MAT220 Lecture 3: Central Tendency and Variation

Contents

In Lecture 1, we introduced the basic language of statistics, including data, variables, populations, samples and sampling methods. In Lecture 2, we studied how to represent data visually using histograms, bar charts, pie charts, time-series graphs and scatterplots. In this lecture, we begin the numerical side of descriptive statistics. Our main goal is to learn how to summarise a data set using a small number of carefully chosen statistics. We will focus on two major ideas: *central tendency*, which describes where the middle of the data is, and *variation*, which describes how spread out the data is.

1 Central Tendency

Throughout this lecture we will be working with quantitative data. We are going to need to talk about this data in a structured way, so we will first introduce some important notation. Throughout all of the formulae used in this lecture, we are going to write the data values using the letter x, labelled with subscripts:

$$x_1,\ x_2,\ x_3,\ x_4,\ \ldots,\ x_n.$$

The subscript simply gives us an easy label to refer to later on. Notice that the labels on the data start at 1 and count up until we reach n. This final number n (whatever it may be) tells us the total size of the data set. When order matters, we will write our data sets from smallest to largest, that is:

$$x_1 \leq x_2 \leq x_3 \leq x_4 \leq \cdots \leq x_n.$$

1.1 The meaning of central tendency

When we collect a lot of data, we often want to summarise it with a single representative number. One way to do this is to create a description of where the "middle" of the data is. This way, we can make claims about where the data points *tend to be*. This is exactly the idea of central tendency.

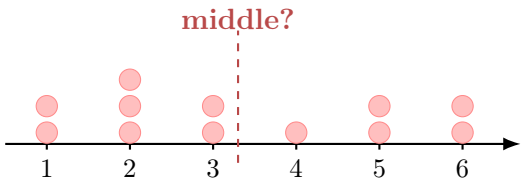
Definition: Central Tendency

Central tendency is the attempt to describe the middle of a data set.

For example, consider the data set

$$(1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6).$$

We can plot this data on a number line by representing each observation as a small ball:



In the diagram, should the middle of the data be this dashed line, or perhaps something else? The task of central tendency is to figure out a meaningful way to find the location of the middle. The intuition is clear, however the mathematical question is less simple, because there are many ways to interpret the word "middle". This is why there are several different measures of central tendency.

1.2 Summation notation

In the formulas we will later introduce, we are going to need to perform some arithmetic operations on data, which means that we also need to figure out how to write things down in an efficient way. To illustrate this, suppose that we were working with a data set of size $n = 1000$ and we wanted to add up all of the values of the data. In this situation, we don't want to keep writing out:

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + \ldots$$

by hand over and over again. To solve this issue, we instead use *summation notation.*

Summation notation

$$\sum_{i=1}^{n} x_i \quad \text{means} \quad x_1 + x_2 + x_3 + \cdots + x_n.$$

The letter i is the **index**. It tells us which term we are currently adding. The lower number tells us where to start, and the upper number tells us where to stop.

As you can see from the above, summation notation is simply an efficient way to write out a really big sum. It should be noted that we can use the notation to write out more than just basic addition of x_i's. For example, we may write:

$$\sum_{i=1}^{n} 2x_i = 2x_1 + 2x_2 + 2x_3 + \cdots + 2x_n.$$

In the above, we see that instead of summing up all the individual x_i values, we sum up all the terms $2x_i$.

Example. Suppose

$$(x_1, x_2, x_3) = (4, 10, 12).$$

Then

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 4 + 10 + 12 = 26,$$

$$\sum_{i=1}^{2} x_i = x_1 + x_2 = 4 + 10 = 14,$$

and

$$\sum_{i=2}^{3} x_i^2 = x_2^2 + x_3^2 = 10^2 + 12^2 = 100 + 144 = 244.$$

Exercise

Suppose

$$(x_1, x_2, x_3, x_4, x_5) = (2, 4, 6, 8, 10).$$

Compute:

(a) $\sum_{i=1}^{4} x_i$

(b) $\sum_{i=3}^{5} x_i$

(c) $\sum_{i=1}^{5} x_i$

(d) $\sum_{i=2}^{3} x_i$

(e) $\sum_{i=1}^{3} \frac{x_i}{2}$

Solution

(a) $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 2 + 4 + 6 + 8 = 20.$

(b) $\sum_{i=3}^{5} x_i = x_3 + x_4 + x_5 = 6 + 8 + 10 = 24.$

(c) $\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 2 + 4 + 6 + 8 + 10 = 30.$

(d) $\sum_{i=2}^{3} x_i = x_2 + x_3 = 4 + 6 = 10.$

(e) $\sum_{i=1}^{3} \frac{x_i}{2} = \frac{x_1}{2} + \frac{x_2}{2} + \frac{x_3}{2} = \frac{2}{2} + \frac{4}{2} + \frac{6}{2} = 1 + 2 + 3 = 6.$

1.3 Three common averages: mean, median and mode

In everyday language, the word "average" is ambiguous. It can refer to several different statistics. We will now review the three most common averages: the **mean**, the **median** and the **mode**. They are all measures of central tendency, but they answer slightly different questions.

Statistic	Basic idea	Main weakness
Mean	The balanced middle of the data.	Sensitive to outliers.
Median	The middle value after sorting.	Does not use the numerical size of every value.
Mode	The most frequent value.	May not be unique, and may not say much for numerical data.

We will now review these one-by-one, starting with the median.

1.3.1 Median

The median is the value that splits the whole data set in half. Once the median is found, at least half of the data lies at or below the median, and at least half lies at or above it.

Definition: Median

The **median** is the middle value of an ordered data set.

In order to find the median, we have to first pay attention to the value of n: if it is odd or even then we will have to perform slightly different operations. The step-by-step process is:

Step 1: Order all values from smallest to biggest.
Step 2: If n is odd, pick the middle entry. This is the $\frac{n+1}{2}$-th entry.
Step 3: If n is even, look at the two middle entries and calculate the number halfway between them.

For example, consider the data set:

$$(1, 2, 3, 4, 5).$$

Here there are 5 data points, i.e. $n = 5$. So, to find the median we add $+1$ to n and then halve it: $\frac{5+1}{2} = 3$. So, we simply go to the 3^{rd} entry in the data set and read off its value. In this case, the median is 3.

Consider now the data set:

$$(2, 3, 4, 5, 6, 7).$$

Here there are now 6 numbers, meaning that $n = 6$ which is even. The median position is $\frac{6+1}{2} = 3.5$. This means that the median will be a number that is precisely halfway between the 3^{rd} and 4^{th} entries in the data set. By inspection, we see that these two entries correspond to the data points 4 and 5. Therefore the median is halfway between these two numbers, i.e. the median is $\frac{4+5}{2} = 4.5$.

1.3.2 Mean

The mean is what we would ordinarily call the "average" of a data set. For a sample

$$x_1, x_2, \ldots, x_n,$$

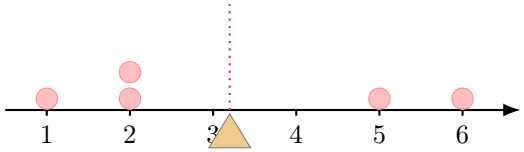
the mean is written $\bar{x}$ and is defined as follows.

Definition: Sample Mean

For a sample $x_1, x_2, \ldots, x_n$, the **sample mean** is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

In other words, the mean $\bar{x}$ is calculated by summing up all the values in the data set, and then dividing by how many there are. Geometrically, we can imagine the mean as a balancing point of the data. If you imagine each data point as a weight sitting on a number line, the mean is the point where the whole system balances. For example, for the data set $(1, 2, 2, 5, 6)$ we can imagine the mean as the balancing point in the middle:



Numerically, in this example we calculate:

$$\bar{x} = \frac{1+2+2+5+6}{5} = \frac{16}{5} = 3.2.$$

One important property of the mean is that the positive and negative deviations from the mean always cancel.

A useful fact about the mean

For any data set,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

In other words, the total signed distance from the data to the mean is always zero.

Keeping with the data set $(1, 2, 2, 5, 6)$ from above, we can verify this fact by hand. Observe that there are three values to the left of the mean $\bar{x} = 3.2$, namely $1, 2$ and 2, and there are two values greater than the mean, namely 5 and 6. Since the values to the left of $\bar{x}$ are smaller than 3.2, the values of $x_i - \bar{x}$ will be negative. In absolute value, the actual distances are

$$(3.2 - 1) + (3.2 - 2) + (3.2 - 2) = 2.2 + 1.2 + 1.2 = 4.6,$$

while for those data points to the right of $\bar{x}$ we have:

$$(5 - 3.2) + (6 - 3.2) = 1.8 + 2.8 = 4.6.$$

These values demonstrate that the sum of the distances to the left of the mean equals the sum of the distances to the right of the mean. This is why it makes sense to call the mean the balanced middle.

Alternatively, this fact about the mean may also be verified algebraically:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) \\ &= x_1 + x_2 + \cdots + x_n - \bar{x} - \bar{x} - \cdots - \bar{x} \\ &= x_1 + x_2 + \cdots + x_n - n \cdot \bar{x} \\ &= \sum_{i=1}^{n} x_i - n \cdot \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i \\ &= 0.\end{aligned}$$

1.3.3 Mode

Simply put, the mode is the most frequently occurring value in the data set.

Definition: Mode

The **mode** is the value that occurs most often in a data set.

In the case that there are competing data points that occur with the same highest frequency, we will say that the data set has no unique mode.

Example. The data set

$$(1, 1, 1, 3, 4)$$

has mode 1, because 1 occurs three times.
The data set

$$(1, 1, 2, 2, 3)$$

has no unique mode, because 1 and 2 both occur twice.

1.4 A worked example

Consider the data set

$$(1, 1, 2, 2, 3, 4, 5).$$

The mean is

$$\bar{x} = \frac{1 + 1 + 2 + 2 + 3 + 4 + 5}{7} = \frac{18}{7} \approx 2.57.$$

There are 7 data points, so the middle entry is the

$$\frac{7 + 1}{2} = 4,$$

so the median is the 4th entry. The 4th entry is 2, so the median is 2.

The values 1 and 2 both occur twice, and no value occurs more often. Therefore, there is no unique mode.

Exercise

Consider the data set

$$(1, 1, 2, 2).$$

Find:

(a) the mean,

(b) the median,

(c) the mode.

Solution

(a) The mean is

$$\bar{x} = \frac{1+1+2+2}{4} = \frac{6}{4} = 1.5.$$

(b) There are 4 data points, so the two middle values are the 2nd and 3rd entries. These are 1 and 2, so the median is

$$\frac{1+2}{2} = 1.5.$$

(c) The values 1 and 2 both occur twice. Therefore, there is no unique mode.

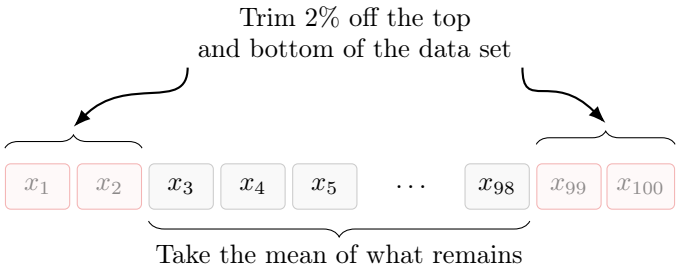
1.5 Trimmed means

The mean uses every value in the data set. This is often useful, but it also allows the mean to be affected by extreme outliers: if we had a single piece of data far off to one side, then the mean would have to compensate a lot in order to maintain its balance. In some situations this feature of the mean might be considered permissible. However, in other situations we might want to intentionally remove the effects of extreme outliers. This motivates the notion of a *trimmed mean*.

Definition: Trimmed Mean

An $x\%$ **trimmed mean** is obtained by removing the lowest $x\%$ and the highest $x\%$ of an ordered data set, and then calculating the mean of the remaining values.

Put differently, we create an $x\%$ trimmed mean by cutting off the bottom $x\%$ and the top $x\%$ of an ordered data set, and then taking the mean of what is left over. For example, if we have an ordered data set of size 100, a 2% trimmed mean would remove the two smallest values and the two largest values. It then takes the mean of the 96 values that are left over:



Trimmed means are especially helpful when trying to average out a collection of human judgements. For example, in figure skating, a panel of judges give scores to an individual performance. If one judge gives an unusually high score and another gives an unusually low score due to inherent bias or human error, then those extreme scores might affect the mean in an unfair way. To help regulate this effect, a performance score can be calculated with a trimmed mean to help reduce the influence of these extreme judgements.

Exercise: Trimmed means

Consider the ordered data set

$$1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24.$$

Here $n = 20$.

(a) Compute the 5% trimmed mean.

(b) Compute the 20% trimmed mean.

(c) Explain why a 50% trimmed mean is not meaningful here.

Solution

(a) Since 5% of 20 is 1, we remove the smallest value and the largest value. The remaining data are

$$1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23.$$

The sum of these values is 149, so the 5% trimmed mean is

$$\frac{149}{18} \approx 8.28.$$

(b) Since 20% of 20 is 4, we remove the lowest four values and the highest four values. The remaining data are

$$4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11.$$

The sum of these values is 84, so the 20% trimmed mean is

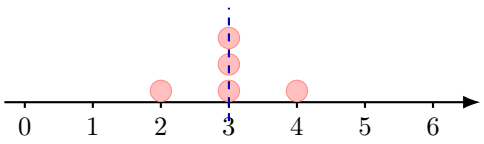
$$\frac{84}{12} = 7.$$

(c) Trimming 50% from the bottom and 50% from the top removes the entire data set. There are no values left to average, so a 50% trimmed mean is not meaningful.

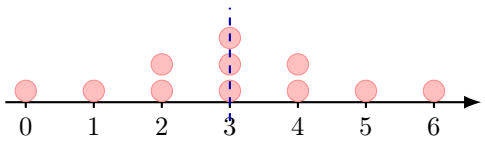
2 Variation

Measures of central tendency allow us to identify where our data tends to be, in relation to a middle. But, central tendency cannot tell us how spread out data is. For example, consider the two data sets drawn below:

$A = (2, 3, 3, 3, 4)$



$B = (0, 1, 2, 2, 3, 3, 3, 4, 4, 5, 6)$



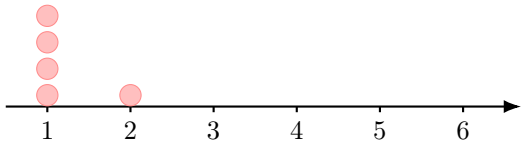
These two data sets have the same mean (both are 3), but data set B is clearly spread out much more than data set A. There is no sense in which an individual description of the centre of data will ever be able to capture the feature of spread. Instead, we need to introduce a new concept: variation.

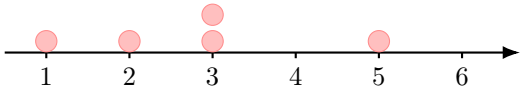
Definition: Variation

Variation describes how spread out the data is.

2.1 Range

Consider the two data sets, $A = (1, 1, 1, 1, 2)$ and $B = (1, 2, 3, 3, 5)$, which are drawn below.





Visually, it is very clear that B is more spread out, and A is more tightly packed. One way to quantify this is by looking at the distance between the largest and smallest value in the data sets: for A, the whole data set is contained between 1 and 2, whereas for B the whole data set spreads out between 1 and 5.

This simple observation motivates our first definition of spread: the *range* of a data set.

Definition: Range

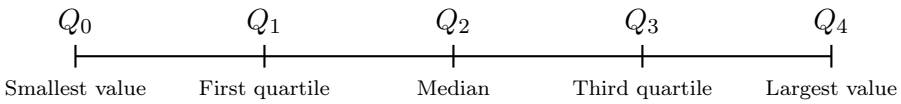
The **range** of a data set is

$$\text{range} = \max(\text{data}) - \min(\text{data}).$$

For example, our data set $A = (1, 1, 1, 1, 2)$ has range $2 - 1 = 1$, and our data set $B = (1, 2, 3, 3, 5)$ has range $5 - 1 = 4$. So, the range confirms what we saw visually: B is more spread out than A.

2.2 Quartiles

The word "quarter" means one fourth of something. In distinction to this, the word *quartile* means a value that helps divide an ordered data set into four quarters. Since there are four quarters in total, we need five lines to divide them up: three in the middle and two at the endpoints:



The five values

$$Q_0, \quad Q_1, \quad Q_2, \quad Q_3, \quad Q_4$$

mark the minimum, first quartile, median, third quartile and maximum.

Quartiles

For an ordered data set:

- Q_0 is the smallest value.
- Q_1 is the median of the bottom half of the data.
- Q_2 is the median of the whole data set.
- Q_3 is the median of the top half of the data.
- Q_4 is the largest value.

Different textbooks sometimes use slightly different conventions for quartiles, especially when the data set has an odd number of values. In this course, we will use the following practical rule: first find the median, then find the median of the lower half and the median of the upper half. If n is odd, we will not include the median in either half when finding Q_1 and Q_3.

2.3 Interquartile range

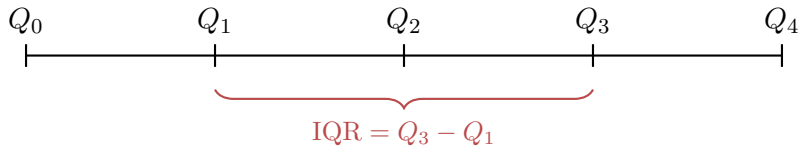
Definition: Interquartile Range

The **interquartile range**, abbreviated **IQR**, is

$$\text{IQR} = Q_3 - Q_1.$$

It measures the spread of the middle half of the data.

Visually, the interquartile range measures this distance:



The IQR is often more resistant to outliers than the range, because it focuses on the middle half of the data and ignores the lower and upper quarters.

2.3.1 Worked example: quartiles and IQR

Consider the ordered data set:

$$(1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24).$$

To calculate the quartiles, we will first find the median. Since there are 20 data points in total, the median will lie halfway between the 10th and 11th entries. These entries are 6 and 7, therefore, the median is $Q_2 = \frac{6+7}{2} = 6.5$. To find the value of Q_1, we need to calculate the median of the bottom half of the data. Writing this out, we are looking for the middle of the data:

$$(1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6).$$

The median of this bottom half lies halfway between 4 and 5, so

$$Q_1 = \frac{4+5}{2} = 4.5.$$

Similarly, the top half of the data is:

$$(7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24).$$

The median of this top half lies halfway between 10 and 11, so $Q_3 = \frac{10+11}{2} = 10.5$.

Putting this all together, the five-number summary is:

$$Q_0 = 1, \qquad Q_1 = 4.5, \qquad Q_2 = 6.5, \qquad Q_3 = 10.5, \qquad Q_4 = 24,$$

and the interquartile range can be calculated as:

$$\text{IQR} = Q_3 - Q_1 = 10.5 - 4.5 = 6.$$

2.4 Variance

The range only uses two numbers: the minimum and the maximum. This makes it quick to compute, but it also makes it blind to most of the data set. The range is also extremely sensitive to outliers. For instance, consider the following two data sets:

$$A = (1, 1, 1, 2, 2, 2, 3, 6) \qquad B = (1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6)$$

Both data sets have the same range: $6 - 1 = 5$. However, visually there is still a sense in which data set A is less spread out than data set B. In data set A most of the data is packed around $1, 2$ and 3, and there is a single outlier that causes the range to equal 5. In contrast, in data set B the data is much more uniformly spread, and there are no outliers.

We have just identified a fundamental flaw with the range of a data set: although it is useful as an initial measure of variation, the range is extremely sensitive to outliers and can therefore miss the

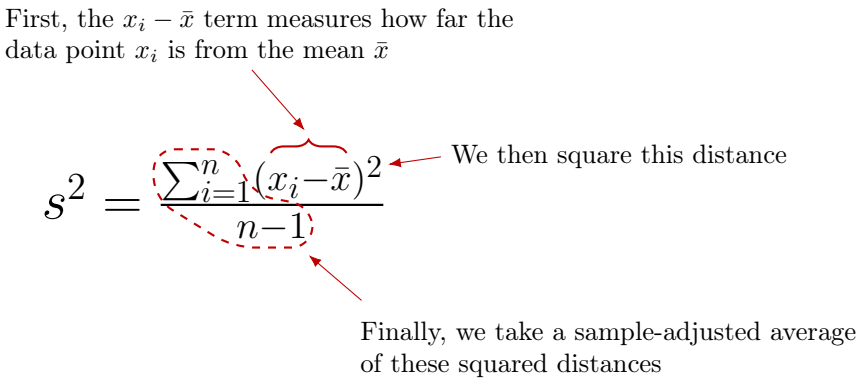
internal structure of a data set. To fix this, we will now introduce a more careful measure of spread known as *variance*.

Definition: Sample Variance

For a sample $x_1, x_2, \ldots, x_n$ with sample mean $\bar{x}$, the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

As you can see, the formula is a lot more complicated than those we have seen before. So, we will break this down step-by-step to explain what s^2 is actually trying to say. Below is an annotation of the formula:



In a nutshell, the variance s^2 measures a sample-adjusted average of the squared distances between each data value and the mean:

1. The expression $x_i - \bar{x}$ measures how far away the point x_i is from the centre $\bar{x}$. It is not quite a distance, because it can take negative values, but is instead called a *deviation*.
2. The expression $(x_i - \bar{x})^2$ then squares all of these deviations.
3. Next, the summation notation $\sum_{i=1}^{n}(x_i - \bar{x})^2$ says to add up all of these squared deviations, i.e.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2.$$

4. Finally, we divide by $n - 1$ to get the sample variance.

We will now elaborate on why *this* is the formula, in particular.

2.4.1 Why do we square the deviations?

It is reasonable to ask why the variance formula squares the deviations. Instead, why not just add up the raw deviations from the mean? In fact, the issue is hidden away in the important property of the mean that we saw earlier:

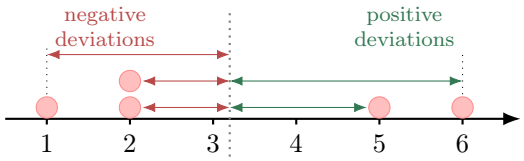
$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

This says that the sum of all the raw deviations will always equal zero. If we were to try to use the raw deviations without squaring them, then we would always get:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{0}{n-1} = 0,$$

no matter what data we started with. For example, consider the data set:

Data: $(1, 2, 2, 5, 6)$



whose mean is $\bar{x} = 3.2$. If we tried to add up the raw deviations, we would get:

$$(1 - 3.2) + (2 - 3.2) + (2 - 3.2) + (5 - 3.2) + (6 - 3.2) = 0.$$

Squaring solves this problem in general. Once we square a deviation, both negative and positive deviations have non-negative squares:

$$(-2)^2 = 4, \qquad 2^2 = 4.$$

This means that the sum is forced to be non-negative, and therefore avoids the problem mentioned above.[1]

2.4.2 Why do we divide by $n - 1$?

Normally, when we take an average, we add values and divide by the number of values. So it is natural to ask why the sample variance divides by $n - 1$ instead of n.

The short answer is that s^2 is actually a **sample statistic** and not a **population parameter**. Usually, we use a sample to estimate the spread of a larger population. When we use the sample mean $\bar{x}$ inside the formula, we have already estimated one quantity from the sample, and this makes the squared deviations slightly too small on average. Dividing by $n - 1$ corrects for this effect.

Bessel's correction

The use of $n - 1$ instead of n in the sample variance formula is called **Bessel's correction**. In this course, the main point is practical:

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

is the standard formula for sample variance.

We will not prove this correction in this course. For now, you should know which formula to use and what it is measuring.

[1] Also, squaring gives especially large weight to deviations with large absolute value.

2.4.3 A worked example: variance

Example. Consider the data set

$$(1, 1, 2, 3).$$

First compute the mean:

$$\bar{x} = \frac{1+1+2+3}{4} = \frac{7}{4} = 1.75.$$

Now compute the deviations and squared deviations:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-0.75	0.5625
1	-0.75	0.5625
2	0.25	0.0625
3	1.25	1.5625

The sum of the squared deviations is

$$0.5625 + 0.5625 + 0.0625 + 1.5625 = 2.75.$$

Since $n = 4$, we divide by $n - 1 = 3$:

$$s^2 = \frac{2.75}{3} \approx 0.917.$$

2.5 Standard deviation

Variance is useful, but it has one awkward feature: it is measured in squared units. For example, if the data is measured in minutes, then the variance is measured in minutes squared. This quantity can be hard to directly relate to our initial data. We can solve this issue by simply taking the square root of the variance, so the result will be in the same units as we started with. This is precisely the definition of the standard deviation.

Definition: Sample Standard Deviation

The **sample standard deviation** is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

The standard deviation is in the same units as the data. Because of this, it is easier to interpret as a distance from the mean.

In the previous example, we found that

$$s^2 \approx 0.917.$$

Therefore,

$$s = \sqrt{0.917} \approx 0.958.$$

So the standard deviation is approximately 0.958.

2.6 Coefficient of variation

Sometimes we want to compare the spread of data sets with very different scales. For example, a standard deviation of 2 is large if the mean is 4, but very small if the mean is 2000. To make the spread *relative* to the size of the data set, we can use the coefficient of variation.

Definition: Coefficient of Variation

For positive ratio-scale data, the coefficient of variation is a unitless measure of relative spread. It is defined as the ratio:

$$\text{CV} = \frac{s}{\bar{x}}.$$

A larger coefficient of variation means more variation compared with the typical size of the values.

This statistic is most useful when the mean is positive and not close to zero. If the mean is very close to zero, dividing by the mean can make the coefficient of variation misleading or unstable.

2.7 Sample statistics and population parameters

As discussed in Lecture 1, inferential statistics often uses sample statistics to estimate population parameters. Because of this, we use different notation for samples and populations.

Quantity	**Sample statistic**	**Population parameter**
Mean	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
Variance	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
Standard deviation	$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

Here, n is the sample size, while N is the population size. Observe that for the sample variance and sample standard deviation we have used Bessel's correction. But we do not need to do that for the population variance or population standard deviation, because we are no longer trying to make an estimate.

3 Box and Whisker Plots

3.1 The five-number summary

A box and whisker plot is a visual representation of the five-number summary.

Five-number summary

The **five-number summary** of a data set is:

$$\text{lowest value}, \quad Q_1, \quad \text{median}, \quad Q_3, \quad \text{highest value}.$$

Equivalently,

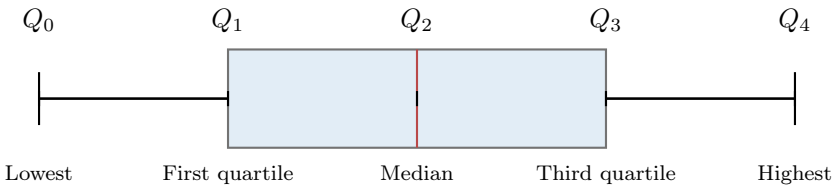
$$Q_0, \quad Q_1, \quad Q_2, \quad Q_3, \quad Q_4.$$

Box and whisker plots are useful because they show both central tendency and variation at the same time. The median shows the middle of the data, the box shows the middle half of the data, and the whiskers show the outer parts of the data.

3.2 Anatomy of a box and whisker plot

In the simple version of a box and whisker plot used in this lecture:

- the left whisker begins at the lowest value Q_0,
- the left side of the box is Q_1,
- the line inside the box is the median Q_2,
- the right side of the box is Q_3,
- the right whisker ends at the highest value Q_4.



The box runs from Q_1 to Q_3. The whiskers run from Q_0 to Q_1 and from Q_3 to Q_4.

Definition: Box and Whisker Plot

In the simple version used in this lecture, a **box and whisker plot** is a graph that displays a data set using its five-number summary: lowest value, Q_1, median, Q_3 and highest value.

3.3 Constructing a box and whisker plot

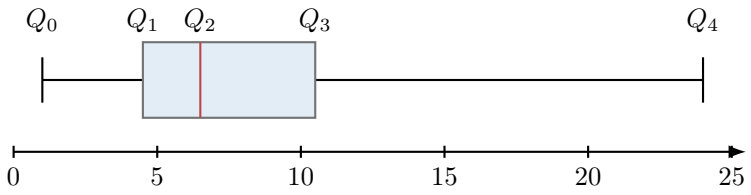
Consider the data set:

$$1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24.$$

We have previously determined the five-number summary:

$$Q_0 = 1, \qquad Q_1 = 4.5, \qquad Q_2 = 6.5, \qquad Q_3 = 10.5, \qquad Q_4 = 24.$$

The resulting box and whisker plot is:



Notice that the right whisker is much longer than the left whisker. This tells us that the upper end of the data is more spread out than the lower end.