Statistics Teaching Note H

Peter Mitchell, University of Sheffield 21st February 2018

Overview of Statistical Tests

1. **What statistical tests do.** Given the variability in any set of data, statistical tests allow us to distinguish interesting variation from uninteresting, background, random variation; if you like, to tell the signal apart from the noise. Interesting variation arises from differences between groups of items in the set of data, or from associations in the data. Differences between groups spring from questions such as: do left-handed cricketers score more runs than right-handed ones, or how do the stone axe-heads of Wales, the Lake District, East Anglia and Kent differ in size? Associations occur when we ask questions about whether colour of hair is related to colour of eyes (including do redheads really tend to have green eyes?), or if the number of mink living on a river is correlated with the number of water voles, and so on. The background variation arises from miscellaneous causes, not relevant to the question of interest.

   In the set of data, interesting and background variation are mixed up in the actual values of the observations. To answer the question about differences or associations we need a method that can separate the two sources of variation and give us a way of assessing whether the difference or association is likely to be genuine or just a chance effect in the data. In principle, statistics is simple: it is about measuring variation, attributing causes, and calculating the probabilities of obtaining particular sets of results by chance.
2. **A dose of scientific philosophy.** In terms of formal logic and scientific method, we set up a **null hypothesis** that there is a specified difference between the groups, or a specified association. The specified difference is often zero difference or zero association but it can be a non-zero difference or some particular association if required. We then attempt to falsify (i.e. to nullify) the hypothesis, using a statistical test to calculate how likely it is that we would get the results we have in the set of data *if* the null hypothesis were true. If it is not very likely then we proceed as if the null hypothesis has been falsified, i.e. we work on the basis that there *is* a difference between the groups, or there *is* an association. We cannot be perfectly sure, but we can be reasonably sure that the difference or association is "true" unless a rather unlikely event has occurred. The unlikely event is the occurrence of a set of data showing the specified difference or association when really there isn't one. We return to this point in §4.
3. **How statistical tests work.** We choose an appropriate statistical test, depending on the question asked and the type of data (see §5 and §6). We calculate the **test statistic** for this test from the data. Now we need to find the **degrees of freedom** (abbreviation df or d.f.); this depends on the number of observations or on the number of groups in the data (see Appendix A, §A1). Then we find out the **probability associated with the calculated test statistic** and its degrees of freedom. A computer package will provide the probability to three or four decimal places but for other calculations we look up the test statistic in tables of critical values to find a range of probability associated with it (see Appendix B). The probability is interpreted to provide the **statistical significance**, i.e. do we regard the difference or association of our initial question as statistically significant or not? The conventional levels of significance are shown in Figure 1.

Threshold probabilities

		P = 0.05 **(1 in 20)**		**P = 0.01** **(1 in 100)**		**P = 0.001** **(1 in 1000)**	
Range of probability	P > 0.05		P < 0.05 (but > 0.01)		P < 0.01 (but > 0.001		P < 0.001
Commonly used phrase	Not significant		Significant		Very significant		Highly significant
Coded with asterisks	NS		*		**		***

Figure 1. Diagram to show how ranges of probability are coded to provide levels of significance. It is set out in the same way as tables of critical values for test statistics, reading columns from left to right with probability becoming smaller. If your probability is exactly on the threshold, then the interpretation is to the left. So $P \geq 0.05$ means not significant, $P < 0.05$ significant; $P \geq 0.01$ significant, $P < 0.01$ very significant, and so on.

4. **Proof and probability in statistical testing.** You *cannot* prove anything with statistics. You *can* compute the probability of obtaining a set of results like this (or more extreme) if the null hypothesis was true. If the probability is small, then you can proceed as if the unlikely event has not occurred, as if the result is genuine. We use certain thresholds of probability, arbitrary probabilities but well-accepted and conventional (see §12). The thresholds separate ranges of probability, as shown in Figure 1, and these ranges are interpreted as levels of statistical significance. See Appendix B for how to use the table of critical values for a test statistic to find the probability associated with your calculated value.

   It is important to realize that the probability associated with the test statistic is *not* the probability that any hypothesis is true (see §11 and Appendix A, §A2). It is the probability of obtaining this set of results or results more extreme than this *if* the null hypothesis *was* true. The true understanding is this. If the null hypothesis is true (i.e. there really is no difference or no association) then we would expect to obtain results like these (or more extreme), with this calculated probability. That is what "significant at P = some value" really means. There is more about P values and statistical significance in §11–§13, but for the moment we turn to how data are collected and used in statistical tests.

   If you obtain nothing else from this module, remember this, shouted here in a box.

You *cannot* prove anything with statistics. You *can* compute the probability of obtaining a set of results like this (or more extreme) if the null hypothesis was true. If the probability is small, then you can proceed as if the unlikely event has not occurred, as if the result is genuine.

5. **Procedure.** Figure 2 summarizes the procedure as a flow chart. Notice the order: you must think of the question first, then find out what data will be required, then decide on which test will be appropriate, then consider the assumptions, and only then collect suitable data. (Well, that is the ideal, not always attained in practice.)

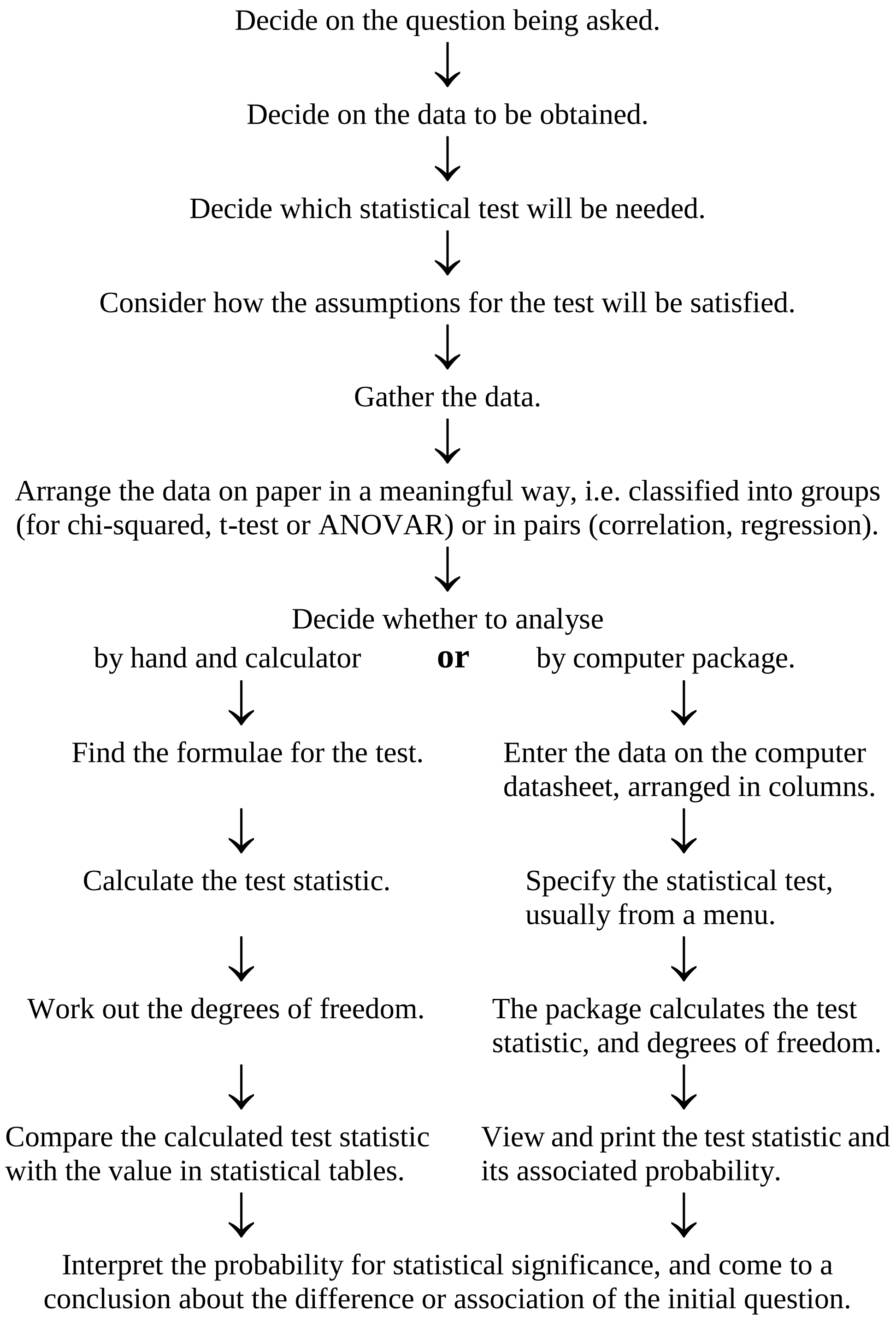

Decide on the question being asked.
Decide on the data to be obtained.
Decide which statistical test will be needed.
Consider how the assumptions for the test will be satisfied.
Gather the data.
Arrange the data on paper in a meaningful way, i.e. classified into groups (for chi-squared, t-test or ANOVAR) or in pairs (correlation, regression).
Decide whether to analyse
by hand and calculator
or
by computer package.
Find the formulae for the test.
Enter the data on the computer datasheet, arranged in columns.
Calculate the test statistic.
Specify the statistical test, usually from a menu.
Work out the degrees of freedom.
The package calculates the test statistic, and degrees of freedom.
Compare the calculated test statistic with the value in statistical tables.
View and print the test statistic and its associated probability.
Interpret the probability for statistical significance, and come to a conclusion about the difference or association of the initial question.

Figure 2. Flow chart of the procedure for statistical testing.

6. **Choosing a test.** Refer to Table 1 to choose a suitable test for your question and data. An outline of each test, including an example, is given on pages 11–16.

Table 1. Guide to statistical tests, arranged by their purpose.

Purpose	**Statistical test**
Describe the data and make simple inferences.	**Descriptive statistics and simple inferences.** Descriptive statistics include mean and standard deviation, and the frequency distribution of the observations. Simple inferences require calculation of standard error, looking up a value of Student's t, and calculation of confidence limits. Descriptive statistics and simple inferences are not statistical tests as such but do use Student's t which is used elsewhere as a test statistic. From the sample of data you can make inferences about the population that the sample represents, from which it was taken.
Look for association.	
Classification of attributes.	**Chi-squared test**, which looks for association between categories, cross-classified into groups, using the actual counts. The test statistic is chi-squared, χ^2.
Between two measured attributes.	**Correlation.** The two attributes must be measurable, continuous quantities that can be plotted on a scatter graph. The test statistic is the correlation coefficient, r.
Compare means.	
Only two groups.	**The t-test.** The means come from the two groups. The test statistic is t, Student's t.
Two or more groups.	**Analysis of variance** (abbreviated to ANOVAR or ANOVA). The means come from different groups of cases (or items or subjects) in the set of data. The test statistic is the variance ratio, F.
Predict one quantity from another.	**Regression.** The two quantities are measurable, continuous attributes, related by a straight line (in simple regression). We obtain the line of best fit to draw on a scatter graph with the equation for the line. Analysis of variance is used to test whether the regression is statistically significant (test statistic is F). You will also see the coefficient of determination, r^2, which indicates the proportion of variation explained by the regression.

7. **A worked example.** References to the procedure in Figure 2 are in italics, and technical terms in statistical testing (mentioned in §3) are in bold. This example is imaginary but plausible.

A researcher suspects that the voles on an island are larger than those on the mainland (*initial question*); they are known to be the same species, and adult males and females do not differ in weight. It will be necessary to catch some voles and weigh them (*data to be obtained*). Variation in vole weights is expected so a statistical test that compares the average weights of island voles and mainland voles is required. Reference to Table 1 shows that the t-test is appropriate for comparing the means of the two groups (*decide on suitable statistical test*) with the **null hypothesis** that the two groups do not differ in mean weight. The *assumptions* are considered for a t-test (p. 14) and used to write the protocol for fieldwork. Traps are set in widely scattered locations in vole habitat on the island and on the mainland. The trapped voles are weighed (*data gathered*) and released. Traps are moved after each successful operation to minimize the chance of catching the same vole again. The data and calculated quantities are given in Table 2 (*data on paper; calculations for test statistic and for degrees of freedom*).

Table 2. Weights of voles captured on an island and on the adjoining mainland, and calculations for the t-test.

	Weight of vole (g)	
	Island	Mainland
	118	119
	126	130
	130	120
	125	123
	120	121
	131	129
	122	114
	124	121
	121	118
		115
		117
n	9	11
mean	124.11	120.63
SD	4.4001	5.1239

Difference between the means 3.4747
Pooled variance = 23.191
SE of the difference = 2.1645
t = 1.605
Degrees of freedom = $n_{island} + n_{mainland} - 2 = 18$
Probability > 0.10 (from statistical table;
in fact P = 0.126 from a computer package)
so not significant

The 95% confidence limits for the *difference* (3.4747 g) are –1.0729 to 8.0223 g. Since the statistical test was not significant at $P \geq 0.05$, the limits include zero.

The researcher concludes that there is no significant difference between the weights of voles on the island and the mainland (t = 1.61, 18 degrees of freedom, P = 0.126; two-sample t-test, homogenous variances). (The sentence above includes the **statistical significance**, the value of the **test statistic**, the **degrees of freedom**, and the **probability associated with the calculated test statistic**. The interpretation is given first and then the evidence for it as the details of the statistical test in brackets. There are several versions of the t-test so for completeness the particular version, here the commonest, is named.)

Points to note. The researcher is sampling from two populations (it is thought) so in each location the traps are widely spaced to minimize capture of related voles which would not be independent observations (families of voles could be consistently smaller or larger than the mean). The procedure to avoid recapture was mentioned above, and this also reduces the chances of capturing a vole related to a previously trapped one. Consequently the observations are random and independent samples from the population, as far as is practicable. This is an important assumption for the t-test. Others are about homogeneity of variance (satisfied here because the standard deviations for the two groups are very close) and Normality (hard to tell with such small samples but each group has a central peak in its frequency distribution even if it is not much like the silhouette of a bell).

Note that we cannot say there is *no* difference in weight: clearly there is, with sampled island voles 3.47 g heavier on average. But we say that this is not significant statistically (our interpretation of the results of the t-test) which means that we will act as if there is no difference. We believe that this observed difference in weight in our data arose by chance alone. Moreover, we have calculated the probability that if there was in fact only one population of voles, spread across the mainland and the island, we could obtain two samples like this with this difference in weight or a more extreme difference. The probability is the one associated with the test statistic, i.e. P = 0.126. Expressed in another way, there is a better than one in ten chance (about one in eight) of obtaining samples like this, that have this difference in weight or a larger one, when there really isn't a difference.

8. **Statistical significance and practical significance.** Statistical significance can be paraphrased as "is it true?" or more correctly "is it likely to be true?". Practical significance is "does it matter?" or "is it worth taking account of, basing decisions on?". Practical significance is the more general term but you will see this kind of significance spoken of in a particular context, e.g. is it educationally significant, or clinically significant (in medical research), or archaeologically significant, or biologically significant?

   Returning to the vole example above, it is possible to calculate how large the samples would need to be for a difference in weight of 4 g to be seen as statistically significant (i.e. with probability associated with the test statistic, t, less than 0.05—see Figure 1). It turns out that with two samples of 13 voles each, assuming the same variability as found before, then a difference of 4 g would be significant (P < 0.05). This is a common finding: larger samples may give a statistically significant result because the larger value of n affects the calculation of standard error and also, through the degrees of freedom, the critical value of the test statistic. If the samples of voles were 22 each from island and mainland then a 4 g difference could be detected with P = 0.01, i.e. it would have to be a hundred to one chance to get such samples if the difference in weight did not really exist.

However, is a difference in weight of 4 g of biological significance, given that it is about 3.3% of the average weight of voles? The weights of small mammals are variable during the day and over a season depending on food availability and temperature, and could easily change by 10% in an individual vole. In this context, the difference of 4 g, even if statistically significant, may have no biological significance from the point of view of assessing whether the voles were better fed on the island than the mainland.

On the other hand, we may be wondering if island voles are heavier than mainland ones, as often found, so we look for evidence that the island population has started to diverge from the mainland one through its isolation. In this case, a difference of 4 g could be biologically significant. A divergence has to start small and become larger, so a statistically significant difference could be seen as identifying the start of the divergence process.

There are no easy answers to these problems about statistical and practical significance. You need to be aware of the topic in case it turns up in your own research in the future. Another way of looking at this problem is the difference between hypothesis testing (to obtain statistical significance) and estimation of size of effect (to assess practical significance—see also §14). For example, in the vole example above, we can focus attention on the size of the difference, where we find that the 95% confidence interval is –1.07 to 8.02 g. These limits include zero so we have to conclude (with 95% limits) that we have not detected a significant size of effect, just as we concluded from the statistical test.

9. **The conclusion of a statistical test.** It is important to write a clear statement of results from a statistical test. This will include the calculated value of the test statistic, the degrees of freedom, the probability associated with the test statistic, and *your* interpretation of that probability as a statistical significance. (These are the terms highlighted in §3, the components of statistical testing.) If it is not evident from the test statistic named, or there are several versions of the test, state which statistical test was used. When you have an exact probability, from a computer package, then quote that in the statement of results. Otherwise, give the value with a < or > sign, as in Figure 1. In statistics, it is sufficient to say, for example, P < 0.05 without specifying also > 0.01 because it is understood that if P was < 0.01 you would have said so.

Example for a chi-squared test. Although the proportion of left-handed males (20%) is greater than for females (10%), this was not statistically significant (χ^2 = 2.550, 1 d.f., P > 0.05). (The symbol, χ^2, is chi-squared, i.e. the Greek letter chi, squared.)

Example for a correlation. There is a significant correlation between monthly totals of rainfall and sunshine (r = –0.699, 10 degrees of freedom, P = 0.0115).

Example for a t-test. The mean interpupillary distance in males was 2.89 mm greater than in females and this was highly significant (t = 5.22, 141 d.f., P < 0.001; two-sample t-test, homogeneous variances). (Since there are several versions of the t-test, it is helpful to specify which one was used.)

10. **Parametric and non-parametric tests.** All the statistical tests given in Table 1, except chi-squared, are called parametric tests because the test statistic is calculated from the actual values of the data, the measured values. There are other tests, called non-parametric, which work on the ranks of the data, i.e. the place in order from smallest to largest, instead of the actual value. On the whole, parametric tests are to be preferred because they use all the information that you have. Non-parametric tests are required when the data really do not conform to certain assumptions necessary for parametric tests to provide correct results, i.e. an accurate value of the probability associated with the test statistic. In this module we have not included non-parametric tests because of the limited time available for teaching statistics. If non-parametric tests turn out to be needed for your research project then you will need to look them up in statistical textbooks, and seek help from a statistician if necessary.
11. **Correct understanding of P values.** It is hard to grasp the correct use of probability (P value) in statistical testing. The definition of P value is: "the probability of the observed data (or data showing a more extreme departure from the null hypothesis) when the null hypothesis is true" (Everitt (1998) Cambridge Dictionary of Statistics). See also Appendix A, §A2.

    The probability associated with the test statistic provides a quantitative measure of the evidence about the null hypothesis. If the probability is small, we are in a position to not accept the null hypothesis because there is a lot of evidence (in the data) against it. "Not accept" is the recommended phrase over "reject" which is too definite. The question is: how small must the probability be? The practice has evolved of using P = 0.05, 0.01 and 0.001 as threshold or boundary values of probability (see Figure 1). Other people may refer to these values of probability as cut-off levels or cut-offs. By using these thresholds we are turning a quantitative measure into a qualitative one, i.e. categories, so that we can make a decision, one way or the other, about whether the result is to be regarded as genuine. (In a similar way, the quantitative score in goals of a football match is converted to a qualitative measure, categories, of win, lose or draw.) The categories or levels of statistical significance (Figure 1) are your interpretation of the P value. Whenever possible, provide the exact probability so that readers can apply different thresholds if they wish (see Appendix B, §B1).
12. **Why use P = 0.05 as the first threshold of statistical significance?** The value of P = 0.05 is entirely arbitrary but has become a well accepted convention. All that can be said to justify 0.05 is that its use over many years in research work in all fields has been successful, on the whole, in identifying interesting results. The one in twenty chance of being misled has not held back progress in research.

    Moreover, the use of P = 0.05, 0.01 and 0.001 thresholds was a necessity in the days before computers. It was not practicable to calculate the exact probability so statistical tables were used (see Appendix B) to find a range of probability for a particular calculated value of the test statistic with its degrees of freedom.

13. **What "not significant" means.** If your calculated probability is 0.05 or greater and you interpret this in the conventional way as not significant, what does this mean about your results? It is better to say that "a difference could not be detected" (a difference between means for example), or that "there was no detectable association" (in a chi-squared test), or "no detectable correlation". This is preferable to "there was no effect" (of experimental treatments or between groups in a survey or relationship between measurements or categories). There could be an effect but we simply failed to find it (as discussed in the vole example, §7). To use the well known adage: absence of evidence is not the same as evidence of absence.

    Not significant is a broad category. It could mean that:

    (a) there is really no effect; or

    (b) there is a small effect but it is lost in the background variability (noise); or

    (c) there is a large effect but not detectable in this set of data because there was also a large amount of background variability.

    This last case can be common in educational, medical or biological research where people and organisms are intrinsically so variable.

    Whenever possible give the exact probability (see Appendix B, §B1). The P values of 0.055 and 0.55 would each be interpreted as not significant but they have different implications. If P = 0.055 then you might think that this is a near miss. If you had more degrees of freedom (from larger samples or more replication) then you might have detected an effect. You might try the survey or experiment again with more degrees of freedom; or you would not be surprised if similar surveys or experiments *had* been successful in detecting an effect. In the case of P = 0.55 you would probably conclude that there was genuinely no effect, or such large amounts of background variation as to completely conceal an effect even with more degrees of freedom.

14. **Effect size versus hypothesis testing.** The procedure for statistical tests has been set out in terms of testing hypotheses (§2–4). It is also important to look at the size of the effect; effect size and hypothesis testing are complementary. This is easily done for the t-test or analysis of variance where we are comparing the means of treatments or groups. We calculate the difference between means and the confidence interval for the difference, as in Table 2, bottom right. The advantage of effect size is that we can see the direction of the effect (island voles heavier) and the size (by 3.5 g), plus make a decision on statistical significance based on the 95% confidence limits (–1.07 to 8.02). Since these limits include zero, we interpret the results as not significant. We assert, in the language of confidence limits, that the true value of the difference lies in the range –1.07 to 8.02 g unless a 1-in-20 chance has occurred. Since this range includes zero, and some values where the difference is in the reverse direction, we conclude that it is not significant.

    Setting out clearly the effect size and its confidence limits also leads to consideration of practical significance. Read §8 again because this is an important topic to understand.

15. **Outlines of statistical tests.** On the six pages following there is a one-page outline of descriptive statistics and simple inferences, chi-squared test, correlation, t-test, analysis of variance, and regression. A standard set of headings is used. The outline provides enough details to enable you to understand what the test does, when and how it is applied, and how the result is interpreted. This will help you to assess use of these tests in the research literature that you read. For your own research you may need further details which can be found in textbooks. To carry out the tests, the computer package MINITAB (available on the university network) is probably the easiest one for beginners to use.

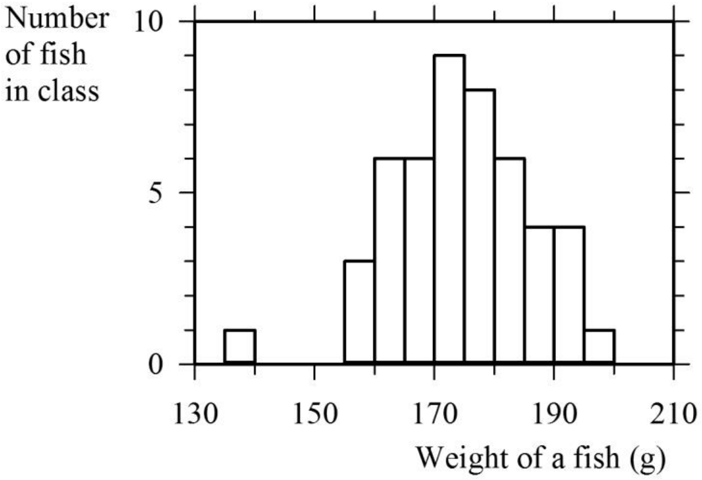
Descriptive Statistics and Simple Inferences

1. **Purpose.** To describe a set of data, and then to make inferences about the population from which the sample of data comes.
2. **Data.** Measurements for a number of cases or items or subjects; a single group.
3. **Null hypothesis.** Not applicable.
4. **Test statistic.** Not applicable, but Student's t is used for calculating confidence limits.
5. **Degrees of freedom.** For a sample of n cases, df = n–1.
6. **Assumptions.** Data from a random and independent sample of the population, if making inferences about the population. Measurements Normally distributed if using confidence limits of the observations (Normal distribution not necessary if using confidence limits of the mean and sample size larger than ten).
7. **Example.** Weights of a cohort of salmon caught on migration shown in a frequency distribution, and with descriptive statistics and simple inferential statistics. In the graph, classes are 135.0–139.9 g, 140.0–144.9 g, and so on.

   Descriptive statistics

   n = 48
   Mean 174.0 g
   Median 174.0 g
   Modal class 170–175 g
   Standard deviation 11.500 g
   Range 136–196 g
   Coefficient of variation
   (SD/mean) 0.066092 or
   6.61%

   Frequency distribution



   Simple inferential statistics
   Standard error 1.6599 g
   df = 47
   Value of t 2.021
   95% confidence limits 170.6–177.4 g (using t for 40 df, the closest value at hand)

   Statement of results. The mean weight of salmon was 174.0 g (SD 11.500 g, n = 48; 95% confidence limits of the mean 170.6–177.4 g).
8. **Variations and elaborations.** Measures of skewness (asymmetry, especially in the tails) and kurtosis (thickness of the tails) can describe a distribution further in terms of departure from Normality. There are other distributions, e.g. Poisson, binomial, which can also be described by mean and standard deviation.
9. **Equivalent non-parametric test.** The median is a measure of central tendency that is obtained from the ranks so is non-parametric; quartiles and interquartile range, or use of centiles, are non-parametric measures of spread of a distribution.

The Chi-squared Test

1. **Purpose.** To look for association between categories, cross-classified into groups.
2. **Data.** Counts of occurrence in the groups, in a cross-classified table of c columns and r rows for the counts (not including marginal columns and rows for descriptors or totals).
3. **Null hypothesis.** No association between categories; cases or items or subjects occur in the groups independently of the categories.
4. **Test statistic.** Chi-squared, χ^2.
5. **Degrees of freedom.** For cross-classified table of c columns and r rows, df = (c – 1) × (r – 1).
6. **Assumptions.** Data from a random and independent sample of the population where factors affecting classification into categories acted uniformly (no identifiable sub-groups where the factors acted differently).
7. **Example.** Are handedness and sex associated? The data are numbers in each group classified by handedness and sex, from undergraduates taking APS240 in October 1993.

		Sex		
		Male	Female	Total
Handedness	Left	10	7	17
	Right	39	63	102
	Total	49	70	119

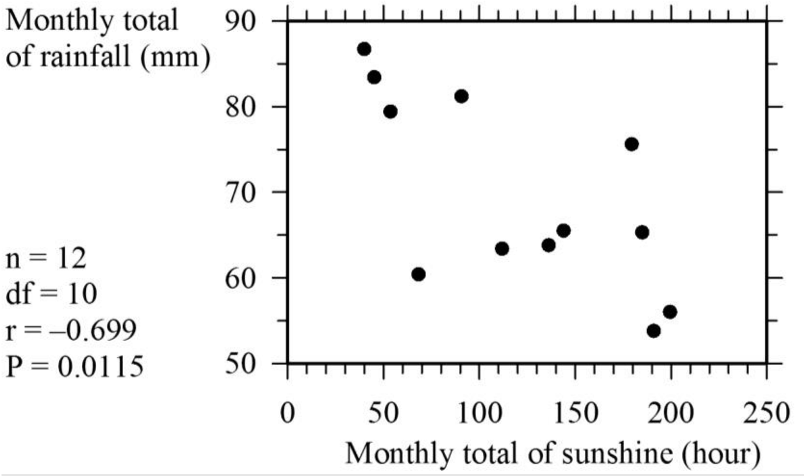
2 × 2 contingency table; df = 1
Chi-squared (χ^2) = 2.550
P = 0.110, NS

Statement of results. Although the percentage of females that is left-handed (10%) is lower than for males (20%), in a set of data of this size (119 individuals) handedness and sex are not associated, they occur independently (χ^2 = 2.550, 1 degree of freedom, P = 0.110).

8. **Variations and elaborations.** If counts are low, especially expected count (see details of method in textbooks) below 5, then beware of modifications (e.g. Yates's correction) to take account of this. The version of the test outlined above is the contingency test. Another version is the goodness-of-fit test where instead of a null hypothesis of no association there is a null hypothesis to test particular frequencies in each group. This hypothesis may come from theory or previous experience. The data are organized in a table with the actual counts in the groups as one column (or row) and the expected counts (from the hypothesis of the frequency in each group) as another column (or row). The degrees of freedom are the number of groups minus one.
9. **Equivalent non-parametric test.** Not applicable: the chi-squared test is a non-parametric test.

Correlation

1. **Purpose.** To look for association (linear correlation) between two measured attributes (Pearson's correlation).
2. **Data.** Two measurements from each case or item or subject, that can be plotted on a scatter graph (does not matter for the statistics which measurement is x). *Always* plot the graph, even if only a sketch.
3. **Null hypothesis.** No (straight-line) association between the two measurements; the points on the graph are a random cloud or occur in a horizontal or vertical line.
4. **Test statistic.** The correlation coefficient, r, measures the strength of association, between 0 for no association to –1 for perfect negative association, to +1 for perfect positive association. Test of significance is usually based on Student's t-test, and tables of critical values of r are available for direct assessment.
5. **Degrees of freedom.** For sample of n cases, df = n–2. Note that there are n *cases* (or items or subjects), each case with two numbers.
6. **Assumptions.** Data from a random and independent sample of the population, where there is an underlying straight-line relationship between the measurements. The measurements are distributed bivariate Normally.
7. **Example.** Scatter graph of sunshine and rainfall recorded at Sheffield (monthly means for the period 1981–2010), and the correlation between them.



Statement of results. There is a moderately strong correlation (r = –0.699) between monthly totals of rainfall and sunshine; the correlation is significant (10 degrees of freedom, P = 0.0115). It is a negative correlation, i.e. high sunshine tends to be associated with low rainfall.

8. **Variations and elaborations.** None: correlation is a simple and fairly crude technique. If there is a statistically significant correlation then it may prompt further investigation and gathering of data, for example to predict one measurement from the other using regression.
9. **Equivalent non-parametric test.** There are two methods of correlation when the data are ordinal: Spearman's and Kendall's rank correlations.

The t-Test

1. **Purpose.** To compare the means of two groups.
2. **Data.** Measurements for the cases or items or subjects in the two groups.
3. **Null hypothesis.** No difference between the means of the two groups; the groups are samples drawn from the same population.
4. **Test statistic.** Student's t.
5. **Degrees of freedom.** For two groups of n_A and n_B cases, df = $n_A + n_B - 2$.
6. **Assumptions.** Data from random and independent samples of the populations, where the measurements have underlying Normal distributions which are equal in variability.
7. **Example.** Is the distance between the eyes different in males and females? The data are the interpupillary distances measured when setting up a binocular microscope by undergraduates taking APS116 in October 2005.

   Summary of data.

	Males	**Females**
Number	53	90
Mean	62.36 mm	59.47 mm
Standard deviation	3.470 mm	3.032 mm

   $n_A = 53$, $n_B = 90$; df = 141
   Student's t = 5.22
   P = 0.00000063, ***

   Statement of results. The difference in interpupillary distance is statistically highly significant (t = 5.22, 141 degrees of freedom, P = 0.00000063; two-sample t-test, homogeneous variances). On average the interpupillary distance is 4.6% smaller in females than in males. The difference is 2.89 mm with 95% confidence limits of 1.79 to 3.99 mm.
8. **Variations and elaborations.** For marked unequal variability of the two groups, especially with small groups or unequal size groups, then use a t-test designed for this, known as separate variances t-test, Behrens–Fisher test, Welch test or Satterthwaite-adjusted t-test, These tests are best carried out with a computer package.

   The confidence limits for the difference between the means can also be calculated, to examine practical significance (and put more emphasis on the size of effects than on hypothesis testing). If the confidence limits include zero then the difference will not be statistically significant (at the same probability, e.g. P = 0.05 for 95% confidence interval).
9. **Equivalent non-parametric test.** The non-parametric test that is analogous to the t-test is called the Mann–Whitney test or U-test or Wilcoxon test or Mann–Whitney–Wilcoxon test.

Analysis of Variance (ANOVAR)

1. **Purpose.** To compare the means of two or more groups.
2. **Data.** Measurements for the cases or items or subjects in the groups.
3. **Null hypothesis.** No difference between the means of the groups; the groups are samples drawn from the same population.
4. **Test statistic.** Variance ratio, F.
5. **Degrees of freedom.** The variance ratio has two degrees of freedom, first for the hypothesis being tested and second for the residual variation, spoken, for example, as "5 on 26 degrees of freedom". The first df is the number of groups minus one in the initial analysis, or the number of means being compared minus one in further analyses. The residual df is calculated from the ANOVAR table; in the simplest example of k groups each with n cases, residual df = k(n – 1).
6. **Assumptions.** Data from random and independent samples of the populations, where the measurements have underlying Normal distributions which are equal in variability (technically, the residuals (observation minus treatment mean) must come from one Normal distribution). Differences between group means arise from a small amount added or subtracted, not so large as to be multiplied or divided by a (mathematical) factor.
7. **Example.** Results from an experiment on growing potatoes with different amounts of fertilizer (imaginary data).

Table of treatment means

Fertilizer (kg/ha)	0	50	100	150	200
Yield (t/ha)	20.5	29.5	37.0	39.5	39.0

Table of Analysis of Variance

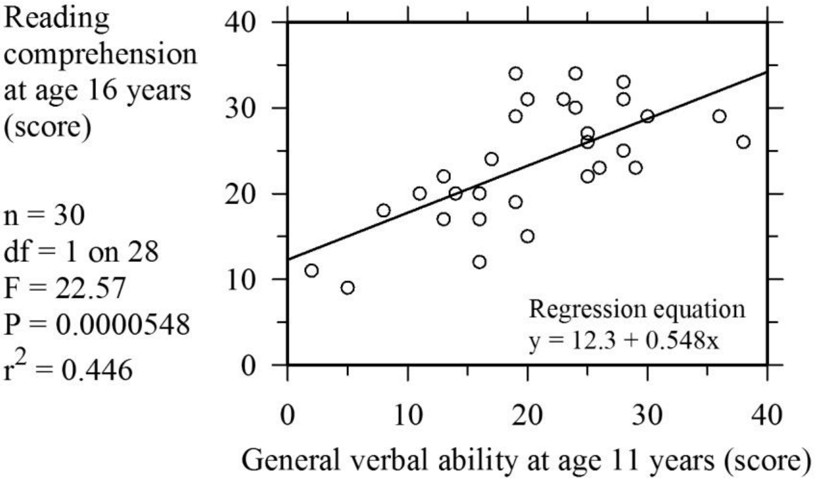
Source	Degrees of freedom	Sum-of-squares	Mean square	Value of F	Probability
Treatments	4	1050.80	262.70	5.53	0.0061 **
Residual	15	713.00	47.53		
Total	19	1763.80			

Statement of results. There is a very significant effect of fertilizer on yield: yield increases with increasing amount of fertilizer applied, especially at lower rates of application (F = 5.53, 4 on 15 degrees of freedom, P = 0.0061).

8. **Variations and elaborations.** Analysis of variance is a very general technique which has many variants for different sorts of experiments and different situations. The aim is always to calculate the background, uninteresting variation correctly and compare it with the variation for which there is an explanation (experimental treatments or groups identified in a survey).
9. **Equivalent non-parametric test.** The non-parametric tests that are analogous to ANOVAR are the Kruskal–Wallis test and the Friedmann test (each for particular versions of ANOVAR).

Regression

1. **Purpose.** To predict one measured attribute from another by fitting a line to the data and using the equation of the line.
2. **Data.** Two measurements from each case or item or subject, that can be plotted on a scatter graph with the measurement to be predicted (response or dependent variable) as y. *Always* plot the graph, even if only a sketch.
3. **Null hypothesis.** The slope of the fitted line is zero, i.e. the predicted y value will be constant whatever the value of x.
4. **Test statistic.** Variance ratio, F, for the statistical significance of the regression. The coefficient of determination, r^2, indicates the proportion of variation explained by the regression.
5. **Degrees of freedom.** The variance ratio has two degrees of freedom, first for the regression, always one df, and second for the residual variation, spoken, for example, as "1 on 26 degrees of freedom". The residual df for a regression is n–2, where n is the number of *cases* (or items or subjects) with paired values (x and y).
6. **Assumptions.** Data from a random and independent sample of the population, where there is an underlying straight-line relationship between the measurements (in the ranges of x and y used; extrapolation beyond these ranges is unreliable). The x values known without error so that all the variability in the point on the graph is in the y value. The variability of the y values is homogeneous (the same all along the x-axis, between smallest and largest value of x) and Normally distributed.
7. **Example.** Regression of score of reading comprehension when aged 16 years on score of general verbal ability when aged 11 years, for 30 children drawn at random from the National Child Development Study (children born in 1958).



Statement of results. The regression of reading comprehension (age 16) on verbal ability (age 11) is highly significant (F = 22.57, 1 on 28 degrees of freedom, P = 0.0000548). The reading comprehension (age 16) can be predicted from the equation y = 12.3 + 0.548x where y is the score for reading comprehension (age 16) and x is the score for general verbal ability (age 11).

8. **Variations and elaborations.** This is simple regression for fitting a straight line. The method can be elaborated for curves of all sorts, and for more than one x value being used to predict y (multiple regression).
9. **Equivalent non-parametric test.** There are non-parametric methods for regression but no simple equivalent for the straight-line regression used here.

Appendix A
Background information

A1. **Degrees of freedom.** Even statisticians have trouble with this term. Brian Everitt (The Cambridge Dictionary of Statistics, 1998) starts his technical definition: "an elusive concept that occurs throughout statistics"! The degrees of freedom indicates how many independent pieces of information there are in the data, or how many opportunities there are for the data to vary. Loosely speaking, the number of degrees of freedom is a method of saying how large the set of data is, in a way that is relevant to the particular statistical test. We are calculating the probability of getting a certain result (our result or a more extreme one) if the null hypothesis is true. Clearly this probability will depend in part on the size of the set of data, in other words, on how many possibilities there are of obtaining this result. The larger the set of data, the more possibilities there are.

A2. **Frequentist and Bayesian statistics.** The classical statistics that we are using is called frequentist because it uses the overall frequency of events as the probability of a single event occurring. That is why the explanations have phrases such as "if we sampled repeatedly" or "if we performed this experiment many times" or "in the long run". It leads to the results of statistic tests being couched in what seems convoluted terms of the probability of this set of data (or one more extreme) occurring if a particular hypothesis was true. That is why it is not possible to assign a probability to any particular hypothesis, only to the data given the null hypothesis.

An alternative approach is to take the set of data as given and find the probability of the hypothesis. This is the province of Bayesian statistics which is slowly becoming more widely used in many research fields. You need to be aware of the Bayesian approach because it may turn up in research that you are reading, or you may find that you need to learn about it and use it yourself. Bayesian statistics requires a deeper knowledge of probability and mathematical logic.

The main reason for knowing of the difference between frequentist and Bayesian statistics is that it helps to understand the exact meaning of the probability produced in statistical testing (§4). Frequentist (classical) statistics provides the probability of obtaining data like the results (or more extreme) given the null hypothesis (usually that there is no difference between means, or no association). In contrast, Bayesian statistics provides the probability of the hypothesis given the data found.

Appendix B
Using statistical tables

B1. **From probability to significance.** Computer packages analysing data provide an exact figure for the probability associated with the calculated test statistic. You can interpret it as a level of statistical significance using Figure 1. You need to think carefully about the range in which your value of probability falls ($P \geq 0.05$, or $P < 0.05$ but ≥ 0.01, and so on). Give the exact probability figure when writing the conclusion of the statistical test, whenever you can, because it is part of the evidence for your conclusion (see §13). Rounding off to three significant figures is a suitable measure of precision for these probabilities.

If you calculate the test statistic by hand and calculator then you can use statistical tables to find the range of probability associated with the value—see below. Once you have the range of probability, it is easy to interpret this as statistical significance using Figure 1. Alternatively, there are resources on websites which will calculate the exact probability given your value of the test statistic and one or two other inputs, typically degrees of freedom.

There are many of these resources and currently I prefer the one at http://danielsoper.com/statscalc3 (look for "Probability (p-Values)" and then for the relevant test statistic). Be sceptical: always check one of these resources with a couple of entries where you know the answer from a table of critical values. For instance, suppose you use Daniel Soper's p-Value Calculator for a Student t-Test. Try entering 9 degrees of freedom and t = 2.262 (from a t-table, column for P = 0.05). Two answers are produced, one-tailed and two-tailed, and the two-tailed value of probability is 0.05001285 which is close to the 0.05 that you are expecting. (We have not used one-tailed statistics in this module—consult statistical textbooks if you wish to find out more). Now we have confidence in the resource, and realize that we will need to use the two-tailed value. For this reason—to check web-resources—knowing how to use statistical tables is still a valuable skill, so read below.

B2. **How to use a table of critical values.** There is a separate table of critical values (as they are called) for each test statistic, often found at the back of statistical textbooks. Beware of statistical tables obtained from websites because they may not be the common ones that you need but more specialized tables. It is safer to use those at the back of statistical textbooks. In particular, you need the table for two-tailed t-tests, not one-tailed, for work in this module.

Start with the column (sometimes a separate table) for P = 0.05. Enter the table using the number of degrees of freedom for the test statistic. (Sometimes df is denoted by ν, the Greek letter nu, lower case. It looks remarkably like a letter v but is actually nu; see here in larger size: ν (nu) versus v.) If the calculated value is larger than the tabulated value (for a given P value and degrees of freedom) then the result is declared statistically significant, at the stated probability. If significant with the P = 0.05 column, you can proceed to the next column for P = 0.01 and then to P = 0.001, to find the lowest probability for your calculated test statistic. This probably sounds obscure but once you have used statistical tables a few times you will get the hang of it.

B3. **Worked example of using a statistical table.** Suppose you have a correlation coefficient, r, of 0.756 with 12 degrees of freedom (from the 14 observations, since df = n–2 for correlation). A portion of the statistical table for the correlation coefficient is shown in Table B1. Find the row for 12 degrees of freedom. The critical value tabulated for $P = 0.05$ is 0.532 and the calculated value (0.756) is much larger than this so we have significance at least at $P < 0.05$. We go further, in this table literally further along the row for 12 degrees of freedom: the critical value for $P = 0.01$ is 0.661, again a hit; but at $P = 0.001$ the table has 0.780 and the calculated value does not exceed this. So the probability associated with this value of r (12 df) is $P < 0.01$ (but not $P < 0.001$) so you can claim that it is very significant, and code as ** (from Figure 1).

Table B1. Part of the table of critical values for the correlation coefficient, r. Extracted from Parker, R.E. (1979) Introductory Statistics for Biology, London (Arnold).

Degrees of freedom	**Probability** **0.05**	**0.01**	**0.001**
1	0.99692	0.999877	0.99999877
2	0.950	0.990	0.999
…			
…			
11	0.553	0.684	0.801
12	0.532	0.661	0.780
13	0.514	0.641	0.760
…			
…			
90	0.205	0.267	0.338
100	0.195	0.254	0.321

P.L. Mitchell, Dept. of Animal and Plant Sciences, and Mathematics and Statistics Help (MASH), University of Sheffield.
21st February 2018. P.L.Mitchell@Sheffield.ac.UK