Statistical analysis 1

Introduction

Scientists use statistics to help them analyse and understand the evidence they collect during experiments. Statistics is an area of mathematics that measures variation in data and the di erences and relationships between sets of data. By using statistics we can examine samples of populations or experimental results and decide how certain we can be about the conclusions we draw from them.

1.1 Mean and distribution

Assessment statements State that error bars are a graphical representation of the variability of data. Calculate the mean and standard deviation of a set of values. State that the term standard deviation is used to summarise the spread of values around the mean and that 68% of all values fall within plus or minus one standard deviation of the mean. Explain how the standard deviation is useful for comparing the means and the spread of data between two or more samples. Deduce the significance of the di erence between two sets of data using calculated values for *t* and the appropriate tables. Explain that the existence of a correlation does not establish that there is a causal relationship between two variables.

Calculating the mean

The **mean** is an average of all the data that has been collected. For example if you measure the height of all the students in your class you could find the mean value by first calculating the sum of all the values and then dividing by the number of values.

$x \frac{\Sigma x}{n}$

A normal distribution

If you measured the height of ten students and plotted the values on a graph with the height on the *x*raxis and the number in each height group on the *y*raxis you would get a result that did not show an obvious trend. If you measured the height of 100 students your graph would begin to look bell shaped with most values in the middle and fewer on either side. As you measured more and more students heights your graph would eventually become a smooth curve as in Figure 1.1 (overleaf). This bellr shaped graph of values is called a **normal distribution**.

Normal distribution curves may be tall and narrow if all the values are close together or flatter and wider when the data are more spread out. The mean value is at the peak of the curve.

Mathematical vocabulary

You will nd it helpful to understand some mathematical vocabulary rst.

x represents a single value for example a person s height *x* 1.7 m *n* represents the total

number of values in a set
x called x bar represents the mean of a set of values Σ represents the sum of the values s represents the standard deviation of the sample ± represents plusrorrminus t

is a value calculated using a formula called the $t$rtest

2

Statistical mean

analysis using standard deviation or the $t$rtest requires a spread of data that is close to a normal distribution. This is why when measuring samples from a population it is best to get as many samples as possible.

Standard deviation

The **standard deviation** shows the spread of all the values around the mean and therefore it has the same units as the values. Value (for example, height)

In a normal distribution 68% of all the values in a sample fall within **Figure 1.1** A normal distribution.

±1 standard deviation of the mean and this increases to 95% within ±2 standard deviations of the mean (Figure **1.2**).

mean mean

68% 1s

68%

95%

2s 95%

Value

Value **Figure 1.2** Two different normal distribution curves for both, 68% of the values fall within 1 standard deviation of the mean, and 95% fall within 2 standard deviations.

These percentages are the same for all shapes of normal distribution Any observable di erence between curves. The standard deviation tells us how much the data spreads out each a characteristic in a species is called

side of the mean that is whether the distribution is tall and narrow or **variation**.

wide and flat and this allows us to compare sets of data.

1 In Figure **1.2**, which sample shows the greatest variation?

Calculating the standard deviation

Check that you know how to

The symbol for standard deviation is s. Remember that s is normally calculate standard deviation on calculated for a sample from the population. Standard deviation is calculated your calculator.

by entering the data into a scientific or graphical calculator or a spreadsheet.

2 The length of the index finger of five students was measured. Calculate the standard deviation.

Finger length /mm

12 18 15 16 14

Using the standard deviation

The standard deviation can be used to give more information about di erences between two sample areas or sets of data that are being studied. We use standard deviation to compare the means and the spread of data in two sets of samples.

For example if a biologist compared the height of pine trees growing on a westrfacing mountain slope with that of trees on an eastrfacing slope data might be recorded as in Table **1.1**.

	Height of westnfacing trees / m 0.5	**ing trees /**
Measurements	16	18
	12	20
	14	19
	13	10
	18	12
	16	9
	16	11
	15	21
Total	120	120
Mean	15	15

Table 1.1 Data recorded for the heights of pine trees on westn and eastnfacing mountain slopes.

Looking at the mean values it seems that the heights of the trees in the two areas are similar. By calculating the standard deviations of the data we can examine the results more closely and see whether this is correct. We can work out the standard deviation of each set of data (using the standard deviation function in a spreadsheet or on a graphical computer) and compare the spread of the data in each case.

The mean values do not show any di erence between the two sets of data but the standard deviation for the westrfacing trees is 1.9 m and that for the eastrfacing trees is 5.0 m. This information tells us that there is much wider variation in the heights of the trees on the eastrfacing slope. A biologist presented with this information would need to consider other factors besides the direction of the slope that may have a ected the height of the trees. Calculation of the standard deviation has given additional information which allows us to think about whether the di erences between two samples are likely to be significant.

Error bars on graphs

Error bars are a way of showing either the range or the standard deviation of data on a graph.

When data are collected there is usually some variability in the values and the error bars extend above and below the points plotted on a graph to show this variability.

3

4

For example Table **1.2** shows data collected on heart rate during exercise. A di erent value is recorded in each trial. For a small number of values (three or four) the mean is plotted and the error bar added to show the highest and lowest values as in Figure **1.3**. This shows the **range** of the values. For a larger number of values (five or more) the standard deviation is calculated and this is shown in the same way (Figure **1.4**).

Trial number Heart rate/
144
beats min ^{1}
142
1 135
2 142
138
3 139
136
mean 139
134
132 **Table 1.2** For a small number of values, only the mean is calculated.
130 **Figure 1.3** Here, the error bar shows the range of the data. The mean value of 139 is plotted with the error bar ranging from the highest value, 142, to the lowest value, 135.

Trial number Heart rate/ beats min ^{1}
1 137
2 141
3 134
4 136
5 140
6 mean 139 138

Figure 1.4 This time, the error bar shows the standard deviation from the mean. The standard deviation 2.6 mean value of 138 is plotted with the error bar showing 1*s*. **Table 1.3** For five or more values, the standard deviation is also calculated.

error bar 140
142
140
138
136
134
132

Signi cance

5% significance means that if an investigation was carried out 100 times and each time there was a di erence then 95 of those di erences are probably due to the factor being investigated and only 5 are probably due to chance.

In Table **1.3** where the mean is 138 and *s* is 2.6 this means that 68% of the values fall within 138 $\pm$ 2.6 that is between 135.4 and 140.6 and 95% of the values fall within 138 $\pm$ 5.2 that is between 132.8 and 143.2.

1.2 The *t*ntest

In order to decide whether the di erence between two sets of data is important or **signi cant** we use the ***t*-test**. It compares the mean and standard deviation of the two sets of samples to see if they are the same or di erent.

A value for *t* is calculated using a statistical formula. We then look up this value in a standard table of *t*values like the one in Table **1.4.** Note that *t* unlike standard deviation does not have units. You do not need to

know the formula for calculating *t* but if you are interested you can find it in the glossary.

There are two important column headings in a table of *t*values: **degrees of freedom** and **signi cance level or probability** .

Probability shows whether chance alone could make a di erence between two sets of data that have been collected. There are four di erent levels of probability shown in Table **1.4.** The most important column to biologists is the one headed 5% or 0.05 . If values fall into this category it means that 95% of the time the di erences between the two sets of values are due to significant di erences between them and not due to chance. These are called the **critical values** . Biologists use the 5% or 0.05 value because living things have natural inbuilt variation that must be taken into account.

vel or

Degrees of freedom	**10% or 0.1**	**5% or 0.05**	**1% or 0.01**	**0.1% or 0.001**
18	1.73	2.10	2.88	3.92
19	1.73	2.09	2.86	3.88
20	1.72	2.09	2.85	3.85
21	1.72	2.08	2.83	3.82
22	1.72	2.07	2.82	3.79
23	1.71	2.07	2.81	3.77
24	1.71	2.06	2.80	3.75
25	1.71	2.06	2.79	3.73
26	1.71	2.06	2.78	3.71
27	1.70	2.05	2.77	3.69
28	1.70	2.05	2.76	3.67
29	1.70	2.05	2.76	3.66
30	1.70	2.04	2.75	3.65
40	1.68	2.02	2.70	3.55

60 1.67 2.00 2.66 3.46

120 1.65 1.98 2.62 3.37

← decreasing significance increasing →

Table 1.4 Table of
*t*nvalues.

Degrees of freedom is calculated from the sum of the sample sizes of the two groups of data minus 2:

degrees of freedom (n_1 n_2) 2

where n_1 is the number of values in sample 1 and n_2 is the number of values in sample 2.

Remember to use the *t*rtest there must be a minimum of 10 to 15 values for each sample and they must form a normal or nearrnormal distribution.

Worked example 1

Increase in height of plants after 30 days/cm 0.5 Two sets of soybean plants were grown with and without the

Sample number Group 1 no fertiliser Group 2 0.1% fertiliser

addition of fertiliser. The heights of

1 10.0 12.5 the plants were measured after 30

2 7.0 13.0 days. Both sets of data formed nearr

3 9.5 13.0 normal distributions so a *t*rtest was carried out.

4 5 8.5 12.5 7.5 15.5 Is there a significant di erence in

6 10.0 12.5 growth between the two sets of plants?

7 8 9.5 10.5 9.5 14.0

9 8.5 10.0

10 8.5 10.5

mean 8.9 12.4

calculated value for *t* 5.96

Step 1 Determine the number of degrees of freedom for the data:

degrees of freedom (10 10) 2

18

Step 2 Go down the degrees of freedom column on the *t*rtable in Table **1.4** to the 18 value. **Step 3** Go across the table and find the critical value of *t* that is the number in the 5% or 0.05 column. In this example it is 2.10. **Step 4** Calculate a value for *t* using the appropriate statistical formula in your calculator or spreadsheet. In this case the calculated value for *t* is 5.96. **Step 5** Compare the calculated value for *t* with the critical value from the table. If the calculated value of *t* is greater than this critical value then there is a significant di erence between the sets of data. If the calculated value of *t* is lower than the critical value then the di erence is due to chance. In this case 5.96 is greater than 2.10 so we conclude that there is a significant di erence between the means. This indicates that the fertiliser may have caused the increase in growth.

Use Table **1.4** to help you answer these questions.

3 In an investigation to compare two groups of plants grown with different levels of minerals, the degrees of freedom (df) was 20 and the calculated value for *t* was 4.02. Was there a significant difference between the two sets of data?

4 In another investigation the body mass of crabs living on a westnfacing shore was compared with that of crabs from an eastnfacing shore. The degrees of freedom was 37 and the calculated value for *t* was 1.82. Was there a significant difference between the two sets of data?

If the calculated value of *t* is close to the critical value the conclusion is less certain than if there is a greater di erence between the values.

Worked example 2 An investigation was carried out to see if light intensity a ected the surface area of ivy leaves. A random sample of 10 leaves was collected from each side of a wall one sunny and the other shaded. The surface area for each leaf was found and *t* was calculated as 2.19.

The *t*rvalue of 2.19 is greater than the critical value of 2.10 for 18df shown in Table **1.4.** This indicates that light intensity does a ect the surface area of these ivy leaves.

The value 2.19 is very close to the critical value of 2.10 and so this conclusion is quite weak. If the calculated value for *t* were much higher say 2.88 we could feel much safer with the conclusion and if it were as high as 3.92 for example we could feel very certain.

5 An investigation was carried out on the effect of pollution on the density of branching coral off the Indonesian island of Hoga. The number of corals found in 9 m^2 was counted in a clean area and in a polluted area. Both sets of data formed nearnnormal distributions so a *t*ntest was carried out.

Sample number	Branching corals / number per 9 m^2	
	Clean area	Polluted area
1	7	6
2	8	6
3	5	5
4	9	4
5	8	6
6	7	5
7	10	7
8	8	4

9 8 7

10 9 5

11 6 6

12 7 6

13 6 8

14 9 4

15 11

16 8

mean 7.9 5.6

calculated value of *t* 4.50

Determine if the pollution has an effect on the density of branching coral.

How certain is your conclusion?

7

Why are statistics important?

In science statistics are often used to add credibility to an argument or support a conclusion. International organisations such as the United Nations collect data on health to ensure aid programmes are properly directed.

Being able to use and interpret statistics is an important skill.

Questions to consider

1 Do you believe the following statements? If not why not?

There is a 75% chance that in a group of 30 people two will have the same birthday.

8 out of 10 dentists recommend Zappo toothpaste.

85% of lung cancers are related to smoking.

2 Can statistics be manipulated to produce misleading

1.3 Correlation and cause

Correlation is one of the most common and useful statistics. It describes the degree of relationship between two variables.

In the last 30 years the number of people taking a holiday each year has increased. In the last 30 years there has also been an increase in the number of hotels at holiday resorts. Plotting this data on a graph and adding a trend line as shown in Figure **1.5** shows a **positive correlation**.

Similarly a graph can be plotted to show annual deaths from influenza and the number of influenza vaccines given. In this case there is a **negative correlation** as shown in Figure **1.6**.

With these examples we might feel safe to say that one set of data is linked to the other and that there is a **causal relationship** because there are more tourists more hotels have been built greater use of the influenza vaccine has resulted in fewer deaths from influenza.

However it is important to realise that just because the graph shows a **trend** it does not necessarily mean that there is a causal relationship. For example plotting the number of people using mobile phones in the last

12
0

10
0

80

6040200 0 10 20 30 40 50 60 70 80 Number of people taking an annual holiday × 1000

Figure 1.5 A positive correlation.

claim
s?

3500
3000
2500
2000
1500
1000
500
0

0 50 100 150 200 250 300 350 400

450 Number of influenza vaccinations given in the community **Figure 1.6** A negative correlation.

10 years against the area of Amazon rainforest cut down would show a positive correlation. But this does not mean that the use of mobile phones has caused rainforest to be cut down nor does it mean that a reduction in rainforest area results in more mobile phone use.

Observations without experiments can show a correlation but usually experiments must be used to provide evidence to show the cause of the correlation.

Endnofnchapter questions

1 An error bar drawn on a graph or chart must always be a representation of:

A the mean B the standard deviation C the variation shown by the data D the *t*rvalue (1)

2 The width of 10 leaves was measured and the values in mm were 12 13 13 14 14 14 14 15 15 16. The mean is 14.0 mm. What is the best estimate of the standard deviation?

A 1 mm B 2 mm C 7 mm D 14 mm (1)

3 Measurements of trunk diameter were taken for 23 trees in one wood and the trunk diameters of 19 trees

were measured in a second wood. If a *t*rtest were carried out the degrees of freedom used would be:

A 23 B 19 C 42 D 40 (1)

9

10

4 A student examined two walls one facing east and the other facing west. He measured the percentage of each wall that was covered with lichens. Sixteen samples from each wall were recorded. The calculated value of *t* was 1.84. Using the *t*rtable (Table **1.4**) the conclusion is:

A degrees of freedom are 16 and there is a significant di erence between the walls B degrees of freedom are 14 and there is no significant di erence between the walls C degrees of freedom are 30 and there is a significant di erence between the walls D degrees of freedom are 30 and there is no significant di erence between the walls (1)

5 1000 bananas were collected from a single plantation and weighed. Their masses formed a normal distribution.

How many bananas would be expected to be within 2 standard deviations of the mean?

A 680 B 950 C 68 D 95 (1)

6 In a normal distribution what percentage of values fall within ±1 standard deviation of the mean and ±2 standard deviations of the mean? (2)

7 The lengths of the leaves of dandelion plants growing on a lawn were measured. The mean length was 35 mm and the standard deviation was 4 mm. A second set of data on dandelion leaf length was collected from a wasteland area some distance away. The mean was 97 mm and the standard deviation 20 mm. What can you say about the di erences in the lengths of the dandelion leaves from the two di erent habitats? (2)

8 Salmon live and reproduce in two rivers in Norway the Namsen and Gaula rivers. Data were collected on the number of eggs laid by the salmon in these two rivers. In the River Namsen the mean number laid was 1200 eggs per salmon and the standard deviation was 45. For the River Gaula the mean was 770 eggs per salmon and the standard deviation was 48. Is there a di erence between the number of eggs laid by the fish in the two di erent rivers? (2)

9 Over a period of 20 years the number of elephants in the Moremi

550 game reserve in Botswana was recorded each year. Data were also collected on the number of fallen and broken trees. The data are shown on the graph on the right.

530
510
a State the trend shown by the graph.
490
470 **b** What can you say about the relationship between the two sets of data?
450
430
410
390
370
0
0
50 100 150 200 Number of elephants
(3)

10 Dung beetles collect fresh dung in which to lay their eggs. There are two groups those that bury dung (buriers) and those that roll away balls of dung (rollers). An investigation was carried out to see if there was any di erence in the quantity of dung removed from a field by these two groups. Both sets of data formed nearrnormal distributions so a *t*rtest was carried out.

Sample number	Mass of dung buried/g 1	Mass of dung rolled away/g 1
1	56	54
2	58	52
3	56	53
4	55	51
5	53	55
6	60	54
7	58	54
8	56	55
9	51	53
10	55	52
11	57	53
12	52	56
13	54	51
14	57	52
15	59	53
mean	55.8	53.2
calculated value of *t*	3.55	

Using Table **1.4** on page **5** what conclusion can you draw concerning any di erence in the mass of dung removed by the two groups of beetles? (3)