

R Module 3



t-tests and related comparisons

Certificate in EnviroStats (Non-Award)

This document is part of an online Certificate in EnviroStats (Non-Award) by the University of Canberra. Course enquiries can be directed to the address below. Expressions of interest in the course can be made online through:

<http://aerg.canberra.edu.au/envirostats>

Copies of this publication are available from:

The Institute for Applied Ecology
University of Canberra ACT 2601
Australia

Email: bernd.gruber@canberra.edu.au, georges@aerg.canberra.edu.au

Copyright © 2011 Arthur Georges, Bernd Gruber [Converted the manuscript from SAS to R]

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, including electronic, mechanical, photographic, or magnetic, without the prior written permission of the author.

R is an open source statistical programme. It is developed by:

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

SPONSORED BY:

Materials development team:

Author:	Arthur Georges, 2002, 2006
Instructional designer:	Peter Donnan, 2002
Editor:	Loretta Barnard, 2002
Graphic Design:	Peter Delgado, 2002
Desktop Publishing:	Kristi McDonald, 2004 Sue Bebbington, 2004
FDDU Project Manager:	Deborah Veness, 2002
Dynamic Web Page Design:	TCNI Software Solutions PO Box 47 LATHAM ACT 2615 Australia

First prepared in January, 2002 for Semester 1, 2002.

Reprinted January 2003 for Semester 1, 2003.

Reprinted January 2004 for Semester 1, 2004.

Reprinted November 2004 for Semester 1, 2005.

Revised and reprinted, June 2006

Reprinted February 2007 for Semester 1, 2007

Revised combined with former SAS Module 3 and converted to R, January 2011

Published by Technology & Educational Design Services

(TEDS)
University of Canberra
ACT 2601, AUSTRALIA

Module 3 T-tests and related two sample comparisons

Key Concepts of Statistical Tests	5
The role of statistics in science	5
Parameter estimation	5
Hypothesis testing	6
Summary	8
Lesson 1: Estimation and Confidence Limits	9
Samples and populations	9
Sampling distributions	11
Confidence limits	13
Lesson 2: Hypothesis testing	15
Rationale	15
Formal procedure	17
One-tailed and two-tailed tests	19
Statistical significance	20
Type I and Type II errors	21
Power of the test.....	22
Importance or strength of result	22
Planning of experiments	23
Interpretation of non-significant results	24
Robustness	26
Degrees of freedom	26
Significance, power, strength – putting it all together	27
Lesson 3: The F-test	29
Comparing two variances	29
Lesson 4: The t-test	33
Comparing two means	33
Other analysis options	36
Wilcoxon rank-sum test	36
Paired t-test	37
The Wilcoxon signed-ranks test	38
Lesson 5: Application	39
Standard deviation or standard error?	39
Confidence limits or the T-test	40
Choosing a statistical test	41
Level of measurement	41
Are the assumptions tenable?	42
Normality.....	42
Homogeneity of variances.....	42
Randomness and independence	43
How can violations be overcome?	43
Reporting the results of a test	46
Summary	47
Lesson 6: Step-through Examples	49
Example 3-1: Copper in Carp Tissue	49
Example 3-2: Lowland Grassland Remnants	55
Example 3-3: Weight loss during incubation	58
The problem.....	58
The data.....	58
Example 3-4: Habitat complexity scores	61

Exercises	63
Exercise 3-1: Macro-invertebrate Abundance	63
Exercise 3-2: Elephant population counts	64
Exercise 3-3: Mercury levels in Bonnethead Sharks	65
Exercise 3-4: Organochlorine and Parasite Load in Gulls	66
Exercise 3-5: Does supplemental feeding deter sap-hungry bears?	69
Exercise 3-6: Should whale surveys be stratified by habitat type?	70
Exercise 3-7: PCBs and Kestrel Reproductive Behaviour	72
Exercise 3-8: Effect of coal ash on oral deformities of tadpoles	74
References	76

Key Concepts of Statistical Tests

The role of statistics in science

In Module 2 (Univariate Descriptive Statistics), the focus of attention was on samples. Powerful statistical tools were introduced for condensing and summarising single samples to more easily convey the essential features of the data in what is said and written. While this is an important aspect of statistical endeavour, much of statistical theory is concerned with a second topic—**statistical inference**.

When we study samples, we are seldom directly interested in them *per se*. We study them to learn something of the population from which the samples were drawn. We infer properties of the entire population, which we have not studied in its entirety, from our detailed knowledge of a sample of observations. The convenience of studying finite samples rather than the population as a whole comes at a cost. Samples, because they are finite and often relatively small, are somewhat akin to a fuzzy snapshot—the general impression of the population is evident, but the sample is an inexact representation. No matter how intensively we study the sample, there will be a level of uncertainty in what we discover, if we try to extrapolate our findings to the entire population. This uncertainty is often referred to as **sampling error**.

Sampling error has important practical consequences namely:

- Sample statistics will typically differ somewhat from the corresponding true values for the entire population. Estimating by how much they differ is a problem addressed under the heading of **parameter estimation**.
- Any two samples, even if taken from identical populations, will differ typically in all of their statistics. Determining whether the observed difference in sample statistics is great enough to conclude that the true population values differ is a problem addressed under the heading of **hypothesis testing**.

Because they deal with making inferences about population parameters on the basis of sample statistics, parameter estimation and hypothesis testing are grouped under the broader heading of statistical inference.

Parameter estimation

In the first category of analysis, parameter estimation, sample statistics are used to estimate the true values of the populations from which the samples are drawn. Say that we wish to know something of

the heights of students on a university campus, for the sake of future planning or the design of lecture theatres and seating. Time or resources may not permit an examination of the entire student population, so we choose instead to select and examine a representative sample of say 100 students. We then measure their heights and calculate the average height of people in the sample to yield a sample mean, \bar{y} .

If we obtained a figure of 170 cm for the mean of our sample, then provided our sample was reasonably large and provided we selected our sample at random to ensure it was representative of the entire population, we can be reasonably sure that the mean height of all students on campus, the true mean μ , is around the value calculated for the sample.

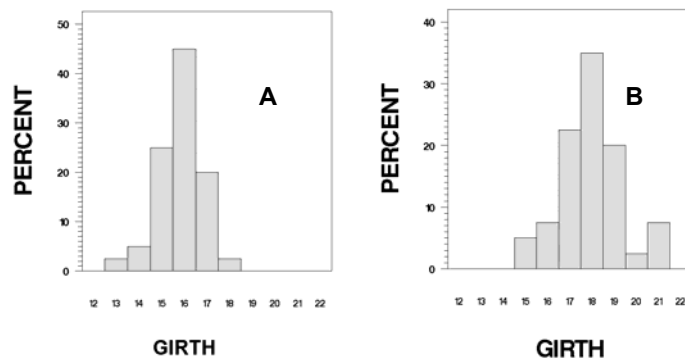
But how sure, and how close is our estimate to the real figure?

Statistical analysis provides a way of answering these questions. We might not know the value of the true population mean, but statistical analysis allows us to determine a range within which we can be 95% sure that it will lie. It allows us to use the statistics calculated from a sample to place bounds, or confidence limits, on the true value for the population from which the sample was drawn.

Hypothesis testing

The most common application of statistics in biology is to test specific scientific hypotheses. Consider a hypothetical example, rather trivial because of the magnitude of the difference between the populations under study. Two forests are of the same age. One forest was treated with fertiliser in a fashion consistent with past practice, the other with a new fertiliser and a new procedure. The Forestry Department is interested in knowing if the new procedure has produced significant increases in yield. The forests were large and it was not feasible to measure every tree, so the forester selected two random samples, each of 100 trees, from each forest. A variable was then selected to measure tree girth, one that would give an indication of the size of trees in each forest and of subsequent yields. Frequency histograms were constructed for each sample, and perused for an indication of whether girths were different in the forest subjected to the new treatment (Figure 3–1).

Figure 3–1.
Frequency histograms for random samples of pine trees ($n=100$) selected from each of two forests of the same age. A: Traditional fertiliser treatment; B: New fertiliser treatment.



You might conclude by simple inspection that the girths of the trees in the two forests are different. The samples are different most definitely, but when taking repeated samples even from the same population, one can expect them to differ to some degree just through chance.

The samples are different, but how can you be sure the two populations are really different?

An hypothesis test allows us to calculate the probability of obtaining two samples as divergent as the above two, on the assumption that they were drawn from two identical populations of trees. If this probability is very low, then the forests are probably actually different. If it is high, then the differences we observe may well have been nothing more than random fluctuations between samples, that is, sampling error.

Clearly, by resolving the ambiguity that arises from sampling error, statistics in general, and hypothesis testing in particular, provide very important tools for science.

To digress for a moment, a test you could use in this instance is the t -test, to compare two means. Applying the t -test to this example, we find that the probability of obtaining two samples as divergent as these two (or worse) from the one population is very small (less than 0.001). We therefore conclude that the two populations (the two forests) differ in reality and that the new fertiliser and procedure are superior.

The case of the two forests was fairly clear-cut and one might argue that the use of statistics was not required to arrive at a sound conclusion. But there are many cases where a decision cannot easily be made without relying on statistical techniques.

In a series of eight trials, specimens of the bush rat, *Rattus fuscipes*, were provided with a choice of two baits to test the rats' preferences for one bait over the other. The results are shown in Table 3–1. In all

but two cases, *R. fuscipes* chose bait A over bait B. Clearly the rats are showing preference for bait A—or are they?

Table 3–1. Bait preference of *Rattus fuscipes*, in an experiment involving eight trials.

	Individual							
	1	2	3	4	5	6	7	8
Bait A	1	1	2	1	1	1	2	1
Bait B	2	2	1	2	2	2	1	2

Before we can come to a conclusion one way or the other, we need to know what probability there was of obtaining the above result had we used rats that were unable to distinguish between baits A and B.

For interest only, the appropriate test to use in this case would be based on the binomial distribution. There are $k = 8$ trials, $P\{\text{choose A}\} = 0.5$, $P\{6 \text{ or more choices of A}\} = 0.145$. This test gives the probability of obtaining the observed result (or one more divergent from expectation), on the assumption that the rats have no preference for one bait over the other. The probability in this instance turns out to be 0.145 (= 14.5%). That is, if the entire experiment of eight trials was repeated 100 times, then a result at least as convincing as the one shown in Table 3–1 would be expected to occur about fourteen times—fourteen times out of 100 using rats unable to distinguish between baits A and B! Most would consider a probability of 14.5% too high to conclude that *Rattus fuscipes* prefers bait A over bait B.

Summary

In summary then, a fundamental *raison d'être* of statistics arises because any two samples, even if taken from the same population, typically will differ in all of their statistics. They will differ because of sampling error, that is, because finite samples will be inexact representations of the populations from which they are drawn. How then can a scientist evaluate the observed difference in means between two samples? They might differ by chance, because any two samples would be expected to differ, or they might differ because the populations from which they were drawn differ. Statistical tests provide a means for rationally deciding between these two possibilities.

A fundamental problem for the scientist: Does an observed difference or trend for the samples reflect a true difference or trend for the populations from which the samples were drawn, or did the observed difference or trend occur by chance alone?

Statistical analysis enables us to assign a measure of reliability to inferences made about the world around us from observations on finite samples. This is not true of science alone. We draw conclusions about the real world based on finite sets of observations throughout

our lives. It is just that in science, we need to assign an objective measure of reliability, in the form of a probability statement, to the inferences we make. Such measures of reliability give us confidence in the conclusions we draw from our data, and provide a formal structure for conveying that confidence to others in what we write.

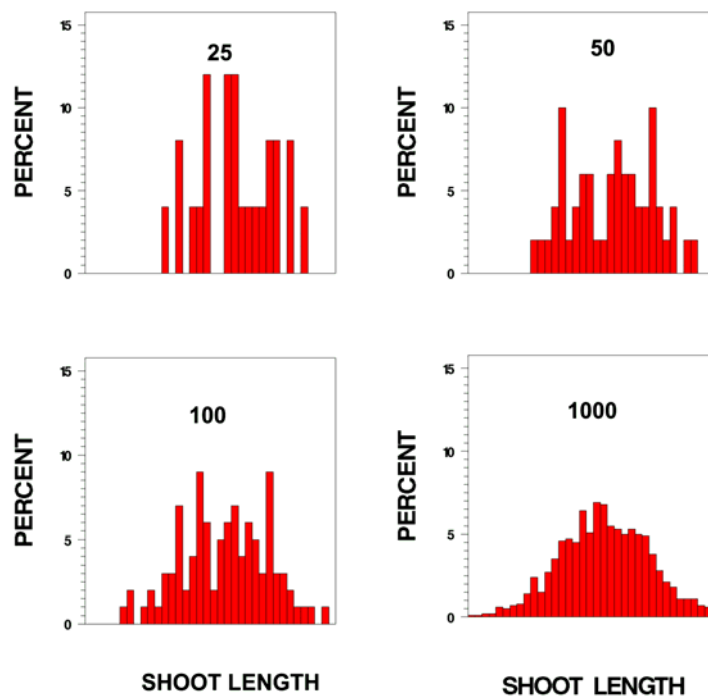
Lesson 1: Estimation and Confidence Limits

Samples and populations

The totality of observations with which we are concerned, whether finite or infinite, constitutes what we call a **population**. Once, the word population referred to observations on people, but today it applies to measurements of any entities of interest, whether it be people, animals, plants or objects. The number of entities that comprise the population is called the size of the population which, in many circumstances may be regarded as infinite.

A **sample** is a subset of the population. A **random sample** is a sample taken in such a way that each element of the population has the same probability of being selected. We take random samples to ensure that our samples are representative of the population from which they are taken, so that what we learn from the study of the samples will be more or less true of the populations themselves.

Figure 3–2.
A series of
samples, of
increasing size,
of
shoots of *Banksia
ericifolia* from the
Jervis Bay
National Park.



Samples should be considered fuzzy snapshots of the populations from which they are drawn, with the degree of fuzziness diminishing as the intensity of sampling or the sample size increases. Consider again the samples of *Banksia* shoots presented in Workbook 2. Figure 3-2 shows the frequency distributions for a series of samples taken from the Jervis Bay study site, increasing in sample size from 25 to 1000. Here we have a series of better and better approximations of the true but unknown frequency distribution for the population being sampled.

The sample frequency distribution approximates the true but unknown distribution for the entire population, and the approximation becomes progressively better as the sample size increases.

A parallel situation arises when we consider sample statistics. We could approach our study of *Banksia ericifolia* by measuring the lengths of each and every shoot in the study site, and then we might summarise the data by calculating the average or mean shoot length.

In this case, we would be examining the entire population of shoots at the study site and the average that we calculate would be called the **population mean**, designated μ . Note that the population mean is a fixed figure characteristic of the population. It is not subject to variation—no matter who calculates such a mean or how many times it is calculated, barring mistakes, the same value will be obtained in each case. As such the population mean is called a **parameter**.

Alternatively, time or resources may not permit an examination of the entire population of shoots, and we might choose instead to select and examine a random sample of say 100 shoots. We could then measure their lengths and calculate the average length of shoots in the sample—the **sample mean**, usually designated \bar{x} or \bar{y} . The difference between the sample mean \bar{y} and the population mean μ is that the sample mean is subject to natural variation or, as it is called, **sampling error**. If we were to repeat our sampling procedure by selecting another 100 shoots, we would obtain a value for \bar{y} that differed from the first one, and if we repeated the procedure once more, a third value would most likely result. For this reason, the sample mean is called a **statistic**.

The statistic \bar{y} is said to **estimate** the population mean μ .

In taking a sample of shoots and calculating their mean length, we have one object in mind. The sample itself is of little interest — we wish to learn something of the population. If we obtained a figure of 29.8 mm for the mean of our sample, then provided our sample was reasonably large and provided we selected our sample at random to ensure it was representative of the entire population, we can be

reasonably sure that the mean length of all shoots at the study site is around the value calculated for the sample.



Key Point

Statistics, calculated from samples, are estimates of true population parameters. The estimation improves as sample size increases.

Sampling distributions

Consider what happens if we replicate our sampling by taking, say, 30 samples of *Banksia* shoots each with a sample size of 50 shoots. For each sample, we can calculate a sample mean, and the values obtained will vary from one mean to another simply by chance, because of sampling error. It is not likely that any one of them will equal the true but unknown population mean, but with sample sizes of 50, they will be pretty close.

It turns out that the means of replicated samples from a normally distributed population, such as the population of *Banksia* shoots, are themselves normally distributed regardless of the sample size. The true mean for this sampling distribution is equal to the mean for the parent population, μ . Its standard deviation is called the standard error, to distinguish it from the standard deviation of the parent population. The standard error of the distribution of replicated

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

means, $\sigma_{\bar{y}}$, and the standard deviation of the parent population, σ , are related by:

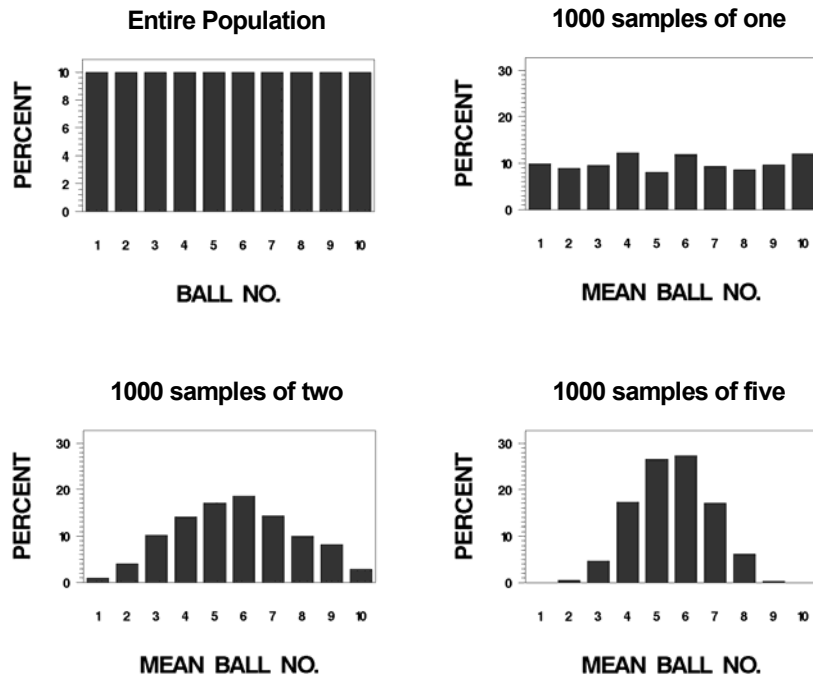
A standard error is the standard deviation of a distribution for a sample statistic, such as the sample mean.

What is even more remarkable is that means for replicated samples taken from populations with distributions of any shape will follow a normal distribution, provided sufficiently large samples are collected. This concept is usually presented in the form of the Central Limit Theorem, which states that:

Sample means drawn repeatedly from a single population with mean μ and standard deviation σ , will be normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$, irrespective of the distribution of the population from which the samples are drawn, provided the size of the samples (n) is large.

Consider a population with a uniform distribution, such as the population represented by wooden balls numbered from 1 to 10 in a bag from which we select balls at random with replacement. The distribution for this population is shown in Figure 3–3.

Figure 3–3. The distribution of means from a series of samples of increasing size taken from a population of numbered balls (top) each with an equal probability of selection.



If we take samples of size 1, then the distribution of sample means will be uniform and very similar to that of the parent population (Figure 3–3), as each possible mean value, 1 to 10, has an equal probability of 1/10 of occurring.

If, however, we take samples of two, the probability of obtaining a mean of 1 will be substantially lowered, because to obtain such a mean, two successive ones will need to be drawn, with a probability of 1/100. Similarly, a mean rounding to 2 can occur with ball combinations 1-2, 2-1, 2-2, 1-3 and 3-1 with a probability of 5/100. The probabilities of means falling at the extremes of the range will be depressed while those centrally will be inflated (Figure 3–3). It is not difficult to convince yourself of the truth of the Central Limit Theorem, and come to understand how it comes about, by perusing the results of sampling from the uniform distribution shown in Figure 3–3.

This knowledge is of great practical value. Because the sampling distribution of the mean is normally distributed under certain conditions, we are able to place objective measures of reliability on the inferences we make about the population mean. In particular, we

can calculate confidence limits for the population mean, using the above knowledge of the distribution of the sample mean.

Confidence limits

If a sample is selected from a normal population or, failing this, if the sample is sufficiently large, say $n \geq 30$, we can establish a confidence interval for the population mean by considering the sampling distribution of \bar{y} . Because the sampling distribution of \bar{y} will be approximately normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$, we have:

$$\Pr\left[\mu - 1.96 \frac{\sigma}{\sqrt{n}} < \bar{Y} < \mu + 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$$

All we are doing here is using knowledge of the normal distribution to say that 95% of sample means will lie between the true population mean and ± 1.96 times the standard error.

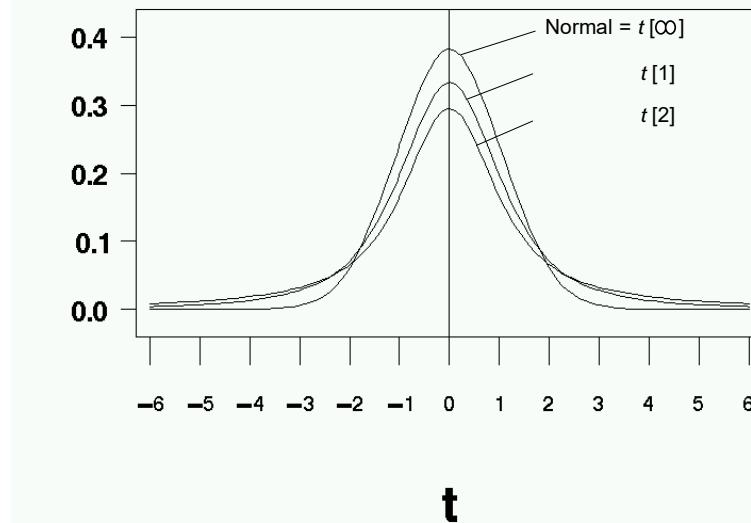
This equation can be rearranged to yield one that is much more useful. By subtracting \bar{y} and μ from each term and multiplying each term by -1, we have:

$$\Pr\left[\bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$$

So now, although the true population mean μ is unknown, we can at least define an interval within which we can be 95% sure it will lie. The catch is that the equation for the interval still contains an unknown, the population standard deviation σ .

If σ is known, the distribution of sample means about the true population mean will be normal, but if σ is only available as an estimate we must call upon another distribution worked out by statisticians — the t -distribution (Figure 3–5).

Figure 3–5.
The t -distribution.



The t -distribution has the following properties:

- The area under the curve = 1, as is the case for all continuous probability distributions.
- It is bell-shaped, symmetrical about the mean, median and mode, which are all of equal value.
- It has a complex formula defined uniquely by three parameters, the mean μ , the standard deviation σ and the sample size n . Note that the shape of the t -distribution depends on the sample size (Figure 3–5), unlike that of the normal distribution.
- It approximates normality as $n \rightarrow \infty$. The approximation is reasonably good for $n > 30$ and can be regarded as exact for $n > 120$.

The boundaries of the interval within which 95% of values lie, vary according to sample size, but have been tabulated for the t distribution by statisticians. These tabulated values replace 1.96 in the equation for the 95% confidence interval.

$$\Pr \left[\bar{Y} - t_{[0.05,2,v]} \frac{S}{\sqrt{n}} < \mu < \bar{Y} + t_{[0.05,2,v]} \frac{S}{\sqrt{n}} \right] = 0.95$$

where $t_{[0.05,2,v]}$ marks the boundary of the region within which 95% of t values would be expected to lie (5% spread over the two tails).

At last we have a formula that can be used. If our sample is selected from a normal population or, failing this, if our sample is sufficiently

large, then we can use attributes from our sample (sample mean, standard deviation and sample size) to determine the interval within which we can be 95% sure the true population mean lies.

This is an immense step forward, given that previously all we could do was make vague unsupported statements to the effect that the population mean was somewhere near our sample mean.



Key Point

Confidence limits provide bounds within which you can be 95% sure, say, that a true population parameter lies.

Lesson 2: Hypothesis testing

Rationale

It is worth repeating a fundamental dilemma faced by scientists in a wide range of disciplines. Any two samples, even if taken from the same population, typically will differ in all of their statistics. How then can a scientist evaluate the observed difference in means between samples taken from two populations? The samples might differ by chance, because any two samples would be expected to differ, or they might differ because the populations from which they were drawn differ.

This dilemma is resolved through the decision-making process of hypothesis testing.

Hypothesis tests rely on a perverse form of logic, pioneered by Euclid in his proof that $\sqrt{2}$ is irrational. Recall that an irrational number is one that cannot be expressed as a conventional fraction p/q , where p and q are whole numbers. Euclid must have been a lateral thinker, because he approached the problem in an odd way. He began by assuming what he set out to disprove, that is, he began by assuming that $\sqrt{2}$ was a rational number. He put:

$$\sqrt{2} = \frac{p}{q}$$

where p and q have no common divisor. It is possible to do this for all rational numbers. By simple rearrangement this gives:

$$p^2 = 2q^2$$

and so p^2 must be even, as it is equal to a number that is a multiple of 2. If p^2 is even, then p must be even. If p is even, it can be written as being equal to $2k$, giving:

$$(2k)^2 = 2q^2$$

or:

$$2k^2 = q^2$$

so q^2 and q must be both even.

If p and q are both even, this is a contradiction of the initial assumption that $\sqrt{2}$ was expressed as a fraction of two whole numbers with no common divisor. Hence the initial assumption of the rationality of $\sqrt{2}$ must be false.

What has Euclid done?

- He wished to prove what he suspected—that $\sqrt{2}$ is irrational (not in Euclid's words of course).
- He began by assuming the opposite—that $\sqrt{2}$ is rational.
- A chain of logical reasoning resting upon this assumption led Euclid to an impossibility.
- Euclid was thus forced to reject his initial assumption and to accept the alternative.

I ask you to hold in your mind for a few minutes, Euclid's lateral approach to a mathematical proof.

Consider for the moment that you are part of a large class of students undertaking a subject in elementary statistics. The lecturer puts a proposition to the class. He has a coin that he will toss repeatedly, and admits that the coin may be double-headed. He offers a small prize, a bag of marshmallows, to the first student to claim that the coin is indeed double-headed, provided they are correct in their assertion. However if they are incorrect, they must pay a penalty of \$300 to their favourite charity.

- The coin is tossed, and it comes up heads. No one hazards a guess, because the cost of being wrong is high (\$300) and the prize is of relatively low value (\$1.65). After all, the probability of getting a head if the coin has heads on one side, tails on the other, is 50:50 or 0.5.
- The coin is tossed again. Heads! No response from anyone in the class. The evidence for bias in the coin is still too weak, because even with an unbiased coin, two heads may be thrown by chance. There is a 1 in 4 probability of this.

- Again—heads! Suspicions are aroused, but still, three heads in succession occurs with a probability of 0.5^3 or 12.5%. Enough to risk losing \$300 in the hope of gaining a packet of marshmallows? I think not.
- Another toss, and another head. There may be someone in the audience willing to take a punt on 0.5^4 or 6.3%, but certainly when a fifth head is thrown, the hands begin to rise. Five heads in a row from an unbiased coin would be expected to occur only 0.5^5 or 3.1% of experiments of this nature.

How is this process related to the proof of Euclid? When confronted with the need to conclude that the coin was double-headed, we automatically work on the assumption that the reverse is true. We ask, if the coin is unbiased, what would be the probability of obtaining the observed results. It is only when the probability is low enough, that we decide that our assumption of no bias is poorly founded and conclude that the coin is double-headed.

Like Euclid, we assume what we hope to disprove, and rely on mathematics to lead us to an unacceptable result. The difference is that Euclid made his decision when he arrived at an impossible result (a contradiction), whereas we make our decision when we arrive at an improbable result.

This is the foundation of hypothesis testing. We begin by making an assumption, called the null hypothesis (the coin is unbiased). We then take the observed result and use mathematical theory to determine the probability of obtaining that result by chance alone, if the null hypothesis is true. If this probability is low, then we have some foundation for deciding that the initial assumption, the null hypothesis, is false.

How small this probability must be before we are willing to reject the null hypothesis will depend on the costs of a decision to reject the null hypothesis when it is true (\$300) relative to the benefits gained by a correct decision (one packet of marshmallows valued at \$1.65). If the penalty had been \$3000, then few hands would have risen on the fifth head in a row. By convention, in general science at least, the probability required to justify a decision to reject the null hypothesis is 0.05 or less.

Formal procedure

Let us now apply this rationale in a more serious vein. A statistical hypothesis is an assumption or statement, which may or may not be true, about one or more populations. If we have taken two independent samples, one from each of two populations, and we wish to know if the means of these two populations differ, we should begin by assuming that they are the same. Such an assumption, or null hypothesis, is written:

$$H_0 : \mu_1 = \mu_2$$

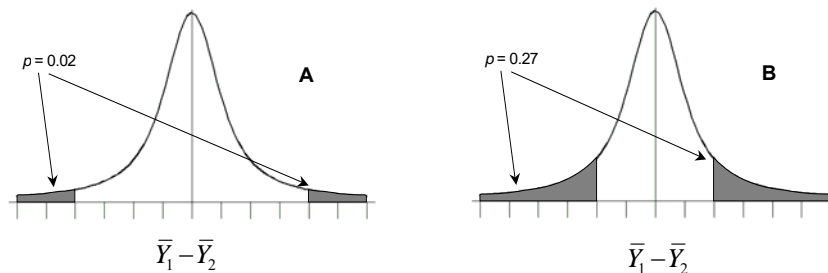
It is the hypothesis that we would often hope to reject, provided there is sufficient evidence to do so. A corresponding alternative hypothesis might be:

$$H_1 : \mu_1 \neq \mu_2$$

Next we ask how sample estimates of these population means might be expected to differ, *if the null hypothesis is true?*

Each null hypothesis has a test statistic associated with it. Clearly, if the null hypothesis is true, then we would expect the sample means of our two samples to be about the same and in fact, on average, we would expect the test statistic $(\bar{y}_1 - \bar{y}_2)$ to be equal to zero (Figure 3–6). Of course in practice, it will hardly ever be exactly equal to zero because of sampling error. Repeated measures of this statistic will vary around the value zero.

Figure 3–6.
A diagram showing how a theoretical distribution, based on the assumption that the null hypothesis is true, is used to decide the validity of that assumption (see text).



Provided you know the mathematical relationship that describes the distribution of the test statistic, you can make informed statements about whether an observed value for the test statistic is probable or improbable, if the null hypothesis is true. This provides us with the foundation for making a decision.

If you find the probability of obtaining the observed value for the test statistic or a more extreme value is very low (< 0.05), then you have a sound basis for rejecting the null hypothesis. In Figure 3–6A, the probability of obtaining a value for the test statistic, or a more extreme value, is 0.02, shown by the shaded area, and the null hypothesis is rejected. If on the other hand, the probability is very high, our observed value of the test statistic may well have occurred by chance, and we have no firm evidence for rejecting the null hypothesis. In Figure 3–6B, the probability of obtaining a value for the test statistic or a more extreme value is 0.27, shown by the shaded area, and we have insufficient evidence to reject the null hypothesis.



Key Point

The end product of an hypothesis test is a value, p , which is the probability of obtaining the observed value of the test statistic by chance alone, if the null hypothesis is true. Small values of p , say $p < 0.05$, provide a sound basis for rejecting the null hypothesis. Large values, say $p > 0.05$, provide insufficient evidence to reject the null hypothesis.

If the value of p indicates that the observed value of the test statistic could not have reasonably occurred by chance (say $p < 0.05$), you reject the null hypothesis. If, on the other hand, the value of p indicates that the value of the test statistic might well have occurred by chance alone, you have insufficient evidence to reject the null hypothesis. You do not accept the null hypothesis because failure to reject it is ambiguous—the null hypothesis might be true, or you may have insufficient data at hand to reject it.

One-tailed and two-tailed tests

Usually when a test is performed, the null hypothesis is an hypothesis of no difference or no trend, for example:

$$H_0 : \mu_1 = \mu_2$$

The alternative hypothesis proposes that there is a difference, for example:

$$H_1 : \mu_1 \neq \mu_2$$

Occasionally, we may have evidence, quite independent of the data we have collected, to believe that a difference in the true population means, if any, is in one direction only. Note that we must have **extrinsic** evidence that the **true** population parameters differ in one direction only. A one tailed test is not appropriate simply because the difference between the samples is clearly in one direction or the other.

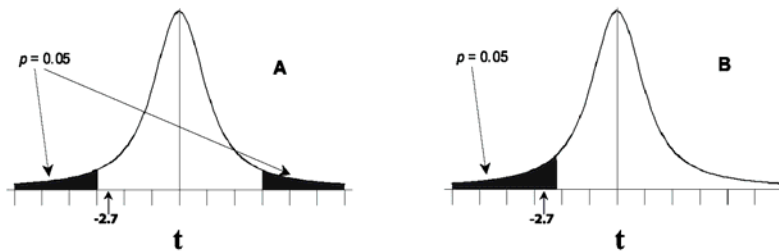
The alternative hypothesis then takes the form:

$$H_1 : \mu_1 < \mu_2 \text{ or } H_1 : \mu_1 > \mu_2$$

Tests involving such alternative hypotheses are referred to as **one-tailed tests**. The probability of obtaining a particular value of the test statistic, or one more extreme, in a one-tailed test is only half the probability that would have been obtained in a two-tailed test (Figure 3–7). Hence, the one-tailed test is more powerful and *a priori* knowledge of the direction of the true difference in population means should be used if available. In the example of Figure 3–7, a

significant result is obtained for the one-tailed test whereas the two-tailed test applied to the same data yielded a non-significant result.

Figure 3–7.
A comparison of the probabilities arising from (A) a two-tailed test ($H_0 : \mu_1 \neq \mu_2$) and (B) a one-tailed test ($H_1 : \mu_1 < \mu_2$).



Both tests are comparing the test statistic -2.7 with the critical values at the 95% level of significance. The two-tailed test yields a non-significant result; the one-tailed test yields a significant result. This illustrates the advantage of using a one-tailed test when extrinsic information on the direction of the difference permits it.

Statistical significance

In hypothesis testing, deciding whether to reject the null hypothesis depends upon a prior decision on what **level of significance** or alpha level to use in the test. We might choose, for example, to reject the null hypothesis if the probability of obtaining the observed results by chance alone is less than 5% or $\alpha = 0.05$. The level set for α depends upon what chance you are willing to take in being wrong when you decide to reject the null hypothesis. In science, we typically set α at 0.05 or 5%, and so are willing to be wrong in rejecting the null hypothesis in 5 of every 100 tests we choose to apply. If the costs of wrongly rejecting a null hypothesis are particularly high, then $\alpha = 0.01$ or a 1% level of significance may be more appropriate. If, however, you are conducting studies where the consequences of falsely rejecting the null hypothesis are not quite so severe, then a significance level of $\alpha = 0.10$ may be appropriate. There is no hard and fast rule on the appropriate level of significance. It depends very much on contemporary practice in your particular discipline.



Key Point

When the null hypothesis is rejected, we say that the result is **statistically significant**

For example, we may state that there is a significant difference between two means. This is shorthand for the statement that the observed difference between sample means is unlikely to have occurred by chance alone, at the accepted level of significance, if the population means are the same.

Type I and Type II errors

You will have gathered from the above discussion that hypothesis testing does not enable the researcher to make a cut and dried decision. A significant result at the 5% level of significance will lead to the decision that the true population means are different. When you make such a decision, you do so in the knowledge that there is a 5% probability that you are incorrect. Nothing is certain.

When a significant result is obtained at the 5% level, there are two possible scenarios. The null hypothesis may be true, and by virtue of the test result and the level of significance set, this is the most probable scenario. The null hypothesis may not be true, despite the outcome of the test, and you will have achieved a false positive result. The experimenter controls the probability of such an adverse outcome, usually at 5%. Your decision is to reject the null hypothesis is sound, on the balance of probabilities, but you may be wrong in doing so.

A **Type I error** (false positive) is made when we reject the null hypothesis when it is true. We might, for example, declare two population means to be different when, in fact, they are not (Table 3–2). Equally, we may err in the other direction, that is, we may accept a null hypothesis when it is false. We might, for example, fail to detect a difference between two population means when, in fact, they are different (Table 3–2). In so doing, we make a **Type II error** (false negative).

Table 3–2.
Outcomes of an
hypothesis test.
Two outcomes
are satisfactory,
the other two
are undesirable.

		DECISION	
		H ₀ rejected	H ₀ accepted
H ₀ is false	Correct decision	False Negative Type II Error ($p = \beta$ usually unknown)	
H ₀ is true	False Positive Type I Error ($p = \alpha$ usually 0.05)	Correct decision	

Type I and Type II errors have the following properties:

- The probability of a Type I error is specified independently of the experiment, by convention. In the sciences generally, it is typically set at $\alpha = 0.05$. In medicine, where the costs of rejecting a true null hypothesis are higher, it may be set higher, say at $\alpha = 0.01$.
- The probability of committing a Type II error, β , cannot be calculated without detailed knowledge of the alternative hypothesis, that is the hypothesis we accept when we reject the null hypothesis.
- If for a fixed sample size, we choose to reduce α from 0.05 to 0.01 to increase our confidence in a significant result, we will be

increasing the value of β . This occurs because the more stringent we make our conditions for rejecting the null hypothesis, the more likely it is that we will not reject a null hypothesis when it is false.

- Increasing the sample size does not affect the probability of a Type I error, but will progressively reduce the probability of a Type II error as the power of the test progressively increases.

Power of the test

Ideally, a test of significance should reject the null hypothesis when it is false. Power is the probability of getting a significant result when the null hypothesis is false. **Power** is defined as $1 - \beta$, where β is the probability of a false negative, accepting the null hypothesis when it is false.

A test becomes more powerful as the data available increases, so power is usually presented as a curve plotted against sample size, a power curve. A more powerful test for a given sample size is sometimes called a more sensitive test or a less conservative test.

Importance or strength of result

The observation that the power of a test to detect a true difference increases with sample size has an important practical consequence. Minuscule differences between parameters, while of no practical consequence, can be eventually detected as significant by a statistical test, provided we have a large enough sample size. So if we have very large samples, a highly significant difference may be so small in magnitude as to be of no biological importance.

Imagine that we are comparing male and female body depth for the introduced carp. Over a period of many years, data accumulates until we have approximately 7000 fish of each sex, and a t-test demonstrates that the difference in body depth of 0.1 mm between males and females is highly significant ($p < 0.0001$). What are we to conclude? The difference is real, but is it likely to be of any biological consequence to the fish?

Hence there are two considerations in hypothesis testing. The first is the probability that the result occurred by chance, that is, whether or not the result is significant in the statistical sense (usually written as $p < 0.05$ or 0.01 or 0.005 etc). The second is the strength of the result, which enables us to gauge whether or not the result is of biological importance. We cannot uncritically accept a result that is highly statistically significant at the 0.001 or 0.00001 level as being an important finding. We must also check the magnitude of the difference or trend, and assess whether it is substantial enough to warrant further attention.

Planning of experiments

The likelihood of obtaining useful results in both parameter estimation and hypothesis testing is greatly influenced by sample sizes. When estimating parameters, sample sizes need to be adequate to ensure that the estimates are sufficiently precise to be useful. In hypothesis testing, sample sizes need to be adequate to be reasonably certain of detecting an important difference when it exists. At the same time, sample sizes should not be so large that the cost of the study becomes excessive, nor is there much value in having weak, unimportant differences becoming highly significant. Planning the intensity of sampling is important in the design of experimental and observational studies.

Optimal sample size for a t-test will be affected by:

- **The size of the smallest difference that it is important to detect.** The smaller the difference, the larger will be the sample size required to detect it, all other things being constant.
- **The variability of the data.** The more variable the data within samples, the more difficult it will be to demonstrate a given difference between samples against the backdrop of that variability.
- **The acceptable probability of detection.** The more certain you want to be of detecting a difference of a given size, the larger will be the samples required to give you that greater certainty.
- **The level of significance of the test.** It will take larger samples to be reasonably sure of detecting a given difference at the 1% level of significance than at the 5% level of significance.

These elements are encapsulated in the formula for the optimal sample size for a t-test. To be 80% sure ($P = 1 - \beta = 0.8$) of detecting a given difference between two means (δ) at the 5% level of significance ($\alpha = 0.05$), you will require a sample size of:

$$n \geq 2 \left(\frac{\sigma}{\delta} \right)^2 (t_{\alpha[v]} + t_{2(1-P)[v]})^2$$

where n is the required size of each sample, σ is the true parametric standard deviation, $\nu = n_1 + n_2 - 1$ degrees of freedom, P is the intended power of the test, $t_{\alpha[v]}$ is the value from a two-tailed t -table with ν degrees of freedom and level of significance α and $t_{2(1-P)[v]}$ is the value from a two-tailed t -table with ν degrees of freedom and level of significance $2(1 - P)$.

You might have noticed also that n is on both sides of the equation, since ν is a function of n . This means that the equation must be used

in iterative fashion, in the same way that Robinson Crusoe made his wheelbarrow. He fashioned a crude wheel with which he made a crude lathe. He used the lathe to make a better wheel, and from there a better lathe to make a still better wheel and so on, until he had a very good wheel for his wheelbarrow. In the formula above, we guess a value of n and use the formula to obtain a better estimate of n , which we again feed into the formula, and so on, until n does not change from iteration to subsequent iteration.

To use this formula, you need also to make some hard decisions. First, you need to decide on what is the smallest difference (δ) upon which you will place some importance. You must decide that differences smaller than that value are of little or no consequence. Second, you need to estimate the parametric standard deviation, σ , and this must be estimated before you collect the data. You can use a ball-park figure based on experience, or you can undertake a pilot study to estimate σ and then the desired sample size before expending resources on the major optimised study. You may find the ratio of δ to σ easier to estimate.

Finally, you need to decide on the risk you are willing to take ($P = 1 - \beta$) in not finding an important difference when it actually exists. There is no general agreement on the value of P . The value of 0.80 seems to have currency in the same way as 0.05 has currency for α . Some would argue for higher values of 0.90 and 0.95, but ultimately it comes down to how important to you it is to detect a true difference of δ if it exists. What risk are you willing to take of missing it?

Regardless of the problems of its computation, the cost savings of this approach can be considerable, either through minimising the risk of undertaking the study only to find that no difference can be demonstrated (when it exists) or by avoiding the expense of collecting more data than is required for success.

The above analysis is sometimes called **prospective power analysis**.

Interpretation of non-significant results

Nothing is certain in statistics. A significant result is ambiguous—the result could be real, or it could be a Type I error—but the risk of an error is quantified. At the 5% level of significance, rejecting the null hypothesis carries with it a 5% chance of being incorrect in making that decision.

There is also ambiguity in a non-significant result—there may well be no difference between samples, or the sample sizes may not be large enough to detect a difference that is there (you make a Type II error). The risk of this type of error cannot usually be quantified, unless the

alternative hypothesis H_1 is known (see above). A non-significant result, therefore, is very difficult to interpret.

Strictly speaking, you would interpret a non-significant result as having failed to demonstrate a difference. Occam's razor rules in favour of the status quo. You cannot firmly conclude that no difference exists, as inadequate sample size might be the villain.

The way out of this dilemma, should you wish to draw inference from a non-significant result, is a **retrospective power analysis**.

In a retrospective power analysis, you ask, given your sample sizes, what might be the smallest difference (δ) you could be reasonably confident of detecting ($P = 1 - \beta = 0.80$). Using S as an estimate of σ , the formula for the smallest difference likely to be detected by a t-test with sample sizes of n , is:

$$\hat{\delta} \geq (t_{\alpha[v]} + t_{2(1-P)[v]}) \sqrt{\frac{2S^2}{n}}$$

If $\hat{\delta}$ is so small as to be of no consequence, then your interpretation of the negative result is acceptable. If, on the other hand, even a large difference would often go undetected with your sample sizes, you have nothing to report except your embarrassment at not having collected more data.

Retrospective power analysis is a controversial area, and the analyses have not adequately been incorporated into statistical packages. Many power analysis algorithms give you the optimal sample size to be reasonably sure of detecting the difference you observed in your analysis, which is about as useful as teats on a bull. What you need is the probability of detecting the smallest important difference, given the current design and sample sizes. Only then can you judge whether a non-significant result derived from inadequate design, or absence of a true effect.

Controversial also is the value chosen for the probability of detecting the difference if it exists. It has been argued that just as a small α (Type I Error) is required to declare a difference to be nonzero, so too a small β (Type II Error) should be required to declare a difference to be zero. We have chosen $P = 1 - \beta = 0.80$ above, which is developing similar currency as $\alpha = 0.05$, the defacto standard for the Type I Error. Cogent arguments can be made for 0.84, 0.90 and 0.95 and corresponding values for β .

Robustness

All tests are derived from statistical theories based on various assumptions. For example, the t-test is based on the assumption that the populations from which the samples are drawn are normally distributed with equal variances and that there has been randomness and independence in sampling. But in the end all we are concerned with is an objective basis for making a decision. The possibility exists that the results of a statistical test may be little affected by moderate violations of the theoretical assumptions. A statistical test is said to be robust with respect to one or more of its assumptions if its predictions are little affected by moderate violations of its assumptions.

Degrees of freedom

The concept of degrees of freedom is important in hypothesis testing, primarily because many of the theoretical distributions used to obtain a p value depend on the number of degrees of freedom. It is a difficult concept to grasp without a full appreciation of the mathematical basis of statistics. For a particular test statistic, the number of degrees of freedom is equal to:

$$df = n - \lambda$$

where n is the number of measurements making up the sample and λ is the number of parameter estimates calculated from the sample and used to calculate the value of the test statistic. The standard deviation

$$S = \sqrt{\frac{(Y - \bar{Y})^2}{n - 1}}$$

has $n-1$ degrees of freedom because there are n values from which we subtract 1 for the sample mean which appears in the formula.

Essentially, ten independent values comprise ten independent pieces of information, because knowledge of one value provides no information per se on any other value. Knowledge of the sample mean uses up one piece of information, because if you know the mean and nine of the sample values, then the tenth sample value is uniquely determined (you have 10 equations in ten unknowns). Similarly, once you specify nine deviations from the sample mean, the tenth is uniquely determined, since the sum of deviations about the mean must be zero. Only nine of the ten deviations are free to vary, that is there are nine degrees of freedom.

Significance, power, strength – putting it all together

A general approach to hypothesis testing is as follows.

- Formulate the research question as a statistical null hypothesis. This is the proposition that you wish to accept or reject. It is your initial assumption.
- Use statistical theory to carry that initial assumption forward to yield a p value. The p value is the probability of getting the observed data if the null hypothesis is true. Additional assumptions are generally necessary to move from the initial null hypothesis to a p value.
- Convince yourself that the additional assumptions made in arriving at the p value are upheld, or take measures to ensure that they are upheld, such as transformation.

If the p value suggests that our observed results are improbable (assuming the null hypothesis is true), then we reject the null hypothesis. We state that we have demonstrated a **statistically significant result**.

If the p value suggests that our observed results are quite probable, even if the null hypothesis is true, then we are unable to reject the null hypothesis. We state that our observed difference is **not statistically significant**.

If a result is statistically significant, there are three possible scenarios:

- The populations really are different, so the outcome of the test is correct. The difference is large enough to be of biological or practical importance and so is scientifically interesting. We **accept the result of the test** and interpret the difference as an important finding.
- The populations really are different, so the outcome of the test is correct. However, the difference is tiny, not large enough to be of any biological or practical importance. We **accept the result of the test** but do not proceed to interpret the difference as if it were important.
- The populations are identical, so there really is no difference. By chance, we have obtained larger values in one sample and smaller values in the other. At the 95% level of significance, such a spurious significant result will occur in 5% of experiments where there really is no difference. Of course we cannot know that our result is spurious, and we **accept the result of the test**. In doing so, we have made a Type I error.

If a result is not significant, then there are two possible scenarios:

- The populations are identical, so the outcome of the test is correct.
- The populations really are different, but we have been unable to demonstrate it. Our samples are too small to yield a reasonable probability of detecting an important true difference.

In practice, we are unable to draw a firm conclusion from a non-significant result without a retrospective power analysis. A power analysis will yield one of two outcomes:

- There is a reasonable probability of detecting an important difference if it exists, given our sample sizes. We have failed to detect such a difference, so **we accept the negative result of the test**. We conclude that there is no difference between the populations and interpret this result accordingly. In so doing, we run a calculable risk of making a Type II error.
- Even a substantial difference is unlikely to be detected, given our sample sizes. **We cannot make a decision** one way or the other. We conclude that we have insufficient evidence to detect a difference, should one exist. We can place no interpretation on the lack of significance from the test.

Of course, all this is contingent on the assumptions of our test being met, or at least not violated to an extent that would invalidate the outcome of the test.

Lesson 3: The F-test

Comparing two variances

Let us see how this general approach of hypothesis testing applies in a specific case, where we wish to know if the difference in two sample variances is sufficient evidence to conclude that two population variances differ. The **F-test** is appropriate for this sort of problem.

Consider an experiment on a species of native waterfowl involving birds both in the wild and in captivity. We wish to determine if the number of eggs produced per clutch is more or less variable among birds held in captivity than for birds in the wild. Wild birds were sampled at random from within the known range of the species, and captive birds were selected at random from within available breeding colonies. Relevant data were collected with the statistics shown in Table 3–3.

Table 3–3.
Summary
statistics for
clutch size in
captive and wild
waterfowl.

	Captive birds	Wild birds
Mean	9.9	10.0
Variance	0.4762	3.3333
Sample Size	7	10

The variance in clutch size differs between our samples of wild and captive birds, but then one would expect the variances of any two samples to differ, even if drawn from the same population. Our problem is to decide whether the observed difference in sample variances occurred simply through chance, or whether it reflects a true difference in the variability of clutch sizes between captive and wild birds. We must conduct an hypothesis test by drawing upon statistical theory, which allows us to estimate the probability of obtaining our observed pair of sample variances by chance alone.

Such theory predicts that:

If s_1^2 and s_2^2 are the variances of independent, random samples of sizes n_1 and n_2 taken from normal populations with an unknown but common variance σ^2 , then the statistic

$$F = \frac{S_1^2}{S_2^2}$$

will follow an F distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom.

Here we have assumed:

- **Randomness:** items in each sample were selected from their respective populations at random, that is, each entity in a population had an equal likelihood of selection.
- **Independence:** each item in a sample was independent of the other items. In sampling a broader population, birds taken from a single flock are unlikely to be independent in any measurement, as they may be closely related. Measuring yields of a crop from adjacent plots in a field may lead to problems of dependence.
- **Normality:** the populations from which the samples are drawn are normally distributed.

We have also assumed that the variances in the clutch sizes for wild and captive birds are the same. This latter assumption is the null hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

It is the assertion that we hope either to accept or reject having performed the F-test. The alternative hypothesis is that the variances are unequal.

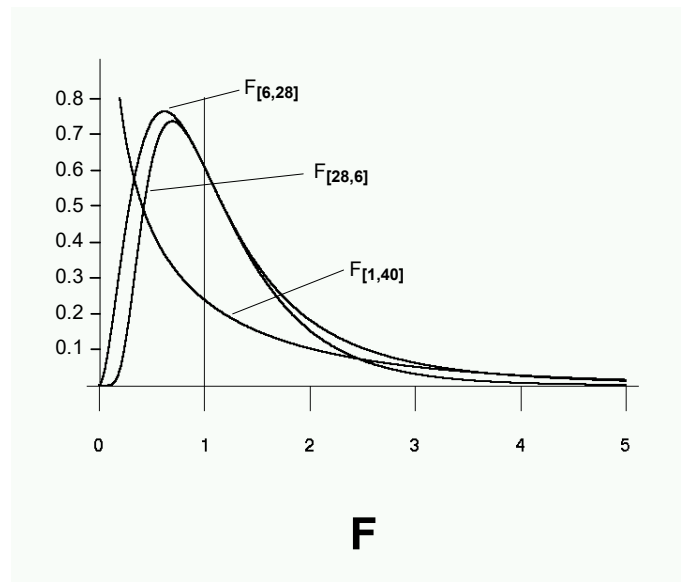
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Under the null hypothesis, the ratio of sample variances will follow an **F distribution**, which has the following properties:

- The area under the curve = 1, as for all continuous probability distributions.
- It is strongly skewed to the right (Figure 3–8).
- The F variable ranges from zero to ∞ .
- It has a complex formula but is uniquely defined by two parameters, v_1 and v_2 , the degrees of freedom associated with each of the two samples.
- The mean is approximately 1, but more exactly $(n_2 - 1)/(n_2 - 3)$.

Figure 3–8.
The F distribution
for various values

of v_1 , the
degrees of
freedom for the
sample with the
larger variance,
and v_2 , the
degrees of
freedom for the
sample with the
smaller variance.



By convention, F is calculated by placing the larger sample variance over the smaller sample variance, when the alternative hypothesis is that the variances are unequal.

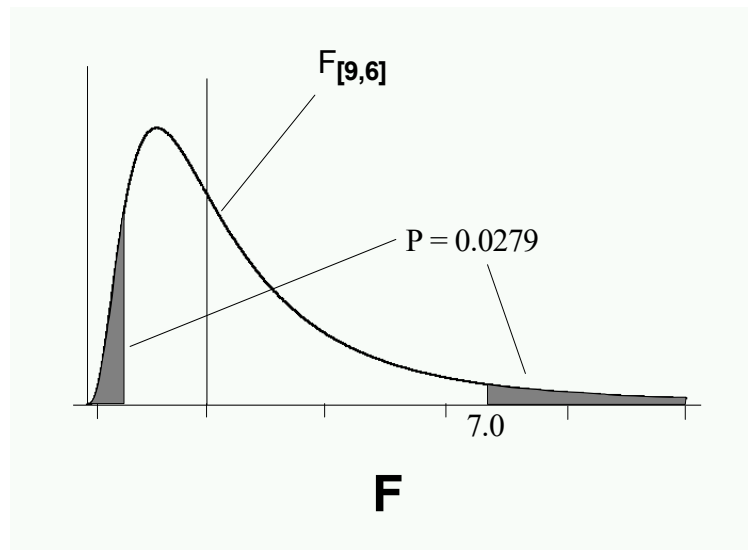
The **F ratio** for our data is given by:

$$F = \frac{S_1^2}{S_2^2} = \frac{3.3333}{0.4762} = 7.0$$

with 9 and 6 degrees of freedom respectively.

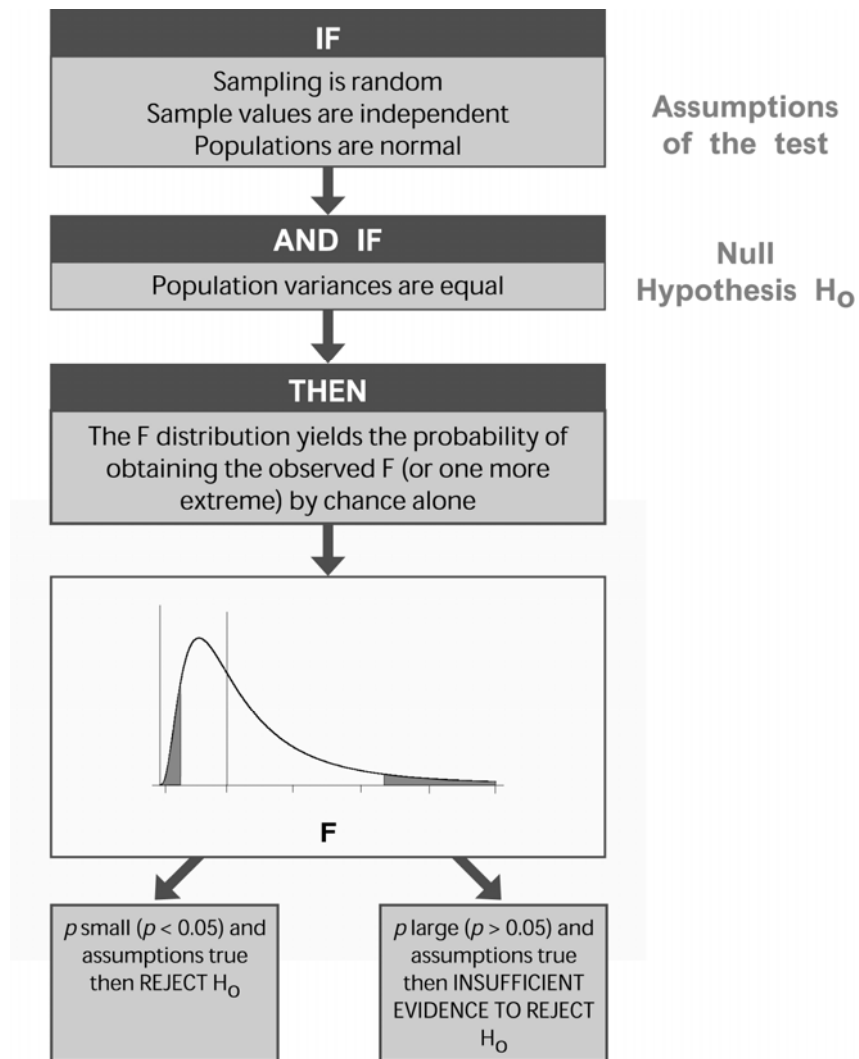
R uses the F distribution with 9 and 6 degrees of freedom to calculate the probability of obtaining an F ratio of 7.0 if the null hypothesis is true (Figure 3–9). This probability is low at 0.0279, certainly less than 0.05. In fact, we could expect variance ratios of 7.0 or greater to occur by chance in only 2.79% of experiments of this nature. So we conclude that there is a significant difference between the true variances in clutch size of wild and captive birds. The clutches of wild birds are more variable than those of captive birds.

Figure 3-9.
A diagram showing how the F distribution is used to determine the probability of obtaining a sample F statistic of 7.0 or one more extreme.



The rationale of the F-test is shown diagrammatically in Figure 3-10.

Figure 3-10.
Rationale of the F-test.



We would summarise the results of the analysis as follows:

The difference in variability in clutch size of wild and captive birds was significant ($F = 7.0$; $df = 9,6$; $p < 0.05$). Clutches of wild birds were more variable than those of captive birds of the same species.

Now that we have established that the difference in variances is probably a true reflection of reality, we might ask why? The captive birds were almost certainly subject to less variable environmental conditions than wild birds. The availability of food and shelter would have been more constant in captivity and captive birds would have been isolated from predators and other potential sources of injury. Hence, for birds in captivity, one would expect the allocation of energy to reproduction to have been much more constant than for wild birds whose energy must be spent in responding to conditions in a much more variable environment.

Lesson 4: The t-test

Comparing two means

A **student's t-test** is used to decide whether two population means can be considered different on the basis of their respective sample means. Consider an example.

In a study of the bearded dragon, *Pogona barbatus*, a herpetologist was interested to know if the mean calorific content of eggs differed between first and second clutches of the season. Lizards found nesting were returned to the laboratory and examined with a laparoscope to determine whether the clutch was the first or second of the season, and the nest was robbed of one egg. Data on calorific content, in calories per egg, are shown for several clutches (Table 3–4).

Table 3–4.
Summary statistics for mean calorific content of yolks of eggs of the lizard *Pogona barbatus*. The data are for first and second clutches, but are not paired.

	First clutch	Second clutch
Mean	643.0	623.6
Variance	1990.05	5532.70
Sample Size	6	13

The mean calorific content of eggs from first and second clutches certainly differ in the samples, by some 19.4 calories. But any two sample means, even for samples taken from the same population, can be expected to differ, perhaps by as much as 19.4 calories. Perusal of

the sample means per se does not permit us to make a decision. We need to apply an hypothesis test—the t-test—to obtain the probability of obtaining a difference of 19.4 calories by chance alone. Let us make the following assumptions:

Randomness: items in each sample were selected from their respective populations at random, that is, each entity in a population had an equal likelihood of selection.

Independence: each item in a sample was independent of the other items.

Equality of variances: the variances of the two populations are equal. Of course the sample variances may differ through sampling error.

Normality: the populations from which the samples are drawn are normally distributed, at least when the samples are small (df).

Let us also assume the null hypothesis:

$$H_0 : \mu_1 = \mu_2$$

Consider the test statistic:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where s_p^2 is a pooled sample estimate of the common population variance, given by:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Statistical theory predicts that the observed t will follow a standard t distribution with a mean of zero, a standard error of about 1 and $m_1 + m_2 - 2$ degrees of freedom.

The characteristics of the t distribution have been described earlier when we dealt with confidence limits.

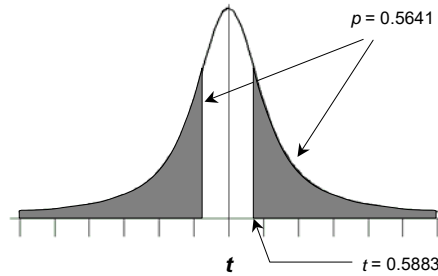
The t value from our data is:

$$t = \frac{643.03 - 623.58}{\sqrt{\frac{(6-1)1990.05 + (13-1)5532.70}{6+13-2} \cdot \left(\frac{1}{6} + \frac{1}{13} \right)}} = 0.5883$$

R uses the t distribution with 17 degrees of freedom to calculate the

probability of obtaining a t value of 0.5883 if the null hypothesis is true (Figure 3–11). This probability of 0.5641 is very high, much greater than 0.05. In fact, we could expect differences in magnitude of 19.4 calories or greater to occur by chance in 56.4% of experiments of this nature, if H_0 is true. We have no evidence to suggest that the null hypothesis is false.

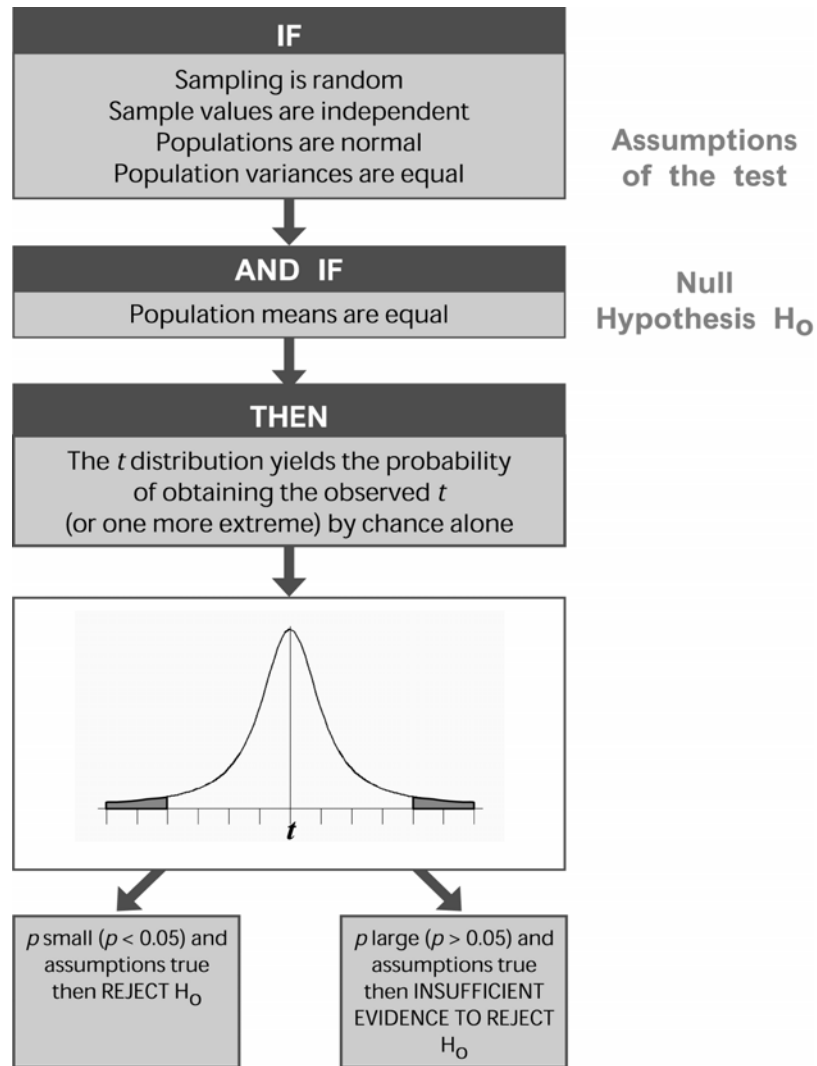
Figure 3–11.
A diagram showing how the t distribution is used to determine the probability of obtaining a sample t value of 0.5883.



We report that we are unable to demonstrate a significant difference in calorific content of eggs between first and second clutches of *Pogona barbatus* ($t = 0.59$, $df = 17$, $p = 0.5641$).

A diagrammatic representation of the rationale of the t -test is presented in Figure 3–12.

Figure 3–12.
Rationale of the
t-test.



Other analysis options

Wilcoxon rank-sum test

When at least ordinal measurement has been achieved, the Wilcoxon rank-sum test may be used to test whether two independent samples have been drawn from the same population. It is an alternative to the *t*-test, and does not have the restrictive assumptions of normality and homogeneity of variances, though it should be realised that the Wilcoxon rank-sum test is not a test of means. Instead it tests whether two independent samples have been drawn from identical populations. The test assumes randomness and independence in sampling.

The test is applied by combining the two samples and then ranking the measurements in order of increasing magnitude. If the populations from which the samples are drawn are identical, we would expect the values of both samples to be randomly spread

through the combined sample. However, if Population 2 has larger values than Population 1, we would expect a tendency for the Sample 2 values to be better represented among the higher ranks of the combined sample. The Wilcoxon rank-sum test measures this tendency, and yields a probability of it occurring by chance alone.

The use of the Wilcoxon rank-sum test is not restricted to non-normal populations. It can be used in place of the t-test when the populations are normal, although it is not as powerful as the t-test for detecting a true difference between populations. The Wilcoxon rank-sum test is typically superior to the t-test for non-normal populations.

Paired t-test

The standard t-test requires that samples are completely independent, that is that knowledge of one sample value provides no information on the value taken by a second value in either sample, with respect to their sample means. The paired t-test is designed to compare samples that are pairwise dependent. We might wish to compare growth rates of an aquatic plant *Vallisneria* sp. before and after administration of a substrate nutrient. A pair of growth measurements is taken from each plant, one measurement before, and one after, administration of the nutrient. The growth of some plants may be more rapid than for others, quite irrespective of our manipulations, and so the before-after measurements will depend, jointly, on which plant is considered.

Alternatively, total filterable phosphorus might be measured at each of ten randomly selected sites in a lake, in two seasons. We might want to know if there are seasonal differences in the levels of total filterable phosphorus. If the same sites are visited in each season, this is a paired experiment with two measurements per site. If on the other hand, the sites were chosen randomly on each sampling occasion, the experiment would not be paired.

With paired samples, the procedure is to calculate the difference between members of each pair, and then to test whether the mean of these differences (as opposed to the difference in means used by the standard t-test) is significantly different from zero.

The paired t-test assumes that:

- The entities selected for repeated measurement are independent and selected at random from a large pool of possible choices. In the above case, we assume that the *Vallisneria* plants are selected at random from a large population of individual plants.
- The differences between repeated measurements are normally distributed, or if not, the sample size is large.

The Wilcoxon signed-ranks test

The Wilcoxon signed-ranks test is a non-parametric alternative to the paired t-test. As with the paired t-test, it relies on calculating the differences between the paired values in the sample, but rather than calculating the mean difference, the differences are first ranked from smallest to largest without regard to sign. The sign of the original difference (+ or -) is attached to each rank, and the positive and negative ranks are separately summed.

If there is no difference between our two populations, the differences observed between pairs would occur by chance alone. Some of the larger ranked differences would be positive and some negative, and the sum of the positive ranks should approximately equal the sum of the negative ranks. However, if there is a true difference between pairs, these two sums will be quite divergent. The Wilcoxon signed-ranks test determines the probability of obtaining the observed difference in summed ranks by chance alone, and so provides a basis for decision.

The Wilcoxon signed-ranks test is considered to be an excellent alternative to the paired t-test because it is almost as powerful in rejecting a false null hypothesis, when the assumptions of the t-test are satisfied. When the assumption of normality is violated, it is usually more powerful than the paired t-test.

Lesson 5: Application

Standard deviation or standard error?

Means reported in scientific papers are often followed by a \pm and a second value. It is not always clear what this second value represents. Some authors report means with the standard deviation, others with the standard error. Which is correct?

There is no simple answer to this question, as it depends upon the context in which the statistics are reported. If the author is describing a sample, with no implication regarding the population from which the sample was drawn, then the mean plus or minus the standard deviation provides the reader with an indication of the average value for the sample and the spread of values about that average. The focus of attention is on the sample itself, not the population from which it was drawn. In presenting a mean and standard deviation, it is implied that the population is normally distributed, for otherwise the standard deviation provides limited information on the spread of values in the sample.

If, on the other hand, the authors are reporting the mean with the implication that it is approximately true of the population from which the sample was drawn, then the mean plus or minus the standard error is appropriate. This provides the reader with an indication of precision of the mean, that is, a range within which he or she can be 67% sure the true population mean lies. If we are simply describing the sample in terms of means and standard deviations, then data yielding

$$30.4 \pm 0.2 \text{ mm}$$

are as informative as data yielding

$$30.4 \pm 9.6 \text{ mm}.$$

If the above data were means and standard errors, then the first set would be far more informative than the second set. In the first instance, we can be 99% sure that the true population mean lies in the range 29.8–31.0 mm (mean \pm 3 SE). In the second instance, the corresponding range is 1.6–59.2 mm, and not very useful.

When reporting means and standard errors in this way, it is assumed that the sampling distribution of the mean is normal. The samples must be large, or if they are small, the population from which the samples were drawn must be normally distributed.

Some authors present the sample mean plus or minus the 95% confidence limits for the population mean, and others use the mean

plus or minus two standard errors as a large sample approximation to the 95% confidence limits. This last approach should be generally avoided.

You should always clearly specify whether you are reporting the standard error, standard deviation or confidence limits with means, usually in the Materials and Methods section of the manuscript. Always report the sample size to enable the reader to convert from one form to the other, and the range is also useful. For example:

30.4 ± 0.2 mm (range = 27.1 – 37.1 mm, $n = 74$)

Confidence limits or the T-test

A common graphical technique for comparing two samples is to construct figures depicting the sample ranges, means and 95% confidence limits (Dice and Leraas, 1936). If the confidence limits do not overlap, then we conclude that the population means are different. If they do overlap, then we have insufficient evidence to support such a conclusion. Unfortunately, the use of confidence limits in this way for comparing two means is extraordinarily conservative, as the example in Table 3–5 shows.

Note that the 95% confidence limits just meet, so our decision on whether the data support a difference in population means is borderline.

On the other hand, applying a t-test to the data yields a sample t of -3.014 with a probability of occurring by chance alone of only 0.0052. The result is highly significant, and demonstrates that the t-test is a far more powerful approach to demonstrating differences between means.

Table 3–5.
Data for two samples. The 95% confidence limits just overlap, whereas the t -test demonstrates a clear difference between population means ($p = 0.0052$).

	Sample A	Sample B
Mean	9.74	14.0
Standard deviation	4.00	4.00
Standard error	1.00	1.00
Sample size	16	16
Range	1.0 – 15.0	4.0 – 21.0
Confidence limits (95%)	7.61 – 11.87	11.67 – 16.13

If a graphical presentation showing the statistical significance of the difference between means is necessary, then a graph of the confidence interval of the difference between means is appropriate, as it has similar power to the t -test.

Choosing a statistical test

Students and researchers alike are often confused by the array of statistical tests that might be used to address a specific hypothesis. We need some rational basis for choosing among them. We need to ask:

- Is the test appropriate for the hypothesis being tested?
- Are the data measured at a level appropriate to the test and the statistics manipulated by that test?

Are the assumptions of the test tenable? If not, is the test robust to the violations of the assumptions?

- Is the test the most powerful appropriate to the problem and the data at hand?

Level of measurement

Level of measurement has a profound influence on the selection of appropriate descriptive statistics (Workbook 2). Since hypothesis tests are conducted on statistics, level of measurement is equally important for choosing an appropriate hypothesis test.

Table 3–6 provides recommendations on the test appropriate to the data at hand. Not all of these tests are covered in this Workbook.

Table 3–6.
Recommended tests for different levels of measurement. Only those tests marked with an asterisk* are covered in this Workbook.

	Independent Samples	Dependent Samples
Nominal	χ^2 Test of association	McNemar test
	Fisher exact test	
Ordinal	Wilcoxon rank-sum test*	Sign test*
Interval and Ratio	Student's t-test *	Paired t-test *
	Welch (Satterthwaite's) approximate t-test *	Wilcoxon signed-ranks test*

Note that because data at a higher level of measurement can always be converted to a lower level, albeit with some loss of information, tests that apply at the nominal and ordinal levels of measurement can also be applied to data measured at the interval and ratio levels. Hence the Wilcoxon rank-sum test can be applied in place of the student's t-test and the Wilcoxon signed-ranks test may be used in place of the paired t-test, but in so doing the measurements or intermediate values must be converted to ranks.

Are the assumptions tenable?

The assumptions of the t-test are quite restrictive, and the difficulty for the student and researcher is summed up quite nicely by Boneau (1960). He writes:

Psychological [and biological] data too frequently have an exasperating tendency to manifest themselves in a form which violates one or more of the assumptions underlying the usual statistical tests of significance. Faced with the problem of analysing such data, the researcher usually attempts to transform them in such a way that the assumptions are tenable, or he may look elsewhere [among the non-parametric alternatives] for a statistical test.

A further difficulty faced by a researcher proposing to do a two-sample comparison is that there are generally insufficient data at hand for a rigorous test of the assumptions of the chosen test. The irony of this is brought home most clearly when we consider the assumption of normality.

Normality

Several tests of normality are available, such as the probability plots and Shapiro-Wilk's tests introduced in Module 2. The Catch-22 is that reasonably large samples are required to be sure of accepting the null hypothesis of normality, when the data are normal, but when you have large samples, the assumption of normality is not at issue because of the central limit effect. Normality is important when the samples are small, but when they are small, tests for deviation from normality are very weak. Often, decisions on whether the assumption of normality is tenable depend on experience with the type of data at hand. Alternatively, the judgment may be based on more extensive studies reported in the literature.

Homogeneity of variances

The assumption of equality of variances can be tested with the two-tailed F-test described earlier. Many researchers routinely apply an F-test of variances before a t-test on means. Use of the F-test in this way has been criticised because it is quite sensitive to violations of the assumption of normality, much more sensitive than the t-test is to either the assumption of normality or the assumption of equality of variances. Applying an F-test before a t-test has been likened to testing the mood of the sea in a rowboat before setting sail in an ocean liner. There are however few alternatives to its use when there are two samples in total.

Randomness and independence

Randomness and independence in sampling are essential if the samples are to be representative of the populations from which they are drawn. There are tests available (eg the runs test) but again, these are useful only if sufficient data are at hand to enable confidence in the decision to accept the null hypothesis of randomness or independence. Both are best ensured by careful experimental and sampling design. If these assumptions are violated, then one has little choice but to discard the data then redesign and repeat the experiment.

How can violations be overcome?

There are several options for responding to perceived violations of the assumptions of normality and homogeneity of variances. You can attempt to manipulate the data in various ways by screening outliers or by transformation so that the assumptions are met. You can trust that the test is robust enough to be little affected, in a practical sense, by the violations. You can choose alternative approximate procedures developed to cater for known violations of the assumptions of the test. Or you can seek other tests in the domain of non-parametric or distribution-free tests, that do not have such restrictive assumptions, though these non-parametric procedures may be less powerful.

Relying upon the robustness of the t-test

Many researchers perform t-tests routinely provided there are no gross and obvious violations of the assumptions, and trust that the t-test is sufficiently robust to withstand any minor violations of the assumptions. Fortunately, several empirical studies have shown that the t-test is robust enough to withstand considerable violations of its theoretical assumptions (eg Lindquist, 1953; Srivastava, 1958; Boneau, 1960). The study of Boneau is worth reading to see how such an empirical study is conducted.

These empirical studies have established the following points:

- If the sample sizes are equal or nearly so, then the t-test is remarkably robust, and
- the larger the samples, the more robust the test.

The latter point is supported by both theory (read around the Central Limit Theorem) and the empirical studies. As a rule of thumb, equal sample sizes of 30 or more are sufficient to overcome all but gross deviations from normality and homogeneity of variances. If, in addition, the parent distributions are symmetrical, much smaller sample sizes are acceptable (say as low as 15). If the variances of those symmetrical populations are equal, then sample sizes as small as five will suffice.

- A combination of unequal sample sizes and unequal variances invalidates the standard student's t-test.

If the larger sample is taken from the population with the larger variance, then the student's t-test will be too conservative. If the larger sample is taken from the population with the smaller variance, then the standard t-test will be too liberal. In either case, Welch (Satterthwaite's) approximate t-test should be used (Satterthwaite, 1946; Steel and Torrie, 1980).

- One-tailed tests are seriously affected by samples drawn from skewed populations, whether or not the sample sizes are equal.

Despite these results, many researchers remain unmoved and choose to routinely apply one of the non-parametric procedures discussed below. The choice is yours.

Turning to non-parametric tests

Tests at the nominal or ordinal level of measurement, such as the Wilcoxon rank-sum test and the Wilcoxon signed rank test, have fewer assumptions than the 'parametric' F and t-tests.

Any test that is appropriate at the nominal or ordinal level can be applied at the interval or ratio level, and because of their fewer assumptions, many researchers prefer to routinely apply non-parametric tests in place of the F- and t-tests. But what are the pros and cons involved in making this decision?

Probability statements obtained from most non-parametric tests are exact probabilities (except in the case of large samples where very good approximations are used), regardless of the shape of the distribution for the population from which the samples are drawn. These exact probabilities apply even for very small samples. In contrast, many parametric tests such as the t-test become more robust only as sample size increases, and if applied to very small samples, they are valid only if the nature of the population distribution is known exactly.

So with fewer assumptions, exact probabilities and a broader range of data types acceptable to the non-parametric tests, why persevere with parametric tests at all? For many, the answer lies in a consideration of power.

There is a trade-off when one decides to use a non-parametric test as an alternative to the t-test. Non-parametric tests such as the Wilcoxon rank-sum test (see Siegel and Castellan, 1988) are generally less powerful than their parametric counterparts. That is, non-parametric alternatives to the t-test will be less likely to detect a true difference between populations than the t-test, in situations where the assumptions of the t-test are upheld. Thus if you play it by the

book, select a non-parametric test when you are unsure of the validity of the assumptions of the t-test, and then get a non-significant result, you find yourself plagued by the nagging doubt that the result might have been significant if a t-test could have been applied.

The difference in power between parametric and non-parametric tests is sometimes over-emphasised. For example, if the Wilcoxon rank-sum test is applied to data that might be properly analysed by the more powerful t-test, its power efficiency approaches 95.5% as sample size increases and is close to 95% even for samples of moderate size. It is therefore an excellent alternative to the t-test and does not have the restrictive assumptions of the t-test. The power of the Wilcoxon signed-ranks test compares equally well with that of the paired t-test.

Caveats for paired comparisons

The paired t-test has three assumptions—first, that the entities selected for repeated measurement are independent; second, that they are selected at random from a large pool of possible choices; and third, that the differences between paired measurements are normally distributed.

Often the paired t-test is applied without recognition of these constraints, as when environmental scientists ‘replicate in time’ to avoid the additional costs of true replication. For example, in a study of the effects of sewage on the abundance of various species of benthic macro-invertebrate, a biologist collected data on the numbers of the mayfly *Baetis* sp. caught in drift nets at various times of the day and night, upstream and downstream of a sewage outlet on a small stream. The data, after a log transformation, are shown in Table 3–7.

Table 3–7. Counts of mayfly nymphs (*Baetis* sp.) caught in drift nets at various times of the day and night, upstream and downstream of a sewage outlet on a small stream.

Time	Upstream	Downstream
14:00	3.33	3.52
16:00	3.22	3.02
18:00	3.07	2.26
20:00	3.31	3.22
22:00	4.20	3.91
Midnight	4.12	4.07
02:00	4.18	4.18
04:00	4.31	3.57
06:00	3.76	3.04
08:00	3.46	3.12
10:00	3.16	3.04
12:00	3.03	3.04

As the observations are paired, one pair per time period, a paired t-test might seem appropriate. However, the times selected for

repeated measurement were systematically, not randomly, chosen. This can present some difficulties.

There is the possibility of a systematic trend in time in the counts of drifting invertebrates. This in itself is not a problem, as the differences between upstream and downstream members of each pair may still behave like random and independent measurements. If however there is a trend in time, and the magnitude or direction of the trend differs between the upstream and downstream sites, then the differences between sites will no longer be independent. A systematic component to the differences, brought about by differing trends through time at the upstream and downstream sites, will also destroy any likelihood of normality of the differences.

Hence, the paired t-test is effective where we have "replication through time" only if any trend through time is the same for both upstream and downstream localities. The true abundance of drifting invertebrates at upstream and downstream localities must differ by a constant magnitude across the times. This is unlikely to be true in general for studies of this kind.

Compensating trends through time for the upstream and downstream sites will severely reduce the power of the paired t-test for detecting an impact, to the delight of those responsible for the discharge. Carried to extreme, we can imagine that there is a steady linear increase in abundance of drifting invertebrates with time at the upstream site, and a reverse but compensating trend at the downstream site. The mean difference between upstream and downstream will be zero, and non-significant. Yet at any one time, there is typically a substantial difference between upstream and downstream sites, and a clear impact of the discharge.

In summary, if we cannot consider the entities to be randomly selected from a large pool of possibilities (we might have two localities sampled at fixed times), then we must assume that any trend across the entities is of an identical nature. If there is any true difference between first and final measurements, then it must be one of constant magnitude only. So-called 'replication through time' should be avoided in studies of the type described above.

Reporting the results of a test

Having deliberated over the correct procedure to apply, and having performed all the calculations and interpreted the outcome of the tests, you would be forgiven for feeling obliged to make your labours apparent to the reader by including full details of the workings involved. You must resist this temptation. In reporting the results of a test, such as a t-test, in a manuscript or report, all that is required is a statement to the effect:

The difference of 1.2 g between mean weights of male and female hatchlings was significant ($t = 2.74$, $d.f. = 29$, $p < 0.05$). The male hatchlings were heaviest.

The p value refers to the probability that the sample t value occurred by chance alone. Had the result been significant at the 1% level, then the statement $p < 0.01$ would have been used. It is customary to report significant results by rounding the exact probability to the next highest value in the set 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001 etc. Exact probabilities are reported when the result is not significant (ie $p = 0.34$).

A description of the statistical procedures used should appear in the Materials and Methods section of the report. For example:

Means were compared using a standard Student's t -test, unless population variances could not be assumed equal, in which case a modified procedure (Snedecor and Cochran 1980:96-98) was employed. Population variances were compared with an F -test. Means are presented with their standard errors, unless otherwise specified.

Summary

You should by now have acquired some important tools for your statistical arsenal. It is possible to:

- construct confidence limits on the true population mean using statistics calculated from a single sample.
- determine the probability that two samples have been drawn from the same population using a variety of procedures depending on whether the samples are independent or pairwise dependent, on the level of measurement chosen when collecting the data, and on whether various assumptions of the tests available are tenable.

You should be aware of key concepts, such as:

- the difference between samples and populations; statistics and parameters.
- the meanings of the terms null hypothesis and alternative hypothesis.
- the meaning of statistical significance and how it is distinct from the strength of the result.
- the distinction between Type I and Type II errors.
- the power of a test, how to estimate an appropriate sample size and when to rely on a negative result.
- the robustness of a test.

It is now appropriate to put this knowledge to use via worked examples and exercises.

Lesson 6: Step-through Examples

Example 3-1: Copper in Carp Tissue

This is an example of a t-test applied where the population variances have been shown to be unequal.

The problem

After reports of people becoming ill on eating European carp, a laboratory decided to formalise procedures for routinely analysing metals in fish tissue. Two methods were considered suitable for the analysis of copper (Cu) in fish tissue – Graphite Furnace Atomic Absorption Spectroscopy (AAS) and Flame AAS.

The trade-offs are that Graphite Furnace AAS has a lower throughput rate and is therefore costly, and replicated measurements obtained by this technique are inherently more variable than Flame AAS. Graphite Furnace AAS is more sensitive however, and can measure lower concentrations of copper than can Flame AAS.

With the anticipated greater variability of the Graphite Furnace measurements in mind, the environmental chemist chose to perform more measurements by this technique than by Flame AAS to balance precision of the estimates of the mean copper determination for the two techniques, regardless of cost.



Note

The foundation for this decision lies in an understanding that the precision of the mean, represented by the standard error, is directly proportional to the standard deviation and inversely proportional to the [root] sample size.

$$SE = \frac{SD}{\sqrt{n}}$$

Hence, we wish to analyse the data to determine first whether the Graphite Furnace technique produces more variable results than the Flame technique, as anticipated. Second, we wish to determine whether the two techniques yield the same determinations. Finally, we need to make a recommendation on which technique to employ for assay of copper in carp, taking into account precision and cost.

The data

The data shown in Table 3-8 are concentrations of copper in the flesh of carp expressed in $\mu\text{g/l}$.

Table 3-8.
Concentrations
of copper in the
flesh of carp
extracted by two
methods of
Atomic
Absorption
Spectroscopy.

FLAME	GRAPHITE FURNACE
25	23
24	18
25	22
26	18
	17
	25
	19
	16

R expects the data in the form of pairs of values. The first is a discrete character variable indicating to which sample the measurement belongs, and the second is the measurement itself. Note that the sample sizes are unequal. The `t.test()` function in R can cope with this possibility.

The data in this case are held in a disk file called CARP.DAT in the data folder, and look like this:

```

FLAME      25
FLAME      24
FLAME      25
FLAME      26
GRAPHITE   23
GRAPHITE   18
GRAPHITE   22
GRAPHITE   28
GRAPHITE   17
GRAPHITE   25
GRAPHITE   19
GRAPHITE   16

```

The first field in the data, containing the character strings FLAME or GRAPHITE, is called a **factor** in the R manuals, but may be variously called a **treatment**, **breakdown variable**, **dummy variable** or **indicator variable** depending on which textbook you read. I prefer the term **factor** to distinguish discrete variables measured on the nominal scale from continuous variables. The second field in the data, containing the measurements of copper concentration, is referred to as the **response variable**.

The Analysis



Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui)



A program needs to be written to read the data in to R and to perform the t-test.

The function that does this for us in R is called `t.test()`. So lets check it by typing

```
> ?t.test
```

Try to understand the help for this function. Again the important sections are: **Description, Usage, Arguments, Value and Examples**. When you read the help carefully you see that there are two ways to specify the data for a t.test in R. First you can use two vectors x and y, which are then used to calculate the t.test. The other way is to use the formula interface, which can be convenient, depending on the way the data are organised. We will try both ways to show how it is done. First thing we have to load the data, by setting the working directory and by using the `read.table()` function.

```
> #set your working directory
> setwd("d:\\bernd\\biometryworkbook\\data")
> carp <- read.table("carp.dat", header=TRUE)
> carp
```

	method	conc
1	FLAME	25
2	FLAME	24
3	FLAME	25
4	FLAME	26
5	GRAPHITE	23
6	GRAPHITE	18
7	GRAPHITE	22
8	GRAPHITE	28
9	GRAPHITE	17
10	GRAPHITE	25
11	GRAPHITE	19
12	GRAPHITE	16

For the first way, we need to have two vectors, one that contain all measurements of concentration for the first method ("FLAME") and one that contains the data for the second method ("GRAPHITE"). There are many ways to get these numbers, here I will use simple indexing.

```
> x <- carp$conc[carp$method=="FLAME"]
> y <- carp$conc[carp$method=="GRAPHITE"]
```

We can check the content of `x` and `y` by typing their names into the R-console.

```
> x
[1] 25 24 25 26

> y
[1] 23 18 22 28 17 25 19 16
```

And now we perform the `t.test`.

```
> t.test(x,y)
```

*Box 3-1
Output of
t.test() for
determination of
copper on fish
tissue using two
analytical
techniques.*

```
Welch Two Sample t-test

data:  carp$conc by carp$method
t = 2.5923, df = 7.988, p-value = 0.03204
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 0.4408144 7.5591856
sample estimates:
 mean in group FLAME mean in group GRAPHITE
                        21                        25
```

The output is shown in box 3.1. So what do we get here. The first line tells us that the Welch test has been used. The Welch test is used if sampling variances are not equal and this is the default option in the `t.test()` function. Check the help `?t.test` and we find under Arguments, that `var.equal=FALSE` is the default. Hang on we would like to test, if this assumption is correct, because otherwise we could have used the standard t-test for equal variances. So let us test for differences in variances.

```
> var(x)
[1] 0.6666667

> var(y)
[1] 17.71429
```

So clearly they seem to be different, a formal test is the F-test (called `var.test()` in R).

```
> var.test(x,y)

F test to compare two variances

data:  x and y
F = 0.0376, num df = 3, denom df = 7, p-value = 0.02121
```

```

alternative hypothesis: true ratio of variances is not
equal to 1
95 percent confidence interval:
 0.006389739 0.550380458
sample estimates:
ratio of variances
 0.03763441

```

So the p-value is 0.021, hence we accept the alternative hypothesis (reject the null hypothesis) of: true ratio of variances is not equal to 1. (The ratio is in fact: 0.037)

So let's have a look at our t.test output of Box 3-1. Here the p-value is about 0.032, which is smaller than 0.05, therefore we reject the null hypothesis, that the difference between means of x and y is zero. We also get the actual mean of x (25) and mean of y(21). You can calculate them by yourself using the `mean()` function. In addition we get a confidence interval for the mean appropriate to the alternative hypothesis. This confidence interval does not include zero, which is another way to demonstrate that the difference in the mean is different from zero.

Okay now let's try the otherway of using the formula specification. Here it is a good idea to look at the Arguments and Example section of the help page to get an idea about the syntax required. Under Arguments we find the following statement:

```

formula a formula of the form lhs ~ rhs where lhs is a numeric
variable giving the data values and rhs a factor with two
levels giving the corresponding groups.

```

So translated into our example we type:

```
> t.test(conc, method, data=carp)
```

The `data=` helps us to use the header names, without attaching our data.frame, a longer version using the \$ sign would have been:

```
> t.test(carp$conc, carp$method)
```

Both versions give the same output as Box 3-1.



Submit the above commands for execution.

Results

In this particular case, we have $F = 0.037$ with 7 and 3 degrees of freedom. The probability of obtaining this result by chance alone if the population variances are equal (H_0 true) is only 0.0212,

somewhat less than 0.05. We conclude that there is good evidence of a difference in the population variances, and cannot therefore perform a Student's t-test, but use the Welch (Satterthwaite's) approximate t-test, which is the default (`var.equal=FALSE`). We have $t = 2.5923$ with 8 degrees of freedom. The probability of obtaining this value by chance alone if the population means are equal (H_0 true) is only 0.0320, less than 0.05. We conclude that there is a true difference in the mean concentration of the chemical as determined by the two methods.

The results section of the report would include:

Graphite Furnace AAS systematically under-estimates by an average of 4 $\mu\text{g/l}$ the concentration of copper when compared to Flame AAS ($t = 2.59$, d.f. = 8, $p < 0.05$). Graphite Furnace AAS yields a larger variance in the determinations and as such is less precise than Flame AAS ($F = 26.57$; d.f. = 7,3; $p < 0.05$).

Discussion

As anticipated from prior knowledge of the two techniques, the results from the Graphite Furnace AAS yielded more variable determinations than those of the Flame AAS. Given that the Graphite Furnace technique costs approximately \$6.00 per determination compared with \$0.50 for Flame AAS, and that for concentrations of 16 to 26 $\mu\text{g/l}$ sensitivity is not an issue, the cheaper, more precise Flame AAS is to be preferred. The observation that the two techniques differed in their mean estimates is of some concern, and the procedures followed in performing both analyses should be examined for potential sources of error.

Example 3-2: Lowland Grassland Remnants

This is an example of a Wilcoxon Rank-Sum test applied where the assumptions of the standard t-test are thought to be untenable.

The problem

In a study of lowland grassland remnants in the Australian Capital Territory, a subsidiary objective was to ascertain if the cover of dominant grass species were different at two grassland sites, Tharwa Road and Dudley Street, Canberra. The dominant grass species under consideration were *Danthonia* spp., *Stipa bigeniculata* and *Bothriocloa macra*. Combined percentage cover was estimated for each of 20 quadrats in the two grassland remnants. Because the data were in the form of percentages, often close to the extreme percentage of 100%, Sarah Sharp suspected that the data would not be normally distributed, and chose to do a non-parametric analysis. Turning to a non-parametric alternative to the t-test is common practice when faced with suspected failure of the assumptions of normality and homogeneity of variances. The R manuals provide for this option in the form of the **Wilcoxon Rank-Sum Test**. This test has few assumptions, namely that sampling is random and the measurements are independent.

If the test is used to compare central tendencies only (equality of medians), then there is the additional assumption that the two distributions have the same shape.

The data

Sarah's data are shown in Table 3-9 and have been converted to a form suitable for R and stored in file GRASS.DAT as follows.

Table 3-9.
Percentage cover, combined for the three dominant grass species, in 20 quadrats in each of two grassland remnants.

SITE 1 Tharwa Road Grassland	SITE 2 Dudley Street Grassland
40 61 55 55 55	95 74 95 90 32
50 60 36 40 40	64 . 95 75 100
30 55 25 20 45	75 80 60 85 90
20 46 35 54 100	85 80 60 57 60

The analysis



Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui)



The following R code is required to read in the data

```
> setwd("d:\\bernd\\biometryworkbook\\data")
> grass <- read.table("grass.dat", header=TRUE)
```

We check if the data are read in correctly.

```
> summary(grass)

      site      cover.perc
DUDLEY:20  55      : 4
THARWA:20  60      : 4
           40      : 3
           95      : 3
           100     : 2
           20      : 2
           (Other):22
```

So there are two variables named `site` and `cover.perc`, which is fine, but it seems to be strange that the summary of `cover.perc` is not showing the mean, min, max etc. So let us further check our data.

```
> str(grass)

'data.frame':  40 obs. of  2 variables:
 $ site      : Factor w/ 2 levels "DUDLEY","THARWA": 2 2 2 2 2 2
 2 2 2 2 ...
 $ cover.perc: Factor w/ 24 levels ".", "100", "20", ...: 9 17 14 14
14 12 16 8 9 9 ...
```

So `grass` is a `data.frame`, but both columns are factors, which is odd. The reason for this can be seen if you have a closer look at the levels of `cover.perc`. The first is ".", which should be the sign for missing data. As we did not specify this in the `read.table()` function R used the default symbol for missing data, which is `NA` and not ".". Hence we have to read in our data again, this time with the option `na.strings="."`.

```
> grass <- read.table("grass.dat", header=TRUE,
na.strings=".")
> summary(grass)
```

```
      site      cover.perc
DUDLEY:20  Min.    : 20.00
THARWA:20  1st Qu.: 42.50
           Median : 60.00
           Mean   : 60.87
           3rd Qu.: 80.00
           Max.   :100.00
           NA's   :  1.00
```

```
> str(grass)

'data.frame':  40 obs. of  2 variables:
 $ site      : Factor w/ 2 levels "DUDLEY","THARWA": 2 2 2
 2 2 2 2 2 2 ...
```



```
$ cover.perc: int 40 61 55 55 55 50 60 36 40 40 ...
```

Now, this looks better, so we can start to do the Wilcoxon test (`wilcox.test()`). A quick look in at the help pages (`?wilcox.test`) reveals that there is a formula version just like in the previous example.

```
> wilcox.test(cover.perc ~ site, data=grass)
```



Submit the above commands for execution.

The `summary()` function yields summary statistics such as means and medians useful for summarising the results. The output of `wilcox.test()` should be as given in Box 3-2.

*Box 3-2
Output of
wilcox.test used
to perform a
Wilcoxon Rank –
Sums Test to
compare percent
cover at two
grassland sites.*

```
data: cover.perc by site
W = 340, p-value = 2.578e-05
alternative hypothesis: true location shift is not equal to 0
```

The part of the printout of greatest interest is the W score and the probability value associated with W ($\text{Prob} >> |W|$). This is the probability of obtaining the observed difference in summed ranks by chance alone. If it is less than 0.05, then there is a significant difference between the two samples. In this case, there is a highly significant difference between the two methods ($W = 340$, $p < 0.0001$).

Results

Sarah summarised her results as follows:

Combined percentage cover of the three dominant native grass species, ranged from 20 to 100 for the Tharwa Road grassland remnant (mean 46.1, $n = 20$) compared to a range of 32 to 100 for the Dudley Street grassland remnant (mean 76.4, $n = 19$). The greater percentage cover at Dudley Street compared to that of Tharwa Road was statistically significant (Wilcoxon Rank-Sum Test, $W = 340$, $p < 0.0001$).

Example 3-3: Weight loss during incubation

This is an example of a paired t-test.

The problem

It has long been known that soft-shelled reptile eggs absorb water from the surrounding substratum during incubation, whereas the opposite is true of hard-shelled birds' eggs. Australian freshwater tortoises lay hard-shelled eggs in a chamber excavated in soil adjacent to the water in which the adults live. An experiment was designed to determine whether the hard-shelled eggs of an Australian species of tortoise gained or lost weight during incubation.

The data

One egg was chosen at random from each of ten clutches, then weighed and placed on moist vermiculite in constant environment chambers. The water potential of the substratum, the humidity and the temperature were all monitored and held constant throughout the experiment. After 14 days, each egg was again weighed. The data are shown in Table 3-10 and can be found in the file eggs.dat in your data folder on your hard disc.

Table 3-10.
Change in
weight of eggs
(grams) of the
Australian
tortoise
Chelodina
longicollis during
incubation under
controlled but
moist conditions.

EGG #	INITIAL WEIGHT	FINAL WEIGHT	DIFFERENCE
1	6.21	6.39	+0.18
2	6.12	6.30	+0.18
3	6.61	6.48	-0.13
4	6.26	6.48	+0.22
5	6.44	6.39	-0.05
6	6.35	6.57	+0.22
7	6.44	6.71	+0.27
8	6.52	6.75	+0.23
9	6.12	6.39	+0.27
10	6.57	6.66	+0.09

To test the hypothesis of a change in egg weight over time we need to perform a paired t-test, because the measurements on each egg are repeated. Although the eggs have been selected at random, we do not have true replication within samples. Observations are matched as pairs.

The analysis



Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui)



The `t.test()` function has been told that that it has to perform a paired t-test. If you check the help pages (`?t.test`) you should be able to find the option to tell the `t.test()` function that it has to calculate a paired version of the test.

```
> setwd("d:\\bernd\\biometryworkbook\\data")
> eggs <- read.table("eggs.dat", header=TRUE,
na.strings=".")
```

Check the data.frame `egg` if it contains the correct data (e.g. `summary()`, `dim()`, `str()`, `names()`). This time the data are organised in two vectors with headings `initials` and `final`, therefore we need to use the vector version of the `t.test()` function.

I hope you had a look at the help pages of `t.test()` and found the following argument `paired`. As a default `paired` is set to `FALSE`, hence to have a paired version of the t-test we set it to `TRUE`.

```
> t.test(eggs$initial, eggs$final, paired=TRUE)
```

The output is shown in Box 3-3.

Box 3-3. Output of `t.test()` with argument `paired=TRUE` for egg weights of *Chelodina expansa*

```
Paired t-test
data:  eggs$initial and eggs$final
t = -3.4178, df = 9, p-value = 0.007654
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -0.24595718 -0.05004282
sample estimates:
mean of the differences
 -0.148
```

Results

The results can be summarised as follows.

There was a significant difference between initial and final egg weights for eggs of *Chelodina longicollis* incubated under constant moist conditions for 14 days (Paired T = -3.42, d.f. = 9, $p < 0.01$). Egg weight increased on average by 0.148 ± 0.043 g (range -0.13 to 0.27 g). It appears that hard-shelled freshwater turtle eggs, like the soft-shelled eggs of turtles of the northern hemisphere, take up water during incubation.

Just for interest, consider what result we would have obtained if the data had been analysed inappropriately as a Students t-test (Box 3-4).

```
> t.test(eggs$initial, eggs$final, paired=FALSE)
```

Box 3-4. Output of `t.test()` with argument `paired=FALSE` comparing the initial and final weights of turtle eggs, without due regard to repeated measurement

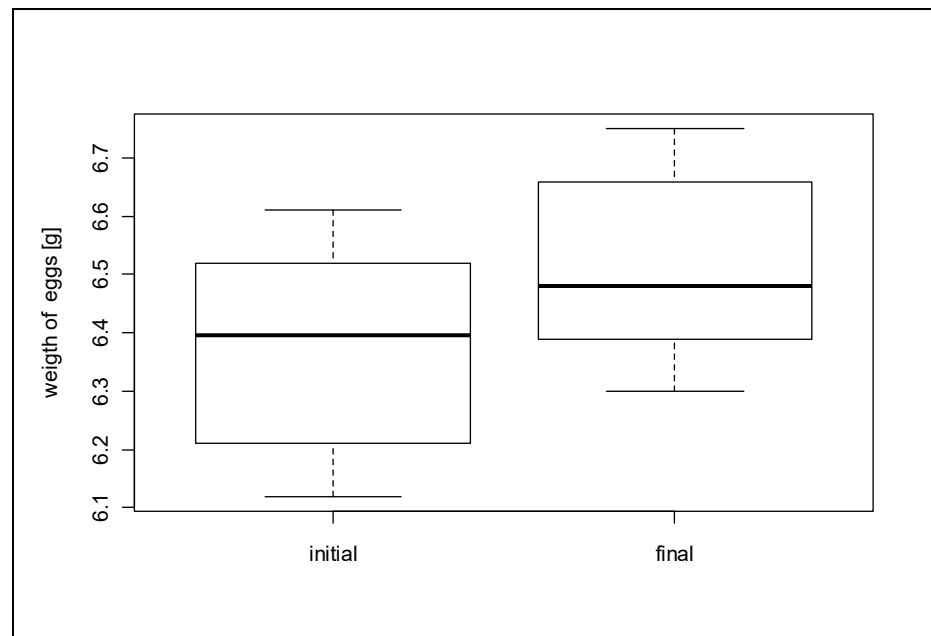
```
Welch Two Sample t-test
data:  eggs$initial and eggs$final
t = -1.975, df = 17.561, p-value = 0.0642
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -0.305719199  0.009719199
sample estimates:
mean of x mean of y
 6.364      6.512
```

In contrast to the results of the paired t-test, the Student's test for independent samples reveals no significant difference in the average weights of eggs before and after 14 days of incubation ($p = 0.0642$). The problem for the Student's t-test in this case is that high variability in the initial weights of eggs is obscuring any differences there might be between initial and final weights, taken pairwise. This problem could be overcome by selecting eggs all with the same initial weight for our experiments, but this would be extremely wasteful of data (many eggs would be rejected), or perhaps impossible. Instead, with the paired t-test, variability in initial weights is eliminated by considering only weight change.

To visualize our finding you could plot a boxplot of the two vectors (Box 3-6).

```
> boxplot(eggs, ylab="weigh of eggs [g]")
```

Box 3-6. Boxplot comparing initial and final weight of turtle eggs



Example 3-4: Habitat complexity scores

This is an example of a Wilcoxon Signed-Ranks test.


The problem

The Student's t-test has a non-parametric alternative in the Wilcoxon Rank-Sum test. Similarly, the paired t-test can be replaced by the Wilcoxon Signed-Ranks test. In the following example, two scientists independently applied Newsome's and Catling's (1979) habitat complexity scoring to the same 17 quadrats to assess the scoring method's reproducibility. Newsome's and Catling's score is a composite value calculated from ratings for five measures – soil moisture content and percentage cover by tree canopy, shrub canopy, ground herbage and rocks/logs. Each of these attributes of a site are scored on an ordinal scale, and the resulting scores are summed for an overall habitat complexity score.


In this analysis, we wish to determine if two trained scientists can reproduce each other's scoring.

The data and analysis

A Wilcoxon Signed-Ranks test is used to determine if the two scientists obtained significantly different scores, as follows:



Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui)



The data are stored in the file `newcat.dat`. Let us read in the data and check how they are organised.

```
> setwd("d:\\bernd\\biometryworkbook\\data")
> newcat <- read.table("newcat.dat", header=TRUE)
> newcat
```

```
      site john ralph
1       A     5     3
2       B     4     3
3       C     6     4
4       D     6     5
5       E     3     3
6       F     2     3
7       G     5     2
8       H     3     3
9       I     1     2
10      J     4     3
11      K     5     2
12      L     4     2
13      M     4     5
```

14	N	7	2
15	O	5	5
16	P	5	3
17	Q	5	1

So this time there are three columns, the first is the site name and then we have the scores of John and Ralph. We need to use these scores and specify that we want to have the `wilcox.test` using a paired set of data. You may have guessed this is the same as in the example before, and that there is an argument `paired` we need to set to `TRUE`.

```
> wilcox.test(newcat$john, newcat$ralph, paired=TRUE)
```

The relevant component of the rather volumous output is contained in the block with heading "Tests for Location" (Box 3-5). The results of the Wilcoxon Signed-Rank test are on the line labelled "Signed Rank".

Box 3-5. Output from PROC UNIVARIATE, used to perform a Wilcoxon Sign-Rank test on habitat scores

```
Wilcoxon signed rank test with continuity
correction
data:  newcat$john and newcat$ralph
V = 94.5, p-value = 0.008407
alternative hypothesis: true location shift is not equal to
0
```

Results

"There was a significant difference in the habitat complexity scores recorded by the two scientists for the same 17 quadrats ($V = 94.5$, $p < 0.01$). Clearly greater effort must be made to ensure comparability between the two before any habitat surveys get underway".

Source

Newsome, A, & Catling, P. (1979). Habitat preferences of mammals inhabiting heathlands of warm temperate coastal, montane and alpine regions of south-eastern Australia pp. 310–316, in *Ecosystems of the World*, Volume 9A, Specht, R. (ed), Amsterdam: Elsevier Publishing.

Exercises

Exercise 3-1: Macro-invertebrate Abundance

A limnologist was interested to know if densities of the benthic invertebrate *Jappa* sp. (Ephemeroptera: Leptophlebiidae) differed between two lakes, Allom and Deepwater. The two lakes were in the same geographical area, but differed in physico-chemical characteristics and the amount and texture of particulate matter that overlaid the littoral substratum. Lake Allom is a perched dune lake, that is, it resides well above the regional water table. Material of terrestrial origin accumulates in perched lakes, driving production. Deepwater Lake is a window lake, that is, it is formed when the lie of the land drops below the regional water table. Production tends to be lower in window lakes.

Ten replicate collections were taken from random locations in each lake using a column sampler, and the animals were returned alive to the field station for sorting and identification. Unfortunately, five collections from Lake Deepwater were destroyed in transit. The data are raw counts of *Jappa* sp. and can be found in the lakes.dat file in your data folder.

Table 3-11.
Counts of benthic
mayfly larvae in
two dune lakes on
Fraser Island,
Queensland.
Missing values
are shown as
periods

	Lake Allom	Deepwater Lake
	4	.
	5	2
	36	.
	15	.
	14	4
	8	.
	14	12
	28	5
	19	7
	15	.

We wish to know if the "standing crop" of *Jappa* sp. differs for the two lakes. You will need to appreciate that counts of benthic invertebrates are notoriously skewed to the right, because the animals tend to aggregate in the environment. Hence some form of transformation will be required to normalize the data prior to application of a parametric test. Select an appropriate transformation and apply it before performing the test.

- Perform a T-Test to compare the standing crop of mayfly larvae in the two lakes, and provide a brief report using the proforma supplied.
- Briefly discuss the assumptions you have made and why they are likely or not likely to be true.

- (c) Name the transformation you have chosen and briefly justify your choice.
- (d) State clearly the null hypothesis you are testing.
- (e) Present the results of the analysis, in the form of an abbreviated printout.
- (f) Summarise your statistical conclusions in a concisely worded paragraph, suitable for inclusion in the Results section of a publication or report.
- (g) Discuss your results in a biological context, as you might when writing the Discussion section of a publication or report.

Exercise 3-2: Elephant population counts

Management of elephant populations, and particularly those in enclosed areas, relies on estimations of population sizes and growth rates (Whitehouse et al., 2001). Techniques used to count elephants include aerial total counts and sample surveys, direct counts from the ground using line-transect sampling and stratification, faecal counts, and intensive ground-based surveys providing total counts by means of registration of individually known animals. Evaluation of alternative techniques is crucial in order to assess the accuracy of results obtained.

Whitehouse et al. (2001) used two methods to count the elephant (*Loxodonta africana africana*) population within the Addo Elephant National Park over a 20 year period. Helicopter surveys provided aerial total counts and intensive ground-based studies provided registration counts, based on individual recognition of the elephants.

Addo Elephant National Park is 60 km from Port Elizabeth in the Eastern Cape Province of South Africa. The elephants are restricted to a fenced area of 103 km², although the entire park currently covers approximately 700 km².

Intensive population monitoring from the ground should, intrinsically, provide a more reliable method of estimating total population size than aerial counts, particularly in small, confined populations. The researchers were interested in comparing the two methods to determine whether the less labour intensive aerial survey method produced adequate results.

Table 3-12.
Ground and aerial
counts of the
African Elephant
*Loxodonta
Africana* from
Addo Elephant
National Park.

YEAR	AERIAL COUNT	REGISTRATION COUNT
May-78	92	96
Jun-79	102	102
May-81	108	108
May-83	116	116
Apr-85	120	131
May-86	118	138
May-87	121	145

Oct-87	135	148
Apr-88	140	152
Apr-89	151	163
Dec-89	162	170
May-90	161	174
Jun-91	173	185
Jun-92	175	195
Jun-93	183	200
Oct-93	194	202
Mar-94	195	208
Apr-94	193	210
Mar-95	212	220
May-95	208	221
Oct-95	209	229
Mar-96	218	236

Read the data from the file “elephant.dat” in your data file folder.

- Analyse the data using a Wilcoxon Signed-Ranks Test to determine if there is a significant difference between the two methods of population determination. Show the relevant section of the output of your analysis here.
- What do you conclude from the comparison? Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Remember to distinguish between the magnitude of the result and its statistical significance.
- What are the management implications of your results?

Exercise 3-3: Mercury levels in Bonnethead Sharks

Florida’s commercial and recreational shark landings represent a significant portion of the total U.S. Atlantic shark landings. Shark landings have increased significantly during the past decade, because human consumption of shark meat has become increasingly acceptable, and especially in Asian markets where the demand for shark fins is very high -- as are the prices paid for them.

Mercury is a toxic metallic element that bioaccumulates in fish tissue, and can therefore represent a major dietary source of mercury in humans. Elevated mercury concentrations in fish have been a growing concern among resource management agencies. Apex predators, particularly long-lived species such as billfishes, tunas, mackerels, and sharks are reported to accumulate relatively high levels of mercury.

The Florida Department of Health and Rehabilitative Services released a Health Advisory Note in 1991 urging limited consumption of all shark species from Florida waters. However, the Health Advisory Note was derived from a limited number of samples taken from retail sources and from studies that lacked important information regarding species, capture location, sex, and size of the sharks examined.

A more detailed study to rectify these shortcomings was later undertaken by Adams and McMichael (1999) who published a detailed analysis on total mercury on four shark species -- bull shark (*Carcharhinus leucas*); blacktip shark (*C. limbatus*); Atlantic sharpnose shark (*Rhizoprionodon terraenovae*); and the bonnethead shark (*Sphyrna tiburo*). All are from the east-central coast of Florida. The data analysed by Adams and McMichael (1999) were augmented as new specimens were collected and tested for mercury

The juvenile/adult bonnethead shark data are provided in the file BONNY.DAT. The first variable in the data file is (sex), taking on the values M for male and F for female. The second variable contains the precaudal lengths (PCL) in mm and the third variable contains mercury concentrations (HG) in parts-per-million.

We are asked to reanalyse their bonnethead data set, with the additional specimens, to answer several questions.

- (a) Read the dataset (BONNY.DAT) into a data.frame suitable for analysis. Undertake an analysis to determine the proportion of sharks with mercury concentrations greater than 0.5 ppm Hg.
- (b) Construct histograms for precaudal length and mercury separately for each sex. What do you conclude? If you were to proceed with a t-test, without transformation, what would be your justification?
- (c) Analyse the data using appropriate two-sample tests to determine if there are significant differences in precaudal length (PCL) or total mercury level (Hg) between the sexes. Show the relevant sections of the output of your analyses here.
- (d) What do you conclude from the comparisons? Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Remember to distinguish between the magnitude of the result and its statistical significance.
- (e) What are the management implications of these results? Feel free to conduct further analyses.

Exercise 3-4: Organochlorine and Parasite Load in Gulls

Pesticide residues from organochlorines are thought to suppress immune function in birds and mammals, but there have been few assessments of this idea in the field. If establishment and survival of parasites is limited by host immunity, we would expect increased parasite intensities in animals with high organochlorine burdens.

Kajetl Sagerup addressed this question in a study of Glaucous Gulls (*Larus hyperboreus*) on Bear Island in the western Barents Sea off the coast of Norway. Forty gulls of similar size and age were collected by trapping or shooting and a liver sample was taken from each bird for analysis of nine selected polychlorinated biphenyls and four chlorinated pesticides. The digestive tracts were removed for counts

of 12 species of intestinal parasite. The birds were weighed using a spring balance.

Before the analysis could proceed, Kajetl needed to determine if there were significant differences in any of the variables with respect to the birds' sexes or method of capture. Kahetl was keen to pool the data across capture method and sex in order to increase sample sizes and the power of subsequent analyses.

The data file gull.dat contains, in this order, the **sex** of the bird (male or female), its **weight** in grams, the **method of capture** (trapped or shot) and concentrations of the following organochlorines (ng/g wet weight):

Chlorinated pesticides:	hexachlorobenzene, oxychlrodane, p,p'DDT, Mirex
Polychlorinated biphenyls:	Total concentration of PCBs 28, 52,99, 101,118, 138, 153, 170, and 180.

Remaining columns of the data file are:

Number of tapeworms (Cestoda)

Number of roundworms (Nematoda)

Number of flukes (Digenea)

Number of thorny-headed worms (Acanthocephala)

Counts for each species: *Cryptocotyle lingua*, *Anomotaenia micracantha*, *Alcataenia dominicana*, *Paricterotaenia porosa*, *Microsomacanthus ductilis*, *Aploparaksis larina*, *Tetrabothrius erostris*, *Anisakis simplex*, *Contracaecum osculatum*, *Paracuaria adunca*, *Stegophorus stellaepolaris*, *Corynosoma strumosum* (12 species).

Use appropriate analyses to determine:

- if male and female gulls of similar age differed significantly in load of any of the organochlorine pollutants, or in parasite load for any of the major parasite categories.
- if gulls caught by shooting versus trapping differed significantly body weight or in load of any of the organochlorine pollutants, or in parasite load for any of the major parasite categories.

You will need to pay particular attention to the assumptions of normality and heterogeneity of variances in selecting an appropriate test as the sample sizes are moderate and unequal.

- (a) Construct histograms and probability plots for each of the organochloride variables, each of the summary variables on parasite loads (cestode, nematode, etc) and for the variable weight. If they are not normally distributed, try a range of transformations. Give a summary of the outcome of this analysis below, and if transformation is not successful on one or more of them, explain why.
- (b) Analyse the data using appropriate two-sample tests to determine if there are significant differences between the two methods of capture or the two sexes as per the objectives outlined above. Treat the tests as independent. Show the relevant sections of the output of your analyses here.
- (c) What do you conclude from the comparisons? Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Remember to distinguish between the magnitude of the result and its statistical significance.
- (d) What advice would you give Kajetl on pooling the data for capture methods or for sex prior to more advanced analysis?

Exercise 3-5: Does supplemental feeding deter sap-hungry bears?

When black bears emerge from their winter dens, they forage ravenously to replenish their depleted body fat reserves. Bears forage then on the new sapwood growth which they harvest by clawing or biting away the outer bark to access the underlying sapwood (xylem and phloem). Damage is concentrated in 15-25 year old stands of managed conifer trees. Stand damage within the affected stands may be extensive because a single bear may peel bark from 50-70 trees a day. Peeling results in partial or complete girdling of the tree, causing death or reduced growth. Stand damage generally declines as summer foods, such as berries, become available.

Stephen Partridge from Washington State University studied whether it was possible to meet the needs of the bears in the spring by providing alternative food sources for hungry bears. They set out feeders that dispensed pelleted food *ad libitum* for the duration of bear activity in spring until natural foods were available in summer. The feeders were established in a timber stand owned by a commercial timber company and control areas were on adjacent lands of a state natural resources agency, in similar habitat and vegetation types. Bears were live-trapped, anesthetized, and weighed. Body condition was determined by techniques of bioelectrical impedance analysis and isotopic water dilution. Bears were recaptured later and remeasured so that gains or losses in mass or fat could be determined.

Table 3-13. Body mass and condition for Black Bears with and without the benefits of supplementary feeding.

Sex	Age	TRT	Mass Change(g/d)	Fat Change (g/d)
Male	Sub	Control	-23.08	.
Female	2	Control	-28.26	-17.71
Female	4	Control	-1.85	-1.56
Female	4	Control	-78.05	-51.66
Female	Adult	Control	190.63	86.37
Male	2	Treatment	145.45	16.65
Male	1	Treatment	197.83	29.32
Male	Sub	Treatment	14.58	3.95
Male	(1-2)	Treatment	127.27	.
Male	4	Treatment	153.52	30.18
Female	3	Treatment	86.76	20.38
Female	2	Treatment	90.91	12.18
Female	2	Treatment	129.79	29.26
Female	14	Treatment	395.24	142.61
Female	11	Treatment	358.21	114.03
Female	11	Treatment	224.19	48.42
Female	(10-14)	Treatment	-76.19	-49.77
Female	9	Treatment	108.57	54.08
Female	Adult	Treatment	111.11	24.81
Female	Adult	Treatment	226.32	108.84

You are asked to analyse the data to determine whether there were differences in mass change (g/day) or fat change (g/day) for bears from the control and treatment areas.

- (a) Analyse the data using appropriate two-sample tests to determine if there are significant differences in mass gain/loss and fat gain/loss between the control and treatment two feeding regimes. Show the relevant sections of the output of your analyses here. Pay attention to the assumptions of your tests.
- (b) What do you conclude from the comparisons? Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Remember to distinguish between the magnitude of the result and its statistical significance.
- (c) What do you conclude from the analysis. Is it worthwhile proceeding to gather data on impact on the conifer forest directly? Do you have any reservations about combining the data across sex or size classes in this analysis?

Exercise 3-6: Should whale surveys be stratified by habitat type?

Sperm whales (*Physeter macrocephalus*) and beaked whales (*Mesoplodon* spp. and *Ziphius cavirostris*) are deep-diving cetaceans that frequent shelf-edge and Gulf Stream waters of the northeast coast of the United States of America. Observers with binoculars can scan the surrounding waters to sight whales as their ship passes along a prescribed course at a set speed.

Gordon Waring from New Mexico State University and his colleagues summarised the number of whale sightings per km of survey during seven summer shipboard surveys (1990, 1991, 1993, and 1995-1998). They then used GIS to determine habitat use based on bathymetric and oceanographic features.

The beaked whales were concentrated at the colder shelf edge, whereas sperm whales were associated with warmer off-shelf water. Sperm whales and beaked whales do not associate closely (insofar as they are not found in mixed social groups), so the data can be analysed separately for each type of whale.

Gordon hypothesised that features of underwater bathymetry, such as the presence of underwater canyons, might influence relative sighting rates.

The data are found in the file WHALE.DAT and comprise variables giving the trip identity, the month of observation, the number of beaked whales sighted per km in canyon habitat and in non-canyon habitat, followed by similar data for sperm whales.

- (a) Construct histograms and probability plots for the sighting data near canyon and non-canyon habitat for sperm whales and for beaked whales. What would you anticipate as an outcome of a comparison of canyon and non-canyon counts for each species?
- (b) Analyse the data using appropriate two-sample test to determine if there are significant differences between the two habitat types (for each type of whale). Treat the tests for each whale species group as separate and independent. Show the relevant sections of the output of your analyses here.
- (c) What do you conclude from the comparisons? Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Remember to distinguish between the magnitude of the result and its statistical significance.
- (d) Discuss what the outcome of the analysis means for future surveys of whales.

Exercise 3-7: PCBs and Kestrel Reproductive Behaviour

Polychlorinated biphenyls (PCBs) bio-accumulate and biomagnify in the higher trophic levels of the food chain. Carnivores such as birds of prey (raptors) are thus potentially susceptible to PCB residues. As reproductive behaviours are under hormonal control, and PCBs are potential endocrine disruptors, the alteration of breeding behaviour or timing may be one mechanism responsible for reproductive failure by raptors. Documented examples of such aberrant behavior includes egg-destroying behavior, decreases in nest defense and nest attentiveness, and modifications of courtship behavior.

Sheri Fisher and her colleagues from the Avian Science and Conservation Centre at McGill University studied courtship behaviour of male and female American kestrels (*Falco sparverius*) after clinical exposure to PCBs. Captive male and female kestrels were randomly assigned to a PCB-exposed group (25 birds of each sex) or control group (25 birds of each sex). Care was taken to ensure that the diets of the control and treatment birds followed a common standard protocol, apart from the addition of PCBs to the diets of the treatment birds. The treatment birds were fed dead day-old cockerels that had previously been injected with an aliquot of 100 μ l of PCB mixture dissolved in safflower oil (4.85 mg PCB/g of oil). Control birds were fed cockerels that had been injected with safflower oil only. After a month on their prescribed diets, the kestrels were paired. All birds were experienced breeders, and none were related.

Birds were introduced and allowed to settle in for two days before formal observation. Pairs of birds were chosen randomly and observed in random order. Observations were repeated on a two day cycle. Each observation period was for 10 minutes and the number of behaviours performed by each kestrel per period was recorded.

The data reside in the file KESTREL.DAT and comprise of a breakdown variable with the values PCB and CONTROL, the pen in which the animals were kept, the repeat in the cycle of observation, and three response variables. The first response variable is the number of sexual displays, the second response variable is the number of flight behaviours and the third response variable is the number of inactive behaviours (sleeping, resting). The response variables are expressed in number of behaviours per minute.

- (a) Construct histograms and probability plots for each of the behavioural variables. If they are not normally distributed, try a range of transformations. Give a summary of the outcome of this analysis below, and if transformation is not successful on one or more of them, explain why.
- (b) Analyse the data using appropriate two-sample tests to determine if there are significant differences between the treatment and control

groups as per the objectives outlined above. Treat the tests as independent of each other. Show the relevant sections of the output of your analyses here.

- (c) What do you conclude from the comparisons? Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Remember to distinguish between the magnitude of the result and its statistical significance.
- (d) What do you see as the major outcomes of this study?

Exercise 3-8: Effect of coal ash on oral deformities of tadpoles

Many studies in toxicology focus on identifying lethal limits of a particular compound or group of compounds in severely stressed habitats. There is no doubt that mortality of organisms can indicate that a habitat is severely polluted. However one cannot assume that a moderately polluted habitat that still hosts an apparently viable populations is necessarily benign. Organisms may be impacted in a sublethal manner if pollutants result in changes to physiology, morphology, or behavior. If such alterations impact an organism's growth or reproduction, then non-lethal levels of pollution might have long-term effects at a population level, by way of effects on the energetics of individuals.

Amphibians are increasingly viewed as bioindicators of environmental stress, since there is widespread concern about population declines of frogs at a global scale. A significant concern in all regards is the widespread source of pollutants. We will consider an example involving the combustion of coal for generation of electricity, which produces fly ash and bottom ash, enriched in trace elements. Ash is disposed of by burial in landfills or more commonly by pumping the slurried waste into open-water settling basins. Frogs from a surrounding wetland often colonise a settling basin, despite the presence of potentially toxic compounds.

The presence and abundance of tadpoles or recently metamorphosed juveniles indicates that conditions within settling basins are not severe enough to result in widespread mortality for developing larvae. However, it is unknown if tadpoles in a basin or wetland experience sublethal effects that could ultimately influence growth, metamorphosis and subsequently reproduction. Rowe et al. (1996) investigated the oral deformities that occurred in bullfrog tadpoles (*Rana catesbeiana*) in ash basins. A working hypothesis was that oral deformities could feasibly inhibit the ability of tadpoles to feed upon periphyton (algal scum) and thereby impact their growth.

Tadpoles have multiple rows of teeth (which appear as comblike structures) and feed by scraping off and ingesting algae. Rowe et al. (1996) used a dissecting microscope to count the number of teeth on the two rows nearest the mouth (anterior row and posterior row) of 100 tadpoles. Fifty tadpoles were collected from an Ash Basin (ASH) and fifty were from a control pond Off-Site (OS).

The data reside in the file `tadpole.dat`, and comprise four columns. The first two columns are counts of teeth in the anterior and posterior tooth rows, respectively, for Ash Basin. The third and fourth columns are counts of teeth in the anterior and posterior tooth rows, respectively, for the control pond.

Analyse data for the anterior and posterior rows separately. Based on the number of teeth in the anterior row or the posterior row, are tadpoles different at the two sites?

- (a) Read the dataset (tadpole.dat) into a R data.frame in a form suitable for analysis. Calculate summary statistics for each variable broken down on location of collection. What is the likely outcome of the analysis, based on your perusal of means and standard errors?
- (b) Analyse the data using appropriate two-sample tests to determine if there are significant differences in mass gain/loss and fat gain/loss between the control and treatment two feeding regimes. Show the relevant sections of the output of your analyses here. Pay attention to the assumptions of your tests.
- (c) What do you conclude from the comparison? Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Remember to distinguish between the magnitude of the result and its statistical significance.
- (d) What are the management implications of these results?

References

- Adams, D. H. and R. H. McMichael Jr. 1999. Mercury levels in four species of sharks from the Atlantic coast of Florida. *Fisheries Bulletin* 97: 372-379.
- Boneau, C (1960). The effects of violations of assumptions underlying the T test. *Psychological Bulletin* 57(1):49-64.
- Boxall, G. D., J. J. Sandberg, and F. J. Kroon. 2002. Population structure, movement and habitat preferences of the purple-spotted gudgeon, *Morgurnda adspersa*. *Marine and Freshwater Research* 53: 909-917.
- Dice, LR & Leraas, HJ (1936). A graphic method for comparing several sets of measurements. *Contributions to Laboratory Vertebrae Genetics* 3:1-3.
- Fisher, S. A., G. R. Bortolotti, K. J. Fernie, J. E. Smits, T. A. Marchant, K. G. Drouillard, and D. M. Bird. 2001. Courtship behaviour of captive American kestrels (*Falco sparverius*) exposed to polychlorinated biphenyls. *Archives of Environmental Contamination and Toxicology* 41:215-220.
- Garcia, C.B and L. O. Duarte. 2002. Consumption to biomass (Q/B) ratio and estimates of Q/B predictor parameters of Caribbean fishes. Naga, *ICLARM Quarterly* 25(2): 19-31.
- Lindquist, EF (1953). *Design and Analysis of Experiments in Psychology and Education*. Boston, Houghton Mifflin.
- Newsome, AE & Catling, PC (1979). Habitat preferences of mammals inhabiting heathlands of warm temperate coastal, montane and alpine regions of south-eastern Australia, pp. 310-316 in *Ecosystems of the World*, Volume 9A RL Specht(ed), Amsterdam, Elsevier Publishing.
- Partridge, S. T., D. L. Nolte, G. J. Ziegtrum, and C. T. Robbins. 2001. Impacts of supplemental feeding on the nutritional ecology of black bears. *Journal of Wildlife Management* 65: 191-199.
- Paukert, C. P. and D. W. Willis. 2001. Comparison of exploited and unexploited yellow perch *Perca flavescens* (Mitchill) populations in Nebraska Sandhill lakes. *Fisheries Management and Ecology* 8: 533-542.
- Rowe, C. L., O. M. Kinney, A. P. Fiori and J. D. Congdon. 1996. Oral deformities in tadpoles (*Rana catesbeiana*) associated with coal ash

deposition: effects on grazing ability and growth. *Freshwater Biology* 36: 723-730.

Sagerup, K, E. O. Henriksen, A. Skorping, J. U. Skaare, and G. W. Gabrielsen. 2000. Intensity of parasitic nematodes increases with organochlorine levels in the glaucous gull. *Journal of Applied Ecology* 37: 532-539.

Satterthwaite, FW (1946). An approximate distribuion of estimates of variance components. *Biometrics Bulletin* 2:110-114.

Seigel, S & Castellan, NJ Jr(1998). *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed, New York, McGraw-Hill Book Company.

Srivastava, ABL (1958). Effect of non-normality on the power function of the t-test. *Biometrika* 46:421-429.

Waring, G. T., T. Hamazaki, D. Sheehan, G. Wood, and S. Baker. 2001. Characterization of beaked whale (Ziphiidae) and sperm whale (*Physeter macrocephalus*) summer habitat in shelf-edge and deeper waters off the northeast US. *Marine Mammal Science* 17(4): 703-717.

Whitehouse, A. M., Hall-Martin, A. J. & Knight, M. H. (2001). A comparison of methods used to count the elephant population of the Addo Elephant National Park, South Africa. *African Journal of Ecology* 39 (2):140-145.