Statistics: Measures of Data

Statistics is the field of collecting, analysing, interpreting, and presenting data. It often involves surveying people or things and asking questions about a particular characteristic or event.

Example 1

A statistic is often phrased like "3.6% of people in Australia are unemployed" or "around 44% of people in Australia went to the cinema between July 2021 and June 2022".

*If you want a better understanding of what "3.6% of people" means, check out the **Statistics: Sampling vs Population** helpsheet.*

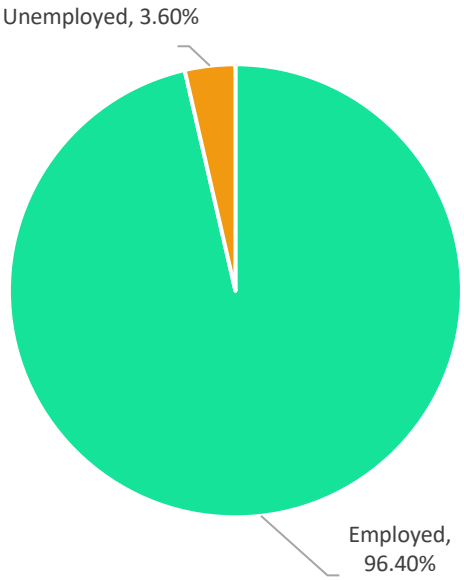

Figure 1 - Unemployment rate.

We can use these measures to inform our decision making.

A government might use the unemployment rate from *Example 1* as evidence that they need to create more jobs. A movie theatre might use attendance rates as evidence that they need to advertise their films more. This is one use of statistics; the other is to make predictions.

Predictions

One aim of the field of statistics is to be able to make accurate assumptions (or, guesses) about people or things based on a representative sample of those people or things.

Once we analyse the data, we can make certain statements about what is most likely to occur for a particular population of people or a certain event. That is, we can take the percentage statistic to be the *chance* that something will occur.

Example 2

Consider the above statistic:

Around 44% of people in Australia went to the cinema between July 2021 and June 2022.

We can then say that we have a 44% *chance*, if we choose someone at random from the Australian population, that they will have been to the cinema in that period. We can also *predict* that there is a 44% chance they have been to the cinema in the last year or so.

*If you want a better understanding of how the concept of "chance" works, check out the **Probability** helpsheet.*

In the next section, we'll look at a more in-depth case study of data collection and the useful things we can do with it.

Case Study

A teacher is trying to illustrate how to survey and collect data to their students. They are university students completing an introductory statistics class. The teacher decides to measure the height of each student, in centimetres (cm), to show off some useful measures of data. The data they collect from the 17 students in the class is as follows (in cm):

140, 172, 148, 153, 148, 173, 173, 172, 164, 172, 164, 173, 173, 184, 139, 159, and 164.

We can represent this data for ease of viewing within *Table 1*, shown on the right:

Note that the student number only represents the order in which we measured the students and shows nothing else about each student.

Some important information to be able to calculate about our sample data are:

- Range, Maximum, and Minimum ("*Measures of spread*")
- Mean (or, average), Median, and Mode ("*Measures of centre*")

Measures of spread describe how widely distributed (spread) our data is. Measures of centre describe the "midpoint" of our data which can be used as a representation of the entire set.

Student	Height (cm)
1	140
2	172
3	148
4	153
5	148
6	173
7	173
8	172
9	164
10	172
11	164
12	173
13	173
14	184
15	139
16	159
17	164

Table 1 – Table showing measured height in centimetres of students in the class.

Important

Consider the phrase "the average person". Usually this refers, not to any specific person, but the "typical" person.

"The average person likes to eat ice cream".

"The average Australian goes to the beach twice a year".

"Average", in an everyday sense, means "normal" or "typical". This does not mean every person behaves the same, but that "most people" would behave in this way "most of the time".

In Statistics, "average" is the same as "mean" and is calculated using a formula.

"The average of our data set is 6".

Let's first look at measures of spread, then at measures of centre.

Test your knowledge!

1. Consider how you might calculate the "minimum" or "maximum" of a data set. Can you think what the two terms might refer to? (Refer to the end to check your answer)

Measures of Spread

As we described above, "measures of spread" help to illustrate how widely distributed our data set is. This can help us to make predictions or judgements about where future data might lie if we ask a new, random person on the street or if we were to re-conduct the survey entirely.

The three important measures of spread we will look at are minimum, maximum, and range:

- Minimum/maximum: the smallest (min) and largest (max) values within your data set.
- Range: the difference between your minimum and maximum values. This describes the span or spread of your data.

Test your knowledge!

2. From our data, let's identify and calculate our measures of spread.

$$\text{Maximum} = 184\text{cm}\ (Student\ 14)$$
$$\text{Minimum} = 139\text{cm}\ (Student\ 15)$$
$$\text{Range} = \text{Max} - \text{Min}$$
$$= ??? - ???$$
$$= ???$$

(Refer to the end to check your answer)

Student	Height (cm)
15	139
1	140
3	148
5	148
4	153
16	159
9	164
11	164
17	164
2	172
8	172
10	172
6	173
7	173
12	173
13	173
14	184

Table 2 – Our original data set, reorganised from smallest to largest.

We can make useful judgements on our data set once we know the max, min, and range. For our case, we can see that the tallest person in our study is $184cm$ tall and that the shortest is $139cm$. This means that the height of students within this survey ranges across $45cm$ as this is the difference between the tallest and shortest.

Example 3

Imagine a government who surveys their population's annual income.

They might find that their lowest paid working citizen is paid $10,000 a year and that their highest paid working citizen is paid $100,000 a year, for a difference (or "*range*") of $90,000 a year.

That government might have a target range of $80,000. They might then use that range as justification to increase the wage of the lowest paid worker to $20,000, reducing the range to fit their target.

Measures of Centre

A measure of centre helps us to describe the "midpoint" of our data. The three measures of centre we will look at are mode, median, and mean:

- Mode: the most common data point, the one that is shown the most.
- Median: the middle data point when the data is ordered from highest to lowest.
- Mean: also called "average", the central value of all data. Calculated by adding all the data values up then dividing that total by the number of data points.

Important

Depending on context, our centres can be used as representative of the entire data set.

You may have heard of a "median house price" for a city or suburb. This price describes the "expected" price of a house for that region.

Let's calculate our measures of centre:

- Mode: this is our "most common" data value. We can count how many times each value appears in our data set. From Table 3, we can see that our mode is $173cm$.
- Median: By ordering our data from lowest to highest and checking the value which is in the "middle", we can see that our mode is $164cm$.
- Mean: We add up all data values from the sample then divide that total by 17, the number of data points. From that, we can see that our average is $163cm$, shown below:

Value	Frequency
139	1
140	1
148	2
153	1
159	1
164	3
172	3
173	4
184	1

Table 3 - Frequency table ("tally") of different data values, showing "173" as the most frequently occurring.

$$\mathbf{139 + 140 + 148 + 148 + 153 + 159 + 164 + 164 + 164 + 172 + 172 + 172 + 173 + 173 + 173 + 173 + 184 = 2271}$$

$$\frac{2271}{173} = 163cm$$

Next, we will look at how to put this data to use in a statistical summary.

Test your knowledge!

3. Our mode is 173cm as it appears 4 times in our data set. If it only appeared twice, it wouldn't be the mode anymore. Can you explain why?
4. Median is used for housing prices instead of mean. Can you explain why? Consider a town where there are 10 houses selling for $100,000 each and 1 house selling for $10,000,000. Why would median be the best measure of centre for that town?

(Refer to the end to check your answers)

Statistical Summaries

Now that we know our measures of centre, as well as our measures of spread, we can start to make statements about our data set that we can use to "predict" what might happen if we conducted our experiment a second time.

Summaries of statistical data usually look something like this:

"*17 students were sampled from a first-year university statistics class and their heights were recorded in centimetres. The tallest student measured 184cm in height while the shortest is 139cm, giving a range of 45. The mean, median, and mode are 163, 164, and 174, respectively. Therefore, the average student in this class is 163cm tall.*"

We can then predict that if we were to pick a student at random from another measured cohort, their height would be *around* 163cm. We can also compare the average height of this set with another sample and make judgements on each cohort based on their differences (e.g., "*Cohort 1 is, on average, taller than Cohort 2*").

Answers

Note: The final answer is displayed in **red.**

1. "Minimum" is the smallest value in our data set. "Maximum" is the largest value in the set. We work these out "by inspection" or by arranging the data in the set from smallest to largest and identifying the first and last value in the set.
2. To find the range, we subtract the Minimum value from the Maximum value.
   That is $184cm - 139cm$, which gives us a range of $45cm$.
3. $174cm$ is the mode as it appears 4 times, which is the most of any value. If it only appeared twice, then we would have two other values tied to be the mode, 164 and 172, with 3 occurrences each. We would call this a "*bimodal*" set.
4. Mean, as a measure of centre, is susceptible to one or two very large or very small numbers "throwing it off". In our example, where 10 houses are selling for $100,000 and 1 house is selling for $10,000,000, the mean price would be $1,000,000 which is not a very good representation of the housing market in that area. The median, however, would be $100,000, which is a great measure – so we can "expect" houses to cost about $100,000 for that area.

Related helpsheets

- Percentages helpsheet
- Fractions helpsheet