# M1.9 – Select and use a statistical test

### Tutorials

Learners may be tested on their ability to select and use:

* the chi squared test ($χ^{2}$) to test the significance of the difference between observed and expected results
* the Student’s *t*-test/ unpaired *t-*test
* the Paired *t-*test
* the Spearman’s rank correlation coefficient.

### Statistical tests

Choosing and using statistical tests can seem daunting at first, but they are very useful tools for analysing data. In simple terms each type of statistical test has one purpose: to determine the probability that your results could have occurred by chance as opposed to representing a real biological effect.

Why do we need statistical tests? As scientists we are interested in finding results that apply as general rules. For example, on average are students in Year 10 taller than students in Year 9? The best and most complete way to do this would be to find every single student across the whole country that is currently in Years 9 or 10 at school and measure every single one. In reality we cannot collect data from every school in the country, it would just take too long. Therefore in this example, and with all experiments, we collect data from a small subset of the population instead (this is our sample). From the sample data (e.g. all of Year 9 and 10 in one school) we infer things about the population as a whole.

Statistical tests allow us to make quantitative statements about the inferences we have made. We can put a number on how confident we are that our conclusion about the whole population is correct based on the sample we have taken.

We will cover four types of statistical test: **the chi squared test, the Spearman’s rank correlation, the Student’s *t*-test /** unpaired *t-*test**, the paired *t*-test**.

The choice of which statistical tests we use on our data depends on the question being asked. So always look at your data and ask yourself whether you can say yes to these questions – the one that fits best tells you which statistical test to perform.

1. Am I looking at frequencies, and whether my observations differ from expected values? For example – count the number of red, purple and white flowers that come from a genetic cross of two purple flowers where I expect a ratio of 1:2:1

Test – **chi squared test**

1. Am I looking at the relationship between two variables?

For example – ice-cream consumption and blood sugar levels, to see if people who eat a lot of ice-cream have higher blood sugar.

Test – **Spearman’s rank correlation**

1. Am I looking at the whether there is a difference in the means between two separate/independent groups?
For example – measuring the heights of men and women to see if there is a difference in the average height by gender

Test – **Student’s *t*-test/** unpaired *t-*test

1. Am I looking at whether there is a difference in the mean between the same group before and after a change?
For example – measuring the cholesterol levels in people before and after switching to a vegetarian diet to see if there is an effect on cholesterol of this dietary change

Test – **Paired *t*-test**

Which of the four tests is most appropriate for answering the example question we had earlier: ‘on average are students in Year 10 taller than students in Year 9?’?

### Hypotheses

Statistical tests allow us to test hypotheses about relationships. With every statistical test we generate two competing propositions:

* the null hypothesis (H0)
* the alternative (H1)

The alternative hypothesis comes from your idea that a particular effect will be present, while the null is simply the opposite, that the effect is absent.

Taking our previous example of height and year group we can generate the following hypotheses:

* H1 : Students in Year 10 are taller on average than students in Year 9
* H0 : On average students in Year 10 do not differ in height from students in Year 9

The reason we have a null hypothesis is because we cannot prove experimental hypotheses but we can reject dis-proven hypotheses. This can be quite confusing but, simply put, it is easier to dis-prove a theory than prove one. If our data gives us the confidence to reject the null hypothesis then this provides support for our alternative hypothesis, but it does not prove it.

Similarly if our statistical test shows no significant effect, we refer to this as failing to reject the null hypothesis. This is the statistics equivalent of using “not guilty” rather than “innocent” in a court verdict; we have not provided the evidence to reject the null hypothesis at this time but it doesn’t preclude changing our minds if more evidence comes to light at a later date.

### Student’s *t-*test / unpaired *t*-test

This is the best test for looking at average differences between independent groups. So this is the test we would use to compare, for example, the average height of children in Year 9 and the average height of children in Year 10. We would take a sample from each of the year groups (one Year 9 class and one Year 10 class) and measure the variable we’re investigating (height) for all the individuals in each sample. Then, on the basis of these measurements, we use the Student’s *t*-test to say whether we can be reasonably confident that there really is a difference in the mean height of all Year 9 children compared to all Year 10 children.

|  |  |
| --- | --- |
| **Year 9** | **Year 10** |
| **Name** | **Height (cm)** | **Name** | **Height (cm)** |
| Connor | 170 | Mick | 178 |
| Tristan | 181 | Keith | 174 |
| James | 178 | Charlie | 173 |
| Brad | 170 | Ronnie | 175 |
| Alana | 174 | Danielle | 178 |
| Este | 180 | Taylor | 178 |
| Tegan | 157 | Jenny | 174 |
| Sara | 157 | Gemma | 163 |
| **Mean** | **170.9** | **Mean** | **174.1** |
| **Standard deviation** | **9.5** | **Standard deviation** | **4.9** |

You can see that the means of our two sample groups **are** different.

|  |  |  |  |
| --- | --- | --- | --- |
| **Mean** | **170.9** | **Mean** | **174.1** |

No-one can say there is **no** difference there.

But we are not interested in the samples. We are interested in using the data from our samples to say things (with confidence) about the **whole population** (in this case all of year 9 and all of year 10). The important thing for us to find out, therefore, is whether the difference we see between the sample means is **significant** – is it big enough (given the size of the sample and how much variation we see in the data) for us to be confident that it reflects a real difference between the two year groups rather than just chance variations in the samples we happen to have picked?

Our **null hypothesis** is that there is no significant difference between the heights of Year 9 and Year 10 children. If this is true the difference between the sample means is **not** because there is really any difference between the means for all Year 9s and all Year 10s. It just arose by chance in the particular samples we took. The difference we see in our samples is not big enough to make us confident in saying that the two year groups really are different. We would say that there is **‘not a significant difference’**.

The **alternative hypothesis** is that the there **is** a significant difference in height between the two whole year groups. In order to reject the null hypothesis we need to identify a ‘significant difference’ between the sample means. The difference is big enough that we can be confident it is telling us there is a real difference between the year groups. We can make a statement such as “on average the students in Year 10 are taller than students in Year 9”.

The statistical test allows us to find out whether we can confidently reject the null hypothesis.

The Student’s *t*-test formula is as follows:



The modulus sign (vertical lines) in the numerator tells us to ignore any minus sign once we have subtracted one mean from the other

Subscript A and B refer to the two groups – Year 9 and Year 10.

$\overbar{x}$ refers to the mean, so that $\overbar{x}$A is the group mean of the Year 9 class

*S* is the standard deviation

n is the sample size

So that

*t* = $\frac{(170.9-174.1)}{\sqrt{\frac{(9.5)^{2}}{8} + \frac{(4.9)^{2}}{8}}}$

*t =* 0.858

To understand what this means we must look up the value in the Student’s *t-*test significance tables.

First calculate the degrees of freedom (df) which is *n-1* for each group:

(8-1)+(8-1) = 14

At 14 df the value of *t* = 0.858 is below the threshold of *t* = 2.15 which is the threshold at which we consider the difference to be significant

**There is no significant difference between the height of students in Year 9 and Year 10 in our dataset.**

Therefore we state that we have failed to reject the null hypothesis

This does **not** mean that we have proved that the mean height of Year 9 and Year 10 children is the same. It means that we have failed to show a significant difference based on the data we have gathered. Perhaps there really is no difference, or perhaps there really **is** a difference but our samples failed to show it (which could be for many reasons but most obviously it could simply be that the samples were not large enough).

Assumptions

When performing a Student’s *t*-test the following things are assumed about the data in order to trust the test result.

* We have two independent groups
* For each group we have taken an unbiased sample and measured the same variable
* The variable is continuous
* The continuous variable is normally distributed for each group
* Each group has approximately equal variances (i.e. similar standard deviations) for this variable
* The sample sizes are roughly equal

### Paired *t*-test

When looking for differences between means in two groups we use a *t*-test. If the two groups are independent of each other we use the unpaired version of this test. However, if the two groups come as related pairs we can use the paired *t*-test, allowing us to identify quite subtle but significant differences that might be missed with the unpaired test. It is essential to understand that the pairing must be done according to some **genuine relationship** between the members of each pair and must always be done based on that relationship **not based on the data gathered**.

For example if we measure a variable such as systolic blood pressure in a set of patients on Monday and then measure the same variable in the same set of patients on Tuesday we have two groups (patients on Monday and patients on Tuesday) and there is a natural pairing across these two groups (data on patient A on Monday will obviously be paired with data on the same patient the next day). This is a prime example where using the paired *t*-test is appropriate.

But beware! You might think that the following scenario would also allow analysis by the paired *t*-test **but it would not**:

* We measure systolic blood pressure in two groups of ten patients.
* We then rank the data in each group from highest to lowest.

Now can we pair up the highest in each group, then pair up second highest and so on? **No!** This is pairing after data gathering and is using the data itself to guide the pairing. Using a paired *t*-test in this case could easily lead us to mistakenly identify a significant difference where none exists.

As an example we will use the paired *t*-test to compare the mean difference in shell size of the same hermit crabs, before and after they are given the opportunity to swap out their shells for one of a range of others. These are measurements on 15 individual crabs measured twice (before and after shell swapping) .

| **Hermit crab** | **Shell size before shell swapping (mm)** | **Shell size after shell swapping (mm)** | **Difference****d** | **How far is difference from mean difference?****(*d*-**$\overbar{d}$**)** | **(*d*-**$\overbar{d}$**)2** |
| --- | --- | --- | --- | --- | --- |
| 1 | 8 | 8 | 0 | -2.2 | 4.84 |
| 2 | 10 | 12 | 2 | -0.2 | 0.04 |
| 3 | 5 | 7 | 2 | -0.2 | 0.04 |
| 4 | 6 | 8 | 2 | -0.2 | 0.04 |
| 5 | 4 | 5 | 1 | -1.2 | 1.44 |
| 6 | 7 | 11 | 4 | 1.8 | 3.24 |
| 7 | 9 | 10 | 1 | -1.2 | 1.44 |
| 8 | 11 | 13 | 2 | -0.2 | 0.04 |
| 9 | 13 | 13 | 0 | -2.2 | 4.84 |
| 10 | 15 | 14 | -1 | -3.2 | 10.24 |
| 11 | 8 | 11 | 3 | 0.8 | 0.64 |
| 12 | 9 | 14 | 5 | 2.8 | 7.84 |
| 13 | 7 | 15 | 8 | 5.8 | 33.64 |
| 14 | 9 | 10 | 1 | -1.2 | 1.44 |
| 15 | 11 | 14 | 3 | 0.8 | 0.64 |
|  |  | **Total** | **33** |  | **70.4** |

$\overbar{d }$**= Mean difference = 33/15 = 2.2**

Calculate for each hermit crab how far the difference in shell sizes is from the mean difference. Square these numbers and add them up to find the total = **70.4**

From this we can calculate the standard deviation of the difference 

$= \sqrt{\frac{70.4}{14}}$ = 2.24

The paired *t-*test is as follows:

= $\frac{2.2 \sqrt{15}}{2.24}$ = 3.8

Now we need to see whether this value of *t* is large enough for us to reject our null hypothesis. We can refer to a critical values table, picking the entry for our desired confidence level (95% or p=0.05) and the correct degrees of freedom. In a paired *t*-test the number of degrees of freedom is n-1.

15 individuals were used so n-1 = 14

The critical value at p = 0.05 for 14 degrees of freedom is 2.145

3.8 > 2.145 so our *t* value is greater than the critical value and we can **reject the null hypothesis** that there is no significant change in average shell size after being given the opportunity to swap shells. Giving hermit crabs the option to change their shells does have an effect on average shell size.

Assumptions

When performing a paired *t*-test the following things are assumed about the data in order to trust the test result.

* We have two groups with some dependency or relationship between specific pairs (one from one group one from the other) (e.g. same subjects measured before and after)
* We have taken an unbiased sample of these pairs and measured the same variable
* The variable is continuous
* The continuous variable is normally distributed for each group with the same variance

### Spearman’s rank correlation coefficient

If we have data on two variables for a set of items and we want to see if these variables are related we can test them for correlation. Correlation comes in two forms:

Positive correlation – as one variable increases in value, so does the other

Negative correlation – as one variable increases in value, the other decreases in value

As an example we will use the Spearman’s rank correlation coefficient to comment on the relationship between the size of a locust and the length of its wings. So in this example the set of items is the locusts in our sample and the two variables we are looking at for each locust are body length and wing length.

When we use the Spearman’s rank coefficient to calculate a correlation, we first have to rank the data for each of the variables.

| **Locust** | **Body length x (mm)** | **Wing length y (mm)** | **Rank x** | **Rank y** | ***d*** | ***d2*** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 15 | 6 | 9 | 10 | -1 | 1 |
| 2 | 10 | 7 | 10 | 9 | 1 | 1 |
| 3 | 80 | 32 | 1 | 1 | 0 | 0 |
| 4 | 45 | 23 | 6 | 5 | 1 | 1 |
| 5 | 53 | 19 | 5 | 6.5 | -1.5 | 2.25 |
| 6 | 62 | 29 | 2 | 2 | 0 | 0 |
| 7 | 35 | 18 | 8 | 8 | 0 | 0 |
| 8 | 41 | 19 | 7 | 6.5 | 0.5 | 0.25 |
| 9 | 58 | 28 | 4 | 3 | 1 | 1 |
| 10 | 61 | 27 | 3 | 4 | -1 | 1 |

If two equal values appear e.g. for Rank y at rank 6, then both are given the rank 6.5 (halfway between rank 6 & 7) and no values are given rank 6 or 7.

Next we calculate the difference between the ranks = *d*  and then square this = *d*2

Then we find the sum of all the *d2* values

∑*d*2 = 1+1+0+1+2.25+0+0+0.25+1+1 = 7.5

Now we can calculate the correlation coefficient using the formula:

$$r\_{s}=1-\frac{6∑d^{2}}{n(n^{2}-1)}=1-\frac{6×7.5}{10(10^{2}-1)}=1-\frac{45}{990}=0.9545$$

An *rs* value of +1 shows perfect positive correlation

An *rs* value of -1 shows perfect negative correlation

An *rs* value of 0 shows no correlation

To find out whether the *rs* value we have calculated is sufficient evidence to reject our null hypothesis we need to refer to the critical values table. Some versions of this table have entries listed according to n, the number of items (locusts in this case, n = 10). Some versions have entries listed by degrees of freedom.

The number of degrees of freedom is:

d.f. = n - 2 = 8

The critical value for the Spearman’s rank correlation coefficient for n = 10 or df = 8 at p=0.05 is 0.6485

Our calculated value for *rs* is therefore greater than the critical value

0.9545>0.6485

Therefore we can reject the null hypothesis that there is no significant correlation between locust body size and wing size. We can accept the alternative hypothesis that there is a significant correlation between locust body size and wing size.

Our *rs* is positive (+0.9545), therefore we have positive correlation: as the size of a locust increases there is a tendency for its wing length to increase.

Remember - correlation does not equal causation – we have shown that these two variables tend to change together, but we have not shown that there is a cause and effect.

Assumptions

When performing a Spearman’s rank correlation the following things are assumed about the data in order to trust the test result:

* We have a set of items and we have data on the same two variables from every one of those items
* Both variables are ordinal (i.e. they can be placed in order (ranked)) or a measurement.

We also assume that there are few (or no) tied ranks. In cases where there are many ties we can correct for this by using a slightly different formula but that is beyond the scope of the A Level Biology maths requirements.

### The chi squared test

When we want to look at distributions of frequencies and whether they differ from expected values we can use the chi squared (χ2) test. Our expected frequencies can be based on previous observations from experiments, or simply an expectation that there should be equal proportions in each category.

For example, we cross two flowers with pink petals – we know that both of these plants are heterozygotes and they carry two co-dominant alleles, one for red petals and one for white petals. We then count the frequency of offspring that develop with either red, white or pink petals.

Our hypothesis is that any differences in the observed numbers of offspring with white, red and pink petals from the expected numbers are due to chance. The null hypothesis is that there is no significant difference between the observed numbers of offspring with white, red and pink petals from the expected numbers. The alternative hypothesis would be that there is a significant difference between the observed numbers of offspring with white, red and pink petals from the expected numbers

We start by working out what our expected frequencies should be. A Punnett square is a good way to do this. In the table below the alleles present in the parental gametes are shown and then, within the outlined section, the four equally likely outcomes of each fertilisation event, giving the alleles present in the offspring and the **resulting appearance**

|  |  |  |
| --- | --- | --- |
|  |  | Parental gamete alleles |
|  |  | Red | White |
| Parental gamete alleles | Red | Red/Red**Red** | Red/White**Pink** |
| White | Red/White**Pink** | White/White**White** |

From this simple table we can see that we expect to see frequencies in the offspring of White:Red:Pink petals at a ratio of 1:1:2.

Let’s say we count a total of 160 offspring from the cross, we can therefore calculate the expected numbers of white, red and pink petals – let’s compare that to the observed numbers in the table below

|  |  |  |
| --- | --- | --- |
|  | **Expected** | **Observed** |
| **White** | 40 | 28 |
| **Red** | 40 | 46 |
| **Pink** | 80 | 86 |

Is there a significant difference between the expected and observed frequencies?

The *χ2*value is calculated as follows: **

*f0* = observed frequencies

*fe* = expected frequencies

*χ2* = $\frac{(28-40)^{2}}{40}+ \frac{(46-40)^{2}}{40}+ \frac{(86-80)^{2}}{80}=4.95$

Once again we must look up this value in the appropriate statistics data table and compare it to the critical value at the appropriate degrees of freedom.

There are 3 offspring types – red, white and pink, so n= 3

Therefore degrees of freedom = n-1 = 2

On the *χ2* table the critical value where p = 0.05 and df = 2 is 5.991

4.95 < 5.991 therefore our *χ2* value does not reach the critical value for significance at 2 degrees of freedom.

Therefore, we cannot reject the null hypothesis: that there is no significant differences between the observed numbers of offspring with white, red and pink petals from the expected numbers.

Assumptions

When performing a Chi-squared test the following things are assumed about the data in order to trust the test result.

* There is a minimum sample size for performing the chi squared test – this is indicated by each expected value in a cell being >5

**Document updates**

 v1.0 April 2017 Original version.

 v1.1 June 2019 Clarification that Student’s *t*-test can be referred to as unpaired *t*-test. Changed the statements for null hypothesis and alternative hypothesis.

 v1.2 January 2022 Corrected the calculation for correlation coefficient on page 9.

V1.3 March 2023 Corrected the calculation for unpaired t-test

V1.4 August 2024 Corrected the answer for unpaired *t*-test to 0.858. Corrected the critical value for chi squared test to 5.991 (as per the Maths Skills Handbook). Changed the wing length for Locust 1 and 2, to reflect rank for ‘y’.

**OCR Resources**: *the small print*OCR’s resources are provided to support the delivery of OCR qualifications, but in no way constitute an endorsed teaching method that is required by the Board, and the decision to use them lies with the individual teacher. Whilst every effort is made to ensure the accuracy of the content, OCR cannot be held responsible for any errors or omissions within these resources.
© OCR 2024 - You can copy and distribute this resource in your centre, in line with any specific restrictions detailed in the resource. Resources intended for teacher use should not be shared with students. Resources should not be published on social media platforms or other websites.

OCR acknowledges the use of the following content: n/a

Please get in touch if you want to discuss the accessibility of resources we offer to support delivery of our qualifications: resources.feedback@ocr.org.uk

We’d like to know your view on the resources we produce. By clicking on ‘Like’ or ‘Dislike’ you can help us to ensure that our resources work for you. When the email template pops up please add additional comments if you wish and then just click ‘Send’. Thank you.

If you do not currently offer this OCR qualification but would like to do so, please complete the Expression of Interest Form which can be found here: [www.ocr.org.uk/expression-of-interest](http://www.ocr.org.uk/expression-of-interest)

Looking for a resource? There is now a quick and easy search tool to help find free resources for your qualification:
[www.ocr.org.uk/i-want-to/find-resources/](http://www.ocr.org.uk/i-want-to/find-resources/)

This resource has been produced as part of our free A Level teaching and learning support package. All the A Level teaching and learning resources, including delivery guides, topic exploration packs, lesson elements and more are available on the qualification webpages.

 If you are looking for examination practice materials, you can find the Sample Assessment Materials (SAMs) on the qualification webpages: [Biology A](http://www.ocr.org.uk/qualifications/as-a-level-gce-biology-a-h020-h420-from-2015/) / [Biology B](http://www.ocr.org.uk/qualifications/as-a-level-gce-biology-b-advancing-biology-h022-h422-from-2015/)

### Produced in collaboration with the University of East Anglia