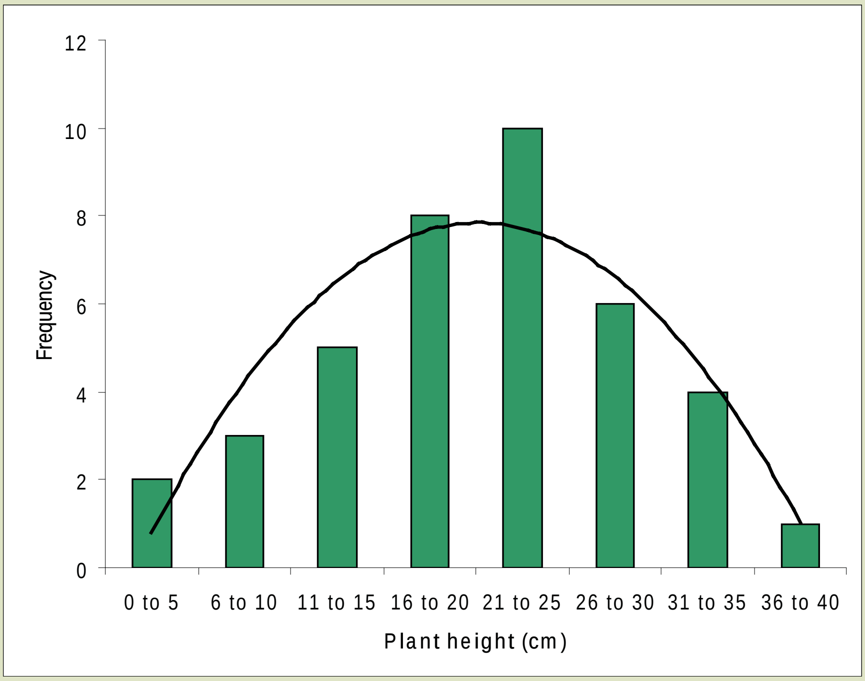
TROPICAL BIOLOGY ASSOCIATION

A simple guide to statistics



This booklet is designed as a 'refresher' to you as to why we do statistics; it provides the background to this question and introduces some of the statistical tests that you might use in your TBA projects. It does not tell you which test is right for your analysis, or what assumptions should be met before the test is valid. This information is provided elsewhere e.g. in the some of the books to be found in the TBA travelling library and you will need to refer to these before analysing data. Most of the tests covered can be used in the Minitab software package that is installed on all TBA laptops and this guide complements the 'Simple guide to Minitab' that will also be made available on your TBA course.

How you use this manual is up to you. You may wish to try some hand worked examples as you work your way through the book or plug some examples into the computer as you go. Or you may prefer to use the Guide as a reference manual, looking up specific tests as you need them. In either case, we hope you find this Guide takes some of the fear out of statistics: computer analysis is a tool like any other, and will take a lot of the hard work out of statistics once you feel you are in charge!

Acknowledgement.
This booklet was adapted for use on the TBA courses with the kind permission of Nigel Mann of the State University of New York and has greatly benefitted from additional input from Rosie Trevelyan (TBA), Kate Lessells of the Netherlands Institute of Ecology and Francis Gilbert of the University of Nottingham. Thanks also to Iain Matthews of the University of St Andrews for comments on an earlier draft.

For any queries concerning this document please contact:
Tropical Biology Association
Department of Zoology
Downing Street, Cambridge
CB2 3EJ
United Kingdom
Tel: +44 (0) 1223 336619
Fax: +44 (0) 1223 336676
e-mail: tba@tropical-biology.org

TROPICAL BIOLOGY ASSOCIATION

A simple guide to statistics

CONTENTS

INTRODUCTION

1.1 WHY STATISTICS ARE NECESSARY:

They allow **degree of objectivity** to be incorporated into assertions.

If statements arising from a study are to be regarded as other than ‘just-so stories’, they must be backed up by statistical data that suggest that **patterns found were not likely to be simply chance events**.

1) You will nearly always get some ‘effect’ by chance

2) Statistical tests allow us to answer the question:
How likely is it that I could have got this result (effect) by chance?
‘How likely is it?’ = ‘What is the probability?’
Statistical tests allow you to calculate this probability: *p*

3) If that probability is **sufficiently small**, then we conclude that the effect is not due to chance and that it’s a real effect
- that the effect is ‘statistically significant’
‘Sufficiently small’: *p* < 0.05
< 5%
< 1 in 20

1.2 WHAT DOES SIGNIFICANCE MEAN?

Often misunderstood!

If the null hypothesis is true, then ***p* = the probability of obtaining the data you actually have**.

- If this probability is less than 0.05 ($p < 0.05$), then we don’t believe the null hypothesis can be true, and we reject it.
- If $p > 0.05$, and we cannot reject the null hypothesis, this does not mean we accept the null hypothesis as true ! We merely have failed to reject it.
- 0.05 is an arbitrary level, but is conventional; two further levels, 0.01 (1 in 100), and 0.001 (1 in 1000).
- *p* is **NOT** the probability of the null hypothesis being true.

It is also important to be aware that:
- with a given size effect, statistical significance increases – *p* decreases – with increasing sample size

- with a given sample size, statistical significance increases – *p* decreases – with increasing ‘effect size’
For example, when tossing a coin and calculating the probability of it landing on ‘heads’: -

	60% Heads	80% Heads	100% Heads
10 throws	$p = 0.75$	$p = 0.057$	$p = 0.027$
100 throws	$p = 0.045$		
1000 throws	$p = 0.001$		

Hence, statistics provide us with an objective way of assessing whether an effect is real, or whether it might just be due to chance; but do not waste time and effort collecting data until $p < 0.000000001$!

1.3 SOME TERMS AND CONCEPTS

- **Data** (singular 'datum') - numbers generated in an experiment/study. Data originate from observations - the measurements, or counts contributed to by each unit in the sample.

1.3.1 Levels of measurement of data

- **Discontinuous** variables/data - usually whole integers, mostly counts, or frequencies of things.
- **Continuous** data - values along a scale, usually measurements (e.g. height, mass, temperature, etc.).

4 levels of measurement:

1. **Nominal/categorical** scale.
   Each observation falls into one of two or more categories e.g. if looking at sex ratio of a species, each individual classified as male, or female. Each individual does not have order of magnitude associated with it.
2. **Ordinal** scale.
   Each observation provides a score, and observations within sample can be ranked from low to high. However, the ordinal numbers do not indicate absolute quantities - intervals between numbers on the scale not necessarily equal. For example, plant species can be ranked on the 'DAFOR' scale, whereby they are classified as dominant, abundant, frequent, occasional or rare. No expectation that dominant organism is, say, 2x more common than abundant species.
3. **Interval** scale.
   Data can be ranked, but now distances between two adjacent points on scale will be the same. Dates and temperature (^{o}C) are on interval scale - there is validity to subtracting one point on the scale from another, to get a measure of the amount of time that has passed, or the change in temperature. However, it is not valid to talk in terms of one point along the scale being 2x, or 3x, more than another - because no absolute 0 in the scales (i.e. May 3rd cannot be expressed as being a certain number of times larger, or later, or whatever, than April 26th!).
4. **Ratio** scale.
   Does have absolute zero. There is meaning to saying that an observation is x times larger, longer, heavier, faster, etc. than another.

As far as statistical test selection goes, interval and ratio scales effectively lead to the same type of test (see below).

1.3.2 Other terms

- **Variable** - used as noun. A characteristic that differs between individuals e.g. size, shape, diet, any biotic or abiotic factor, including behaviour.

In a study, variables are measured, or controlled for. In an experiment, the experimenter usually manipulates one, or more, variables to see effect on another.

The variable manipulated, or controlled, is the **independent variable**. It is also known as the **predictor** variable i.e. the hypothesized (predicted) effects influencing the dependent variable.

Effect is measured on a **dependent variable** (a study tests whether the scores for latter are dependent on scores of former). The dependent variable is also described as the **response** variable; it is what the prediction relates to and the variable that changes in response to the hypothesized effects.

e.g. looking for a relationship between group size of red colobus monkeys and home range in Kibale Forest, Uganda:

- group size is independent variable
- home range is the dependent variable.

In a different study, a connection might be sought between forest type and size of red colobus groups - now group size becomes the dependent variable.

- **Sample** - a subset of the population that contributes to the analysis in the study; - an assumption is that the sample is representative of that population.

Sampling, of course, must be done in a **random** fashion, so that no bias occurs in the data. When you want to generalise, you should take care to randomise treatments and/or samples properly (i.e. all have an equal chance of being selected). The example below illustrates the types of sampling bias that might occur in a study of bird behaviour in Kirindy Forest, Madagascar.

Random sampling:

Souimanga sunbird — *Chadzia* — *Hildegardia*

Does the diurnal pattern of foraging by Souimanga sunbirds differ between *Chadzia* and *Hildegardia?*

Possible sampling biases:
- Places where data collected
- Person collecting the data vs time of day, day of project or plant species
- Time of day or tree species vs day of project

Figure 1
Potential sampling biases in a field study.

- **Independence of observations** in a sample is often assumed in statistical procedures. Each measurement should be independent of all others, or if not, the non-independence must be specified and accounted for in the design of the study and the analysis.

i.e. the value of different observations should not be inherently linked to one another; without due care, non-independence is especially likely where groups, broods, or litters, of animals are being studied.

- **Replication** is required in virtually every type of study and is essential to avoid the problem of non-independence of data. The examples below and overleaf illustrate the nature of the potential problem.

Question: Do male millipedes have longer legs than female millipedes?

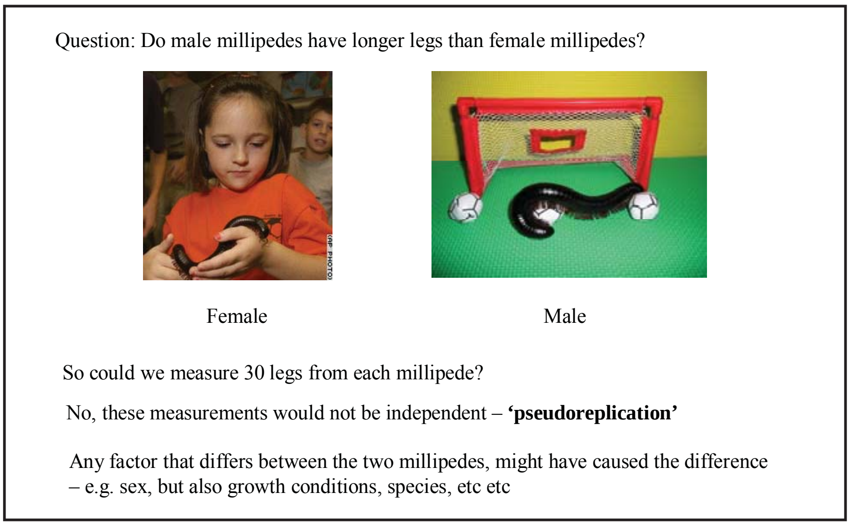
Female — Male

Could we measure just one leg from each millipede?

No, we need to **replicate** the measurements

Figure 2a
How do we ensure replication?

Figure 2b
How do we avoid pseudoreplication?.



Question: Do male millipedes have longer legs than female millipedes?

Female

Male

So could we measure 30 legs from each millipede?

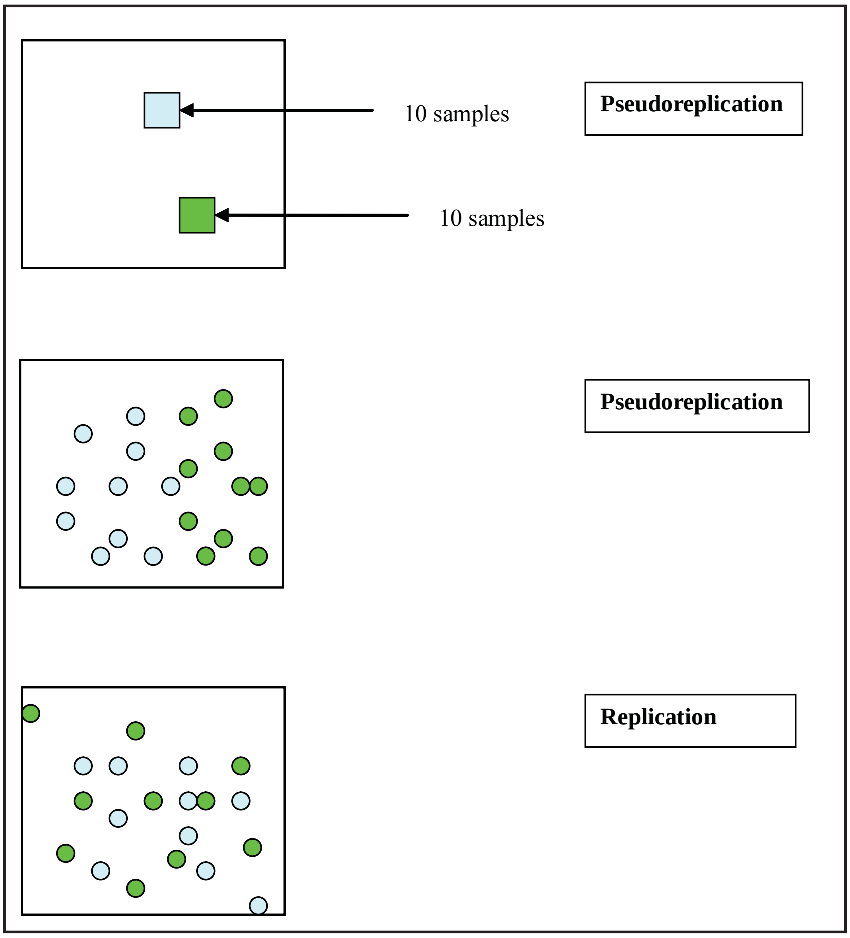
No, these measurements would not be independent – **'pseudoreplication'**

Any factor that differs between the two millipedes, might have caused the difference – e.g. sex, but also growth conditions, species, etc etc

To avoid the problem of pseudoreplication in this example, measure legs from 30 individual female and male millipedes.

- **Psuedoreplication** demonstrates instances where 'treatments are not replicated or replicates are not statistically independent'.

Figure 3
Project design to avoid pseudoreplication.



- **Observer bias** must also be avoided - conscious or subconscious selection, or rejection, of data, for example, that help to validate an earlier prediction.

STATISTICS

Two broad categories of statistics:

- **descriptive** (or summary) stats - used to organise, summarise and describe measures of sample.
- **inferential** (deductive, or analytical) stats - to infer, or predict, population parameters (i.e. to make statements, using probability, about general patterns, based around the sample measures).

2.1 DESCRIPTIVE STATISTICS

Includes measures of average, and variance around them (which give an idea of the spread of data). Averages are measured using **mean, median or mode** and are measures of **central tendency** of data - a single measure, close to centre of distribution of observations, representative of the whole.

- **Mean**

Use population mean, μ = (Σx)/N

or (usual case), an approximation to this, $\bar{x} = \frac{\sum x}{n}$

(NB. $\bar{x}$ = ‘x-bar’).
N = number of items/observations in population
n= number of items/observations in sample
x = each observation.

The mean is strongly affected by extreme results i.e. one very large value can make the mean higher than all other values in a sample; in this case, the mean would not be a very representative average score.

In such cases, it is better to use:

- **Median**

The ‘middle value’, when observations listed in rank order. As it is an ordinal statistic, not all values need to be known – e.g. enough to know that absent values are above certain number (the median life span of 10 pet rabbits can be calculated even if 3 are still alive). Medians can be worked out when data fall into classes.

- **Mode**

In a frequency distribution the mode is the class containing most values.
A frequency distribution may be bimodal, or multimodal. In such cases, usually necessary to carry out separate analyses on the discrete population categories within which the data *are* symmetrical. Mode is most often used as a quick and easy approximate measure of central tendency.

2.1.1 Measuring variability

Populations vary in characteristics, hence the need for statistics - if there was no variation, one value alone would tell us all about that character for all individuals. Giving a measure of the average is not enough - we also need to know something about the variability within samples. Such measures include range, standard deviation and variance.

- **Range** (max – min)

Takes account of two most extreme observations - it is a subtraction of lower from higher.

- **IQR**

A quartile is any one of the 3 values that divide a data set into 4 equal parts; each part representing 1/4th of the sorted sample population. The inter quartile range is a measure of statistical dispersion, and is equal to the difference between the third and first quartiles. As 25% of the data are less than or equal to the first quartile and the same proportion are greater than or equal to the third quartile, the IQR will include about half of the data. The IQR has the same units as the data and because it uses the middle 50%, is not affected by outliers or extreme values. The IQR is also equal to the length of the box in a box plot.

- **Standard deviation**

Most widely used measure of variability.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Can also be denoted by σ.

If working by hand, use:

$$s = \sqrt{\left(\frac{\sum x^2}{n - 1}\right) - \bar{x}^2}$$

(*n*-1) as denominator, not just *n*, for small samples (<30). Has effect of increasing the standard deviation, ‘s’. Described as being like a tax - we are only able to use sample mean and not true mean, so cannot be sure of our data set: hence standard deviation with (n-1) degrees of freedom. As general rule, (n-1) calculator button should always be used in calculations.

- **Variance**

Is the square of the standard deviation, and is important in various statistical calculations.

2.1.2 The confidence in our estimates

Intuitively, larger samples should give better estimate of population mean than smaller samples. Some samples will give larger mean estimate and smaller standard deviation, just by chance. The array of sample means are thus expected to form a normal distribution around the ‘true’ mean:

standard deviation of sample means = **standard error of mean**.

From properties of normal distribution:

68% of large number of sample means fall within ± 1 S.E. of the population mean (μ).

Similarly, the population mean (which we are trying to approximate), has a 68% likelihood of falling within ± 1 S.E. of the sample mean, $\bar{x}$.

S.E. can be calculated from $\frac{s}{\sqrt{n}}$ where *n* = # of observations.

(Note: as *n* gets larger, S.E. gets smaller - i.e. with larger sample size, we become more confident of mean estimate).

e.g. If mean is 70 and S.E. is 2.1, we can be 68% confident that population mean is between 70 ± 2.1.
- more informative than standard deviation, as it is usually the population mean that interests us.

To raise confidence limit to 95%, then use S.E. x 1.96*.

= **95% confidence interval** around true mean.

For above example, 95% confidence interval around the mean

= 70 ± (1.96 x 2.1)

So 95% sure that true mean between 65.884 and 74.116.

This only valid with samples > 30 observations.
With smaller samples, where we cannot be so confident about sample standard deviation, the z-score of 1.96 is not used:
- instead a t-score is used, the size of which is dependant on sample size. Found in table, against appropriate number of degrees of freedom (n-1).

* = the z-score when using a 95% confidence level. Z – scores are also known as standard scores and are calculated by subtracting the population mean from an individual raw score and dividing the difference by the population SD.

2.1.3 Graphing means

As well as plotting means on, best also to incorporate standard error bars, or (even better) bars signifying 95% confidence limits around sample mean, within which true mean lies. If confidence limits around sample means in the different conditions do not overlap, we can be reasonably sure that means are genuinely different – i.e. they are effectively from different populations. If medians are used, 95% confidence bars can also be drawn, but in this case S.E. is not used.

2.2 INFERENTIAL STATISTICS

Statistically significant outcome: when event occurs, whose probability is below certain threshold. The threshold is usually *p*<0.05, but may be P<0.01, or lower (see above). Lower thresholds are usually required when it is more important not to make mistake; by arguing that an effect occurred when really the difference was due only to chance (e.g. when testing drugs against placebos).

Science progresses through testing of hypotheses. A hypothesis might be that calling frogs are predated more by bats than are non-callers.

The **experimental hypothesis** (H_1) - such an effect *will* occur, and the means of two samples, for example, will be different.

The **Null hypothesis** (H_0) - conditions being looked at, or tested, will have *no effect*. (Null = 'nothing').

Outcome of an experiment - based around either accepting or rejecting null hypotheses; ideas can't be proven - just rejected! Hence, **science procedes conservatively**.

A currently held belief is essentially just that - an explanation that has proved to fit best with the data, but maybe only until another better explanation comes along.

2.2.1 One-tailed and two-tailed tests

H_0: $\mu 1 = \mu 2$
Usually H_1 states only that $\mu 1 \neq \mu 2$.

If H_0 rejected, conclude that sample with larger mean has been drawn from population with larger mean - i.e. the difference is 'real' and not just an effect of chance.

Two-tailed test - where no *a priori* prediction has been made about which mean is the larger. Used in almost all cases.

One-tailed test - where H_1 predicts that $\mu1 > \mu2$. Less stringent than two-tailed, so statistically significant difference more likely to be found.

It is tempting to switch to one-tailed test in analysis to obtain significant result, where result would not be significant with two-tailed test. This is cheating!
- decision to use one-tailed test must be made before analysis, and for sound reasons (not just a 'hunch').

e.g. where a one-tailed test might be appropriate: testing of new pain killer against a placebo. Experiment should reveal whether it helps the symptom, or alternatively, has no effect at all. Because of *prior knowledge* of its benign nature, it is not feasible that the drug will make the problem worse, so a one-tailed test is fine.

Type 1 and type 2 errors

- **Type 1: to falsely reject H_0.** Can be reduced by setting a lower threshold for significance. At p = 0.05, type one errors occur 5x out of 100.

- **Type 2: to falsely accept H_0.** Often occurs with small samples, where the data are too few to have much chance of discovering an underlying significant effect.

2.2.2 Differences or trends

May be concerned with either:

- **differences between two or more groups**

Here, groups could be sexes, age categories, experimental versus control conditions, happy versus sad people, etc.

Or:

- **trends between variables**.

In this case, looking for relationship between two more-or-less continuously varying measures. For example, is there a relationship between height and weight, the amount of a drug and its effect, etc.? Such data can be graphed as a scatter plot and may be positive, negative or no correlation. A regression equation can be fitted to the plot, and its statistical significance measured. At its simplest, the data can be tested to see if they form a pattern significantly different than a straight horizontal line (the H_0 - i.e. altering independent variable has no effect on dependent variable).

2.3 CHOOSING THE APPROPRIATE STATISTICAL TEST

One of the most crucial skills that must be learnt by a biologist. Researchers must know how to select appropriate test from many available. Even the laborious calculations have now largely been cut out due to computer packages, so carrying out tests is very quick and errors in test selection are easily made.

2.3.1 Parametric and non-parametric tests

Parametric - used for analysing data that obey certain assumptions.
- data sets should be **normally distributed**, with same **variance**.
- data must also only consist of actual observations, not percentages, ratios, etc.

Such tests are used to compare means. For example:

- matched and unmatched T-tests
- F-tests
- analysis of variance (ANOVA)
- Pearson correlation coefficient.

NB. These tests are not dealt with in the remainder of this guide but will be incorporated in a future TBA Skills Series booklet.

Non-parametric tests - no such assumptions.

Often most appropriate with biology field data, where samples collected are too small for us to be confident about shape of their distribution*, or where there are often reasons that cause data to be actually skewed. Can be used to analyse percentages, ratios, ranked data.

Tests compare medians. For example:

- Wilcoxon matched-pairs test
- Mann-Whitney U-test
- Kruskal-Wallis
- Spearman Rank correlation.

The remainder of this guide deals with these tests and others (binomial test & chi-squared) that you are most likely to use during your TBA project. Parametric tests and GLM (general linear models) are dealt with in separate guides.

* See the TBA ‘Simple guide to Minitab’ to find tests for normal distribution and variance.

2.3.2 The binomial test

Used where data fall into one of two categories, and want to ask if the distributon is random. Such as, given a choice, do sunbirds at Amani prefer to approach red or purple flowers?

If only three birds used in test, and red is chosen by each, we can ask what the probability is of getting this outcome, i.e. of ‘r r r’. We need to work out what all possible outcomes are:

r r r	r r p	r p r	r p p
p p p	p p r	p r p	p r r

So the probability of (r r r) is 1/8 = 0.125 and therefore H_0 cannot be rejected.

If 20 trials were carried out (each with a different bird - each trial must be completely independent), and red selected 15 times, we should again list all the possible ways in which we could obtain an outcome as, or more, extreme than this, and hence again come up with a probability that the result was down to just chance.

However, such a process is laborious. Binomial theory provides a formula that can be used instead:

$$p = \frac{k!}{x!(k-x)!} \times px \times q(k-x)$$

where

p = probability of a particular combination
k = number of trials/events
x = stated number of a particular outcome (e.g. red)
p = probability of a particular outcome (e.g. red)
q = probability of the other outcome (e.g. purple)

So for the 20 trials above,

$$p = \frac{20!}{15!(20 - 15)!} \text{ x } 0.515 \text{ x } 0.5(20 - 15) = 0.015$$

So, the probability of obtaining this outcome is 15/1000 = 0.015 and H_0 can be rejected.

2.3.3 Chi-squared test (χ^2)

This can be used with the same sort of data as a binomial test, but the latter is better (more powerful), when there are only two categories for observations.

With χ^2, expected values have to be created, assuming H_0. For example, if 16 male sunbirds were observed feeding on red flowers in a sample of 24, we can ask if this could be expected to occur just by chance:

	Males	Females
Observed:	16	8
Expected:	12	12

To calculate deviation from expected pattern, subtract the two expected values from the observed values, and then these differences could be added together:

(16-12) + (8-12) = 0.

But this will always equal 0, so gets us nowhere. Therefore, to overcome the counterbalancing effects of positive and negative signs, ‘squares’ of the differences are added together:

$(16\text{-}12)^2 + (8\text{-}12)^2 = 32$.

The problem now is that this doesn’t take into account sample size: a difference of 4 from the expected value, when that value is a low number like 12, is far more significant than difference of 4 when the expected is 1200. So the formula must take sample size into account and does so by dividing each square of the difference by the expected value, before adding together. This gives us the chi-squared formula:

$$\chi^2 = \cdot\frac{\text{(observed - expected)}^2}{\text{expected}}$$

χ^2 table to find out associated probability:

p:	0.50	0.20	0.10	0.05	0.01
χ^2:	0.86	1.64	2.71	3.84	6.64

For the example:
χ^2 = 2.67, so probability of obtaining outcome by chance is between 0.10 and 0.20.
i.e. H_0 not rejected.

This application of χ^2 can be used for multiple categories - e.g if there are 4 categories, then expected number of outcomes in each is 0.25 multiplied by the sample size.

If the χ^2 test is used when there are just 2 categories, then stricly **Yate's correction** should be applied: involves subtracting 0.5 from the numerator in equation, before squaring. The subtraction is made from the absolute value of the difference between O and E (i.e. the sign is ignored, if negative).

So:

$\chi^2 = \cdot\frac{(|O - E| - 0.5)^2}{E}$ where just two categories.

2.3.4 Chi-squared contingency tables

In the above example, observed frequencies were distributed across one row of categories. However, sometimes observations within each sampling unit can fall into one of several categories as well and this will produce a two-way contingency table.

For example, a researcher wants to know if there is a significant difference between the number of fights a male fiddler crab wins when within its territory and when off its territory.

	wins	losses	totals
On territory	10	10	20
Off territory	10	30	40
totals	20	40	60

To calculate expected frequencies needed for χ^2 test:

- look on lower totals row, taking the sample of fights as a whole (i.e. with the assumption that it makes no difference whether on or off territory).

- you can see that 1/3 of all fights are won, and 2/3 are lost. The expectation, if H_0 correct, is that the 1/3 : 2/3 ratio will apply to both on and off territory.

So expected values are:

	wins	losses	totals
on	20x1/3	20x2/3	20
off	40x1/3	40x2/3	40
totals	20	40	60

In the general table below, the expected values are:

a	(a+b)(a+c)/(a+b+c+d)
b	(a+b)(b+d)/(a+b+c+d)
c	(a+c)(c+d)/(a+b+c+d)
d	(b+d)(c+d)/(a+b+c+d)

		total
a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d

In a further example, after day in the field, students were scored as having either wet or dry clothes, and a link was sought between this and the class they belonged to:

	WET		**DRY**		**TOTALS**
Ecologists	5	*6.25*	25	*17.5*	30
Primatologists	15	*9.58*	31	*26.8*	46
Totals	20		56		96

Expected results in italics.

$$\chi^2 = \frac{(5\text{-}6.25)^2}{6.25} + \frac{(25\text{-}17.5)^2}{17.5} + \frac{(15 - 9.58)^2}{9.58} + \frac{(31 - 26.8)^2}{26.8}$$

$$= 7.19$$

Gives significant result at $p < 0.01$.

However, we must be careful in drawing a conclusion that primatologists get wetter than ecologists! Were the data independent? They are not if all primatologists were out together in a group, so were acting as a unit, not as separate individuals. If this was the case then the analysis would be invalid.

In all examples above, the **degrees of freedom** is 1. This is calculated as (number of rows - 1) x (number of columns - 1).

Degrees of freedom

= number of observations in a particular test which can take on any value. One degree of freedom lost for every fixed value (e.g. if we have the mean for a set of data, then knowledge of all values but one will dictate the missing value).

2.3.5 Mann-Whitney U test

Tests for difference between medians of 2 samples, where no logical connections between any point in one column and a specific point in the other. As with most non-parametric tests, Mann-Whitney assesses the ranks of the observations within each sample.

Example:
In a study looking at territory size of little greenbuls inhabiting both the small and large forest patches in and around Kibale, the following data were collected:

Small forest	Rank	Large forest	Rank
5	1	9	5.5
7	4	15	10.5
12	8.5	18	12
15	10.5	25	13
6	2.5	6	2.5
9	5.5	12	8.5
10	7		
Total	39		52

If no difference between the samples, then each column would show equal spread of low and high ranks.

The test statistic U is calculated with the formula:

$$\mathbf{U_1 = n_1n_2 + \frac{n_2(n_2+1)}{2} - R_2}$$

where n_1 = number of observations in first column
n_2 = number of observations in second column
R_2= sum of the ranks in the second column

Then calculate $\mathbf{U_2 = n_1n_2\text{-}U_2}$

Whichever is lower, U_1 or U_2, then looked up in table to determine whether a significant result at sample size of n_1 and n_2.

In the example above, $U_1 = (7x6) + \frac{6(6+1)}{2} - 52 = 11$

and $U_2 = (7x6) - 11 = 31$

Looking up 11 in the table, with samples of 6 and 7: probability of obtaining this result by chance is > 0.05. Critical value is 6, hence H_0 cannot be rejected.

Note: In Minitab the test statistic is given as 'W'.

2.3.6 Kruskal-Wallis test

Medians of several samples can be compared using several Mann-Whitney U tests, but such multiple comparisons run the risk of making type 1 errors. A significant result is accepted at probability of $p = 0.05$, so there is always a 1 in 20 chance of getting the conclusion wrong; as more tests that are done on the data, it becomes more likely that such a mistake will be made.

K-W test overcomes this problem by analysing in one step whether differences occur between the medians of several samples (the parametric ANOVA test does same, but considers means of several samples).

In K-W, there must be 5+ observations in each sample, but samples do not have to be of equal size.

ALL data are ranked together, irrespective of which column they are in. Ranks for each column then added and squared, and each value of R_2 is then divided by respective value of n.

The data are then entered into the following formula:

$$K = \left(\sum\left(\frac{R^2}{n}\right).\frac{12}{N\left(N+1\right)}\right) - 3(N+1)$$

K is then looked up in a χ^2 table, under (# of samples – 1) degrees of freedom.

Note that if the overall result is significant, you can only be really confident that the samples with the smallest and largest rank sums are significantly different from one another. To be certain where the differences lie, you would need to carry out a *post hoc* test (see standard statistics texts for details).

2.3.7 Wilcoxon Matched-Pairs test

Used when observations in one sample (data column) have a natural ‘partner’ in the other i.e. the medians of two *matched* samples are being compared. It may be that the matching stems from individuals being each tested in two different experimental conditions.

During the test each value in one column is subtracted from corresponding value in other. If H_0 correct, then number and magnitude of positive differences between pairs ≈ equal negative differences.

First the subtractions are done; clearly the pairs of data should occur on the same row. Then differences are ranked according to absolute values (i.e. ignoring signs). Next, the sum of ranks of -ve & +ve differences added up.

The smaller of the two rank sums is the test statistic, T.

The value of T is compared to the critical value for the appropriate sample size in the Wilcoxon table. If equal to, or lower than, the tabulated value, then the result is significant, and H_0 can be rejected.

The sample size for the test is the number of pairs for which difference between them ≠ 0. A sample of 6+ is necessary; if >30, use paired T-test (where your data are normally distributed).

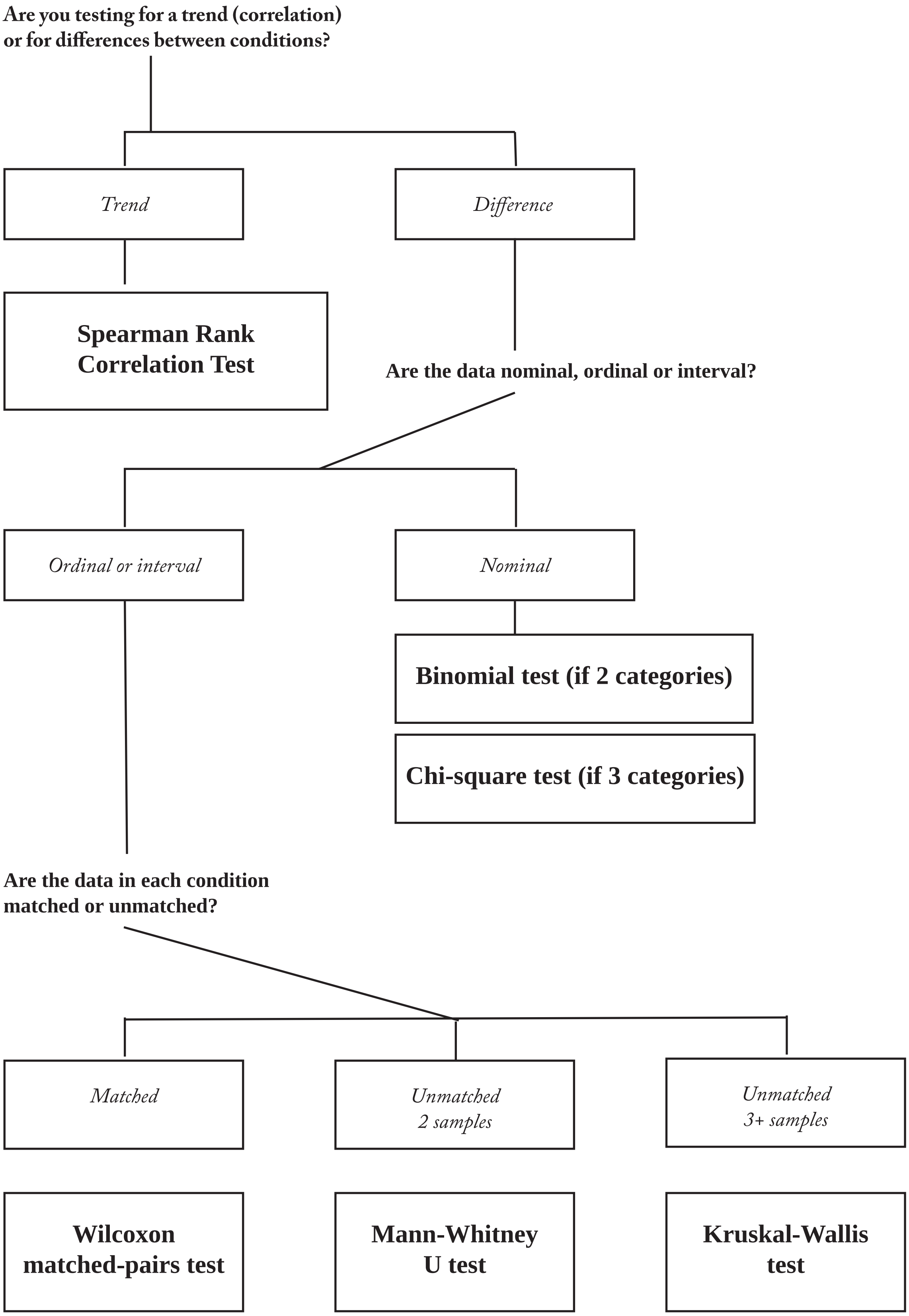
Example:
A study carried out to see if male birds in territories provided with extra food allocate more of their time to singing.

Male	Songs per hour		Diff.	Rank	+ve ranks	-ve ranks
	Provisioned	Not provisioned				
1	18	13	+5	5	5	
2	16	14	+2	3	3	
3	19	12	+7	7	7	
4	10	12	-2	3		3
5	11	10	+1	1	1	
6	20	12	+8	8	8	
7	12	12	0	-		
8	32	26	+6	6	6	
9	15	13	+2	3	3	
Total					33	3

Critical value (T) in Wilcoxon table for sample size of 8 is **3**, at $p = 0.05$.

This is same as that of the smallest rank sum, so H_0 is rejected - there *was* more singing when birds had extra food supplied in their territories.

2.4 SELECTING THE APPROPRIATE TEST (NON-PARAMETRIC DATA)



Are you testing for a trend (correlation) or for differences between conditions?

Trend

Spearman Rank Correlation Test

Difference

Are the data nominal, ordinal or interval?

Ordinal or interval

Nominal

Binomial test (if 2 categories)

Chi-square test (if 3 categories)

Are the data in each condition matched or unmatched?

Matched

Unmatched 2 samples

Unmatched 3+ samples

Wilcoxon matched-pairs test

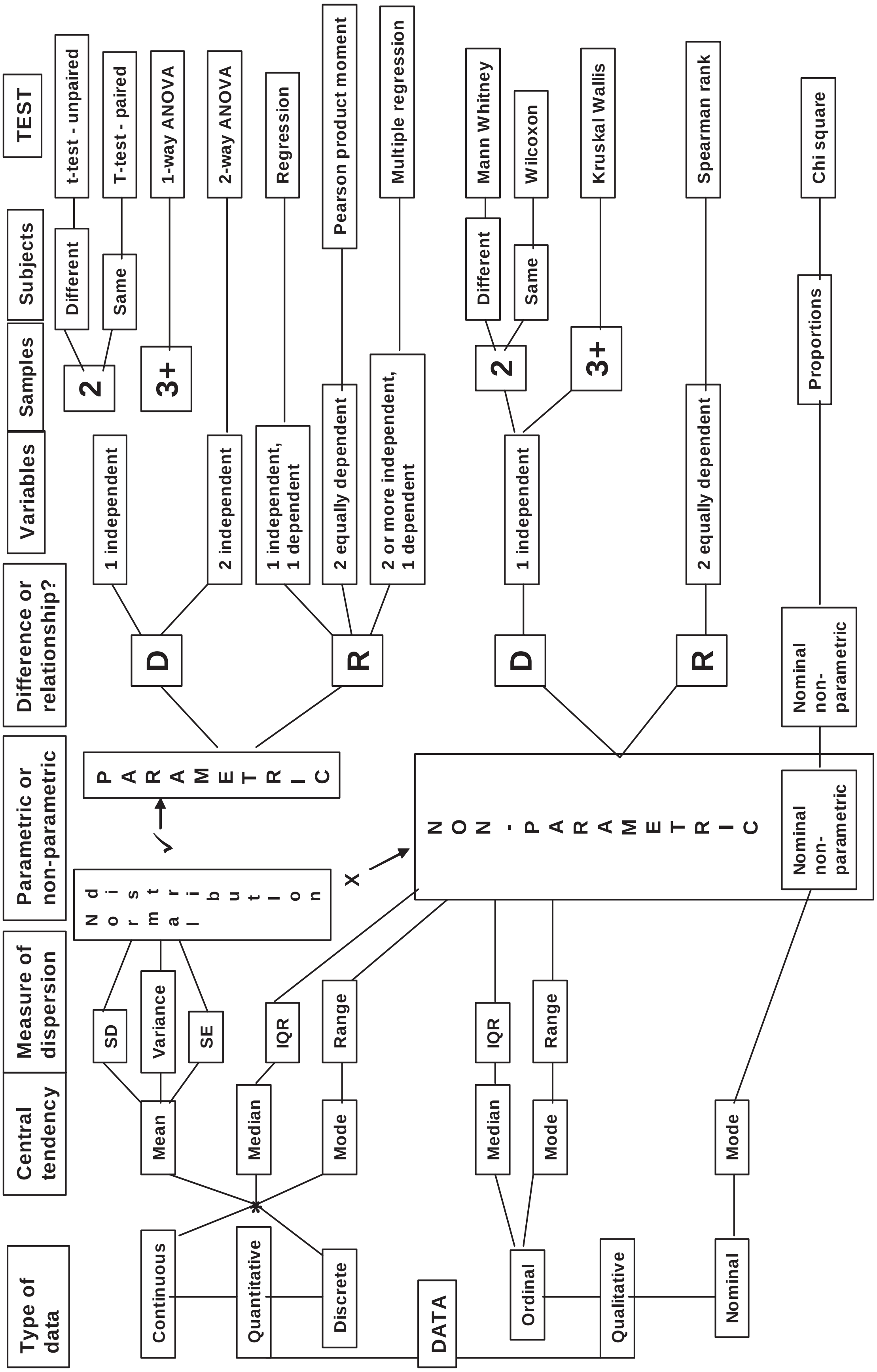
Mann-Whitney U test

Kruskal-Wallis test

95% confidence limits around median values

Sample size	r (for P = approx. 95%)
2	-
3	-
4	-
5	-
6	1
7	1
8	1
9	2
10	2
11	2
12	3
13	3
14	3
15	4
16	4
17	5
18	5
19	5
20	6
21	6
22	6
23	7
24	7
25	8
26	8
27	8
28	9

r = number of values in from the the extremes of the data set. The values at these points, above and below the median, represent the 95% confidence limits of the median.


Type of data
Central tendency
Measure of dispersion
Parametric or non-parametric
Difference or relationship?
Variables
Samples
Subjects
TEST
DATA
Continuous
Quantitative
Discrete
Ordinal
Qualitative
Nominal
Mean
Median
Mode
Median
Mode
Mode
SD
Variance
SE
IQR
Range
IQR
Range
Normal distribution
✓
X
PARAMETRIC
NON-PARAMETRIC
Nominal non-parametric
Nominal non-parametric
D
R
D
R
1 independent
2 independent
1 independent, 1 dependent
2 equally dependent
2 or more independent, 1 dependent
1 independent
2 equally dependent
Proportions
2
3+
2
3+
Different
Same
Different
Same
t-test - unpaired
T-test - paired
1-way ANOVA
2-way ANOVA
Regression
Pearson product moment
Multiple regression
Mann Whitney
Wilcoxon
Kruskal Wallis
Spearman rank
Chi square

TROPICAL BIOLOGY ASSOCIATION

Skills Series

This **statistics guide** was developed to complement the teaching on the Tropical Biology Assocation's field courses. These ecology and conservation field courses are based in East Africa and Madagascar. They are a tool to build capacity in tropical conservation. Lasting one month, the courses provide training in current concepts and techniques in tropical ecology and conservation as well skills needed for designing and carrying out field projects. Over 120 conservation biologists from both Africa and Europe are trained each year.

Tropical Biology Association

The Tropical Biology Association is a non-profit organization dedicated to providing professional training to individuals and institutions involved in the conservation and management of tropical environments. The TBA works in collaboration with African institutions to develop their capacity in natural resource management through field courses, training workshops and follow-up support.

European Office
Department of Zoology
Downing Street
Cambridge CB2 3EJ
United Kingdom
Tel. + 44 (0)1223 336619
Fax + 44 (0)1223 336676
email:tba@tropical-biology.org

African Office
Nature Kenya
PO BOX 44486
00100 - Nairobi, Kenya
Tel. +254 (0) 20 3749957
or 3746090
email:tba-africa@tropical-biology.org

www.tropical-biology.org