INDEX

BIOSTATISTICS
TOPIC 6: ANALYSIS OF DIFFERENCES
I. TWO-GROUP COMPARISONS

IN GOD WE TRUST; ALL OTHERS MUST USE DATA.

I. INTRODUCTION

Before venturing into the central theme of this topic, let us have a few discussions of the nature of scientific research. Some people are proud and arrogant that they know so much. In fact, the less we know, the more certain we are in explanations; the more we know, the more we realise our limitations. Socrate used to say: "I know only one thing - that I do not know". It is not surprised that, John Maddox, the editor of *Nature,* recently remarked in Sydney that "life is still a mystery". I do not think this is a pessimistic comment, but rather a recognition of complexity of life.

From a mathematical point of view, the phenomenon world is nothing more than a set of relations. Everything is conditioned, relative and interdependent. One of the first great principles of population genetics is that the phenotype is the resultant of the individual's genotype and the environment in which that individual develops and lives its life. The phenotype can thus be altered by both change in the genotype and change in the environment.

Therefore, to understand or to explain the world phenomenon, we need to formulate hypotheses. For every phenomenon. we investigate, we must have at least one, numerically precise, statistical hypothesis. Sometimes, there are a number of alternative predictions we can make and each of these must be clearly distinguished before starting the research. This enables us to decide beforehand how we will choose between them when the results are obtained.

It is probably reasonable to say that the acme of scientific method is experimentation. From an abstract theory or concept, a prediction is drawn and an experiment is set up to discover whether this prediction is true (borne out) or not. If the prediction is in the way we expect, we have added some confirmation to the theory, ***but by***

no means proved it to be true (you may consult some philosophical books to see my point - we will discuss this later). There are a number of explanations possible for any observation. Consequently, we can never be sure that the explanation with which we started out is that which must apply in the particular circumstances of one experiment. If we believe that an observation or some observations prove an abstract hypothesis to be true, we commit the fallacy of confirming a consequent in hypothetical argument. A good theory or hypothesis is one which generates a number of different predictions and it becomes ever more confirmed when each of these is verified. Even, when all are verified it may still be false, since some other explanations are still possible, because as discussed earlier, life is a set of interdependent relations. When a number of alternative explanations have been given for a class of events, we generally prefer that which has the wider domain of implication. If the domains are equal, we prefer the more elegant theory. This amounts to saying that scientific explanations are limited by our human capacity to produce them, but this is usually adequate for most of us.

Now, we will see how statistical laws can help us to make our scientific judgement.

II. HYPOTHESIS TESTING

Once a sample is taken, it is usually characterised by one or more sample statistics. The purpose of hypothesis testing is to use these statistics and our knowledge of statistical distribution to make inferences about the population from which the sample is drawn.

A hypothesis, in this case, is a statistical statement that is to be rejected or not rejected. Hypothesis can be formulated about means, variances, differences of means, variances or medians etc.

There are two hypotheses in any statistical test. The first and most important is called H_0 - the null hypothesis. The second is called alternative hypothesis and is denoted by H_1. For example,

a test of two simple hypotheses is $H_0: \mu = 0$ and $H_1: \mu \neq 0$;
a test of one simple and one composite hypothesis is: $H_0: \mu = 0$ and $H_1: \mu > 0$.

To accept H_0, the result of the statistical test must be some number which falls into the *acceptance region*. Any other value in the *critical region,* as shown in the following figure and required the rejection of H_0. For example, if the true mean of a normal distribution is $\mu = 100$ and we hypothesise $H_0: \mu = 100$ or $H_1: \mu \neq 100$, two values x_1 and x_2 must be determined to separate the acceptance and critical region.

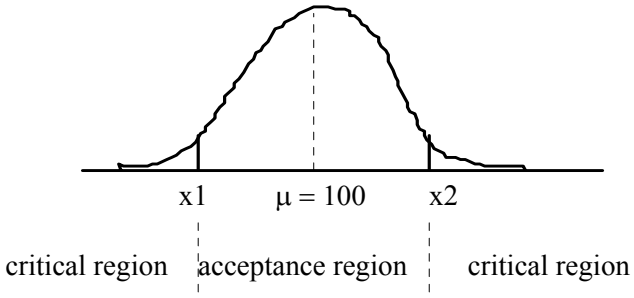

Figure 1: Acceptance and critical region of hypothesis testing.

2.1. TYPE OF ERRORS AND THEIR PROBABILITIES

No statistical hypothesis is ever impossible, it is merely more or less improbable. We must decide before an experiment how improbable H_0 should be for us to reject it. The selection of a rejection area for H_0 is not dictated by the science of statistics, it is a matter of policy for the empirical scientist using statistical methods. If the probability that H_0 is true is very small, we will reject it and put the faith in one of the alternatives. The rejection (or critical) area of the sampling distribution, under the null hypothesis H_0, is defined by a cut-off point which is symbolised by α. The conventional critical value for α are 0.05, 0.01 or 0.001 (5%, 1% or 0.1%) significance level. Thus, if the probability of H_0 being true is less than or equal to α, we reject it; otherwise, we accept it. Therefore, α is the probability of rejecting H_0 , while it is true. This is also called **type I error**.

But, either H_0 or H_1 must be true in reality, we can also make another error of accept H_0 while it is false. This is called **type II error (β level)**.

Reality	Decision	
	Reject H_0	Accept H_0
H_0 is true	Type I error (α)	Correct
H_0 is false	Correct	Type II error (β)

One may also represent this graphically as follows:

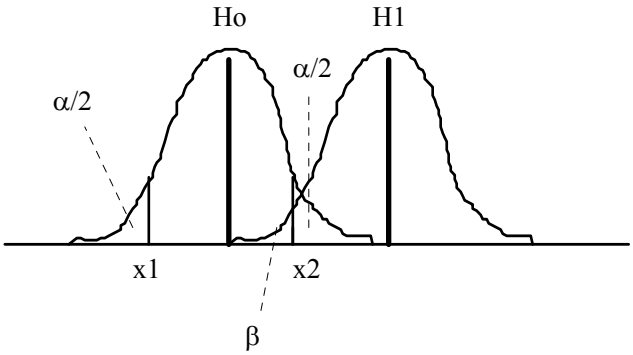

Figure 2: Type I and type II errors.

2.2. ONE-SIDED VERSUS TWO-SIDED HYPOTHESIS ?

If H_1 involves a non-equal relation, for instance, H_0: $\mu = 0$ versus H_1:$\mu \neq 0$, no direction is specified, so the significance area is equally divided between the two tails of the testing distribution in a fashion similar to that of shown in Figure 2. This is called a **two-sided** or **two-tailed** test. If, however, it is known that the parameter can go in only one direction, i.e. H_0: $\mu = 0$ versus H_1:$\mu > 0$ or H_0: $\mu = 0$ versus H_1:$\mu < 0$, the statistic is an **one-sided** or **one-tailed** test.

But the world does not always work that way. One is tempted to gauge the p-value of the test to satisfy one's assumption, therefore, the issue of one-sided or two-sided test is a controversial one. You may care to read the following note from a leading British medical statistician about this issue. It must be noted that we do not have to take his opinion, because, as I said, the issue is arguable in both directions.

2.3. AN EXAMPLE

Let us now take a concrete example. Suppose that we have carried a research into bone loss and found the mean and standard deviation of rate of bone loss (% per year) in femoral neck in 5 subjects were: -1.20 g/cm^2 and 0.8 g/cm^2. The question is that "is it reasonable to say that the rate of bone loss was significantly different from zero (no loss) ?". In statistical language, this question could be translated as:

$$H_0: \mu = 0$$
$$\text{versus } H_1: \mu \neq 0 \ (\mu < 0 \text{ or } \mu > 0)$$

This is a two-sided hypothesis. Now, for $n = 5$, the standard error of the rate of change is:

$$SE = 0.8 / \sqrt{5} = 0.36$$

Because the sample size is small, we can not use the normal distribution, but have to use the *t* distribution (see appendix), to work out the confidence interval of the observed rate of change. Now, if we are prepared to "commit" 1% level of type I error, for two-sided test, the confidence interval around the mean would be (1 - 0.01/2) = 0.995. As can be seen from the *t* distribution, the critical value of for *t* with 0.975 and (5-1) = 4 degrees of freedom is 4.604.

In the observed data we have:

$$t = (-1.2 - 0) / 0.36 = 3.33$$

which is less than the expected *t* distribution. We conclude that the difference (-1.2%) was not statistically significantly different from zero at 1% level. In other words, the observed percent change is within the 99.5% confidence interval around zero.

Is it statistically significantly different from zero at 5% level ?

III. GENERAL PRINCIPLE OF ANALYSIS OF DIFFERENCES BETWEEN TWO GROUPS

3.1. In previous topics, we mentioned that for a normal random variable X with mean $\bar{x}$ and standard deviation s, we would expect that 95% of the values of X will lie between $\bar{x}$-2s and $\bar{x}$+2s. So, for any value, say x_i such that $|x_i - \bar{x}|$ is greater than 2s (or absolute of $\left(\frac{x_i - \bar{x}}{s}\right)$ is greater than 2), we would conclude that x_i is significantly "abnormal". Abnormal should be understood as outside the expected range in a certain probability.

3.2. (a) For a random variable X whose individual values $x_1, x_2, \ldots, x_n$ which were sampled from a population with mean μ_x and variance σ_x^2,. The sample mean and variance of X are:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(b) Similarly, suppose that we have a random variable Y with individual values $y_1, y_2, \ldots, y_m$ of sizes m sampled from a population with mean μ_y and variance σ_y^2. The sample mean and variance of Y are:

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

and

$$s_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$$

(c) Suppose that we want to test the hypothesis of $\mu_x = \mu_y$ against the alternative hypothesis of $\mu_x \neq \mu_y$ ($\mu_x < \mu_y$ or $\mu_x > \mu_y$). The hypotheses can be equivalently stated as (μ_x - μ_y) = 0 versus (μ_x - μ_y) ≠ 0. The most obvious measure of difference is simply ($\bar{x}$ - $\bar{y}$) (the sample mean difference). Although conceptually extremely simple, the mean difference has the disadvantage that its interpretation depends on the unit of measurement as well as on the variability within each group. For instance, we do not know whether a mean difference of 15 is "large" unless we relate this figure to the variability in some way. Therefore, we prefer the **standard distance** D

which is defined as the absolute value of the mean difference divided by the standard deviation of the difference such that $D = \frac{|\bar{x} - \bar{y}|}{s(\bar{x} - \bar{y})}$. Similar to point 1, if $(\bar{x} - \bar{y})$ is more than twice the standard deviation of $(\bar{x} - \bar{y})$, we would conclude that μ_x is significantly different to μ_y. In other words, if $\frac{|\bar{x} - \bar{y}|}{s(\bar{x} - \bar{y})} > 2$, we would reject the null hypothesis.

So, the problem reduces to the finding an expected value and variance for the differences between sample means, $\bar{x}$ and $\bar{y}$.

3.3. It can be shown that, the expected value of $\bar{x}$ and $\bar{y}$ are μ_x and μ_y, respectively, which are simply the expected value of the random variables *X* and *Y*. That is:

for *X*, we have: $E(\bar{x}) = \mu_x$

and $\text{var}(\bar{x}) = \frac{\sigma_x^2}{n}$.

i.e. $\text{SD}(\bar{x}) = \frac{\sigma_x}{\sqrt{n}}$

and for *Y*, we have: $E(\bar{y}) = \mu_y$

and $\text{var}(\bar{y}) = \frac{\sigma_y^2}{m}$

i.e. $\text{SD}(\bar{y}) = \frac{\sigma^2}{\sqrt{m}}$

3.4. It can also be shown that the *difference* between *X* and *Y* are normally distributed, with expected value of:

$$E(\bar{x} - \bar{y}) = \mu_x - \mu_y \qquad [1]$$

and variance $\text{var}(\bar{x} - \bar{y}) = \text{var}(\bar{x}) + \text{var}(\bar{y})$

$$= \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} \qquad [2]$$

i.e $\text{SD}(\bar{x} - \bar{y}) = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$ [3]

Remember that the standard deviation is measured in the same unit as the mean, hence D is dimensionless. It does not depend on the unit of measurement: it does not matter whether we measure in millimetre or in metre. It is therefore possible to compare standard distances irrespective of the scales used. as we mentioned in point 2(c) i.e. $D = \dfrac{|\bar{x} - \bar{y}|}{SD(\bar{x} - \bar{y})}$

The main issue here is the estimation of the standard deviation of $(\bar{x} - \bar{y})$. We will consider this in case-by-case basis as follows.

IV. DIFFERENCES BETWEEN MEANS: INDEPENDENT SAMPLES

4.1. NORMALLY DISTRIBUTED DATA

In the same setting as the problem in section I. That is, we have a sample of *n* and *m* values x_i and y_i which were drawn from two populations with mean μ_x and μ_y, and variances σ_x^2 and σ_y^2, respectively. Suppose further that we have a sample means $\bar{x}$ and $\bar{y}$ and variances s_x^2 and s_y^2. To test the hypothesis of

$$H_o: \mu_x = \mu_y$$

against $$H_a: \mu_x \neq \mu_y$$

we will use the statistics in [1] and [3]. However, to simplify the issue further, we would like to assume that the population variances are equal, i.e. $\sigma_x^2 = \sigma_y^2 = \sigma^2$, then [3] can be reduced to:

$$\text{SD}(\bar{x} - \bar{y}) = \sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)} = \sigma\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)} \qquad [4]$$

The issue is now to estimate the average variance of the two samples, which is:

$$\boxed{s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{(n-1)+(m-1)} = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

i.e $$\boxed{s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}} \qquad [5]$$

With s^2 is as an unbiased estimate of σ^2. Then [4] can be written as::

$$\text{SD}(\bar{x} - \bar{y}) = s\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)} \qquad [6]$$

Hence, the standard distance in [1] becomes:

$$\boxed{D = \frac{\bar{x} - \bar{y}}{SD(\bar{x} - \bar{y})}} = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} \qquad [7]$$

which is distributed according to the *t* distribution with *n*+*m*-2 df.

Example 1: Stegman et al. (JBMR, August 1992) studied the association between ultrasound measurement of bone quality and fracture. Bone quality was measured by apparent velocity of ultrasound (AVU) in m/s. In 37 women with no fracture, mean AVU was 1850 m/s with standard deviation of 59 m/s. In 10 women with low trauma fracture, mean AVU was 1782 m/s and standard deviation of 89 m/s. The authors concluded "these initial results show that those with low trauma fractures have significantly lower AVU than those without". Verify this conclusion.

The difference in AVU between those fracture and non-fracture was 68 m/s. However, as mentioned earlier, we do not know whether this difference is substantial until the variability of the measurement is taken into account. Now, the estimate of common variance for the two groups of patients, by [5], is:

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2} = \frac{36(59)^2 + 9(89)^2}{37+10-2} = 4369 \text{ m}^2/\text{s}^2$$

i.e $s = \sqrt{4369} = 66.1$ m/s.

Then, the standard deviation of the difference between means, by [6], is:

$$\text{SD(diff)} = 66 \times \sqrt{\frac{1}{37} + \frac{1}{10}} = 23.5 \text{ m/s}$$

Now, the absolute difference was 68 m/s, which is nearly three times (=68/23.5) the standard deviation of the difference, hence it seems that the conclusion is true. Let us check our finding in terms of probability.

We learnt from the previous topic that the for a moderate sample size such as in this study, it would probably be reasonable to assume that the statistics defined in [7] have the *t* distribution with 37+10-2 = 45 degrees of freedom, which gives an expected value of 1.99 at 5% significance level (see Table of *t* distribution). However, the observed distance is 68/23.5 = 2.89 which is much higher than the

expected value. We conclude that the observed difference is beyond expected by chance alone, if the null hypothesis is true. We call this "statistically significant" difference. //

4.2. CONFIDENCE INTERVAL

The results of the t-test above are only part of the analysis, since it gives no indication of the size of any possible treatment effect. To do this, we have to present an estimate of the treatment effect, together with some measure of precision. Commonly, this is done by presenting treatment means with one of the following:

(a) Standard error of the mean. This is not particularly useful as we are interested in comparison of means, not in their individual values. For example, 1850 $\pm$ 59 m/s versus 1782 $\pm$ 89 m/s;

(b) Standard error of the difference in means e.g. 68 $\pm$ 23.5 m/s.

(c) A confidence interval for the difference in means. This is probably most useful, and gives far more information than the results of a hypothesis test. Consider two hypothetical studies of the same unit of measurements, where the 95% confidence intervals of the difference are:

(i)	-0.2 to 0.3	m/s
(ii)	-2.0 to -3.0	m/s
(iii)	-0.2 to 15	m/s

All give a "non-significant" result for the t-test (since the CI includes zero). In (i) it is clear that any difference is less than 0.3, too small to be of interest. In (ii) the interval is very wide (the experiment was imprecise); there may be no treatment effect, but may be a difference as large as 3. In (iii), although the difference includes zero but the trend of difference leans toward the positive direction, which clearly

shows a lack of sample size or accuracy of treatment effect. The three studies lead to very different conclusions, yet results of the significance test are the same.

Example 1 (cont):

We can use our knowledge gained in Topics 2 and 5, to construct a 95% confidence interval (CI) of difference by using the expected t value (in this case, 1.99 ≅ 2 for simplicity in calculation!). In our case, the 95% CI of the difference is: $68 \pm 2(23.5)$ = 21 m/s to 115 m/s. What this means is that if we keep sampling patients repeatedly from this population, 95% of the times, we would expected the observed differences in AVU between fracture and non-fracture patients lie between 21 m/s to 115 m/s - a pretty clear and convincing difference. Notice that the CI did not include a zero value. //

4.3. Assumptions And the Case of Unequal Variances

The t-test procedure as illustrated in Example 1 is based on a number of assumptions. The first and most critical one is that the two samples are independent. Practically, this means that the two samples are drawn from two different populations and in set language (Topic 3) that the elements of sample 1 are unrelated to the elements of sample 2. If this assumption is not held, then the t-test above is inappropriate.

The second assumption that we make is that the samples are drawn from normally distributed population. Fortunately, this assumption is less critical. The reason is that for modest sample sized samples, the Central Limit Theorem (Topic 5) applies and the sampling distribution for the sample means are approximately normal. If, however, the population is known to be non-normally distributed, then the non-parametric statistic of Wilcoxon Rank Sum (presented next section) will be used instead of the t-test.

The third and final assumption is that the two population variances are equal. For now, just examine the sample variances to see that they are approximately equal, later we will give a test for this assumption. Many efforts have been made to investigate the effect of deviations from the equal variance assumption on the t

methods for independent samples. The general conclusion is that for equal sample sizes, the population variances can differ by as much as a factor of 3 (i.e $\sigma_1^2 = 3\sigma_2^2$) and the t methods will still apply. This is remarkable and provides a convincing argument to use equal sample sizes. When the sample sizes are different, the most serious case is when the smaller sample size is associated with the larger variance. In this situation and in others where the sample variances (s_1^2 and s_2^2) suggest that $\sigma_1^2 \neq \sigma_2^2$, there is an approximate t test using the statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad [8]$$

with the number of df given by:

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)c^2 + (1 - c)^2 (n_1 - 1)} \qquad [9]$$

where $$c = \frac{s_1^2 / n_1}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(Welch, 1938).

4.4. NON-NORMALLY DISTRIBUTED DATA I: RESPONSES AFFECT MULTIPLICATIVELY

In biological science, several measurements have at least one of the following properties: (i) mean values are more sensibly compared in terms of their ratios than in terms of differences; (ii) the standard deviation is proportional to the mean; and (iii) the measurements have a log-normal distribution (i.e. if X is log-normally distributed then log(X) will be normally distributed). In these cases, it is necessary to transform data before a formal statistical test of significance can be carried out.

Example 2: The following data represent lysozyme levels in the gastric juice of 29 patients with peptic ulcer and of 30 normal controls. It was interested to know whether lysozyme levels were different between two groups.

Lysozyme levels in the gastric juice of two groups of subjects.

Group A (n=29)		Group B (n=30)	
0.2	10.4	0.2	5.4
0.3	10.9	0.3	5.7
0.4	11.3	0.4	5.8
1.1	12.4	0.7	7.5
2.0	16.2	1.2	8.7
2.1	17.6	1.5	8.8
3.3	18.9	1.5	9.1
3.8	20.7	1.9	10.3
4.5	24.0	2.0	15.6
4.8	25.4	2.4	16.1
4.9	40.0	2.5	16.5
5.0	42.2	2.8	16.7
5.3	50.0	3.6	20.0
7.5	60.0	4.8	20.7
9.8		4.8	33.0
Mean	$\bar{x}_1 = 14.31$		$\bar{x}_2 = 7.68$
SD	$s_1 = 15.74$		$s_2 = 7.85$

Firstly, let us apply the method in Example 1 to test for the difference. In this method, the observed difference between the two groups is $\bar{x}_1$ - $\bar{x}_2$ = 6.63 with pooled standard deviation of:

$$s = \sqrt{\frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}} = \sqrt{\frac{28(15.74)^2 + 29(7.85)^2}{29+30-2}} = 12.37$$

then the standard deviation of ($\bar{x}_1$ - $\bar{x}_2$) is $6.63\sqrt{\frac{1}{29}+\frac{1}{30}} = 3.22$

and the standardised distance is t = 6.63 / 3.22 = 2.06 which is significantly greater than the expected t-value of 2.00 with 57 df (from the Table of t distribution). We would conclude that the two groups are different in lysozyme levels.

However, a close inspection of the data reveals that (i) the standard deviation in group A is much higher than that in group B and (ii) the standard deviations vary systematically with the mean. For group A, the ratio of $s_1/\overline{x}_1$ = 15.74/14.31 = 1.10 and group B, $s_2/\overline{x}_2$ = 7.85/7.68 = 1.02. This simple calculation suggest that the data are not normally distributed and the above result (hence, conclusion) is not reliable. The data suggest a logarithmic transformation. The log-transformed data of the above table is as follows:

Group A (n=29)		Group B (n=30)	
1.61	2.34	-1.61	1.69
-1.20	2.39	-1.20	1.74
-0.92	2.42	-0.92	1.76
0.10	2.52	-0.36	2.01
0.69	2.79	0.18	2.16
0.74	2.87	0.41	2.17
1.19	2.94	0.41	2.21
1.34	3.03	0.64	2.33
1.50	3.18	0.69	2.75
1.57	3.23	0.88	2.78
1.59	3.69	0.92	2.80
1.61	3.74	1.03	2.82
1.67	3.91	1.28	3.00
2.01	4.09	1.57	3.03
2.28		1.57	3.50
Mean	$\overline{x}_1$ = 1.92		$\overline{x}_2$ = 1.41
SD	s_1 = 1.48		s_2 = 1.32

then, the pooled standard deviation of two groups is:

$$s = \sqrt{\frac{28(1.48)^2 + 29(1.32)^2}{29+30-2}} = 1.40$$

then the standard deviation of $(\bar{x}_1 - \bar{x}_2)$ is equal to $1.40\sqrt{\frac{1}{29}+\frac{1}{30}} = 0.365$

And the t-statistic is: $t = (1.92\text{-}1.41) / 0.365 = 0.51 / 90.365 = 1.40$, which is less than the expected value of 2 (with 57 df). Furthermore, the 95% confidence interval of differences between the two groups is 0.51-2(0.365) = -0.22 to 0.51+(2(0.365) = 1.24. Both calculations consistently suggest that the differences in lysozyme levels between the two groups is not statistically significant.

CONVERSION OF UNIT OF MEASUREMENTS

It may be noticed here that the confidence interval in log lasozyme is not informative per se, because the logarithm of lysozyme is not understandable unit of measurement. Its importance resides in the fact that it is easily translated into CI for the ratio of the two underlying mean levels. This is left as an exercise for the reader.

This example illustrates the importance of examining assumptions prior to any statistical analysis. //

4.5. NON-NORMALLY DISTRIBUTED DATA II: PROPORTIONS

Example 3: The following data are adapted from the results of a randomised study comparing two methods for training patients with senile dementia to care for themselves. After two weeks of training, each patient was presented with 20 tests involving activities of daily living (unlocking a door, tying one's shoe laces, etc.) and the proportion of tests that were successful was recorded.

The proportions of successful tests out of 20 attempted (X) for two groups of patients with senile dementia.

Group A (n=11)	Group B (n=8)
0.05	0
0.15	0.15

	0.35	0
	0.25	0.05
	0.20	0
	0.05	0
	0.10	0.05
	0.05	0.10
	0.30	
	0.05	
	0.25	
Mean	0.164	0.044
SD	0.112	0.056

If we apply the t-statistic in Example 1, we have the following results:

$$s = \sqrt{\frac{10(.112)^2 + 7(.056)^2}{10 + 7 - 2}} = 0.093$$

and the t-value is $t = \dfrac{0.164 - 0.0444}{0.093\sqrt{\dfrac{1}{11} + \dfrac{1}{8}}} = 2.78$

with 17 df, which is significant at the 5% level.

Before reaching a definite conclusion, let us examine the data a bit closely. In this data set, the two standard deviations (SD) differ by a factor of 2, and it is noteworthy that the group with the smaller mean proportion has the smaller SD. This even make intuitive sense that the proportion can not be less than zero! Indeed, the proportion data present a problem in the normal t-statistic, in that the variance is finite and mean dependent, since the maximum proportion must be equal to 1. For instance, it could be shown that when the proportion increases to 0.5, the SD is also expected to increase. The SD is expected to decline as the proportion approaches 1.

A method that may, in general, be expected to rectify the untoward consequence of unequal variance (heteroscedasticity) when the response variable is a proportion, say *p*, is to transform *p* by the arcsin (angular) transformation: Let

$$A = arcsin\sqrt{p}$$

A is the angle whose sine is the square root of p. It can be shown that A is effectively linear function of p for proportions in the interval from 0.25 to 0.75. In fact, over that interval, A is approximately equal to 0.285 + p. Therefore, in practice, if p is within the range 0.25 to 0.75, the arcsin transformation will be ineffective in reducing any inequality in variance. The arcsine transformation is very effective in $p < 0.25$ or $p > 0.75$.

You will be asked to perform, a t test on the transformed data. Note that the two standard deviations of the transformed data are nearly equal.

4.6. NON-NORMALLY DISTRIBUTED DATA III: COUNTS DATA

Example 4: The data in the following table represent the numbers of oral lactobacilli in the saliva of 7 subjects who had been vaccinated with heat-killed bacilli and six controls.

Counted numbers of oral lactobacilli in the saliva of two groups of subjects.

	Group A (n=7)	Group B (n=6)
	7925	3158
	15643	3669
	17462	5930
	10805	5697
	9300	8331
	7538	11822
	6297	
Mean	10710.0	6434.5
SD	4266.4	3218.8

Based on these data, the value of the t-ratio (standardised distance) for comparing the two means is 2.01 with 11 df, which is not significant at 5% level (expected t value of 11 df is 2.20).

We see that the two SDs are not too unequal, however, the variability is greater in group A, the group with larger mean. Furthermore, the two SDs seem to be proportional to the square roots of the means: $4266.4/\sqrt{10.710} = 41.2$ which is very close to the ratio in group B, $3218.8/\sqrt{6434.5} = 40.1$.

When, as here, the standard deviation is roughly proportional to the square root of the mean, the square roots transformation usually succeeds in equalising the SDs.

You are asked to confirm that the t-value for the square roots transformed data is 2.15. The difference between two groups still does not attain a statistical significance at 5% level, but at least this failure can not be attributed to an attenuating effect of unequal variability on the value of *t*. //

4.7. NON-NORMALLY DISTRIBUTED DATA IV: TIME TO OCCURRENCE OF AN EVENT

Example 5: The data in the following table are from, a randomised study comparing the effects of several combinations of poisons and treatments on the survival times of animals.

Group	Values	Mean	SD
A	4.3, 4.5, 6.3, 7.6	5.675	1.567
B	9.2, 6.1, 4.9, 12.4	8.150	3.363

The method in Example 1, when applied to this data set, yields a t-ratio of 1.33, which is well below the expected value of 2.447 (6 df). The conclusion is, therefore, there was no statistical significance between two groups.

However, the data show that two SDs seem to vary systematically with the means. Specifically, SD is proportional to $\bar{x}^2$. For example $s/\bar{x}^2 = 1.567 / (5.675)^2 = 0.048$, which is equivalent to $3.363/(8.15)^2 = 0.05$. What this means in practice is that, the *reciprocal transformation*, $Y = 1/X$, is appropriate. Fortunately, the reciprocal transformation has physical meaning when, as here, the response variable is in units of time. If the response variable is the time until death or some other event, the reciprocal is related to the death rate or more generally to the rate at which the event occurs.

You will be asked to perform a t-test to the reciprocals of the measurements in the data.

4.8. NON-PARAMETRIC ANALYSIS OF UNPAIRED DATA: THE WILCOXON RANK SUM TEST.

The two sample t test of the previous section was based on several assumptions as described in II(C). There is, however, an alternative test procedure that requires less stringent assumptions. This test, called the Wilcoxon's Rank Sum (WRS) test, is discussed here.

The assumptions for this test are that we have two independent random samples taken from two populations. The WRS test provides a procedure for testing that two populations are identical but not necessarily normal. Since the two populations are assumed to be identical under the null hypothesis, independent random samples from the respective populations should be similar. One way to measure the similarity between the samples is to jointly rank (from lowest to highest) the measurements from the combined samples and examine the sum of the ranks for measurements in sample 1 (or, equivalently, sample 2). Under the null hypothesis of identical populations, the sum of the ranks for a sample will be proportional to the sample size. We let T denote the sum of the ranks for sample 1. Intuitively, if T is extremely small (or large), we would have evidence to reject the null hypothesis that the two populations are identical.

Under the null hypothesis and according to the general principle of section I, the statistic T, will have a sampling distribution with mean and variance given by:

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \qquad [10]$$

and $$\sigma_T^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}. \qquad [11]$$

If both sample sizes are 10 or larger, the sampling distribution of T is approximately normal with mean 0 and variance 1 (standardised normal distribution).

The theory behind the WRS test assumes that the population distributions are continuous, so that there is zero probability that any two observations are identical. In practice, there will often be ties - two or more observations with the same value. For these situations, each observation in a set of tied values receives a rank score equal to the average of the ranks for the set. For example, if two observations are tied for the rank 3 and 4, each is given a rank of 3.5; the next higher value receives a rank of 5, and so on. When there are ties, there is a correction, for the variance formula. Then σ_T^2 is:

$$\sigma_T^2 = \frac{n_1 n_2}{12}\left[(n_1 + n_2 + 1) - \frac{\sum t_i (t_i^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}\right] \qquad [12]$$

where t_i denotes the number of tied ranks in the ith group.

From a practical standpoint, however, unless there are many ties, the correction will have very little impact on the value of σ_T^2.

Example 6: The following data are dissolved oxygen measurements (in ppm) collected from 12 samples in lake A and another 12 samples in a lake B. It is of interest to know whether the distributions of measurements in lakes A and B are identical.

Lake A: 11.0, 11.2, 11.2, 11.2, 11.4, 11.5, 11.6, 11.7, 11.8, 11.9, 11.9, 12.1
Lake B: 10.2, 10.3, 10.4, 10.6, 10.6, 10.7, 10.8, 10.8, 10.9, 11.1, 11.1, 11.3

To apply the Wilcoxon's Rank Sum test, we firstly jointly rank the combined sample of 24 measurements by assigning the rank of 1 to the smallest, and so on. When two or more measurements are the same, we assigned all of them a rank equal to the average of the ranks they occupy.

Value	Rank	Rank1	Ties
10.2	1	1	1
10.3	2	2	1
10.4	3	3	1
10.6, 10.6	4.5	4	2
10.7	6	5	1
10.8, 10.8	7.5	6	2
10.9	9	7	1
11.0	10	8	1
11.1, 11.1	11.5	9	1
11.2, 11.2, 11.2	14	10	3
11.3	16	11	1
11.4	17	12	1
11.5	18	13	1
11.6	19	14	1
11.7	20	15	1
11.8	21	16	1
11.9, 11.9	22.5	17	2
12.1	24	18	1
Sum	216		

For all groups with $t_i = 1$, there is no contribution for the variance σ_T^2. Thus, we need only be concerned with t_i =2, 3. Then [10] and [12] becomes:

$$\mu_T = \frac{n_1(n_1+n_2+1)}{2} = \frac{12(12+12+1)}{2} = 150$$

and

$$\sigma_T^2 = \frac{n_1 n_2}{12}\left[(n_1+n_2+1) - \frac{\sum t_i\left(t_i^2-1\right)}{(n_1+n_2)(n_1+n_2-1)}\right]$$

$$= \frac{12(12)}{12}\left[25 - \frac{6+6+6+24+6}{24(23)}\right]$$

$$= 298.9$$

$$\sigma_T = 17.29$$

And the standardised distance is:

$$z = (216 - 150) / 17.29 = 3.82$$

which exceeds 1.645 (at 5% level); we conclude that the dissolved oxygen measurements are different between lakes A and B.

Notice the variance without correcting for ties is:

$$\sigma_T^2 = \frac{n_1 n_2}{12}(n_1 + n_2 + 1) = \frac{12(12)(25)}{12} = 300.$$

which is not appreciatively different to 298.9. //

V. DIFFERENCES BETWEEN MEANS: PAIRED SAMPLES

5.1. THE PAIRED T-TEST

There is a fairly popular type of experimental design in which two treatments are applied to the *same* subject in two different periods. The statistical solution to this design is called the paired t-test (as oppose to the unpaired t-test presented in Example 1).

Example 7: The following table presents data from a study to compare two treatments (A and B) in the clotting times of plasma (in minutes) for 8 patients.

ID	Treatment A	Treatment B	Difference
1	10.6	10.2	0.4
2	9.8	9.4	0.4
3	12.3	11.8	0.5
4	9.7	9.1	0.6
5	8.8	8.3	0.5
Mean	10.24	9.76	0.48
SD	1.73	1.76	0.084

Do the data present sufficient evidence to claim for treatment effect?

We note that the difference between the two treatments is rather small (|10.24 - 9.76| = 0.48), considering the variability of the data and the small number of measurements involved. At first glance, it would seem that there is little evidence to indicate a difference between the population means, a conjecture that we may check by the method outlined in Example 1 (unpaired t-test). In this method, the pooled estimate of the common variance is $s^2 = 1.748$ (i.e. $s = 1.32$), and the standard error of the difference is $1.32\sqrt{\frac{1}{5}+\frac{1}{5}} = 0.83$. The calculated value of t is then 0.48/0.83 = 0.57, which is much lower than its expected value of 2.306 (8 df). We would be tempted to conclude that there was no significant difference between the two treatments.

A second glance at the data reveals a marked inconsistency with this conclusion. We note that the clotting times of plasma in treatment A is larger than treatment B for *each* of the five patients. These differences, recorded at 0.48, on average.

Suppose that we were to use y, the number of times that treatment A is larger than treatment B, as a test statistic as was done in Binomial distribution. Then the probability that treatment A would be larger than B, assuming no difference between the time, would be $p = 0.5$, and y would be a binomial random variable. If the null hypothesis were true, the expected value of y would be $np = 5(0.5) = 2.5$. If we choose the most extreme values of y, $y = 0$ and $y = 5$, as the rejection region for a two-tailed test, then $\alpha = p(0) + p(5) = 2(0.5)^5 = 0.0625$. We would then reject the null hypothesis. Certainly, this is the evidence to indicate that a difference exists in the mean clotting time of the two treatments.

What went wrong ?

The explanation for this seemingly inconsistency is quite simple. The *t* test described earlier is not the proper test to be used for this kind of study design. We mentioned in section II(c) that one of the assumptions of the t-statistic is that two samples are *independent* and random. Certainly, the independence requirement was violated by the manner in which the experiment was conducted. The pairs of measurements for a particular patients are definitely related. A glance at the data will show that the time readings are of approximately the same magnitude for a particular patient but vary from one patient to another. This is, of course, exactly what we might expect.

The proper analysis of data would utilise the five *difference* measurements (column 3) to test the hypothesis that the average difference is equal to zero, or equivalently, to test the hypothesis that $\mu_D = \mu_A - \mu_B = 0$ against the hypothesis that $\mu_D \neq 0$. Now the standard deviation of the differences is 0.084, i.e. standard error = $\frac{0.084}{\sqrt{5}} = 0.037$. The standardised distance is then 0.48 / 0.037 = 12.8, which is much higher than expected value of 2.776 (with 4 df, at 5% significance level). Furthermore, the 95% confidence interval of the difference is: 0.48 - 2.776(0.037) =0.38 to 0.48 + 2.776(0.037) = 0.58 minutes. We conclude that treatment A has a significantly higher clotting time than treatment B. //

5.2. NON-PARAMETRIC ANALYSIS OF PAIRED DATA: THE WILCOXON SIGNED RANK TEST

This test makes use of the sign and the magnitude of the rank of the differences between pairs of measurements, provides an alternative to the paired t-test as presented above. The formal idea for Wilcoxon Signed Rank test is that the population distribution of differences is symmetrical about *D*; the test is sensitive to the distribution of differences being shifted to the right or left of *D*. In most case *D* is 0; otherwise, we subtract *D* from every measurement and proceed as if $D = 0$. The test uses the non-zero differences ranked in absolute value from lowest to highest. If two or more measurements have the same nonzero difference (ignoring sign) we assign each difference a rank equal to the average of the occupied ranks. The appropriate sign is then attached to the rank of each difference.

By defining n = the number of pairs of observations with a nonzero difference
$T+$ = the sum of the positive ranks.
$T-$ = the sum of the negative ranks.
T = the smallest of $T+$ and $T-$ ignoring their signs.

then the mean and standard deviation of the rank is:

$$\mu = \frac{n(n+1)}{4} \qquad [13]$$

and

$$\sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}} \qquad [14]$$

If we there are ties, the standard deviation would become:

$$\sigma = \sqrt{\frac{1}{24}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_i t_i(t_i-1)(t_i+1) \right]} \qquad [15]$$

where t_i denotes the number of tied ranks in the ith group.

The statistic $z = \dfrac{T - \dfrac{n(n+1)}{4}}{\sigma}$ is approximately normally distributed with mean 0 and variance of 1.

Example 8: Two different drugs were compared on each of 10 different patients in terms of the time (in minutes) to reach maximum concentration. The data are as follows:

ID	Drug A	Drug B	Difference
1	312	346	-34
2	333	372	-39
3	356	392	-36
4	316	351	-35
5	310	330	-20
6	352	364	-12
7	389	375	14
8	313	315	-2
9	316	327	-11
10	346	378	-32

To apply the Wilcoxon Signed Rank test, we firstly rank the absolute values of the n=10 differences. The appropriate sign is then attached to each rank, as follows:

ID	Difference	Rank of absolute difference	Rank with appropriate sign
1	-34	7	-7
2	-39	10	-10
3	-36	9	-9
4	-35	8	-8
5	-20	5	-5
6	-12	3	-3
7	14	4	4
8	-2	1	-1
9	-11	2	-2
10	-32	6	-6

The sum of positive and negative ranks are as follows:

T+ = 4,
T- = -7 + (-10) + . . . + (-6) = -51

Thus, T - the smallest of T+ and T- ignoring the sign, is 4.

For a two-tailed test with n=10 and α=0.05, we see from the Wilcoxon Table that we will reject the null hypothesis if T is less than or equal to 8. Or alternatively, we calculate the standardised distance z as follows:

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{n(n+1)(2n+1)/24}} = \frac{4 - \frac{10(11)}{4}}{\sqrt{10(11)(21)/24}} = \frac{-23.5}{9.81} = 2.39,$$

which is greater than 1.96 (at 5% level of significance). Thus, we conclude that the two drugs have different times to reach maximum plasma concentration. //

VI. DIFFERENCES BETWEEN MEDIANS

Recall that the median is a measure of central tendency, in which half of the observations are less than and half of the observations is exceeding it.

6.1. TEST STATISTIC FOR DIFFERENCE BETWEEN TWO MEDIANS

One can test the null hypothesis that two samples came from a population with the same median by the *median test* (Mood 1950. Introduction to the Theory of Statistics. McGraw-Hill, New York 394-395 pp). The procedure is to set up a table as in the following example, and then apply the Chi-square or Fisher's exact test.

Example 9: Data on body sway were collected for two sample of subjects, the number of subjects who were higher or lower than the median were as follows:

Number	Sample 1	Sample 2	Total
Above median	6	6	12
Not above median	3	8	11
Total	9	14	23

Chi square statistic $\chi^2 = 0.473$, which is lower then the expected value of 5.02 (with 1 df), we conclude that the medians of two samples are equivalent.

This conclusion is consistent with a Fisher's exact test (which we will introduce in the next section).

$$\text{Fisher's exact test} = \frac{(12!11!9!14!)/23!}{6!6!3!8!} = 0.18657 \qquad //$$

6.2. CONFIDENCE INTERVAL FOR A MEDIAN

To find a confidence interval for a population median, we first need to calculate the following quantities:

$$r = \frac{n}{2} - \left(N \times \frac{\sqrt{n}}{2} \right) \quad \text{and} \quad s = 1 + \frac{n}{2} + \left(N \times \frac{\sqrt{n}}{2} \right)$$

where n is the sample size; N is the appropriate value from the standard normal distribution. Then round r and s to the nearest integers. The n sample observations need to be ranked in increasing order of magnitude and the rth to sth in the ranking determine the CI for the population median. This approximation is satisfactory for most sample size. The exactly method based on the binomial distribution can be used instead as shown in the example below.

Example 10:

Suppose that the median systolic BA among 100 patients was 146 mmHg. Using the above formula we have

$$r = \frac{100}{2} - \left(1.96 \times \frac{\sqrt{100}}{2} \right) = 40 \text{ and } r = 1 + \frac{100}{2} + \left(1.96 \times \frac{\sqrt{100}}{2} \right) = 61$$

From the *original data*, the 40th observation in increasing order is 142 mmHg and the 61st is 150 mmHg. Therefore, the 95% CI for the population median is 142 mmHg to 150 mmHg. //

6.3. CONFIDENCE INTERVAL FOR DIFFERENCE BETWEEN TWO MEDIANS

Let $x_1, x_2, \ldots, x_n$ represent the n observations in a sample from one population and $y_1, y_2, \ldots, y_m$ the m observations from a second population, where both populations are thought not to come from normal populations. The difference between the two population medians or means is estimated by the median of all possible $n \times m$ differences $(x_i - y_j)$ for $i = 1,,n$ and $j = 1,,m$.

For studies with small sample size, the CI is calculated based on the following statistic:

$$K = W - \frac{n(n+1)}{2}$$

where W is the percentile distribution of the Mann-Whitney test statistic or of the equivalent Wilcoxon two sample test statistic. The Kth smallest to the Kth largest of the $n \times m$ are the required CI. Values for K for a given m and n is given in the appendix:

The CI for the difference between the two population medians is also derived through these $n \times m$ differences. For studies with each sample size > 20, we can calculate CI as follows:

$$K = \frac{nm}{2} - \left(N \times \sqrt{\frac{nm(n+m+1)}{12}} \right)$$

rounded up to the next integer value, where N is the appropriate value from the standard normal distribution (for example, 1.96 for 95% CI).

Example 11: Consider the data on the globulin fraction of plasma (g/l) in two groups of 10 patients as follows:

Group 1:	38	26	29	41	36	31	32	30	35	33
Group 2:	45	28	27	38	40	42	39	39	34	45

The computations are made easier if the data in each group are first ranked into increasing order of magnitude and then all the differences for group 1 - group 2 calculated as in the following table:

	26	29	30	31	32	33	35	36	38	40
27	-1	2	3	4	5	6	8	9	11	14
28	-2	1	2	3	4	5	7	8	10	15
34	.	.	.							
38										
39										
39										
40										

42
45
45 -19 -16 -15 -14 -13 -12 -10 -9 -7 -4

The estimate of the difference in population medians is now given by the median of these differences. From 100 differences in this table, the 50th percentile is -6 g/l and the 51st is -5 g/l, so the median is -5.5 g/l.

To calculate the 95% CI for the difference in population medians, the value of K is found to be 24 for n=10 and m=10. The 24th smallest difference is -10 g/l and the 24 largest difference is +1 g/l. //

VII. DIFFERENCES BETWEEN VARIANCES AND COEFFICIENTS OF VARIATION

7.1. DIFFERENCES BETWEEN TWO VARIANCES

One of the major applications of a test for equality of population variances is for checking the validity of the assumption (that is $\sigma_1^2 = \sigma_2^2$) for a two-sample t-test. First we hypothesise that two populations of measurements that are normally distributed. We are interested in comparing the variance of populations 1 and 2 as σ_1^2 and σ_2^2, respectively. We denote their respective sample estimates as s_1^2 and s_2^2.

When the independent samples have been drawn from the respective populations, the ratio $F = \frac{s_1^2}{s_2^2} / \frac{\sigma_1^2}{\sigma_2^2} = \frac{s_1^2 \sigma_2^2}{s_2^2 \sigma_1^2}$ possesses a probability distribution in repeated sampling referred to as an F distribution (Topic 5). Under the hypothesis of $\sigma_1^2 = \sigma_2^2$, the statistic becomes $F = s_2^2/.s_1^2$ or $F = s_1^2/s_2^2$. (depend whether $s_1^2 < s_2^2$ .or $s_1^2 > s_2^2$) is the test statistic with $n_1 - 1$ and $n_2 - 1$ df.

Example 2 (Continued):

In this data set, for a sample of 29 patients, the standard deviation of lysozyme is 15.74 (s_1^2 = 247.7) and for a sample of 30 patients the standard deviation is 7.85 (s_2^2 = 61.62).

To test whether the two variances are different, we used the statistic described as above i.e. F = 247.7 / 61.62 = 4.02. Now, the expected value of the F variate with 28 and 29 df is 1.84. Since the observed F is much larger than its expected value, we conclude that the two variances are, indeed, different. //

7.2. DIFFERENCES BETWEEN TWO COEFFICIENTS OF VARIATION

Recall from Topic 3 that a coefficient of variation (CV) is defined as the ratio of standard deviation (s) over the sample mean ($\bar{x}$), i.e. CV = $s/\bar{x}$.

Now, suppose that we have data from two samples of subjects, in which two coefficients of variation are obtained. Lewontin (1966) has shown that the variance ratio

$$F = \frac{\left(s^2_{\log}\right)_1}{\left(s^2_{\log}\right)_2}$$

can be used analogously to the ratio of two variances above, to test for difference between two coefficients of variation. Notice that F is distributed with $n_1 - 1$ and $n_2 - 1$ df. In this statistic, $\left(s^2_{\log}\right)_1$ refers to the variance of the logarithmic data for sample 1, and $\left(s^2_{\log}\right)_2$ refers to variance of the logarithmic data for sample 2.

Unfortunately, we are faced with the requirement of the variance ratio test that the two underlying distributions be normal (or nearly normal). Thus, this test must be applied with caution, for if the two sets of sample data are, in fact, from normal populations, the logarithms of the data *will not* be normally distributed; and the requirement here is that the logarithm be normally distributed.

VIII. DIFFERENCES BETWEEN TWO PROPORTIONS

8.1. THE T-TEST FOR DIFFERENCES BETWEEN TWO PROPORTIONS

UNPAIRED SAMPLES

Many experiments involve the comparison of two proportions, which could be considered as binomial parameters. For comparisons of this type, we assume that independent random samples are drawn from two binomial populations with unknown parameters designated by π_1 and π_2. If y_1 is the number of successes observed for the random sample of size n_1 and y_2 is number of successes observed for the random sample of size n_2, then the point estimate of π_1 and π_2 are: $p_1 = \frac{y_1}{n_1}$ and $p_2 = \frac{y_2}{n_2}$, respectively.

We learned earlier (Topic 4) that the variance of p_1 and p_2 are $SD(p_1) = \sqrt{\frac{p_1(1-p_1)}{n_1}}$ and $\text{var}(p_2) = \frac{p_2(1-p_2)}{n_2}$. It follows that their respective standard deviation is $SD(p_2) = \sqrt{\frac{p_2(1-p_2)}{n_2}}$.

Also, since the two samples are assumed to be independent, the expected value of the difference is

$$E(p_1 - p_2) = \pi_1 - \pi_2 \qquad [16]$$

and the variance of the difference is

$$\text{var}(p_1 - p_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2},$$

i.e.
$$SD(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}. \qquad [17]$$

Thus, according to the general principle presented in section 1, to test for difference between two proportions we need to calculate the standardised distance:

$$z = \frac{p_1 - p_2}{sd(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

However, under the hypothesis that $\pi_1 = \pi_2 = \pi$, we can estimate π by a $p = \frac{y_1 + y_2}{n_1 + n_2}$

and hence the standard deviation of the difference is:

$$\text{var}(p_1 - p_2) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2},$$

i.e. $$SD(p_1 - p_2) = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \qquad [18]$$

then the standardised distance becomes:

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad [19]$$

It follows that 95% confidence interval of $(\pi_1 - \pi_2)$ can be constructed by the statistic: $(p_1 - p_2) \pm 1.645 \times SD(p_1 - p_2)$.

In summary:

	Population	
Characteristics	1	2
Population proportion	π_1	π_2
Sample size	n_1	n_2
Number of successes	y_1	y_2
Sample proportion	$p_1 = \frac{y_1}{n_1}$	$p_2 = \frac{y_2}{n_2}$
Variance	$p_1(1-p_1)$	$p_2(1-p_2)$

Example 12: In a recent opinion poll of 200 people, it was found that 58 out of 100 people interviewed said they would vote for Paul Keating, while 46 people out of another 100 people interviewed said they would vote for John Hewson. Is it true that Paul Keating has a higher electoral appeal than John Hewson or the difference was just due to chance fluctuation?

We use the Binomial theory to answer this question. Let the proportion of people who said they would vote for Keating be $p_1 = 0.58$ and for Hewson $p_2 = 0.46$. Of course, these are only estimates, because we do not know the true proportion of voters for the two leaders π_1 and π_2. Under the hypothesis of no difference e.g. $\pi_1 = \pi_2 = \pi$, we can estimate π by $p = (58+46)/200 = 0.52$. Then, the standard deviation of the difference $(p_1 - p_2)$ is

$$\mathrm{SD}(p_1 - p_2) = \sqrt{0.52(1-0.52)\left(\frac{1}{100}+\frac{1}{100}\right)} = 0.07 .$$

The standardised distance between the two population proportions is then

$$z = \frac{0.58 - 0.46}{0.07} = 1.7 .$$

Since this distance is higher than the expected value of 1.645 (from the standardised normal distribution table), we conclude that Paul Keating has a better chance of election than John Hewson. //

PAIRED OR MATCHED SAMPLES

As we can see from the above example, we have two separate groups of subjects, which can be regarded as independent samples. Sometimes, we observe the proportion of an attribute from the same group of subjects in two different occasions or two matched groups; this is called paired samples. Difference between two proportions in this design can be tackled by a statistic called the **McNemar's test**.

Example 13: The following data represent results of 32 subjects showing numbers with + or without (-) sleeping difficulties among marijuana users and matched controls.

	Marijuana		
	+	-	Total
Control			
+	$n_{11} = 4$	$n_{12} = 9$	$n_{1.} = 13$
-	$n_{21} = 3$	$n_{22} = 16$	$n_{2.} = 19$
Total	$n_{.1} = 7$	$n_{.2} = 25$	$n_{..} = 32$

The comparison of paired proportions is based on the frequencies of pairs with different outcomes. The McNema's test is given by:

$$\chi^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}} \qquad [20]$$

which is referred to the Chi squared distribution with 1 df.

In our example, the actual value of χ^2 is $\frac{(|9-3|-1)^2}{9+3} = \frac{25}{12} = 2.08$, which is less than its expected value under the χ^2 distribution with 1 df (5.02). We conclude that there was no difference between the two groups with respect to sleeping habits. //

8.2. MEASURE OF ASSOCIATION: THE FISHER'S EXACT TEST

The Binomial based test as introduced in Example 12 is powerful when the sample size is reasonably large. In fact, we will learn that when the sample size is small and hence the normal approximation is not very accurate, the test can be unreliable.

When sample size is small, another test of association based on the hypergeometric probability distribution should be used; it is exact probability, and hence called Fisher-Irwin's exact test, so named after the two prominent statistician in this century.

Consider the following typical setting which most in epidemiological studies often result in. We restrict our attention to the four-fold table in which the frequencies $n_{1.}$, $n_{2.}$, $n_{.1}$, $n_{.2}$ are fixed at the observed values.

	Characteristics B		
	B	Not B	Total
Characteristics A			
A	n_{11}	n_{12}	$n_{1.}$
Not A	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

The exact test consists of evaluating the probability associated with all possible 2x2 tables which have the same row and column totals as observed data, making the assumption that the null hypothesis is true. The null hypothesis is that the row and column variables are unrelated.

Under this restriction, the exact probabilities associated with the cell frequencies n_{11}, n_{12}, n_{21}, n_{22} may be derived from the hypergeometric probability distribution as follows:

$$P(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n_{..}!n_{11}!n_{12}!n_{21}!n_{22}!} \qquad [21]$$

This is called the Fisher-Irwin "exact" test statistic for examining the above probability.

Example 14: Consider the following data on the number of subjects with a certain disease, classified by sex. It is interested to assess the exact probability for each cell in the table and hence the association between sex and the disease.

	Sex		
	Males	Females	Total
Disease			

Yes	2	3	5
No	4	0	4
Total	6	3	9

The exact probability associated with the table is $\frac{5!4!6!3!}{9!2!3!4!0!} = 0.119$. //

8.3. MEASURE OF ASSOCIATION IN PROSPECTIVE STUDY: THE RELATIVE RISK

In a prospective study, groups of subjects are followed up to see whether an outcome of interests occurs. Many clinical trials and longitudinal studies are of this design; so are too observational studies where it is impossible to randomise the feature of interest such as BMD. We can assess the association between risk factors and an outcome by calculating the proportion of an outcome for each risk group and then contrast them in a ratio. We call this the **relative risk (RR)**.

The data of this design can be summarised as follows:

	Outcome			
	Yes	No	Total	Proportion
Risk factors				
Yes	n_{11}	n_{12}	$n_{1.}$	$p_1 = n_{11} / n_{1.}$
No	n_{21}	n_{22}	$n_{2.}$	$p_2 = n_{21} / n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$	

Then the relative risk is defined by:

$$\text{RR} = \frac{p_1}{p_2} = \frac{n_{11} / n_{1.}}{n_{21} / n_{2.}} \qquad [22]$$

The standard deviation of ln(RR) is given by:

$$\text{SD(RR)} = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1.}} + \frac{1}{n_{21}} - \frac{1}{n_{2.}}} \qquad [23]$$

Example 15: The following data represent a longitudinal study in which 283 subjects were followed up for 5 years. The number of fractures classified by baseline bone mineral density (BMD) were as follows:

	Outcome			
	Fracture	No fracture	Total	Proportion
Baseline BMD				
Low	15	90	105	0.143
High	8	170	178	0.045
Total	23	260	283	

Using the above statistic, the relative risk is:

$$RR = \frac{.143}{.045} = 3.18$$

and $\ln(RR) = \ln(3.18) = 1.156$

The SD of ln(RR) is: $SD(\ln RR) = \sqrt{\frac{1}{15} - \frac{1}{105} + \frac{1}{8} - \frac{1}{178}}$

$= 0.42$

Then 95% interval for ln(RR) is :

$1.156 - 2(0.42)$ to $1.156 + 2(0.42)$

<=> 0.316 to 1.996

i.e the 95% CI for RR is:

1.37 to 7.36 //

8.4. MEASURE OF ASSOCIATION IN CROSS-SECTIONAL STUDY: THE ODDS RATIO

In case-control or retrospective studies, subjects are selected based on the outcome (as oppose to prospective studies where subjects are selected based on the risk

factors or the characteristic defining the groups). In retrospective, we can not measure the risk of the outcome because of the ways the subjects were sampled. Furthermore, because we can get any value of risk we want by varying the numerator and denominator (number of cases and control) that we choose to study, and so, the relative risk as presented earlier is not a valid test.

We need the a method of calculations based on within each group. Here we consider a very popular test which was originally proposed in the 1950s for 2x2 tables that are not a function of Chi square statistic; it is called the *odds ratio*.

We will study this test by using the following example.

Example 16: Consider the following hypothetical, cross-sectional data on the association between .maternal age and birth weight.

	Weight		
	Low	Normal	Total
Maternal Age			
Young	20	80	100
Mature	30	270	300
Total	50	350	400

The significance of the association between maternal age and birth weight may be assessed by means of the standard Binomial test.

Frequently, one of the two characteristics being studies is antecedent to the other. In this example we are considering maternal age is antecedent to birth weight. A measure of the risk of experiencing the outcome under study for the young mothers is presented as follows:

$$\Omega_{young} = \frac{P(Low|Young)}{P(Normal|Young)}$$

We could estimate this risk by using our observed data as follows:

$$\text{O(young)} = \frac{20/100}{80/100} = \frac{20}{80} = 0.25.$$

Thus, for every 4 births weight normal to young mothers, there is one abnormal birth weight.

Similar, a measure of the risk of experiencing the outcome under study for the old mothers is presented as follows:

$$\Omega_{old} = \frac{P(Low|Old)}{P(Normal|Old)}$$

which is estimated by:

$$\text{O(old)} = \frac{30}{270} = 0.11$$

that is, the odds that an old mother will deliver an offspring abnormal weight is 0.11

The two odds may be contrasted to provide a measure of association as follows:

$$\text{OR} = \frac{O_{young}}{O_{old}} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \qquad [24]$$

In our example the odds ratio is

$$\text{OR} = \frac{20 \times 270}{30 \times 80} = 2.25$$

indicating that the odds of a young mother delivering an offspring with abnormal birth weight are 2.25 times those for an old mother.

The standard deviation of OR is given by:

$$\begin{aligned} \text{SD(OR)} &= OR \times \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \qquad [25] \\ &= 2.25 \times \sqrt{\frac{1}{20} + \frac{1}{80} + \frac{1}{30} + \frac{1}{270}} \\ &= 0.71 \end{aligned}$$

The 95% CI of the odd ratio is then: $2.25 \pm 1.96(0.71)$. //

8.5. MEASURE OF ASSOCIATION IN COMPARATIVE STUDY: RELATIVE DIFFERENCE

In comparative clinical trials, treatments are assigned to subjects at random. The measure of association is sometimes hampered by the fact that subjects are allowed to prematurely withdraw from the study for various (including ethical) reasons. The following example presents a measure of association using the idea of **relative difference.**

Example 17: Suppose that the data in the following table resulted from a trial in which one treatment was applied to a sample of n_1=80 patients randomly selected from a total of n=150 patients and the other was applied to the remaining n_2=70 patients.

	No. of patients	Proportion improved
Treatment 1:	80 (n_1)	0.60 (p_1)
Treatment 2:	70 (n_2)	0.80 (p_2)
Total	150 (n)	0.69 (p)

For this data, the statistical significance of the difference between the two improvement rates can be tested using the binomial theory as described earlier (Example 9). In that method we have: $z = \dfrac{0.8 - 0.6}{\sqrt{0.69(1-0.69)\left(\dfrac{1}{80} + \dfrac{1}{70}\right)}} = 2.18$, which indicates a significant difference at the 0.05 level.

Now, if we consider the simple difference between p_2 and p_1 ($d = p_2 - p_1 = 0.2$) which implies that every 100 patients given the first treatment, an additional 20 would have been expected to improve had they been given the 2nd treatment. The estimate standard deviation of d is $\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}} = 0.07$. An approximate

95% CI for the difference underlying the rates of improvement is 0.20 $\pm$ 1.96(0.07 or between 0.06 and 0.34.

But what we have are just sample data, we do not know the true rate of improvement in either treatment group. Let π_1 be the proportion improving in the population of patients who are given the 1st treatment and π_2 for the same indicator for treatment 2. Let *f* denote the proportion of patients, among those failing to respond to the 1st treatment, who would be expected to respond to the second treatment. It is then assumed that

$$\pi_2 = \pi_1 + f(1-\pi_1)$$

that is, the improvement rate under the 2nd treatment is equal to that under the first plus an added improvement rate which applies only to patients who fail to improve the 1st treatment. In other words,

$$f = \frac{\pi_2 - \pi_1}{1-\pi_1}$$

this is called the *relative difference* (RD), which can be estimated by:

$$\text{RD} = \frac{p_2 - p_1}{1-p_1} \qquad [26]$$

The standard deviation of RD is approximately (Sheps 1959):

$$\text{SD(RD)} = \frac{1}{1-p_1}\sqrt{\frac{p_2(1-p_2)}{n_2} + (1-RR)^2 \times \frac{p_1(1-p_1)}{n_1}}$$

Walter (1975) showed that, more accurate inferences about *f* could be made by taking log(1-*RR*) as normally distributed with mean of log(1-*f*) and standard deviation of

$$SD\left[\log(1-RD)\right] = \sqrt{\frac{p_2}{n_2(1-p_2)} + \frac{p_1}{n_1(1-p_1)}}. \qquad [27]$$

For the data in this example, we have:

$$\text{RD} = \frac{0.8 - 0.6}{1 - 0.6} = 0.5$$

(implying that for every 100 patients who fail to improve under the 1st treatment, 50 would be expected to improve under the 2nd treatment). The SD of ln(1-RD) is

$$SD\left[\log(1 - RD)\right] = \sqrt{\frac{0.8}{70(0.2)} + \frac{0.6}{80(0.4)}} = 0.28$$

Then, 95% CI for log(1-*f*) is -0.69 $\pm$ 1.96$\times$0.28 = -1.24 to -0.14.. By taking antilog, this interval is equal to 0.13 to 0.71. //

8.6. MEASURE OF AGREEMENT: THE KAPPA STATISTIC

Sometimes, a patient is diagnosed by two investigators, but using exactly the same *qualitative* scale of measurement as mild, moderate, severe, etc. If the diagnosis is repeated in several patients, the results can be summarised in 2x2 table as follows:

	Investigator A			
	1	2	3	Total
Investigator B				
1	n_{11}	n_{12}	n_{13}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

Of course, the table could be expanded easily to accommodate more categories, but for the purpose of illustration, it is presented with 3 categories.

Obviously, the proportion of agreement is equal to the sum of diagonal cells divided by total sample size:

$$p = \frac{n_{11} + n_{22} + n_{33}}{n_{..}} \qquad [28]$$

But we can see that this statistic is inadequate as a measure of reliability, because it may be that some agreements could be purely due to chance. In fact, the overall proportion of agreement expected by chance alone is, say,

$$p_{chance} = \frac{n_{1.}n_{.1} + n_{2.}n_{.2} + n_{3.}n_{.3}}{(n_{..})^2} \qquad [29]$$

So, a better measure of agreement than p alone is (p - p_{chance}), that is how much agreement exists beyond the amount expected by chance alone. But we want to have an index with maximum value of 1 to indicate a perfect agreement and close to 0 for a poor agreement. The kappa statistic is built based on this concept and is given by:

$$\kappa = \frac{p - p_{chance}}{1 - p_{chance}} \qquad [30]$$

the standard deviation of κ is:

$$SD(\kappa) = \sqrt{\frac{1}{n_{..}(1 - p_{chance})^2}\left(p_{chance} + p_{chance}^2\right) - \sum_i p_{i.}p_{.i}(p_{i.} + p_{.i})}$$

However, a very good approximate value for SD(κ) can be used:

$$SD(\kappa) = \sqrt{\frac{p(1-p)}{n(1 - p_{chance})^2}} \qquad [31]$$

Example 18: Suppose that 100 patients were assessed by two investigators on the severity of adverse reaction. The results are as follows:

	Investigator A			
	Mild	Moderate	Severe	Total
Investigator B				
Mild	20	12	8	40
Moderate	2	15	13	30
Severe	8	2	20	30

Total	30	29	41	100

The observed proportion of agreement is:

$$p = \frac{20+15+20}{100} = 0.55$$

And the proportion of agreement expected by chance is:

$$p_{chance} = \frac{(40\times 30)+(30\times 29)+(30\times 41)}{100^2} = 0.33$$

The Kappa statistic (κ) is then:

$$\kappa = \frac{0.55-0.33}{1-0.33} = 0.328$$

With approximate SD $SD(\kappa) = \sqrt{\frac{0.55(1-0.55)}{100(1-0.33)^2}} = 0.074$

95% CI of κ is then: 0.328 - 2(0.074) to 0.328 + 2(0.074)

<=> 0.18 to 0.476

There is no golden rule of interpretation of κ , however, the following guidelines may be helpful:

<0.20:	Poor agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Good agreement
0.81 - 1.00	Very good agreement

IX. DIFFERENCES BETWEEN INDICES

Example 19: Consider the following data which represent the number diversity of food items in the diet of cities A and B. The aim is to test whether the diets in these city are different.

Diet Item	A (f_1)	B (f_2)
Oak	47	48
Corn	35	13
Black berry	7	8
Beech	5	.
Cherry	3	
Pine	.	23
Grape	.	11
Other	2	2
Total	99	105

The Shannon's index of diversity (H) and its variance for city A ($n_1 = 99$) is:

$$H_1 = \frac{n \log n - \sum f_i \log(f_i)}{n_1} = 0.5403$$

and
$$s_1^2 = \frac{\sum f_i (\log f_i)^2 - \left(\sum f_i \log f_i \right)^2 / n}{n^2} = 0.001376$$

For City B, the respective statistic is:

$$H_1 = 0.6328$$

and $s_2^2 = 0.000969.$

The standard deviation of difference between two indices is:

$$SD(H_1 - H_2) = \sqrt{s_1^2 + s_2^2}$$
$$= \sqrt{0.001376 + 0.000969}$$
$$= 0.0484$$

and the standardised distance is:

$$t = \frac{H_1 - H_2}{SD(H_1 - H_2)} = 1.91.$$

with $df = \dfrac{\left(s_1^2 + s_2^2\right)^2}{\dfrac{\left(s_1^2\right)^2}{n_1} + \dfrac{\left(s_2^2\right)^2}{n_2}} = 196$, the expected t value is 1.972.

Since the observed t value is less than its expected value, we conclude that the two diversity indices are not statistically different. //

X. SOME COMMENTS AND REFLECTION

10.1. INTERPRETATION OF P VALUES.

The following comments are extracted from D.G Altman's publications.

P value abounds in medical research papers, so it is essential to understand precisely what they mean, and also what they do not mean. The P value is the **probability of having observed our data (or more extreme data) when the null hypothesis is true.** For example, in a clinical study this statement refers to the observed difference between the treatment groups. We are therefore relating our data to the likely variation in a sample due to chance when the null hypothesis is true in the population.

We have seen that samples give results that differ from what is true in the population, and that the variability among samples decreases as the sample size increases. It was seen in previous discussion that these facts are taken into account when test statistics, and hence P values, are calculated.

The interpretation of a P value is problematic. If we carry out a clinical trial to compare two treatments and get a "large" value of P, say greater than 0.2, then we can say that data such as ours could occur often when the null hypothesis is true - that is, the two treatments are equally effective. Conversely, if P value is very small, say less than 0.001, then the null hypothesis appears implausible because our data could hardly ever arise purely by chance when the null hypothesis is true. We can therefore feel confident that the null hypothesis is **not** true and one treatment is better than the other. Between these two extremes lies a grey area, but conventionally a cut-off is chosen and if P is smaller than the cut-off value, the null hypothesis is rejected. The test of the null hypothesis is therefore whether or not P lies below the chosen cut-off point.

Although the choice of the cut-off is arbitrary, in practice in most cases we use 0.05. In other words, an outcome that could occur less than one time in 20 when the null hypothesis is true would lead to the rejection of the null hypothesis. In this formulation, when we reject the null hypothesis we accept a complementary

alternative hypothesis, which in the clinical trial example, is that the two treatments are not equally effective. If the P value exceeds the critical point we do not reject the null hypothesis. However, we can not say that we believe the null hypothesis is true, but only that there is not enough evidence to reject it. This is a subtle but important distinction.

A common misinterpretation of the P value is that it is the probability of the data having arisen by chance, or equivalently, that P value is the probability that the observed effect is not a real one. The distinction between this incorrect definition and the true definition given earlier is the absence of the phrase **when the null hypothesis is true.** The omission leads to the incorrect belief that it is impossible to evaluate the probability of the observed effect being a real one. The observed effect in the sample is genuine, but we do not know what is true in the population. All we can do is with this approach to statistical analysis is to calculate the probability of observing our data (or more unlikely data) when the null hypothesis is true.

10.2. TYPE I AND TYPE II ERRORS AGAIN

The use of cut-off for P leads to treating the analysis as a process for making a decision. Within this framework, it is customary (but unwise) to consider that a statistically significant effect is a real one, and conversely that a non-significant result indicates that there is no effect. Forcing a choice between significant and non-significant obscures the uncertainty present whenever we draw inferences from a sample. When we construct a confidence interval the uncertainty is shown explicitly, but with a hypothesis test, it is implicit, and may easily be overlooked.

Two possible errors can be made when using P value to make a decision. Firstly, we can obtain a significant result, and thus reject the null hypothesis, when the null hypothesis is in fact true. This is called a type I error, and may be thought of as "false positive" result. Alternatively, we may obtain a non-significant result when the null hypothesis is not true, in which case, we make a type II error, which can be thought of as a "false negative" result. The probability of type I and type II errors are sometimes called the alpha (α) and beta (β), respectively.

The value of alpha is determined in advance, usually at 5%. The value of beta depends upon the size effect that one is interested in, and also the sample size. More often we talk about the power of a study to detect an effect of a specified size, where the power is 1-beta. A wide confidence interval is an indication of low power. We will return to this aspect of sample size in a later topic in this course.

10.3. ONE SIDED OR TWO-SIDED P VALUE: REVISITED?

Extreme results can occur by chance equally often in either direction, which we allow for by calculating a two-sided P value. In the vast majority of cases this is the correct procedure. In rare cases, it is reasonable to consider that a real difference in the opposite direction must be due to chance. Here the alternative hypothesis is restricted to an effect in one direction only, and it is reasonable to calculated one-sided P value by considering only one tail of the distribution of the test statistic. For a test statistic with the Normal distribution the usual two-sided 5% cut-off point is 1.96, whereas a one-sided 5% cut-off is 1.645. The difference is not particularly large but can lead to a different interpretation in relation to fixed levels of statistical significance.

One-sided tests are rarely appropriate. Even when we have strong prior expectations, for example that a new treatment can not be worse than an old one, we can not be sure that we are right. If we could be sure we would not need to do an experiment! If it is felt that a one-sided test really is appropriate, then this decision must be made **before the data are analysed**, it must not depend on what the results were.

XI. APPENDIX

Value of K for finding approximate 95% CI for differences in population medians of two unpaired samples with sample sizes n and m from 5 to 20.

	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	3	4	6	7	8	9	10	12	13	14	15	16	18	19	20	21
6	4	6	7	9	11	12	14	15	17	18	20	22	23	25	26	28
7	6	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35
8	7	9	11	14	16	18	20	23	25	27	30	32	35	37	39	42
9	8	11	13	16	18	21	24	27	29	32	35	38	40	43	46	49
10	9	12	15	18	21	24	27	30	34	37	40	43	46	49	53	56
11	10	14	17	20	24	27	31	34	38	41	45	48	52	56	59	63
12	12	15	19	23	27	30	34	38	42	46	50	54	58	62	66	70
13	13	17	21	25	29	34	38	42	46	51	55	60	64	68	73	77
14	14	18	23	27	32	37	41	46	51	56	60	65	70	75	79	84
15	15	20	25	30	35	40	45	50	55	60	65	71	76	81	86	91
16	16	22	27	32	38	43	48	54	60	65	71	76	82	87	93	99
17	18	23	29	35	40	46	52	58	64	70	76	82	88	94	100	106
18	19	25	31	37	43	49	56	62	68	75	81	87	94	100	107	113
19	20	26	33	39	46	53	59	66	73	79	86	93	100	107	114	120
20	21	28	35	42	49	56	63	70	77	84	91	99	106	113	120	128

XII. EXERCISES

1. In a study of the annual change in whole body bone mass in Black and White American children, Nelson et al. (JBMR 1994) reported the following baseline results:

	Black (n=226)	White (n=137)
Age	8.9 (0.64)	9.0 (0.58)
BMC	966.2 (265)	875.5 (228)

Is the difference in BMC significant? Construct a 95% confidence interval for the difference.

2. Since osteocalcin (OC) is non-normally distributed, therefore a ln(OC+1) transformation is necessary. The following data represent the level of ln(OC+!) in two groups of subjects classified according to their VDR genotype.

	BB	bb
n	30	38
ln(OC+1)	2.18 (0.70)	1.80 (0.69)

Perform a t-test on the data and construct a 95% difference between BB and bb subjects in *osteocalcin* (not ln(OC+!)) levels.

2. Consider the data in Example 2 (page 9) where the differences in logarithm of lysozyme between group A and group B was 0.51 and its 95% CI was -0.22 to 1.24. Convert this difference into the original unit of measurement.

3. Perform a t-test on the arcsin transformation of the data in Example 3 (page 12).

4. Perform a t-test on the square root transformation of the data in Example 4 (page 14).

5. Perform a t-test on the reciprocal transformation of the data in Example 5 (page 16).

7. The table below shows the concentrations of antibody to type III group B streptococcus (GBS) in 20 volunteers before and after immunisation.

ID	Antibody concentration	
	Before Immun.	After immun.
1	0.4	0.4
2	0.4	0.5
3	0.4	0.5
4	0.4	0.9
5	0.5	0.5
6	0.5	0.5
7	0.5	0.5
8	0.5	0.5
9	0.5	0.5
10	0.6	0.6
11	0.6	12.2
12	0.7	1.1
13	0.7	1.2
14	0.8	0.8
15	0.9	1.2
16	0.9	1.9
17	1.0	0.9
18	1.0	2.0
19	1.6	8.1
20	2.0	3.7

(a) The comparison of the antibody levels was summarised in the report of this study as "$t = 1.8$; $p>0.05$". Comment on this result.
(b) What method would be more appropriate to analyse these data? Analyse the data with the appropriate method and comment on the result.

8. The effect of benzedrine on the heart rate of dogs in beats/min was examined in an experiment on 14 dogs chosen for the study. Each dog was to serve as its own

control, with half of the dogs assigned to receive Benzedrine during the first study period and the other half assigned to receive a placebo. All dogs were examined to determine the heart rates after 2 hours on the medication. After two weeks in which medication was given, the regimens for the dogs were switch for the second study period. The dogs previously on Benzedrine were given the placebo while the others received Benzedrine. Again heart rates were measured after 2 hours.

The results are as follows:

Dog	Placebo	Benzedrine
1	259	258
2	271	285
3	243	245
4	252	250
5	266	268
6	272	278
7	293	280
8	296	305
9	301	319
10	298	308
11	310	320
12	286	293
13	306	305
14	309	313

9. Patients with chronic renal failure undergoing haemodialysis were divided into groups with low or normal plasma heparin cofactor II (HC II) levels. Five months later the acute effects of haemodialysis were divided by analysing plasma samples taken before and after haemodialysis. As dialysis increases total protein concentration in plasma, the ratio HC II to protein was calculated, with the results shown in the following table:

Group 1 (low)		**Group 2 (normal)**	
before	after	before	after
1.41	1.47	2.11	2.15

1.37	1.45	1.85	2.11
1.33	1.50	1.82	1.93
1.13	1.25	1.75	1.83
1.09	1.01	1.54	1.90
1.03	1.14	1.52	1.56
0.89	0.98	1.49	1.44
0.86	0.89	1.44	1.43
0.75	0.95	1.38	1.28
0.75	0.83	1.30	1.30
0.70	0.75	1.20	1.21
0.69	0.71	1.19	1.30

The authors of this study analysed by separate paired Wilcoxon tests on the data for each group, giving $P<0.01$ for group 1 and $P > 0.05$ for group 2, then concluded "HC II activity increases in group 1 but not in group 2". This statement was wrong, why ? Carry out a better analysis of the data.

10. Sixty-five pregnant women at a high risk of pregnancy-induced hypertension participated in a randomised controlled trial comparing 100 mg of aspirin daily and matching placebo during the third trimester of pregnancy. The observed rates of hypertension are shown in the following table:

	Apirin	Placebo	Total
Hypertension	4	11	15
No hypertension	30	20	50
Total	34	31	65

Do these data suggest that daily aspirin reduces the risk of hypertension in the last trimester of pregnancy?

11. There may be a remedy for baldness - at least that is what million of men hope, if the FDA approves Upjohn's minoxidil for such a use. Minoxidil was investigated in a large, 27-centre study where patients were randomly assigned to receive topical

minoxidil or an identical placebo. Ignoring the centre-to-centre variation, the preliminary results are as follows:

	Sample	% with new hair growth
Minoxidil	310	32
Placebo	309	20

Are the difference statistically significant ?

12. A study was carried out to see if patients whose skin did not respond to dinitrochlorobenzene (DNCB), a contact allergen, would show an equally negative response to croton oil, a skin irritant. The following table shows the results of simultaneous skin reaction tests to DNCB and croton oil in 173 patients with skin cancer:

	DNCB		
	+ve	-ve	Total
Croton oil			
+ve	81	48	129
-ve	23	21	44
total	104	69	173

(a) The authors reported "no correlation" between the two tests. Carry out an analysis appropriate to the clinical question posed.
(b) The results of DNCB test were compared for patients with different stages of cancer, as shown in the following table:

	Stage of skin cancer			
	I	II	III	Total
DNCB reaction				

+ve	39	39	26	104
-ve	13	19	37	69
total	52	58	63	173

Is DNCB reactivity to stage of cancer in these patients ?

13. A study was made of 65 patients who had received or were receiving sodium aurothiomalate as a treatment for rheumatoid arthritis. The aim was to examine the possibility that toxicity to sodium aurothiomalate (SA) might be linked to sulphoxidation capacity, as assessed by the sulphoxidation index (SI). Value of SI>6 were taken as indicating impaired SA. The data were given as follows:

	Major adverse reaction (toxicity)		
	Yes	No	Total
Impaired sulphoxidation			
Yes	30	9	39
No	7	19	26
Total	37	28	65

The authors wrote: "The incidence of impaired sulphoxidation in patients showing SA toxicity (30/37); 81.0%) was significantly greater than in the group without adverse reaction (9/28; 32%) (Chi square test = 27.6; $p<0.001$). Similarly, the incidence of toxicity was significantly increased in those with impaired sulphoxidation (30/39; 76.9%) compared to those with extensive sulphoxidation (7/26; 26.9%) (Chi square test = 36.2; $p<0.001$)".

(a) Why can't both of the above Chi squared tests be correct ?
(b) Carry out a Chi squared test of the data in the table and compare your answer with the two results in the above paragraphs.

14. A case-control study was carried out to investigate the aetiology of acoustic neuromas. Men aged 25-29 at the time of diagnosis who were residents in Los Angeles County were eligible for inclusion. A total of 118 men were identified who

were alive and able to be interviewed. Twenty eight patients were not interviewed because the physician refused permission (12), the patient chose not to participate (9), or the patients could not be located (7). For 86 of the remaining patients the investigators identified and interviewed a neighbourhood control of the same race and within five years of age.

Both members of each case-control pair were interviewed in the same manner by the same interviewer to obtain information about various life experiences. Exposure to loud noise at work was of particular interest. Overall, 58 cases and 46 controls had had some exposure to loud noise at work. There were 20 case-control pairs for which the case but not the control had had such exposure, and 8 pairs where the control but not the case had had some exposure.

(a) Carry out an appropriate analysis to compare the proportion exposed cases and controls;
(b) Calculate the odds ratio for acoustic neuroma associated with exposure to loud noise at work.

15. In a clinical trial in which a total 100 patients are allocated to two treatment groups by simple randomisation. Show that the probability that the difference between the numbers of patients in the two treatment groups exceeds 20 is about 5%.

16. If two studies' results yield $P<0.001$ and $P=0.02$
(a) Is it true that the former has found a stronger effect than the latte ?
(b) If the two studies are identical with the above P values, what are the possible explanation for the large difference.

17. A controlled trial was performed to compare the corticosteroid prednisole and placebo in patients with chronic active hepatitis positive for hepatitis B surface antigen. In response to a letter criticising the analysis the author wrote: "The one-sided test was used in the calculations, since in a previous analysis major complications were encountered significantly more frequently in the steroid-treated group" Is this a valid justification for one-sided test? If not, why not?