

Statistics information

Statistical techniques

In A2 investigations, you must plan how you are going to analyse the results. This must involve a statistical analysis of the data. The method of analysis must be appropriate for the results that you intend to record. Before you decide on the method of analysis ensure that you understand the statistical techniques available.

Need for statistical analysis

It has already been established that biological data are highly prone to variation. Random errors or variability in the biological material will cause data to deviate from an expected or true value. Statistics is the use of mathematical methods to describe data and to determine the probability of events, such as whether differences are due to random factors.

The statistical technique that is appropriate will depend upon what you are dealing with:

- **Measured data** — where you have obtained results by measuring something (say, rate of reaction), repeated the results and calculated a **sample mean**. Note that even if you repeated results in a laboratory a number of times (e.g. five repeats for each pH in the investigation of fungal amylase activity) the replicates still represent a sample of what you could potentially have done, say hundreds of repeats; and the calculated mean represents a sample mean.
- **Frequency data** — where you have made **counts** of something, say the number of germinated seeds out of a total number used, or the number of woodlice in either side of a choice chamber, one side of which is dark and the other light.

Statistical analysis of sample means

The number of measurements that can be made is limited (for example, by time constraints or availability of equipment) and so the measurements represent a sample. All the measurements that might be made represent the population (whether real, such as a population of wild garlic plants, or imaginary, such as all the measurements of amylase activity that could potentially be taken). The sample provides an estimate for the population. There are two potentially important properties that summarise the sample of data collected:

- The **sample mean** (symbol \bar{x}) — a measure of central tendency or average. (There are other measures of central tendency: the mode — the most numerous value; and the median — the value midway between the highest and the lowest. But the mean is the most helpful.)

- The **standard deviation** — a measure of the variability (or spread or dispersion) of the data. Where the variability of the population is estimated, this is denoted by the symbol $\hat{\sigma}$.

A sample mean (\bar{x}) provides an estimate of the mean of the population from which the sample has been drawn. The population mean is referred to as the true mean (and is denoted by μ). How reliable is the sample mean, that is, how close does it lie to the true mean? The reliability depends on two things:

- the sample size, n
- The variability of the data, as measured by $\hat{\sigma}$ (the estimate of the standard deviation)

Since sample means vary, it is important to consider how good an estimate any one sample mean might be. A sample mean will be more likely to lie close to the true mean if the sample size is large and the standard deviation is small. This is measured in a statistic called the **standard deviation of the mean** (also called **standard error of the mean**) with the symbol $\hat{\sigma}_{\bar{x}}$:

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}^2}{n}}$$

The standard deviation of the mean is a measure of how much the sample means would on average differ from the population mean. A small $\hat{\sigma}_{\bar{x}}$ value, relative to the magnitude of the sample mean, indicates a reliable sample mean — that is, one that lies close to the population mean.

95% confidence limits

Even better is to estimate the boundaries within which the true mean might lie. Since it is not possible to have absolute certainty, a convention of 95% probability has generally been accepted in biology. 95% confidence limits are provided by:

$$\bar{x} \pm t(\hat{\sigma}_{\bar{x}})$$

where t is determined from a table of t values at $p = 0.05$ and $n - 1$ degrees of freedom.

95% confidence limits are also plotted on graphs. Figure 35 shows a bar graph for the mean height of pea seedlings in the light and in the dark. The 95% confidence limits are also shown. These limits set the boundaries within which there is a 95% probability of the true mean lying. You can see that, in this example, the limits for the results in the light and the dark *do not overlap*. This suggests that samples are significantly different — the difference between the sample means is real and not due to chance. This decision can only truly be made by undertaking a t -test. However, using 95% confidence limits is the only way in which you can make decisions about significant differences when you are comparing more than two sample means.

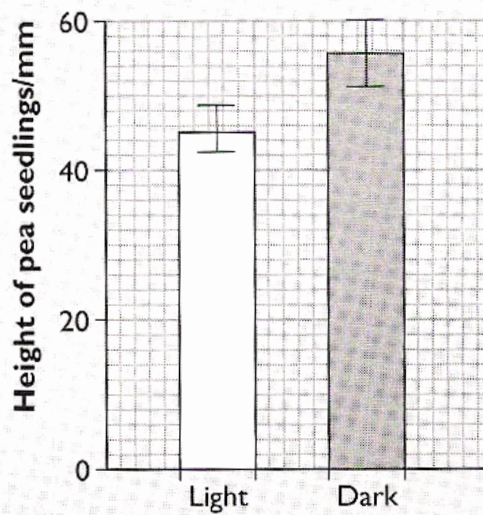


Figure 35 A comparison of the height of pea seedlings grown in the light and in the dark — means and 95% confidence limits plotted

t-test

The *t*-test (sometimes called Student's *t*-test) is a strong statistical procedure for comparing two sample means. Remember that sample means are expected to differ. The *t*-test allows you to determine if the difference is significant, that is, not just due to random factors or chance. The formula for calculating *t*, in terms of $\hat{\sigma}_x$, is given as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\hat{\sigma}_{x_1}^2 + \hat{\sigma}_{x_2}^2}}$$

A starting point in statistical tests is to establish a **null hypothesis** (given the symbol H_0). This is generally stated, for the *t*-test, as: 'The difference between the sample means for [insert the two variables] is simply due to random factors, and is not significant', or 'There is no significant difference between the sample means for [insert the two variables]'.

Tip You must not state that there is 'no difference'. Of course there is a difference — sample means vary. What you must state is that any difference is not significant or is due to chance.

Having carried out a *t*-test, you have your calculated *t*-value. You also know the degrees of freedom for comparing the two samples ($n_1 + n_2 - 2$). You are now in a position to use a *t*-table to determine the probability of the null hypothesis being true. Look across the row for the relevant degrees of freedom (go to the next lower value if the exact value is not included in the table) and find where your calculated *t*-value fits between the *t*-values in the table. Now look up and read off the two *p* values at the top. These are the *p* values within which the probability of the null hypothesis being true lies. A value of $p < 0.05$ means that there is a probability of less than five times in a hundred of the null hypothesis being true, and so H_0 is rejected and a significant difference is concluded. Obviously there is a chance of being wrong (up to five times in a hundred). However, different significance levels are recognised, and you should always quote them. Table 10 shows the different significance levels.

Table 10 Different significance levels in statistical tests

Probability of H_0 being accepted (p value)	Asterisked p value and outcome of test	Significance level
greater than 0.05	$p > 0.05$, accept H_0	No evidence of significant difference
between 0.05 and 0.01	$p < 0.05^*$, reject H_0	Significantly different, 95% level
between 0.01 and 0.001	$p < 0.01^{**}$, reject H_0	Highly significantly different, 99% level
less than 0.001	$p < 0.001^{***}$, reject H_0	Very highly significantly different, 99.9% level
The number of asterisks following the p value denotes the level of significance as indicated above.		

You must remember that a t -test is not an end in itself. It is simply a tool that allows you to make a decision about significant difference. If you do find a significant difference then you must use your biological understanding to suggest an explanation.

Tip You should have practised the use of a calculator to determine statistical parameters. The following procedure is suggested:

- 1 Switch on your calculator, go to statistical mode and enter your data. (Before entering data, you may need to clear the memory of any previous data entered.)
- 2 Determine n — you know this, but this is a check that all the data have been entered.
- 3 Determine \bar{x} — you require this, but this is also a check that the data have been correctly entered.
- 4 Determine $\hat{\sigma}$ — on most calculators this will be denoted by σ_{n-1} (though it may be $x\sigma_{n-1}$, s or sx).
- 5 Square the above (to get $\hat{\sigma}^2$), divide by n (to get $\hat{\sigma}_x^2$) and then root this value to determine $\hat{\sigma}_x$ — this can be used either in calculating confidence limits or in the t -test.

Note that the statistics sheets used in A-level biology are provided on the CCEA biology microsite — at the back of the specification. You should download them. These show the equations for confidence limits and the t -test both in terms of $\hat{\sigma}$ and $\hat{\sigma}_x$. You may be provided with, say, $\hat{\sigma}_x$ in a question in Unit A2 2. Make sure that you use the appropriate equation.

Statistical analysis of frequencies – the chi-squared (χ^2) test

In some investigations the DV is a count or frequency (that is, a number of items). This could be the number of fruitfly phenotypes in a genetic cross, or the number of woodlice in a choice chamber one side of which is light and the other dark, or the number of earthworms in a series of fields. The numbers

counted (referred to as the **observed frequencies**, O) will differ at random from those expected on the basis of a reasoned hypothesis. That hypothesis will depend on the investigation: in a genetic cross it may be a 3:1 ratio; in the earthworm counts in, say, three fields, equal counts in the fields might be expected. This allows you to calculate the **expected frequencies** (E). The χ^2 test allows you to decide whether or not the observed frequencies deviate significantly from those expected. The formula for χ^2 is:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

You must remember that *the sum of the observed frequencies must equal the sum of the expected frequencies*. Again, a null hypothesis (H_0) is set up. A generalised one would be: 'Any differences between the observed frequencies and the expected frequencies are due to chance alone and are not significant'. However, H_0 should be stated specifically for the investigation that you are carrying out:

- For a genetic cross — *The numbers of wild-type and vestigial-winged flies only differ from those expected on the basis of a 3:1 ratio as a result of random factors.*
- For the numbers of earthworms in three fields — *The numbers of earthworms in the three fields are equal and any deviation of the observed counts from this is not significant.*

You use a χ^2 table to determine the probability of this H_0 being true. Look across the row for the relevant degrees of freedom, $n - 1$ (where n is the number of categories), find where your calculated χ^2 value fits between neighbouring tabular χ^2 values; look up and record the two corresponding p values. These are the p values within which the probability of H_0 being true lies.

Tip When quoting p values always ensure that the chevrons are the right way round. Check by ensuring that the larger number is shown as being greater than the lesser number, e.g. $0.05 > p > 0.01$.

Note that the chi-squared test can only be used with raw data (the counts); it cannot be used with processed data such as means or percentages.

Statistics questions

Table 11 The statistical parameters for an investigation into the effect of pH on fungal amylase — the sample size (n) is 5 and, for 95% confidence limits, $t = 2.776$

pH	Mean time taken for starch digestion (\bar{x})/s	$\hat{\sigma}$	$\hat{\sigma}_{\bar{x}}$	95% confidence limits
4.4	780	95.0	42.5	780 ± 118 (662 to 898)
5.1	630	61.3	27.4	
5.6	380	57.9	25.9	380 ± 72 (308 to 452)
6.0	92	14.5	6.5	92 ± 18 (74 to 110)
6.4	260	52.3	23.4	
7.2	580	48.3	21.6	580 ± 60 (520 to 640)
7.6	725	84.5	37.8	725 ± 105 (620 to 830)

Complete the missing confidence limits using the formula:

$$\bar{x} \pm t(\hat{\sigma}_{\bar{x}})$$

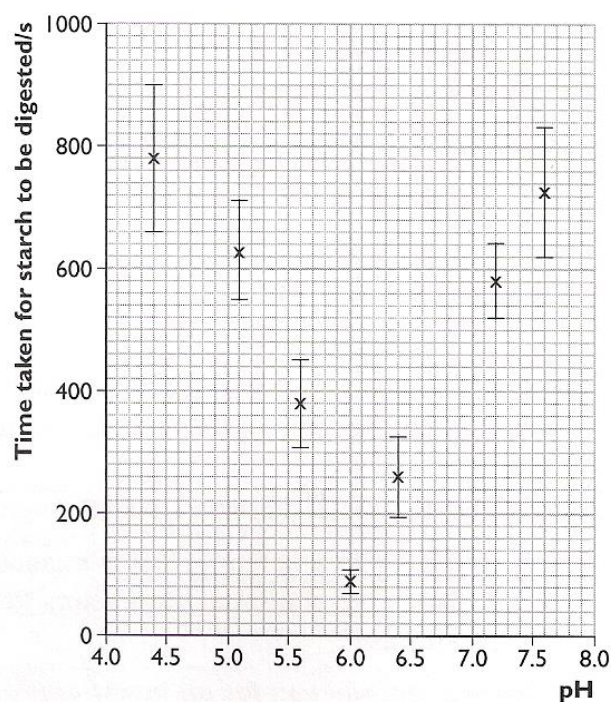


Figure 37 The effect of pH on the activity of fungal amylase — mean time taken for starch digestion and 95% confidence limits plotted

Which pHs provided a significantly different result to that of pH 4? _____

Table 12 The statistical parameters for an investigation into the effect of bile salts on the action of lipase

	Rate of lipase activity/ 10^{-3} s^{-1}	
	Treated — solution of bile salts added	Control — distilled water added (in place of bile salts)
\bar{x}	5.32	4.07
$\hat{\sigma}$	0.811	0.909
$\hat{\sigma}_x$	0.256	0.287
95% confidence limits	5.32 ± 0.58 (4.74 to 5.90)	4.07 ± 0.65 (3.42 to 4.72)

The means and confidence limits can also be plotted on a bar chart to help in assessing the reliability — see Figure 38.

Even though 95% confidence limits may be compared, the appropriate — and stronger — statistical test is the *t*-test. For this you will need to show your calculation. For the investigation described:

H_0 : The activity of lipase treated with a solution of bile salts (measured as the treatment mean) is not significantly different from its activity with distilled water added (the control mean).

Calculate *t* using the formula and complete the gaps in the box below:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

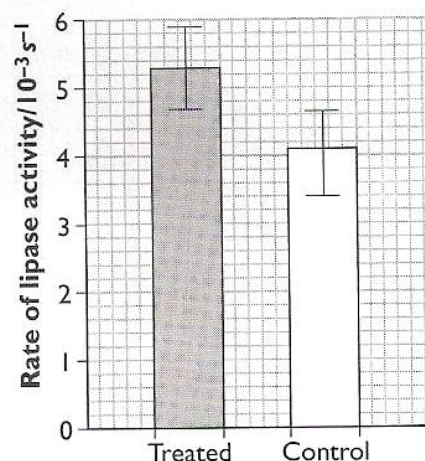


Figure 38 A bar chart comparing the effect of bile salts on the activity of lipase; the treatment included a solution of bile salts and in the control this was replaced with distilled water

Calculated *t* = _____
 Sample size: $n_1 = 10$, $n_2 = 10$
 Degrees of freedom = 18
 _____ > *p* > _____

Since *p* is less than 0.05, the null hypothesis is rejected. There is a highly significant difference (at the 99% level) between the treatment and control means.

Looking at Figure 38 we can see that the activity of lipase is improved by the addition of bile salts. In your 'interpretation', you then enter into a discussion as to why this might be the case.

Tip The rigour of the *t*-test can be seen in the statistical analysis above. A comparison of the confidence limits suggests that there is (just) no significant difference between the means — see Figure 38. However, the more appropriate *t*-test concludes a highly significant difference between the means.

Using the chi-squared (χ^2) test

In an investigation into habitat preferences of woodlice, 20 woodlice are released into a choice chamber, one half of which is open to the light while the other side is darkened. After 20 minutes in the chamber, counts are taken of 8 on the light side and 12 on the dark side. The experiment is repeated 10 times (with different woodlice) and the results summed; in total 80 are found on the light side, 120 on the dark side. The results for 200 woodlice should be more reliable than for 20.

In this situation, the χ^2 test is applicable. The test should start with a null hypothesis:

Write a suitable null hypothesis:

The expected values are calculated — if there is no preference for either condition, and a total of 200 woodlice were released, then each side would be expected to have 100 woodlice. The calculation of χ^2 is shown below:

Categories	Observed (O)	Expected (E)	(O - E)	(O - E) ²	$\frac{(O - E)^2}{E}$
Light side	80	100			
Dark side	120	100			

Calculate Chi-squared using the formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Since p is less than 0.05, the null hypothesis is rejected. There is a highly significant (p is less than 0.01) deviation from the expected values. In this investigation, woodlice have shown a preference for dark conditions.

Note that if only the first set-up had been used (8 in the light side and 12 in the dark) the χ^2 value would have been 0.8, and $0.5 > p > 0.1$, so H_0 would have been accepted (there is no significant difference).

Table 1: Student's t values

d.f.	$p = 0.1$	0.05	0.02	0.01	0.002	0.001
1	6.314	12.706	31.821	63.657	318.31	636.62
2	2.920	4.303	6.965	9.925	22.327	31.598
3	2.353	3.182	4.541	5.841	10.214	12.924
4	2.132	2.776	3.747	4.604	7.173	8.610
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.895	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318
13	1.771	2.160	2.650	3.012	3.852	4.221
14	1.761	2.145	2.624	2.977	3.787	4.140
15	1.753	2.131	2.602	2.947	3.733	4.073
16	1.746	2.120	2.583	2.921	3.686	4.015
17	1.740	2.110	2.567	2.898	3.646	3.965
18	1.734	2.101	2.552	2.878	3.610	3.922
19	1.729	2.093	2.539	2.861	3.579	3.883

Table 2: χ^2 values

d.f.	$p = 0.900$	0.500	0.100	0.050	0.010	0.001
1	0.016	0.455	2.71	3.84	6.63	10.83
2	0.211	1.39	4.61	5.99	9.21	13.82
3	0.584	2.37	6.25	7.81	11.34	16.27
4	1.06	3.36	7.78	9.49	13.28	18.47
5	1.61	4.35	9.24	11.07	15.09	20.52
6	2.20	5.35	10.64	12.59	16.81	22.46
7	2.83	6.35	12.02	14.07	18.48	24.32
8	3.49	7.34	13.36	15.51	20.09	26.13
9	4.17	8.34	14.68	16.92	21.67	27.88
10	4.87	9.34	15.99	18.31	23.21	29.59
11	5.58	10.34	17.28	19.68	24.73	31.26
12	6.30	11.34	18.55	21.03	26.22	32.91
13	7.04	12.34	19.81	22.36	27.69	34.53
14	7.79	13.34	21.06	23.68	29.14	36.12