Statistics

Statistics is a discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data.

Descriptive statistics

Descriptive statistics are brief descriptive coefficients (a numerical or constant quantity placed before and multiplying the variable in an algebraic expression (e.g. *4* in $4x^{y}$))that summarize a given data set, which can be either a representation of the entire or a sample of a population.
Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).

Measures of Central Value

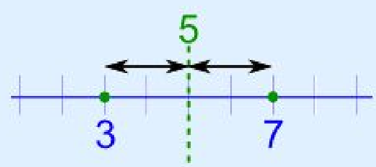
When you have two or more numbers it is nice to find a value for the "**center**".

2 Numbers

With just 2 numbers the answer is easy: go half-way between.

Example: what is the central value for 3 and 7?

Answer: Half-way between, which is 5.



You can calculate it by adding 3 and 7 and then dividing the result by 2:

→ (3+7) / 2 = 10/2 = 5

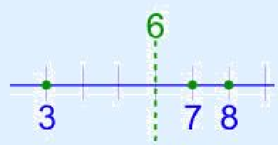
3 or More Numbers

We can use that idea of "adding then dividing" when we have 3 or more numbers:

Example: what is the central value of 3, 7 and 8?

Answer: You calculate it by adding 3, 7 and 8 and then dividing the results by 3 (because there are 3 numbers):

$(3+7+8) / 3 = 18/3 = 6$



Notice that we divide by 3 because we have 3 numbers ... very important!

The Mean

The mean is the **average** of the numbers.

It is easy to calculate: **add up** all the numbers, then **divide by how many** numbers there are.

In other words it is the **sum** divided by the **count**.

Example 1: What is the Mean of these numbers?

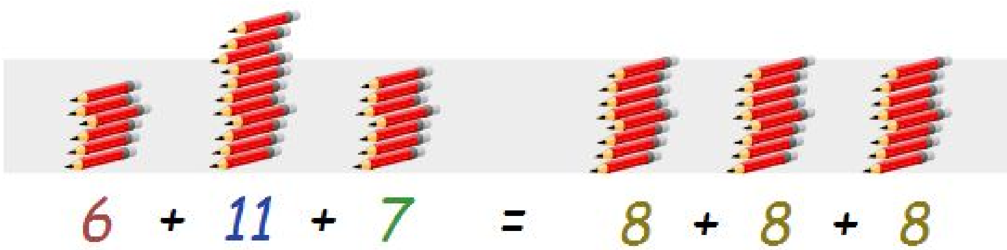
6, 11, 7

- Add the numbers: **6 + 11 + 7 = 24**
- Divide by *how many* numbers (there are 3 numbers): **24 / 3 = 8**

The Mean is 8

Why Does This Work?

It is because 6, 11 and 7 added together is the same as 3 lots of 8:



It is like you are "flattening out" the numbers

Example 2: Look at these numbers:

3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

The sum of these numbers is 330

There are fifteen numbers.

The mean is equal to 330 / 15 = 22

The mean of the above numbers is 22

Negative Numbers

How do you handle negative numbers? Adding a negative number is the same as subtracting the number (without the negative). For example 3 + (−2) = 3−2 = 1.

Knowing this, let us try an example:

Example 3: Find the mean of these numbers:

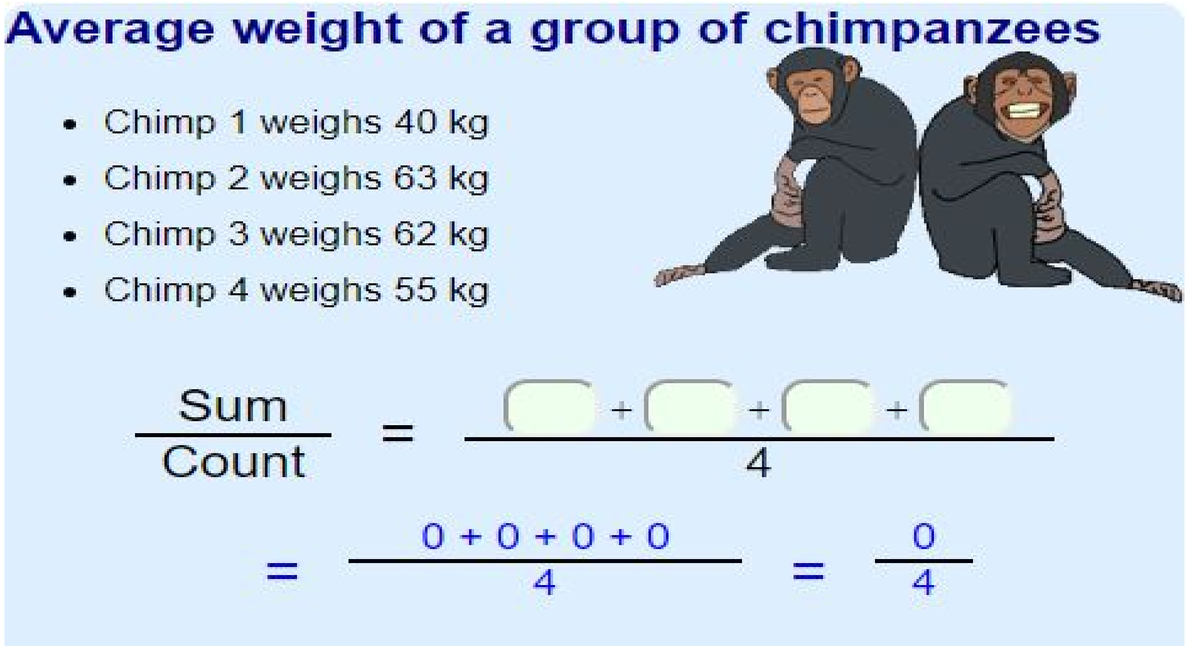
$$3, -7, 5, 13, -2$$

- The sum of these numbers is $\mathbf{3 - 7 + 5 + 13 - 2 = 12}$
- There are **5** numbers.
- The mean is equal to $\mathbf{12 \div 5 = 2.4}$

The mean of the above numbers is 2.4

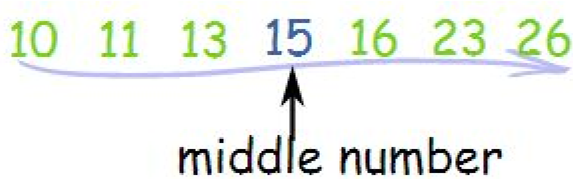
Here is how to do it one line:

$$\text{Mean} = \frac{3 - 7 + 5 + 13 - 2}{5} = \frac{12}{5} = \mathbf{2.4}$$



Median Value

The Median is the ***"middle"*** of a sorted list of numbers.



How to Find the Median Value

To find the Median, place the numbers in **value order** and find the **middle**.

Example: find the Median of 12, 3 and 5

Put them in order:

3, 5, 12

The middle is **5**, so the median is **5**.

Example:

3, 13, 7, 5, 21, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

When we put those numbers in order we have:

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 39, 40, 56

There are **fifteen** numbers. Our middle is the **eighth** number:

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 39, 40, 56

The median value of this set of numbers is **23**.

(It doesn't matter that some numbers are the same in the list.)

Two Numbers in the Middle

BUT, with an **even amount of numbers** things are slightly different.

In that case we find the **middle pair** of numbers, and then find the value that is **halfway** between them. This is easily done by adding them together and dividing by two.

Example:

3, 13, 7, 5, 21, 23, 23, 40, 23, 14, 12, 56, 23, 29

When we put those numbers in order we have:

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56

There are now **fourteen** numbers and so we don't have just one middle number, we have a **pair of middle numbers**:

3, 5, 7, 12, 13, 14, **21**, **23**, 23, 23, 23, 29, 40, 56

In this example the middle numbers are **21 and 23**.

To find the value halfway between them, add them together and divide by 2:

21 + 23 = 44
then **44 ÷ 2 = 22**

So the **Median** in this example is **22**.

(Note that 22 was not in the list of numbers ... but that is OK because half the numbers in the list are less, and half the numbers are greater.)

Where is the Middle?

A quick way to find the middle: **count how many numbers, add 1 then divide by 2**

Example: There are 45 numbers

45 plus 1 is 46, then divide by 2 and we get **23**

So the median is the **23rd number** in the sorted list.

Example: There are 66 numbers

66 plus 1 is 67, then divide by 2 and we get **33.5**

33 and a half? That means that the **33rd and 34th** numbers in the sorted list are the two middle numbers.

So to find the median: add the **33rd and 34th** numbers together and divide by 2.

Question

What is the median of the numbers 4, 2, 11, 6, 2, 9 ?

Answer

Put the numbers in order first: 2, 2, **4, 6**, 9, 11

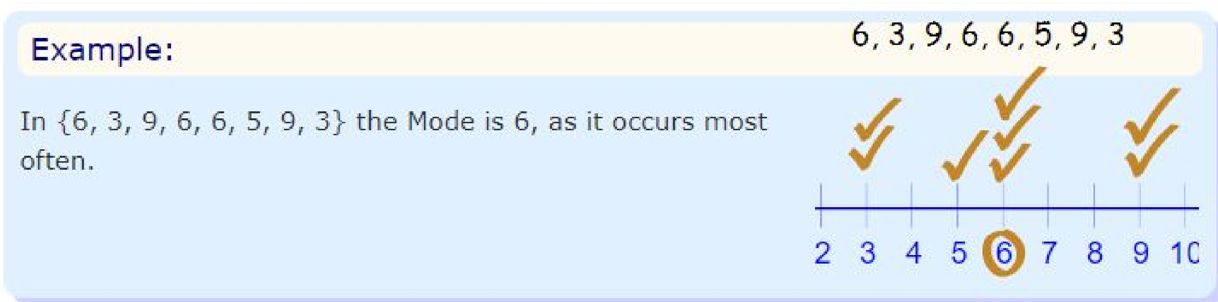
There are two numbers in the middle: 4 and 6.

The average of 4 and 6 is (4+6)/2 = 10/2 = 5

So the median is 5

The Mode

"The mode is simply the number which appears **most often**."



Finding the Mode

To find the mode, or modal value, it is best to put the numbers **in order**. Then **count** how many of each number. A number that appears **most often** is the **mode**.

Example:

3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

In order these numbers are:

3, 5, 7, 12, 13, 14, 20, **23, 23, 23, 23**, 29, 39, 40, 56

This makes it easy to see which numbers appear **most often**.

In this case the mode is **23**.

Another Example: {19, 8, 29, 35, 19, 28, 15}

Arrange them in order: **{8, 15, 19, 19, 28, 29, 35}**

19 appears twice, all the rest appear only once, so **19 is the mode**.

How to remember? Think "mode is most"

More Than One Mode

We can have more than one mode.

Example: {1, 3, 3, 3, 4, 4, 6, 6, 6, 9}

Example: {1, 3, 3, 3, 4, 4, 6, 6, 6, 9}

3 appears three times, as does 6.

So there are two modes: at **3** and **6**

Having two modes is called **"bimodal"**.

Having more than two modes is called **"multimodal"**.

Grouping

In some cases (such as when all values appear the same number of times) the mode is not useful. But we can **group** the values to see if one group has more than the others.

Example: {4, 7, 11, 16, 20, 22, 25, 26, 33}

Example: {4, 7, 11, 16, 20, 22, 25, 26, 33}

Each value occurs once, so let us try to group them.

We can try groups of 10:

- 0-9: **2 values** (4 and 7)
- 10-19: **2 values** (11 and 16)
- 20-29: **4 values** (20, 22, 25 and 26)
- 30-39: **1 value** (33)

In groups of 10, the "20s" appear most often, so we could choose **25** (the middle of the 20s group) as the mode.

You could use different groupings and get a different answer.

Grouping also helps to find what the typical values are when the real world messes things up!

Example: How long to fill a pallet?

Philip recorded how long it takes to fill a pallet in minutes:

{35, 36, 32, 42, 58, 56, 35, 39, 46, 47, 34, 37}

It takes longer when there is break time or lunch so an average is not very useful.

But grouping by 5s gives:

- 30-34: **2**
- 35-39: **5**
- 40-44: **1**
- 45-49: **2**
- 50-54: **0**
- 54-59: **2**

"35-39" appear most often, so we can say it normally takes **about 37 minutes** to fill a pallet.

Question

For the numbers 13, 16, 12, 11, 8, 14, 12 and 18

Answer

The mean = (13+16+12+11+8+14+12+18) ÷ 8 = 104 ÷ 8 = 13

To easily find the median and mode, arrange the numbers in order first:

8, 11, 12, 12, 13, 14, 16, 18

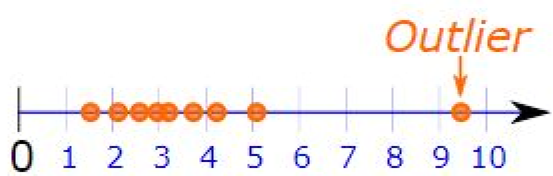
There are two "middle numbers", so the median is the

average of 12 and 13 = (12 + 13) ÷ 2 = 25 ÷ 2 = 12.5

And 12 occurs most often so the mode is 12

Therefore mean > median > mode

Outliers



Outliers are values that **"lie outside**" the other values.

They can change the mean a lot, so we can either not use them (and say so) or use the median or mode instead.

Example: 3, 4, 4, 5 and 104

Mean: Add them up, and divide by 5 (as there are 5 numbers):

(3+4+4+5+104) / 5 = **24**

24 does not represent those numbers well at all!

Without the 104 the mean is:

(3+4+4+5) / 4 = **4**

But please tell people you are not including the outlier.

Median: They are in order, so just choose the middle number, which is **4**:

3, 4, **4**, 5, 104

Mode: 4 occurs most often, so the Mode is **4**

3, **4, 4**, 5, 104

Harmonic Mean

The harmonic mean is:

the reciprocal of the arithmetic mean of the reciprocals

("Reciprocal" just means $\frac{1}{\text{value}}$)

The formula is:

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \dots}$$

Where **a,b,c,...** are the values, and **n** is how many values.

Steps:

- Calculate the reciprocal (1/value) for every value.
- Find the average of those reciprocals (just add them and divide by how many there are)
- Then do the reciprocal of that average (=1/average)

Example: What is the harmonic mean of 1, 2 and 4?

The reciprocals of 1, 2 and 4 are:

$$\frac{1}{1} = 1, \quad \frac{1}{2} = 0.5, \quad \frac{1}{4} = 0.25$$

Now add them up:

$$1 + 0.5 + 0.25 = 1.75$$

Divide by how many:

$$\text{Average} = \frac{1.75}{3}$$

The reciprocal of that average is our answer:

$$\text{Harmonic Mean} = \frac{3}{1.75} = \mathbf{1.714} \text{ (to 3 places)}$$

Another way to think of it

We can rearrange the formula above to look like this:

$$\frac{n}{\text{Harmonic Mean}} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \ldots$$

It is *not* easy to use this way, but it does look more "balanced" (**n** on one side matched with n **1**s on the other, and the mean matched with the values too).

Application OF Harmonic Mean

Harmonic means are often used in averaging things like rates (e.g., the average travel speed given a duration of several trips).

The weighted **harmonic mean** is used in finance to average multiples like the price-earnings ratio because it gives equal weight to each data point.

Conclusion

There are other ways of measuring central values, but **Mean, Median and Mode** are the most common.

Use the one that best suits your data. Or better still, use all three!