

EMPL4005

INFORMATION, DATA AND ANALYSIS

Descriptive Statistics

2. Dispersion

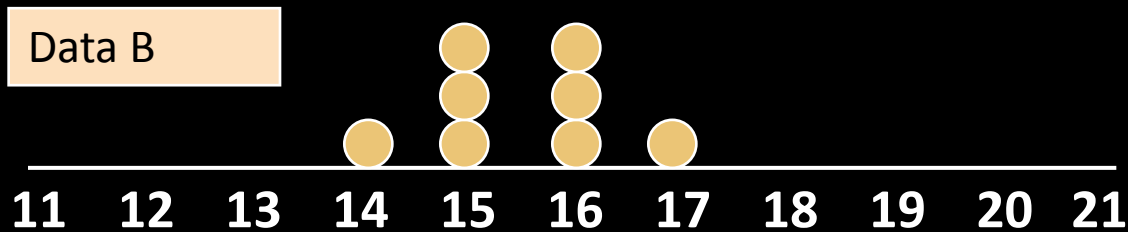
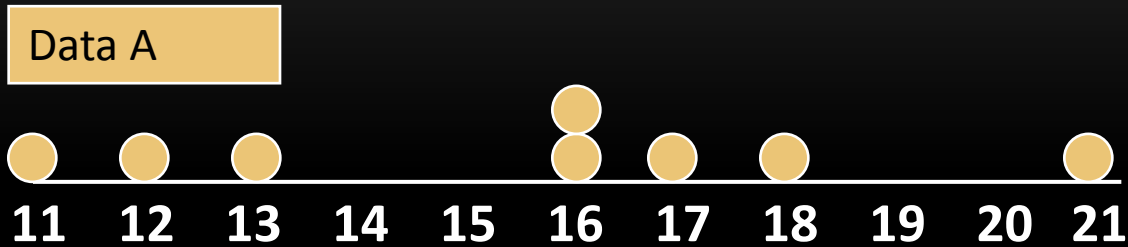
Dispersion

- What we mean by **dispersion** is how far have individual data values strayed from the mean.
- We will look at:
 - Range
 - Quartiles
 - Standard deviation (and variance)

Why use dispersion?

- A measure of **central tendency** (mean, mode, median) only tells us part of the story about our data.
- We lose useful information about how the data is distributed or spread.

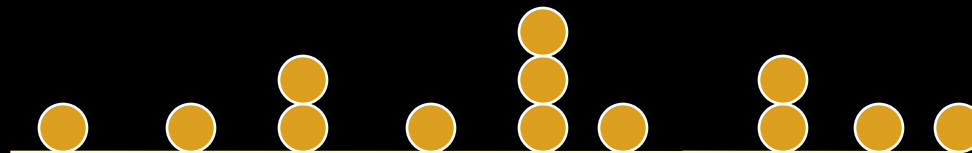
Different distributions with the same mean



The Range

- The **range** summarises the overall spread of the data. The data is spread between the minimum and maximum values.
- The range is the simplest measure of variation.
- It equals the difference between the largest and the smallest values in a set of data:

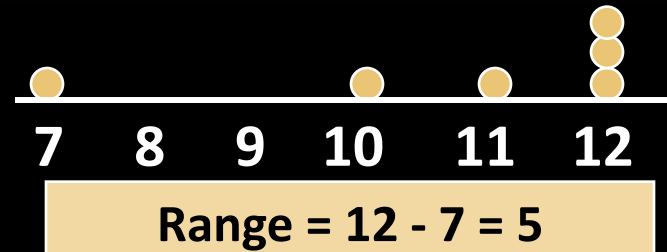
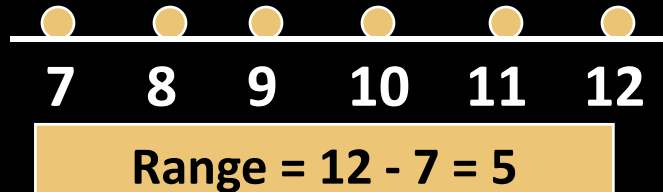
$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$



$$\text{Range} = 14 - 1 = 13$$

Disadvantages of the Range

- It ignores the way in which data are distributed:
- Also, it is sensitive to outliers (extremes):



1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,5

$$\text{Range} = 5 - 1 = 4$$

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,120

$$\text{Range} = 120 - 1 = 119$$

Calculate the range

a.

7 9 8 11 4

b.

21.49 31.50 11.35 13.80 15.47 18.37 38.16
18.47 17.48 18.61 9.77 20.72 22.47 12.46

Answers

a.

- Largest value = 11
- Smallest value = 4
- Range: largest value - smallest value
- Range = $11 - 4 = 7$

b.

Largest value = 38.16

Smallest value = 9.77

Range: largest value - smallest value

Range = $38.16 - 9.77 = 28.39$

Quartiles

- The problem with the range is its lack of detail. The use of **quartiles** provides more information.

Quartiles

- The *first quartile* is the value $\frac{1}{4}$ of the way through the data
- The *second quartile* is the value half way through data (same as the **median value**).
- The *third quartile* is the value $\frac{3}{4}$ of the way through the data.
- The *fourth quartile* is the upper value of the range.

Interquartile Range

- We can eliminate some outlier problems by using the **inter-quartile range**
- Eliminate some high and low valued observations and calculate the range from the remaining values

$$\text{Interquartile range} = 3^{\text{rd}} \text{ quartile} - 1^{\text{st}} \text{ quartile} = Q_3 - Q_1$$

Example

9 5 3 10 14 6 12 7 14

Step 1

Arrange the data in ascending order

3 5 6 7 9 10 12 14 14

Step 2

Find the median (2nd quartile)

3 5 6 7 9 10 12 14 14

Step 3

Find the median of the lower half of the data set (Q1).

3 5 6 7 9 10 12 14 14 = 5.5

Step 4

Find the median of the upper half of the data set (Q3).

3 5 6 7 9 10 12 14 14 = 13

Interquartile Range

- The inter-quartile range is the difference between the first quartile to the third quartile and represents the central 50% of the data.
- In our example, the inter-quartile range is
 - $Q3 - Q1$
 - $13 - 5.5 = 7.5$

Calculate the Interquartile Range

a.

9.77 11.35 12.46 13.80 15.47 17.48 18.37
18.47 18.61 20.72 21.49 22.47 31.50 38.16

b.

9.77 11.35 12.46 13.80 15.47 17.48 18.37
18.47 18.61 20.72 21.49 22.47 38.16

Answers

Step 1

Arrange the data in ascending order

9.77 11.35 12.46 13.80 15.47 17.48 18.37
18.47 18.61 20.72 21.49 22.47 31.50 38.16

Step 2

Find the median (2nd quartile)

9.77 11.35 12.46 13.80 15.47 17.48 18.37
18.47 18.61 20.72 21.49 22.47 31.50 38.16

(Add 18.37 + 18.47 then divide by 2 = 18.42)

Answers

Step 3

Find the median of the lower half of the data set (Q1)

- 9.77 11.35 12.46 13.80 15.47 17.48
- Add $12.46 + 13.80$ together then divide by 2 = 13.13

Step 4

Find the median of the upper half of the data set (Q3)

- 18.61 20.72 21.49 22.47 31.50 38.16
- Add $21.49 + 22.47$ together then divide by 2 = 21.98
- The inter-quartile range is
 - $Q3 - Q1$
 - $21.98 - 13.13 = 8.85$

Standard Deviation and Variance

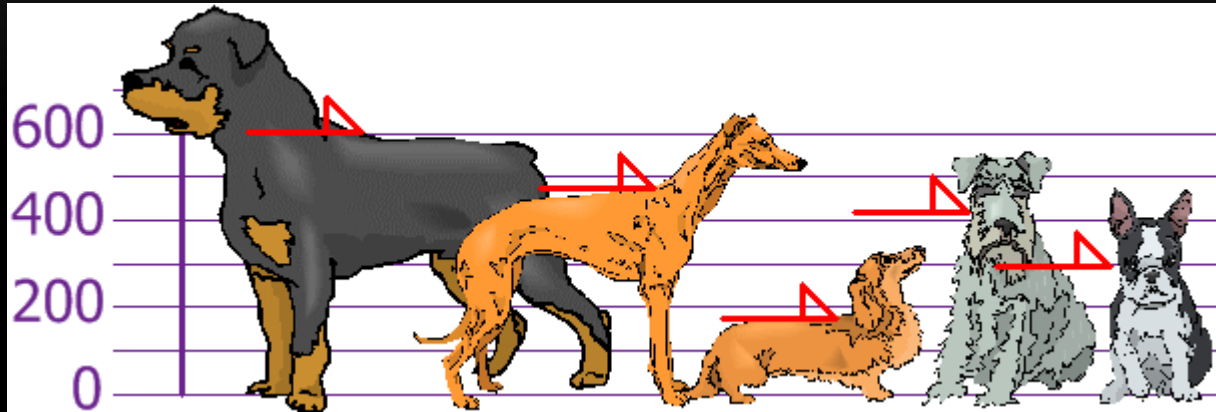
- The Standard Deviation is a measure of how spread out numbers are.
- Its symbol is σ (the Greek letter sigma)
- The formula is the **square root** of the **Variance**.
- So now you ask, "What is the Variance?"

Variance

- The Variance is defined as:
- The average of the **squared** differences from the Mean.
- To calculate the variance follow these steps:
- Work out the Mean (the simple average of the numbers)
- Then for each number: subtract the Mean and square the result
(the *squared difference*).
- Then work out the average of those squared differences.

Example (1)

- You and your friends have just measured the heights of your dogs (in millimetres):



The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

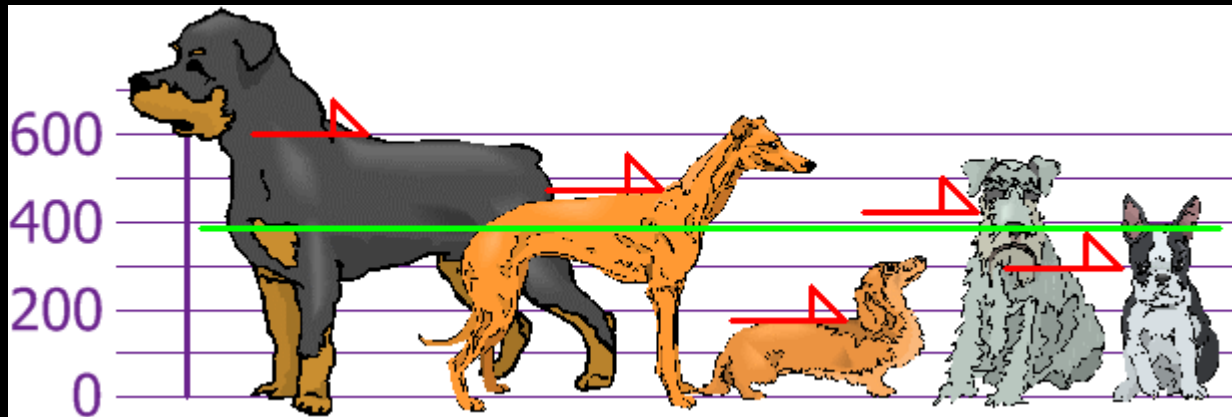
Find out the Mean, the Variance, and the Standard Deviation.

Your first step is to find the Mean:

Example (2)

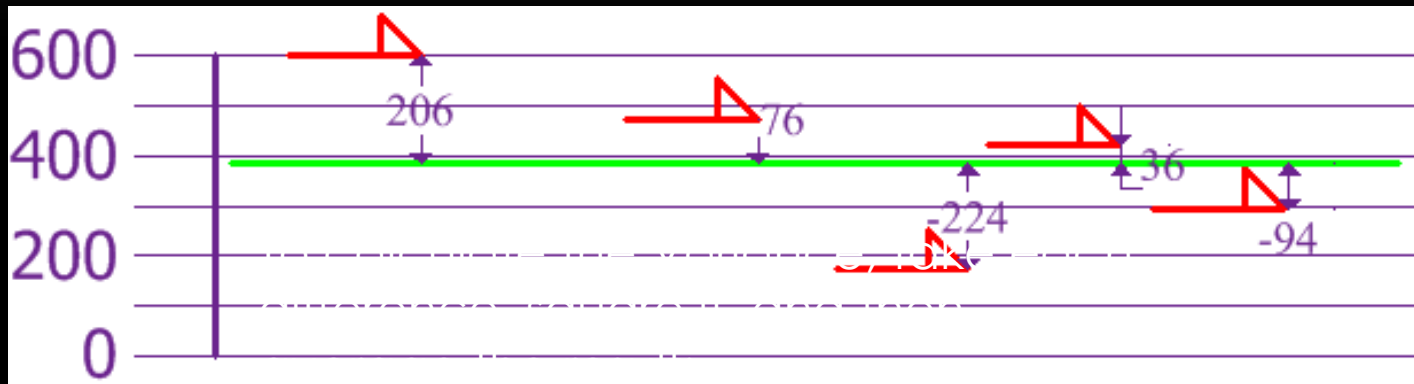
Mean =	$600 + 470 + 170 + 430 + 300$	=	1970	= 394
	5		5	

So the mean (average) height is 394 mm.
Let's plot this on the chart:



Example (3)

Now, we calculate each dog's difference from the Mean:



To calculate the Variance, take each difference, square it, and then average the result:

Example (4)

Variance

$$\begin{aligned}\text{Variance: } \sigma^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\ &= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5} \\ &= \frac{108,520}{5} = 21,704\end{aligned}$$

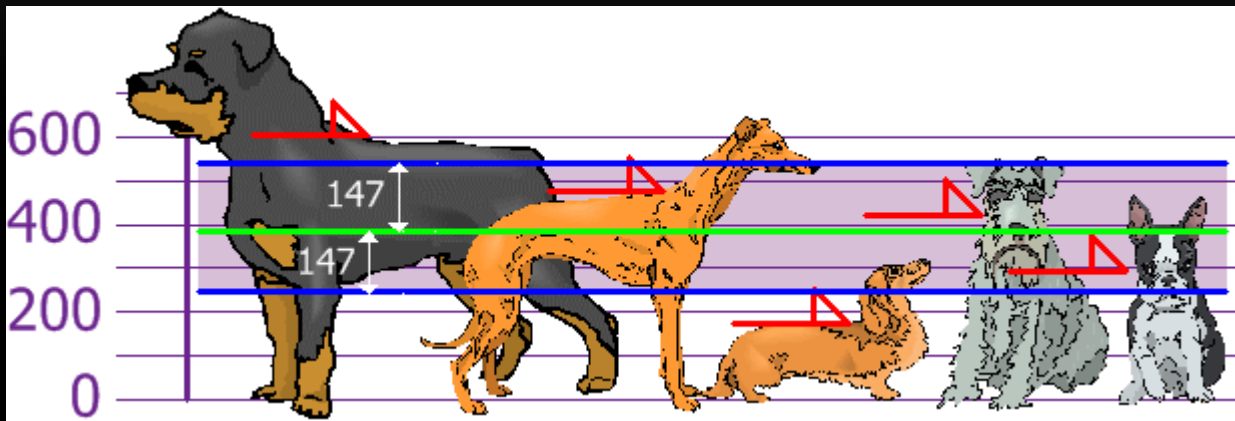
So, the Variance is **21,704**.

And the Standard Deviation is the square root of Variance, so:

Standard Deviation: $\sigma = \sqrt{21,704} = 147.32... = 147$ (to the nearest mm)

So why is this useful?

- Now we can show which heights are within one Standard Deviation (147mm) of the Mean:



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small. Rottweilers **are** tall dogs. And Dachshunds **are** a bit short ... but don't tell them!

Sample Data

- Our example was for a **Population** (the 5 dogs were the only dogs we were interested in).
- But if the data is a **Sample** (a selection taken from a bigger Population), then the calculation changes.
- When you have "N" data values that are:
- **The Population**: divide by **N** when calculating Variance (like we did)
- **A Sample**: divide by **N-1** when calculating Variance

Sample Data

- All other calculations say the same, including how we calculated the mean.
- Example: if our 5 dogs were just a **sample** of a bigger population of dogs, we would divide by **4 instead of 5** like this:
- Sample Variance = $108,520 / 4 = 27,130$
- Sample Standard Deviation = $\sqrt{27,130} = \mathbf{164}$ (to the nearest mm)
- Think of it as a "correction" when your data is only a sample.

Standard Deviation Calculator

- If you are not so hot with using formulae, the wonderful website (cited below) that has helped us demystify statistics (a bit!) also has a **Standard Deviation Calculator** that you can use to work out the calculations for you.
- Enter your data set (the numbers you want to analyse) and watch it calculate in real time.
- [Standard Deviation Calculator](http://www.mathsisfun.com/data/standard-deviation.html) (bookmark this link!)



Summary

- The **range** of a data set is the difference between the smallest and largest values.
- The **variance** of the data summarises the **squared deviation** of each data value from the **mean**.
- The **standard deviation** of a data set is the squared root of the variance.

Some extra help.....

- Here are a few good links to help you get started with Excel:
- [Excel Tutorials for Beginners](#)
- [How to use Formulae in Excel](#)
- [Excel Handbook 1](#)
- [Excel Handbook 2](#)

This website is a great resource for all levels of ability and prior knowledge.

- [Using and Handling Data](#)