The conclusion is:

$$\text{Excluding the median (only method):} \quad Q_1 = 2.5 \quad Q_2 = 5.5 \quad Q_3 = 7 \quad Q_4 = 10$$

[7] There are four method to calculating quartiles, Excel uses a complex method whereas the lectures notes uses a simpler one

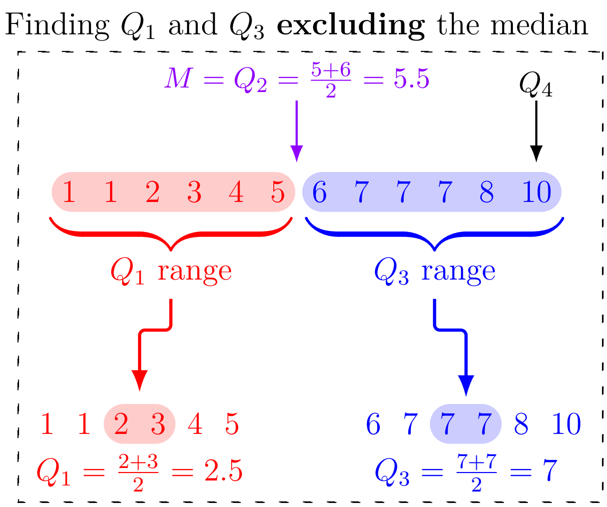

Finding Q_1 and Q_3 **excluding** the median
$M = Q_2 = \frac{5+6}{2} = 5.5$
Q_4
1 1 2 3 4 5
6 7 7 7 8 10
Q_1 range
Q_3 range
1 1 2 3 4 5
$Q_1 = \frac{2+3}{2} = 2.5$
6 7 7 7 8 10
$Q_3 = \frac{7+7}{2} = 7$

Figure 1.13: Dealing with even number of elements and dividing the ranges of Q_1 and Q_3 evenly excluding the median

1.13.3 Interquartile Range and Outliers

Let us define what is an interquartile range, minor outlier and major outlier.

Definition 1.30: Interquartile and minor/major outliers

- Interquartile range (IQR) is the amount of spread of the middle %50 of the data values. It is given by
$$IQR = Q_3 - Q_1.$$
- A *minor* outlier is any value below $Q_1 - (1.5 \times IQR)$ or above $Q_3 + (1.5 \times IQR)$.
- A *major* outlier is any value below $Q_1 - (3 \times IQR)$ or above $Q_3 + (3 \times IQR)$.

Example 1.19: IQR and minor/major outliers of Section 1.13.1 excluding median

Recall that we found the following using the exclusion of the median method:

$$Q_1 = 2, \quad Q_2 = 5, \quad Q_3 = 7, \quad Q_4 = 10$$

- The interquartile is $Q_3 - Q_1 = 7 - 2 = 5$
- The minor outliers range is
$$\begin{aligned} Q_1 - (1.5 \times IQR) &= 2 - (1.5 \times 5) = -5.5 \\ Q_3 + (1.5 \times IQR) &= 7 + (1.5 \times 5) = 14.5 \end{aligned} \tag{21}$$
Any data point that is less than -5.5 or larger than 14.5 is considered as a minor outlier
- The major outliers range is
$$\begin{aligned} Q_1 - (3 \times IQR) &= 2 - (3 \times 5) = -13 \\ Q_3 + (3 \times IQR) &= 7 + (3 \times 5) = 22 \end{aligned} \tag{22}$$
Any data point that is less than -13 or larger than 22 is considered as a major outlier

Example 1.20: IQR and minor/major outliers of Section 1.13.1 including median

Recall that we found the following using the inclusion of the median method (which, coincidentally, has the same IQR and minor/major outliers ranges as the example in Section 1.13.2):

$$Q_1 = 2.5, \quad Q_2 = 5, \quad Q_3 = 7, \quad Q_4 = 10$$

- The interquartile is $Q_3 - Q_1 = 7 - 2.5 = 4.5$
- The minor outliers range is

$$\begin{array}{rcccl} Q_1 - (1.5 \times IQR) & = & 2.5 - (1.5 \times 4.5) & = & -4.25 \\ Q_3 + (1.5 \times IQR) & = & 7 + (1.5 \times 4.5) & = & 13.75 \end{array} \tag{23}$$

  Any data point that is less than -4.25 or larger than 13.75 is considered as a minor outlier
- The major outliers range is

$$\begin{array}{rcccl} Q_1 - (3 \times IQR) & = & 2.5 - (3 \times 4.5) & = & -11 \\ Q_3 + (3 \times IQR) & = & 7 + (3 \times 4.5) & = & 20.5 \end{array} \tag{24}$$

  Any data point that is less than -11 or larger than 20.5 is considered as a major outlier

1.14 Five-Number Summary (Boxplot)

The five-number summary or a boxplot (also known as whisker plot) is a visualize representation of the data values that has five component: minimum value, Q_1, median (which is also Q_2), Q_3 and the maximum value, which is Q_4. Let us plot the boxplot of the following example, where it is given that the minimum value is 3, maximum is $Q_4 = 20$, Q_1 is 7, median is 11 and Q_3 is 14 in Figure 1.14.

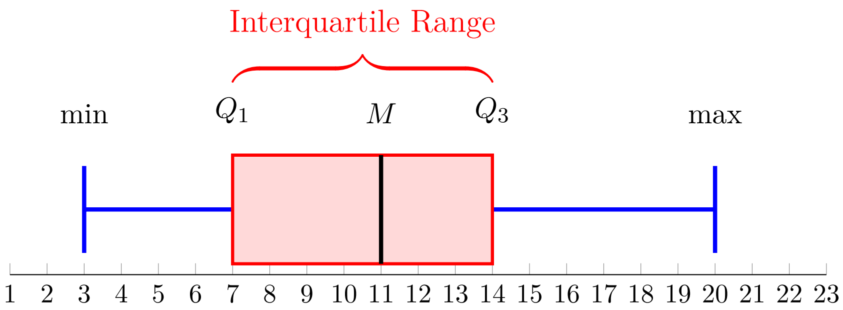


Figure 1.14: A boxplot which illustrates a minimum value is 3, maximum is 20, Q_1 is 7, median (which is also Q_2) is 11 and Q_3 14. The boxplot is horizontal but can also be vertical. The brace with the text 'interquartile range' is added for clarification, it is not a part of boxplots.

We now present some of the common Excel statistical functions in Table 1.9.

Metric	Formula	Explanation
Mean	`=AVERAGE(A2:A10)`	Select the range of values
Median	`=MEDIAN(A2:A10)`	Select the range of values
Mode	`=MODE(A2:A10)`	Select the range of values
Q_3	`=QUARTILE(A2:A10, 3)`	Place this in `F2` and `3` means 3rd quartile
Q_3	`=QUARTILE.EXC(A2:A10, 3)`	Similar to above but the median is not included in calculation (use this if you want to get the same values as Excel's boxplot summary)
Q_1	`=QUARTILE(A2:A10, 1)`	Place this in `G2` and `1` means 1st quartile
Q_1	`=QUARTILE.EXC(A2:A10, 1)`	Similar to above but the median is not included in calculation (use this if you want to get the same values as Excel's boxplot summary)
IQR	`=F2 - G2`	Place this in `H2`
Lower minor bound	`=G2 - (1.5 * H2)`	Lower minor bound
Upper minor bound	`=G2 + (1.5 * H2)`	Upper minor bound
Lower major bound	`=F2 - (3 * H2)`	Lower major bound
Upper major bound	`=F2 + (3 * H2)`	Upper major bound
Sample SD	`=STDEV.S(A2:A10)`	The sample standard deviation
Population SD	`=STDEV.P(A2:A10)`	The population standard deviation
Sample Variance	`=VAR.S(A2:A10)`	The sample variance
Population Variance	`=VAR.P(A2:A10)`	The population variance

Table 1.9: The typical Excel functions we might encounter with the assumption that our data starts at cell `A2` and end at `A10`.