# Statistics in Clinical Trials

## Course Booklet

**Dr Nick Fieller**

**Probability & Statistics, SoMaS**

**University of Sheffield**

***visiting***

UNIVERSITY
OF TAMPERE

2012

# Contents

# Statistical Methods in Clinical Trials

## 0. Introduction

### 0.1 Books

Altman, D.G. (1991) *Practical Statistics for Medical Research*. Chapman & Hall

Andersen, B. (1990) *Methodological Errors in Medical Research.* Blackwell

Armitage, P., Berry, G. & Matthews, J.N.S. (2002) *Statistical Methods in Medical Research (4th Ed.).* Blackwell.

Bland, Martin (2000) *An Introduction to Medical Statistics (3rd Ed).* OUP.

Campbell, M. J. & Swainscow, T. D. V. (2009) *Statistics at Square One (11th Ed).* Wiley-Blackwell

**Campbell, M. J. (2006) *Statistics at Square Two (2nd Ed).* Wiley-Blackwell**

† Julious, S. A. (2010) *Sample Sizes for Clinical Trials, CRC Press.*

Kirkwood, B. R. & Stone, J.A.C. (2003) *Medical Statistics (2nd Ed).* Blackwell

Campbell, M. J., Machin, D. & Walters, S. (2007) *Medical Statistics: a textbook for the health sciences. (4th Ed.)* Wiley

Machin, D. & Campbell, M. J. (1997) *Statistical Tables for the Design of Clinical Trials. (2nd Ed.)* Wiley

**Matthews, J. N. S. (2006), *An Introduction to Randomized Controlled Clinical Trials. (2nd Ed.)* Chapman & Hall**

Pocock, S. J. (1983) *Clinical Trials, A Practical Approach.* Wiley

Schumacher, Martin & Schulgen, Gabi (2002) *Methodik Klinischer Studien.* Springer. (In German)

Senn, Stephen (2002) *Cross-over Trials in Clinical Research.* Wiley

★ Senn, Stephen (2003) *Dicing with Death: Chance, Risk & Health.* CUP

The two texts which are highlighted cover most of the Clinical Trials section of the Medical Statistics module; the first also has material relevant to the Survival Data Analysis section.

† Indicates a book which goes considerably further than is required for this course (Chapter 5) but is also highly relevant for those taking the second semester course MAS6062 Further Clinical Trials.

★ Indicates a book which contains much material that is relevant to this course but it is primarily a book ***about*** Medical Statistics and is strongly recommended to those planning to go for interviews for jobs in the biomedical areas (including the pharmaceutical industry)

### 0.2 Objectives

The objective of this course is to provide an introduction to some of the statistical methods and statistical issues that arise in medical experiments which involve, in particular, human patients. Such experiments are known collectively as ***clinical trials***.

Many of the statistical techniques used in analyzing data from such experiments are widely used in many other areas (e.g. $\chi^2$-tests in contingency tables, t-tests, analysis of variance). Others which arise particularly in medical data and which are mentioned in this course are McNemar's test, the Mantel-Haenszel test, logistic regression and the analysis of crossover trails.

As well as techniques of statistical analysis, the course considers some other issues which arise in medical statistics — questions of ethics and of the design of clinical trials.

## 0.3 Organization of course material

The notes in the main Chapters 1 – 7 are largely covered in the two highlighted books in the list of recommended texts above and are supplemented by various examples and illustrations. These range from simple 'quick problems' to more substantial exercises*.* These task sheets are designed for you to test your own understanding of the course material. If you are not able to complete the simpler problems then you should go back to the lecture notes (and other course material) and re-read the relevant section (and if necessary re-read again & …). Solutions will be provided to these on the course web pages in due course.

Lectures will consist of introducing the material covered in these notes, filling in details of items such as **R** implementation (including specific commands and menu choices), demonstrating computer analyses and going through key parts of the various example sheets. The lectures will be based on PowerPoint presentations and copies of the slides will be made available on the course webpage at:

http://www.nickfieller.staff.shef.ac.uk/tampere12/index.html

These will be placed there at some time after the lecture. Any typing (or other) mistakes in the notes or the exercises and solutions that are brought to my attention will be noted and corrected in a ***Corrections and Clarifications*** section on this page.

The lectures will not necessarily follow precisely what appears in the notes. In some places the lecture slides will follow the notes very closely, repeating some of the examples. In other places the slides will present material in a slightly different way and possible different order and there will be examples and further details in the lectures than are covered in these notes.

## 0.4 A Note on R, S-PLUS and MINITAB

The main statistical package for this course is **R.** It is very similar to the copyright package S-PLUS and the command line commands of S-PLUS are [almost] interchangeable with those of **R.** Unlike S-PLUS, **R** has only a very limited menu system which covers some operational aspect but no statistical analyses. A brief guide to getting started in **R** is available from the course homepage.

R is a freely available programme which can be downloaded over the web from http://cran.r-project.org/ or any of the mirror sites linked from there for installation on your own machine. It is available on University networks. **R** and S-PLUS are almost identical except that **R** can only be operated from the command line apart from operational aspects such as loading libraries and opening files. Almost all commands and functions used in one package will work in the other. However, there are some differences between them. In particular, there are some options and parameters available in **R** functions which are not available in S-PLUS. Both S-PLUS and **R** have excellent help systems and a quick check with help(*function*) will pinpoint any differences that are causing difficulties. A key advantage of **R** over S-PLUS is the large number of libraries contributed by users to perform many sophisticated analyses.

These are updated very frequently and extend the capabilities substantially. If you are considering using the techniques outside this course (e.g. for some other substantial project) then you would be well advised to use **R** in preference to S-PLUS. Command-line codes for the more substantial analyses given in the notes for this course have been tested in **R.** In general, they will work in S-PLUS as well but there could be some minor difficulties which are easily resolved using the help system.

MINITAB is package with a very flexible menu system and a full command-line facility. Some examples of MINITAB code and output are given in the notes since some of those taking the course are already familiar with the package.

## 0.5 Data sets

Data sets used in this course are available on the course web pages in R format. Other formats for some data sets may be available by direct request

### 0.5.1 R data sets

Those in **R** are given first and they have extensions .Rdata; to use them it is necessary to copy them to your own hard disk. This is done by using a web browser to navigate to the course web, clicking with the right-hand button and selecting ‘save target as…’ or similar which opens a dialog box for you to specify which folder to save them to. Keeping the default .Rdata extension is recommended and then if you use Windows explorer to locate the file a double click on it will open **R** with the data set loaded and it will change the working directory to the folder where the file is located. For convenience all the **R** data sets for Medical Statistics are also given in a WinZip file.

**NOTE: It may not be possible to use a web browser to locate the data set on a web server and then open R by double clicking.** The reason is that you only have read access rights to the web page and since **R** changes the working directory to the folder containing the data set write access is required.

### 0.5.2 Data sets in other formats

Most of the data sets are available in other formats (Minitab, SPSS etc) on request.

## 0.6 R libraries required

Most of the statistical analyses described in this booklet use standard functions but some may require use of the `MASS` package and others. This will be indicated on each occasion.

The `MASS` library is installed with the base system of **R** but you may need to install other packages before first usage.

## 0.6 Outline of Course

1. Background:– historical development of statistics in medical experiments. Basic definitions of placebo effect, blindness and phases of clinical trial.

2. Basic trial analysis:– ‘parallel group’ and ‘in series’ designs, factorial designs & sequential designs.

3. Randomization:– simple and restricted, stratified, objectives of randomization.

4. Size of trial:– sample sizes needed to detect clinically relevant differences with specified power.

5. Multiplicity and interim analyses:– multiple significance testing and subgroup analysis, Bonferroni corrections.

6. Crossover trials:– estimation and testing for treatment, period and carryover effects.

7. Binary responses:– matched pairs and McNemar’s test, logistic regression.

# 1. Background and Basic Concepts

## 1.1 Definition of Clinical Trial (from Pocock, 1983)

***Any form of planned experiment which involves patients and is designed to elucidate the most appropriate treatment of future patients under a given medical condition***

Notes:

(i) Planned experiment (not observational study)

(ii) Inferential Procedure — want to use results on limited sample of patients to find out best treatment in the general population of patients who will require treatment in the future.

## 1.2 Historical Background

(see e.g. Pocock Ch. 2, Matthews Ch. 1)

**1537**: Treatment of battle wounds:

Treatment A: Boiling Oil [standard]

Treatment B: Egg yolk + Turpentine + Oil of Roses [new]

New treatment found to be better

**1741**: Treatment of Scurvy, HMS Edinburgh:

Two patients allocated to each of (1) cider; (2) *elixi vitriol*; (3) vinegar; (4) nutmeg, (5) sea water; (6) oranges & lemons

(6) produced "the most sudden and visible good effects."

Prior to 1950s medicine developed in a haphazard way. Medical literature emphasized individual case studies and treatment was copied:— unscientific & inefficient.

Some advances were made (chiefly in communicable diseases) perhaps because the improvements could not be masked by poor procedure.

Incorporation of statistical techniques is more recent.

e.g. MRC (*Medical Research Council in the UK*) Streptomycin trial for Tuberculosis (**1948**) was first to use a ***randomized control***.

MRC cancer trials (with statistician Austin Bradford-Hill) first recognizably modern sequence — laid down the [now] standard procedure.

## 1.3 Field Trial of Salk Polio Vaccine

In 1954 1.8 million young children in the U.S. were in a trial to assess the effectiveness of Salk vaccine in preventing paralysis/death from polio (which affected 1 in 2000).

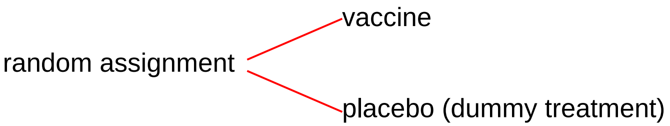
Certain areas of the U.S., Canada and Finland were chosen and the vaccine offered to all 2nd grade children. Untreated 1st and 3rd grade children used as the control group, a total of 1 million in all.

Difficulties in this ‘observed control’ approach were anticipated:

(a) only volunteers could be used – these tended to be from wealthier/better educated background (i.e. *volunteer bias*)

(b) doctors knew which children had received the vaccine and this could (subconsciously) have influenced their more difficult diagnoses (i.e. a problem of *lack of blindness*)

Hence a further 0.8 million took part in a randomised double-blind trial simultaneously. Every child received an injection but half these did not contain vaccine:


random assignment

vaccine

placebo (dummy treatment)

and child/parent/evaluating physician did not know which.

## Results of Field Trial of Salk Polio Vaccine

| Study group | Number in group | Number of cases | Rate per 100 000 |
|---|---|---|---|
| ***Observed control*** | | | |
| **Vaccinated 2nd grade** | 221 998 | 38 | 17 |
| **Control 1st and 3rd grade** | 725 173 | 330 | 46 |
| **Unvaccinated 2nd grade** | 123 605 | 43 | 35 |
| ***Randomized control*** | | | |
| **Vaccinated** | 200 745 | 33 | 16 |
| **Control** | 210 229 | 115 | 57 |
| **Not inoculated** | 338 778 | 121 | 36 |

Results from second part conclusive:

(a) incidence in vaccine group reduced by 50%

(b) paralysis from those getting polio 70% less

(c) no deaths in vaccine group (compared with 4 in placebo group)

Results from first part less so – it was noticed that those 2nd grade children NOT agreeing to vaccination had lower incidence than non-vaccinated controls. It could be that:

(a) those 2nd grade children having vaccine are a self-selected high risk group

or

(b) that there is a complex age effect

Whatever the cause, a valid comparison (treated versus control) was difficult. This provides an example of ***volunteer bias.***

Thus, this study was [by accident] a comparison between a randomized controlled double-blind clinical trial and a non-randomized open trial. It revealed the superiority of randomised trials which are now regarded as essential to the definitive comparison and evaluation of medical treatments, just as they had been in other contexts (e.g. agricultural trials) since ~1900.

## 1.4 Types of Trial

Typically a new treatment develops through a research programme (at a pharmaceutical company) who test MANY different manufactured/synthesized compounds. Approximately 1 in 10,000 of those synthesized get to a clinical trial stage (initial pre-clinical screening through chemical analysis, preliminary animal testing etc.). Of these, 1 in 5 reaches marketing.

The 4 stages of a [clinical] trial programme after the pre-clinical are:–

Phase I trials: Clinical pharmacology & toxicity concerned with drug safety — not efficacy (i.e. not with whether it is effective). Performed on non-patients or volunteers. Aim to find range of safe and effective doses. investigate metabolism of drugs.

n=10 – 50

Phase II trials: Initial clinical investigation for treatment effect. Concerned with safety & efficacy for patients. Find maximum effective and tolerated doses. Develop model for metabolism of drug in time.

n= 50 –100

Phase III trials: Full-scale evaluation of treatment comparison of drug versus control/standard in (large) trial:

n= 100 – 1000

Phase IV trials: Post-marketing surveillance: long-term studies of side effects, morbidity & mortality.

n= as many as possible

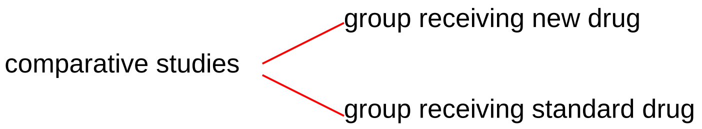
### 1.4.1 Further notes:

Phase I: First objective is to determine an acceptable single drug dosage, i.e. how much drug can be given without causing serious side effects — such information is often obtained from dosage experiments where a volunteer is given increasing doses of the drug rather than a pre-determined schedule.

Phase II: Small scale and require detailed monitoring of each patient.

Phase III: After a drug has been shown to have some reasonable effect it is necessary to show that it is better than the current standard treatment for the same condition in a large trial involving a substantial number of patients. (‘Standard’: drug already on market, want new drug to be at least equally as good so as to get a share of the market). Generally these will be ***superiority trials*** designed to test whether the new treatment is superior to another. However, there are other possibilities: ***non-inferiority trials*** (to test whether a new treatment is no worse [within a specified margin] than another and ***bioequivalence trials*** (to test whether a new treatment is equivalent in effect [within specified margins] to another). This course is concerned with ***superiority trials*** only.

**Note**: Almost all [Phase III] trials now are randomized controlled (comparative) studies:



comparative studies

group receiving new drug

group receiving standard drug

To avoid <u>bias</u> (subconscious or otherwise), patients must be assigned at <u>random</u>.

(Bias:– May give very ill people the new drug since there is no chance of standard drug working or perhaps because there is more chance of them showing greater improvement, e.g. blood pressure — those with the highest blood pressure levels can show a greater change than those with moderately high levels).

The comparative effect is important. If we do not have a control group and simply give a new treatment to patients, we cannot say whether any improvement is due to the drug or just to the act of being treated (i.e. *the placebo effect*). Historical controls (i.e. look for records from past years of people with similar condition when they came for treatment) suffer from similar problems since medical care by doctors and nurses improves generally.

In an early study of the validity of controlled and uncontrolled trials, Foulds (1958) examined reports of psychiatric clinical trials:

- in 52 ***uncontrolled*** trials, treatment was declared 'successful' in in 43 cases (83%)

- in 20 ***controlled*** trials, treatment was 'successful' in only 5 cases (25%)

This is **SUSPICIOUS**.

Beware also of ***publication bias***:– only publish 'results' that say new drug is better, when other studies disagree. Also concern from conflicts of interest — see §1.8 Publication Ethics.

There is also concern that pressures of publication will influence what is published from a trial, e.g. what is declared to be the ***primary response*** may show little effect and instead some different measure is chosen, evidence suggests this happens in about a third of published trials [Mathieu, S. et al., 2009. *Comparison of Registered and Published Primary Outcomes in Randomized Controlled* Trials. JAMA, 302(9), 977-984] and indeed fewer than half of registered trials recorded what the primary outcome was.

The sources of information for this study are ***registers of clinical trials*** such as http://clinicaltrials.gov/ and **PubMed** (http://www.ncbi.nlm.nih.gov/sites/entrez/). However, not all trials are registered though In 2005, the International Committee of Medical Journal Editors announced they would only publish trials that had been registered. Nevertheless Only 20% of **all** cancer trials are published http://theoncologist.alphamedpress.org/cgi/content/abstract/theoncologist.2008-0133v1 (The Oncologist, 15 September 2008) and only 6% of cancer trials run by *commercial industry* are published

## 1.5 Placebo Effect

One type of control is a placebo or dummy treatment. This is necessary to counter the *placebo effect* — the psychological benefit of being given any treatment/attention at all (used in a comparative study)

## 1.6 Blindness of trials

Using placebos allows the opportunity to make a trial *double blind* — i.e. neither the patient nor the doctor knows which treatment was received. This avoids bias from patient or evaluator attitudes.

*Single blind* — <u>either</u> patient <u>or</u> evaluator blind

In organizing such a trial there is a coded list which records each patient's treatment. This is held by a co-ordinator & only broken at analysis (or in emergency).

Clearly, blind trials are *only sometimes possible*; e.g. cannot compare a drug treatment with a surgical treatment.

## 1.7 Ethical Considerations

Specified in Declaration of Helsinki (1964+ammendments) consisting of 32 paragraphs, see http://www.wma.net/e/policy/b3.htm.

Ethical considerations can be different from what the statistician would like.

e.g. some doctors do not like placebos — they see it as preventing a possibly beneficial treatment. (¿How can you give somebody a treatment that you know will not work?). Paragraph 29 and the 2002 Note of Clarification concerns use of placebo-controlled trials.

There is competition between individual and collective ethics — what may be good for a single individual may not be good for the whole population.

It is agreed that it is **unethical** to conduct research which is badly planned or executed. We should only put patients in a trial to compare treatment A with treatment B if we are genuinely unsure whether A or B is better.

An important feature is that patients must give their consent to be entered (at least generally) and more than this, they must give informed consent (i.e. they should know what the consequences are of taking the possible treatments).

In the UK, local ethics committees monitor and 'licence' all clinical trials — e.g. in each hospital or in each city or regional area.

It is also unethical to perform a trial which has little prospect of reaching any conclusion, e.g. because of insufficient numbers of subjects — see later — or some other aspect of poor design.

It may also be unethical to perform a trial which has many more subjects than are needed to reach a conclusion, e.g in a comparative trial if one treatment proves to be far superior then too many may have received the inferior one.

## 1.8 Publication Ethics

See BMJ Vol 323, p588, 15/09/01. (http://www.bmj.com/)

Editorial published in all journals that are members of the International Committee of Medical Journal Editors (BMJ, Lancet, New England Journal of Medicne, … ).

Concern at articles where declared authors have

- not participated in design of study
- had no access to raw data
- little role in interpretation of data
- not had ultimate control over whether study is published

Instead, the sponsors of the study (e.g. pharmaceutical company) have designed, analysed and interpreted the study (and then decided to publish).

A survey of 3300 academics in 50 universities revealed 20% had had publication delayed by at least 6 months at least once in the past 3 years because of pressure from the sponsors of their study.

Contributors must now sign to declare:

- full responsibility for conduct of study
- had access to data
- controlled decision to publish

## 1.9 Evidence-Based Medicine

This course is concerned with ‘Evidence-Based Medicine (EBM) or more widely ‘Evidence-Based Health Care’. The essence of EBM is that we should consider critically all evidence that a drug is effective or that a particular course of treatment improves some relevant measure of well-being or that some environmental factor causes some condition. Unlike abstract areas of mathematics it is never possible to prove that a drug is effective, it is only possible to assess the strength of the evidence that it is. In this framework statistical methodology has a role but not an exclusive one. A formal test of a hypothesis that a drug has no effect can assess the strength of the evidence against this null hypothesis but it will never be able to prove that it has no effect, nor that it is effective. The statistical test can only add to the overall evidence.

### 1.9.1 The Bradford-Hill Criteria

To help answer the specific question of causality Austen Bradford-Hill (1965) formulated a set of criteria that could be used to assess whether a particular agent (e.g. a medication or drug or treatment regime or exposure to an environmental factor) caused or influenced a particular outcome (e.g. cure of disease, reduction in pain, medical condition)

These are:–

- Temporality (effect follows cause)
- Consistency (does it happen in different groups of people – both men and women, different countries)
- Coherence (do different types of study result in similar conclusions – controlled trials and observational studies)
- Strength of association (the greater the effect compared with those not exposed to the agent the more plausible is the association)
- Biological gradient (the stronger the agent the greater the effect – does response follow dose)
- Specificity (does agent specifically affect something directly connected with the agent)
- Plausibility (is there a possible biological mechanism that could explain the effect)
- Freedom from bias or confounding factors (a confounding factor is something related to both the agent and the outcome but is not in itself a cause)
- Analogous results found elsewhere (do similar agents have similar results)

These 9 criteria are of course inter-related. Bradford-Hill comments "none of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be regarded as a *sine qua non*', that is establishing every one of these does not prove cause and effect nor does failure to establish any of them mean that the hypothesis of cause and effect is completely untrue. However, satisfying most of them does add considerably to the evidence.

## 1.10 Summary & Conclusions

- Clinical trials involve human patients and are ***planned*** experiments from which wider ***inferences*** are to be drawn

- Randomized controlled trials are the only effective type of clinical trial

- Clinical Trials can be categorized into 4 phases

- Double or single blind trials are preferable where possible to reduce bias

- Placebo effects can be assessed by controls with placebo or dummy treatments where feasible.

- Ethical considerations are part of the statisticians responsibility

# 2. Basic Trial Analysis

## 2.1 Comments on Tests

Before considering some basic experimental designs used commonly in the analysis of Clinical Trials there are two comments on statistical tests. The first is on the general question of whether to use a one- or two-sided tests, the other is when considering use of a t-test whether to use the separate or pooled version and what about testing for equality of variance first?

### 2.1.1 One-sided and two-sided tests

Tests are usually two-sided unless there are very good prior reasons, not observation or data based, for making the test one-sided. If in doubt, then use a two-sided test.

This is particularly contentious amongst some clinicians who say:–

> "I ***know*** this drug can *only possibly* lower mean systolic blood pressure so I must use a one-sided test of $H_0$: $\mu = \mu_0$ *vs* $H_A$: $\mu < \mu_0$ to test whether this drug works."

The temptation to use a one-sided test is that it is more powerful for a given significance level (i.e. you are more likely to obtain a significant result, i.e. more likely to 'shew' your drug works). The reason why you should not is because if the drug actually *increased* mean systolic blood pressure but you had declared you were using a one-sided test for lower alternatives then the rules of the game would declare that you should ignore this evidence and so fail to detect that the drug is in fact deleterious.

One pragmatic reason for always using two-sided tests is that all good editors of medical journals would almost certainly refuse to publish articles based on use of one-sided tests, (or at the very least question their use and want to be assured that the use of one-sided tests had been declared in the protocol [see §4] in advance (with certified documentary evidence).

A more difficult example is suppose there is suspicion that a supplier is adulterating milk with water. The freezing temperature of watered-down milk is lower than that of whole milk. If you test the suspicions by measuring the freezing temperatures of several samples of the milk, should a one- or two-sided test be used? To answer the very specific question of whether the milk is being adulterated by water you should use a one-sided test but what if in fact the supplier is adding cream?

In passing, it might be noted that the issue of one-sided and two-sided tests only arises in tests relating to one or two parameters in only one dimension. With more than one dimension (or hypotheses relating to more than two parameters) there is no parallel of one-sided alternative hypotheses. This illustrates the rather artificial nature of one-sided tests in general.

Situations where a one-sided test is definitely called for are uncommon but one example is in a case of say two drugs A (the current standard and very expensive) and B (a new generic drug which is much cheaper). Then there might be a proposal that the new cheaper drug should be introduced unless there is evidence that it is very much worse than the standard. In this case the model might have the mean response to the two drugs as $\mu_A = \mu_B$ and if low values are 'bad', high values 'good' then one might test $H_0$: $\mu_A = \mu_B$ against the one-sided alternative $H_A$: $\mu_A > \mu_B$ and drug

B is introduced if ***$H_0$ is not rejected***. The reason here is that you want to avoid introducing the new drug if there is even weak evidence that it is worse but if it is indeed preferable then so much the better, you are using as powerful a test as you can (i.e. one-sided rather than the weaker two-sided version). However, this example does raise further issues such as how big a sample should you use and so on. The difficulty here is that you will proceed provided there is absence of evidence saying that you should not do so. A better way of assessing the drug would be to say that you will introduce drug B only if you can shew that it is no more than K units worse than drug A. So you would test $H_0$: $\mu_A - K = \mu_B$ against $H_A$: $\mu_A - K < \mu_B$ and only proceed with the introduction of B if ***$H_0$ is rejected in favour of the one-sided alternative*** (of course you need good medical knowledge to determine a sensible value of K). This leads into the area of *non-inferiority trials* and *bioequivalence* studies which are beyond the scope of this course but will be considered in the second semester course MAS6062 Further Clinical Trials.

### 2.1.2 Separate and Pooled Variance t-tests

This is a quick reminder of some issues relating to two-sample t-tests. The test statistic is the difference in sample means scaled by an estimate of the standard deviation of that difference. There are two plausible ways of estimating the variance of that difference. The first is by estimating the variance of each sample separately and then combining the two *separate estimates.* The other is to pool all the data from the two samples and estimate a common variance (allowing for the potential difference in means). The standard deviation used in the test statistic is then the square

root of this estimate of variance. To be specific, if we have groups of sizes $n_1$ and $n_2$, means $\bar{x}_1$ & $\bar{x}_2$ and sample variances $s_1^2$ & $s_2^2$ of the two samples then the two versions of a 2-sample t-test are:

(i) separate variance: $t_r = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, where the degrees of freedom r is safely taken as min{$n_1$,$n_2$} though S-PLUS, MINITAB and SPSS use a more complicated formula (the Welch approximation) which results in fractional degrees of freedom. This is the default version in **R** (with function `t.test()` and MINITAB but not in many other packages such as S-PLUS.

(ii) pooled variance: $t_r = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

where r = ($n_1$+$n_2$ – 2).

This version assumes that the variances of the two samples are equal (though this is difficult to test with small amounts of data). This is the default version in S-PLUS.

We will primarily use the first version because if the underlying populations variances are indeed the same then the separate variance estimate is a good [unbiased] estimate of the common variance and the null distribution of the separate variance estimate test statistic is a t-distribution with only slightly more degrees of freedom than given by the Welch approximation in the statistical packages so resulting in a test that is very slightly conservative and very slightly less powerful. However, if you use the pooled variance estimate when the underlying population variances are unequal then the resulting test statistic has a null distribution that can be a long way from a t-distribution on ($n_1$+$n_2$–2) degrees of freedom and so potentially produce wrong results (neither generally conservative nor liberal, neither generally more nor less powerful, just incorrect). Thus it makes sense to use the separate variance estimate routinely unless there are very good reasons to do otherwise. One such exceptional case is in the calculation of sample sizes [see §5.3] where a pooled variance is used entirely for pragmatic reasons and because many approximations are necessary to obtain any answer at all and this one is not so serious as other assumptions made.

The use of a separate variance based test statistic is only possible since the Welch approximation gives such an accurate estimate of the null distribution of the test statistic and this is only the case in two sample univariate tests. In two-sample multivariate tests or in all multi-sample tests (analysis of variance such as ANOVA and MANOVA) there is no available approximation and a pooled variance estimate has to be used.

### 2.1.2.1 Test equality of variances?

It is natural to consider conducting a preliminary test of equality of variances and then on the basis of the outcome of that decide whether to use a pooled or a separate variance estimate. In fact SPSS automatically gives the results of such a test (Levene's Test — a common alternative would be Bartlett's) as well as both versions of the two-sample t-test with two p-values, inviting you to choose. The arguments against using such a preliminary test are (a) tests of equality of variance are very low powered without large quantities of data — appreciate that a non-significant result does not mean that the variances truly are equal only that the evidence

for them being different is weak (b) a technical reason that if the form of the t-test is chosen on the basis of a preliminary test using the same data then allowance needs to be made for the conditioning of the t-test distribution on the preliminary test, i.e. the apparent significance level from the second test (– the t-test) is wrong because it does not allow for the result of the first (– test of equality of variance). You should **definitely not** do both tests and choose the one with the smaller p-value [*data snooping*], which is the temptation from SPSS. In practice the values of the test statistics are usually very close but the p-values differ slightly (because of using a different value for the degrees of freedom in the reference t-distribution). In cases where there is a substantial difference then the 'separate variance' version is always the correct one.

Thus the general rule is 'always use a separate variance test' noting that in S-Plus the default needs to be changed.

## 2.2 Parallel Group Designs

Compare k treatments by dividing patients at random into k groups — the $n_i$ patients in group i receive treatment i.

Group

| 1 | 2 | 3 | . . . . . . . . . . | k |
|---|---|---|---|---|
| X | X | X | . . . . . . . . . . | X |
| X | X | X | . . . . . . . . . . | X |
| • | • | • | | • |
| X | • | • | | • |
| | X | • | | X |
| | | • | | |
| | | X | | . |

Number in group:- $n_1$ $n_2$ $n_3$ . . . . . . . . . . . $n_k$ ; $\Sigma n_i$= N

EACH PATIENT RECEIVES 1 TREATMENT

Often $n_i$=n with n×k=N (i.e. groups the same size), but not necessarily, e.g.

treatment 1 = placebo; $n_1 = 10$
treatment 2 = drug A; $n_2 = 20$
treatment 3 = drug B; $n_3 = 20$

with difference between A & B of most interest and 'hopefully' differences between drug and placebo will be 'large'.

**Note**: Comparisons are 'between' patients

Possible analyses:

| | 2 groups | >2 groups |
|---|---|---|
| Normal data: | t-test | 1-way ANOVA |
| Non-parametric: | Mann-Whitney | Kruskal-Wallis |

## 2.3 In series designs

Here each patient receives all k treatments in the same order

Treatment

| | | 1 → | 2 → | 3 → | . . . . . . . . . . → | k |
|---|---|---|---|---|---|---|
| | 1 | X → | X → | X | . . . . . . . . . . → | X |
| | 2 | X → | X → | X | . . . . . . . . . . → | X |
| | • | • | • | • | | • |
| patient | • | • | • | • | | • |
| | • | • | • | • | | • |
| | • | • | • | • | | • |
| | n | X → | X → | X | . . . . . . . . . . → | X |

Problem: Patients are more likely to enter the trial when their disease is most noticeable, and hence more severe than usual, so there is a realistic chance of a *trend* towards improvement while on trial regardless of therapy,
i.e. the later treatments may *appear* to be better than the earlier ones.

In most cases, patients differ greatly in their response to any treatment and in their initial disease state. So large numbers are needed in parallel group studies if treatment effects are to be detected.

However there is much less variability between measurements taken on the same patient at different times. Comparisons here are 'within' patients.

Advantages:

1. Patients can state 'preferences' between treatments
2. Might be able to allocate treatments simultaneously e.g. skin cream on left and right hands

Disadvantages

1. Treatment effect might depend on when it is given
2. Treatment effect may persist into subsequent periods and mask effects of later treatments.
3. Withdrawals cause problems
   (i.e. if a patient leaves before trying all treatments)
4. Not universally applicable,
   e.g. drug treatment compared with surgery
5. Can only use for short term effects

Possible analyses:

| | 2 groups | >2 groups |
|---|---|---|
| Normal data: | paired t-test<br>(on differences) | 2-way ANOVA |
| Non-parametric: | Wilcoxon signed<br>rank test | Friedman's test |

### 2.3.1 Crossover Design

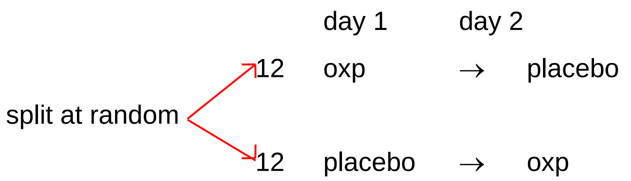
Problems with 'period' or 'carryover' or 'order' can be overcome by suitable design; e.g. crossover design. Here patients receive all treatments, but not necessarily in the same order. If patients crossover from one treatment to another there may be problems of feasibility and reliability.

For example, is the disease sufficiently stable and is patient co-operation good enough to ensure that all patients will complete the full course of treatments? A large number of dropouts after the first treatment period makes the crossover design of little value and it might be better to use a between-patient analysis (i.e. parallel group) analysis of the results for period 1 only.

Example 1 (from Pocock, p112)

Effect of the drug oxprenolol on stage-fright in musicians.

N = 24 musicians, double blind in that neither the musician nor the assessor knew the order of treatment.


day 1 day 2

split at random

12 oxp → placebo

12 placebo → oxp

Each musician assessed on each day for nervousness and performance quality.

Can produce the data in the form

| Patient | Oxp | Plac | Difference | |
|---|---|---|---|---|
| 1 | $x_1$ | $y_1$ | $x_1 - y_1$ | |
| 2 | $x_2$ | $y_2$ | $x_2 - y_2$ | use |
| .. | .. | .. | ........... | paired |
| .. | .. | .. | ........... | t-test |
| 24 | $x_{24}$ | $y_{24}$ | $x_{24} - y_{24}$ | |

More typically design is

washout → treatment → washout → treatment

| | |
|---|---|
| A | B |
| B | A |

(where ‘washout’ is a period with no treatment at all)

Aside: paired t-test is a one-sample t-test on the differences

$$t_{n-1} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_d^2}{n}}}$$

where $s_d$ is the standard deviation of the ***differences***,

i.e. of the n values $(x_{1,i}–x_{2,i})$, i=1,2,…,n

Example 2:—

Plaque removal of mouthwashes

Treatments A — water
B — brand X
C — brand Y

| Patient | order of treatment 1 | 2 | 3 |
|---|---|---|---|
| 1 | A | B | C |
| 2 | A | C | B |
| 3 | B | A | C |
| 4 | B | C | A |
| 5 | C | A | B |
| 6 | C | B | A |

(and perhaps repeat in blocks of six patients)

Note: If it is not possible for each patient to have each treatment use *balanced incomplete block designs*.

## 2.4 Factorial Designs

In some situations, it may be possible to investigate the effect of 2 or more treatments by allowing patients to receive combinations of treatments

|  |  | drug A |  |
|---|---|---|---|
|  |  | NO | YES |
| drug B | NO |  |  |
|  | YES |  |  |

'NO' = placebo

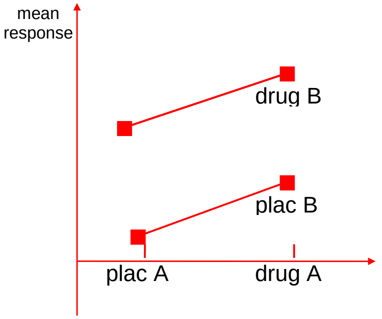
Suppose we had 40 patients and allocated 10 at random to each combination, then overall 20 have had A and 20 have had B.

Compare this with a parallel group study to compare A and B (and a placebo), then with about 40 patients available we would have 13 in each group (3x13 ≈ 40).

This factorial design might lead to more efficient comparisons, because of 'larger' numbers.

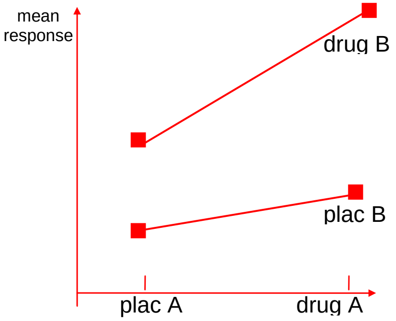
Obviously not always applicable because of problems with **interactions** of drugs, but these might themselves be of interest.

## Types of interaction



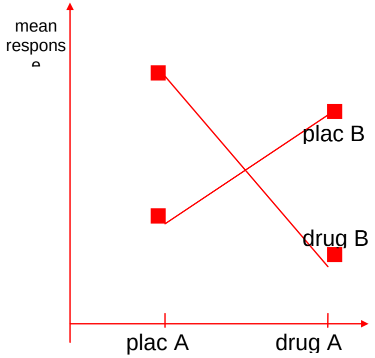
lines parallel ⇒ <u>no interaction</u>

Drug A increases response by same amount irrespective of whether patient is also taking B or not



<u>quantitative interaction</u>

the effect of A is more marked when patient is also taking B



<u>qualitative interaction</u>

A increases response when given alone, but decreases response when in combination with B

## 2.5 Sequential Designs

In its simplest form, patients are entered into the trial in pairs, one receives A, the other B (allocated at random). Test after results from each pair are known.

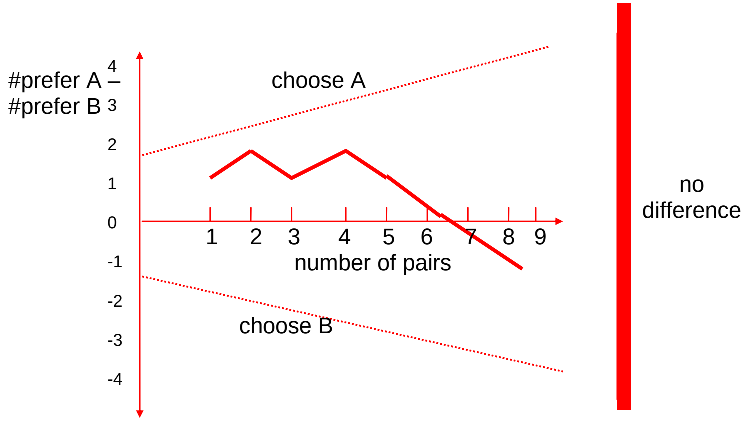
e.g. simple preference data (i.e. patient says which of A or B is better)

| pair | 1 | 2 | 3 | 4 | 5 | 6 | 7 . . . . . . |
|---|---|---|---|---|---|---|---|
| preference | A | A | B | A | B | B | B . . . . . |

.

need 'boundary stopping rules'

e.g .



**Advantages**

1. Detect large differences quickly
2. Avoids ethical problem of fixed size designs (no patient should receive treatment known to be inferior) — but does complicate the statistical design and analysis

**Disadvantages**

1. Responses needed quickly (before next pair of patients arrive)
2. Drop-outs cause difficulties
3. Constant surveillance necessary
4. Requires pairing of patients

## 2.6 Summary & Conclusions

- ‘Always’ use two-sided tests, not one-sided. One-sided tests are almost cheating.
- ‘Always’ use a separate variance t-test.
- Never perform a preliminary test of equality of variance.
- Parallel group designs — different groups of patients receive different treatments, comparisons are ***between*** patients
- In series designs — all patients receive all treatments in sequence, comparisons are ***within*** patients
- Crossover designs — all patients receive all treatments but different subgroups have them in different orders, comparisons are ***within*** patients
- Factorial designs — some patients receive combinations of treatments simultaneously, difficulties if ***interactions***, (quantitative or qualitative), comparisons are ***between*** patients but more available than in series designs
- Sequential designs — suitable for rapidly evaluated outcomes, minimizes numbers of subjects when clear differences between treatments
- Efficient design of clinical trials is a crucial ethical element contributed by statistical theory and practice

# 3. Randomization

## 3.0 Introduction

To avoid **bias** in assigning patients to treatment groups, we need to assign them at random. We need a randomization list so that when a patient (eligible!) arrives they can be assigned to a treatment according to the next number on the list. There are other reasons for using random allocations of treatment groups to subjects, the most important of which is to provide a basis for use of statistical tests and in particular the use of parametric procedures derived from the Normal distribution (e.g. t-tests and $\chi^2$-tets) but the detailed justification of this is well beyond the scope of this course. Provided the total number of possible allocations is 'fairly large' and the actual one used is randomly selected from amongst these then Normal-based tests will be a good approximation to the ideal 'randomization tests' (again beyond the scope of this course). Here 'fairly large' might be more than a hundred or so (but preferably much more). Problems do arise with restricted randomization methods (such as 'minimization', see below) where there can be surprisingly few possible allocations. In the simple example used as an exercise in the course there are 12 subjects but in fact only 4 possible allocations using minimization to select one of these randomly to balance the various prognostic factors whereas with total random allocation into equally sized groups there are 64 possible allocations.

## 3.1 Simple randomization

For a randomized trial with two treatments A and B the basic concept of tossing a coin (heads=A, tails=B) over and over again is reasonable but clumsy and time consuming. Thus people generate random numbers in a statistical computer package (or use tables of random numbers instead. For the purposes of illustration a list of 'random digits' will be used but then some guidance on using **R** will be outlined.

Using the following random digits throughout as an example (Neave, table 7.1, row 26, col 1)
3 0 4 5 8 4 9 2 0 7 6 2 3 5 8 4 1 5 3 2 . . . .

<u>Ex 3.1</u>

12 patients, 2 treatments A & B

Assign 'at random'

e.g. decide 0 to 4 → A
5 to 9 → B

$\Rightarrow$ A A A B B A B A A B B A

Randomization lists can be made as long as necessary & one should make the list **before** the trial starts and make it long enough to complete the whole trial.

Ex 3.2

With 3 treatments A, B, C

decide 1 to 3 → A

4 to 6 → B

7 to 9 → C

0 → ignore

⇒ A B B C B C A C B A A B

In double blind trials, the randomization list is produced centrally & packs numbered 1 to 12 assembled containing the treatment assigned. Each patient receives the next numbered pack when entering the trial. Neither the doctor nor the patient knows what treatment the pack contains — the randomization code is ‘broken’ only at the end of the trial before the analysis starts. Even then the statistician may not be told which of A, B and C is the placebo and which the active treatment.

Disadvantages:– may lack balance (especially in small trials)

e.g. in Ex 3.1 7A’s, 5B’s

in Ex 3.2, 4A’s, 5B’s, 3C’s

Advantage:– each treatment is completely unpredictable, and probability theory guarantees that in the long run the numbers of patients on each treatment will not be substantially different.

## 3.2 Restricted Randomization

### 3.2.1 Blocking

Block randomization ensures equal treatment numbers at certain equally spaced points in the sequence of patient assignments. Each random digit specifies what treatment is given to the next block of patients.

In Ex 3.1 (12 patients, 2 treatments A & B)

0 to 4 → AB

⇒ AB AB AB BA BA AB BA

5 to 9 → BA

In Ex 3.2 (3 treatments A, B & C)

1 → ABC

2 → ACB

3 → BAC

4 → BCA

5 → CAB

6 → CBA

7,8,9,0 → ignore

⇒ BAC BCA CAB BCA

Disadvantage:– This blocking is easy to crack/decipher and so it may not preserve the double blinding.

With 2 treatments we could use a block size of 4 to try to preserve blindness

Ex 3.3

1 → AABB<br>
2 → ABAB<br>
3 → ABBA<br>
4 → BBAA<br>
5 → BABA<br>
6 → BAAB<br>
7,8,9,0 → ignore

⇒ ABBA BBAA BABA

Problem:– at the end of each block a clinician who keeps track of previous assignments could predict what the next treatment would be, though in double-blind trials this would not normally be possible. The smaller the choice of block size the greater the risk of randomization becoming predictable.

A trial without 'stratification' (i.e. all patients of the same 'type' or category) should have a reasonably large block size so as to reduce prediction but not so large that stopping in the middle of a block would cause serious inequality.

In stratified randomization one might use random permuted blocks for patients classified separately into several types (or strata) and in these circumstances the block size needs to be quite small.

### 3.2.2 Unequal Collection

In some situations, we may not want complete balanced numbers on each treatment but a fixed ratio.

e.g. A Standard<br>
B New ← need most information on this

decide on a fixed ratio of 1:2 ⇒ need blocking

Reason:– more accurate estimates for effects of B; A variation probably known reasonably well already if it is the standard.
Identify all the 3!/2! possible orderings of ABB and assign to digits:

1 to 3 → ABB<br>
4 TO 6 → BAB<br>
7 TO 9 → BBA<br>
0 → ignore

⇒ABB BAB BAB BBA

### 3.2.3 Stratified Randomization

(Random permuted blocks within strata)

It is desirable that treatment groups should be as similar as possible in regard of patient characteristics:

<u>relevant patient factors</u>

e.g. age sex stage of disease site

(<50,>50) (M,F) (1,2,3,4)(arm,leg)

Group imbalances could occur with respect to these factors: e.g. one treatment group could have more elderly patients or more patients with advanced stages of disease. Treatment effects would then be *confounded* with age or stage (i.e. we could not tell whether a difference between the groups was because of the different treatments or because of the different ages or stages).

Doubt would be cast on whether the randomization had been done correctly and it would affect the credibility of any treatment comparisons.

We can *allow for* this at the analysis stage through regression (or analysis of covariance) models, however we could *avoid* it by using a stratified randomization scheme.

Here we prepare <u>a separate randomization list for each stratum.</u>

e.g. (looking at age and sex) 8 patients available in each stratum

| | | |
|---|---|---|
| <50, M | A B B A | B B A A |
| ≥ 50, M | B A B A | B A A B |
| <50, F | A B A B | B A A B |
| ≥ 50, F | A B A B | A B B A |

so as a new patient enters the trial, the treatment assigned is taken from the next available on the list corresponding to their age and sex.

### 3.2.4 Minimization

If there are many factors, stratification may not be possible. We might then adjust the randomization *dynamically* to achieve balance, i.e. ***minimization*** (or adaptive randomization). This effectively balances the marginal totals for each level of each factor — however, it looses some randomness. The method is to allocate a new patient with a particular combination of factors to that treatment which 'balances' the numbers on each treatment with that combination. See example below.

Ex 3.5 Minimization (from Pocock, p.85)

Advanced breast cancer, two treatments A & B, 80 patients already in trial. 4 factors thought to be relevant:–

'performance status' (ambulatory/non-ambulatory),
'age' (<50/≥ 50),
'disease free-time' (<2/≥ 2 years),
'dominant lesion' (visceral/osseous/soft tissue).

Suppose that 80 subjects have already been recruited to the study. A new patient enters the trial who is **ambulatory**, **<50**, has ≥ **2 years** disease free time and a **visceral** dominant tissue. To decide which treatment to allocate her to, look at the numbers of patients with those factors on each treatment: suppose that of the 80 already in the study, 61 are ambulatory, 30 of whom are on treatment A, 31 on B; of the 19 non-ambulatory 10 are on A and 9 on B. Similarly of the 35 aged under 50 18 are on A and 17 on B, etc. (the complete set of numbers in each category are given in the table below). We now calculate a 'score':

| **Factors** | **A** | **B** | **next patient** |
|---|---|---|---|
| performance status: | | | |
| ambulatory | 30 | 31 | ← |
| non-ambulatory | 10 | 9 | |
| age: | | | |
| <50 | 18 | 17 | ← |
| ≥ 50 | 22 | 23 | |
| disease free-time: | | | |
| <2 years | 31 | 32 | |
| ≥ 2 years | 9 | 8 | ← |
| dominant lesion: | | | |
| visceral | 19 | 21 | ← |
| osseous | 8 | 7 | |
| soft tissue | 13 | 12 | |

To date, A score = 30 + 18 + 9 + 19 = 76
B score = 31 + 17 + 8 + 21 = 77

⇒ put patient on A
(to balance up the scores)

(if scores equal, toss a coin or use simple randomization)

Unlike other methods of treatment assignment, one does not simply prepare a randomization list in advance. Instead one needs to keep a continually and up-to-date record of treatment assignments by patient factors. Computer software is available to help with this (see §3.5).

Problem:– one possible problem is that treatment assignment is determined solely by the arrangement to date of previous patients and involves no random process except when the treatment scores are equal. This may not be a serious deficiency since investigators are unlikely to keep track of past assignments and hence advance predictions of treatment assignments should not be possible.

Nevertheless, it may be useful to introduce an element of chance into minimization by assigning the treatment of choice (i.e. the one with smallest sum of marginal totals or 'score') with probability p where p > ½ (e.g. p= ¾ might be a suitable choice).

Hence, before the trial starts one could prepare 2 randomization lists. The first is a simple randomisation list where A and B occur equally often for use only when the 2 treatments have equal scores, the second is a list in which the treatment with the smallest score occurs with probability ¾ while the other treatment occurs with probability ¼. Using a table of random numbers this is prepared by assigning S (=Smallest) for digits 1 to 6 and L (=Largest) for digits 7 or 8 (ignore 9 and 0).

### 3.2.4.1 Note: Minimization/Adaptive

Note that some authors use the term Adaptive Randomization as a synonym for minimization methods but this is best reserved for situations where the ***outcomes*** of the treatment are available before the next subject is randomised and the randomization scheme is adapted to incorporate information from the earlier subjects.

## 3.3 Why Randomize?

1. To safeguard against selection bias
2. To try to avoid accidental bias
3. To provide a basis for statistical tests

## 3.4 Historical/database controls

Suppose we put all current patients on new treatment and compare results with records of previous patients on standard treatment. This use of historical controls avoids the need to randomize which many doctors find difficult to accept. It might also lessen the need for a placebo.

Major problems:–

- Patient population may change (no formal inclusion/exclusion criteria before trial started for the historical patients)
- Ancillary care may improve with time ⇒ ‘new’ performance exaggerated.

Database controls suffer from similar problems.

We cannot say whether any improvement in patients is due to drug or to act of being treated (placebo effect). It may be possible to use a combination of historical controls supplemented with [a relatively small number of] current controls which serve as a check on the validity of the historical ones.

## 3.5 Randomization Software

A directory of randomisation software is maintained by Martin Bland at:

http://www-users.york.ac.uk/~mb55/guide/randsery.htm

This includes [free] downloadable programmes for simple and blocked randomization, some commercial software including add-ons for standard packages such as STATA, and links to various commercial *randomization services* which are used to provide full blinding of trials.

This site also includes some useful further notes on randomization with lists of references etc.

R, S-PLUS and MINITAB provide facilities for random digit generation but this is less easy in SPSS.

### 3.5.1 Randomization in R

In **R** the basic command is `sample(.).` Type `help(sample)` to find full details. Here some llustrations:

```
> x<- c(0:9)
> x
 [1] 0 1 2 3 4 5 6 7 8 9
> sample(x)
 [1] 6 3 1 7 5 4 9 8 0 2      permutation
> sample(x,4)
[1] 3 1 6 7      subsample without replacement
> sample(x,4,replace=TRUE)
[1] 0 9 0 7  subsample with replacement
> sample(x,20,replace=T)
 [1] 3 8 1 4 0 9 4 7 5 1 6 4 2 3 1 8 3 3 7 0
```

```
> z<-c(rep("A",5),rep("B",5),rep("C",5))
> z
 [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C"
"C" "C" "C" "C"
> sample(z)
 [1] "B" "A" "A" "A" "C" "C" "B" "B" "C" "A" "B"
"C" "B" "A" "C“
> sample(c(rep("A",4),rep("P",2)))
[1] "A" "A" "P" "A" "P" "A"
```

and considerably more advanced but very concise (use `the help()` system to find out what each bit does)

```
> lapply(rep(list(LETTERS[1:3]),4),sample)
[[1]]
[1] "A" "C" "B"

[[2]]
[1] "A" "C" "B"

[[3]]
[1] "B" "A" "C"

[[4]]
[1] "B" "A" "C"

>
```

and

```
>
matrix(apply(matrix(c("A","B","C"),3,4),2,sample)
,1,3*4)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[,10] [,11] [,12]
[1,] "A"  "B"  "C"  "C"  "A"  "B"  "C"  "A"  "B"
"C"   "A"   "B"
>
>
```

## 3.6 Summary and Conclusions

Randomization

- protects against accidental and selection bias
- provides a basis for statistical tests (e.g. use of normal and t-distributions)

Types of randomization include

- simple (but may be unbalanced over treatments)
- blocked (but small blocks may be decoded)
- stratified (but may require small blocks)
- minimization (but lessens randomness)

Historical and database controls may not reflect change in patient population and change in ancillary care as well as inability to allow for placebo effect.

# 4. Size of the trial

## 4.1 Introduction

What ***sample sizes*** are required to have a good chance of detecting clinically relevant differences if they exist?

**Specifications required**

*[0. main purpose of trial]*

1. main outcome measure (e.g. $\mu_A$, $\mu_B$ estimated by $\overline{X}_A, \overline{X}_B$)

2. method of analysis (e.g. two-sample t-test)

3. result given on standard treatment (or pilot results)

4. how small a difference is it important to detect? ($\delta=\mu_A - \mu_B$)

5. degree of certainty with which we wish to detect it

   (power, $1-\beta$)

**Note**

- 'non-significant difference' is not the same as 'no clinically relevant difference' exists.
- mistakes can occur:

Type I: false positive; treatments equivalent but result significant ($\alpha$ represents risk of false positive result)

Type II: false negative; treatments different but result non-significant ($\beta$ represents risk of false negative result)

## 4.2 Binary Data

Count numbers of ‘**S**uccesses’ & ‘**F**ailures’, and look at the case when there are equal numbers on standard and new treatments:

| | S | F | Σ |
|---|---|---|---|
| standard | $x_1$ | $n–x_1$ | n |
| new | $x_2$ | $n–x_2$ | n |

<u>Model:</u> $X_1 \sim B(n,\theta_1)$ and $X_2 \sim B(n,\theta_2)$ (binomial distributions), where $X_1$ and $X_2$ are the numbers of success on standard and new treatments.

<u>Hypotheses:</u> $H_0$: $\theta_1 = \theta_2$ vs. $H_1$: $\theta_1 \neq \theta_2$

(i.e. a 2-sided test of proportions)

<u>Approximations:</u> Take Normal approximation to binomial:

$$X_1 \sim N(n\theta_1, n\theta_1(1–\theta_1)) \text{ and } X_2 \sim N(n\theta_2, n\theta_2(1–\theta_2))$$

<u>Requirements</u>: take $\alpha$ = P[type I error] = level of test = 5%

and $\beta$ = P[type II error] = 1 - power at $\theta_2$=10%

Suppose standard gives 90% success and it is of clinical interest if the new treatment gives 95% success (or better), i.e.

$\theta_1 = 0.9$

$\theta_2 = 0.95$ (i.e. a 5% improvement)

1– $\beta$ = $\gamma$ is the power of the test and we decide we want $\gamma(0.95)=0.9$ (so we want to be 90% sure of detecting an improvement of 5%)

We have $(X_2/n – X_1/n) \sim N(\theta_2–\theta_1, [\theta_2(1–\theta_2)+ \theta_1(1–\theta_1)]/n)$

since $\text{var}(X_2/n – X_1/n) = \text{var}(X_2/n)+\text{var}(X_3/n)$

$$= \theta_2(1–\theta_2)/n + \theta_1(1–\theta_1)/n$$

so the test statistic is:

$$\frac{\left(\frac{X_2}{n}-\frac{X_1}{n}\right)-0}{\sqrt{\text{var}\left(\frac{X_2}{n}-\frac{X_1}{n}\right)}} \sim N(0,1) \text{ under } H_0: \theta_1 = \theta_2$$

and we will reject $H_0$ at the 5% level if

$$\left|\frac{x_2}{n}-\frac{x_1}{n}\right| > 1\cdot 96\sqrt{\frac{2\times 0\cdot 9\times 0\cdot 1}{n}}$$

(remembering $\theta_1=\theta_2=0.9$ under $H_0$)

The power function of the test is

P[reject $H_0$ | alternative parameter $\theta_2$]

= $\gamma(\theta_2)$ = P{|$X_2$/n –$X_1$/n|>1.96√(2x0.9x0.1/n)| $\theta_1$=0.9,$\theta_2$}

and we require $\gamma(0.95)$ = 0.9

[Note that for $\theta_2$=0.95, var($X_2$/n)=0.95(1–0.95)/n but var($X_1$/n)=0.9(1–0.9)/n since $\theta_1$=0.9 still]

Now

$\gamma$(0.95)=1–P{|$X_2$/n–$X_1$/n|≤1.96√(2x0.9x0.1/n)|$\theta_1$=0.9,$\theta_2$=0.95}

$$=1-\left[\Phi\left\{\frac{1.96\sqrt{2\times.9\times.1/n}-0.05}{\sqrt{\frac{.95\times.05}{n}+\frac{.9\times.1}{n}}}\right\}-\Phi\left\{\frac{-1.96\sqrt{2\times.9\times.1/n}-0.05}{\sqrt{\frac{.95\times.05}{n}+\frac{.9\times.1}{n}}}\right\}\right]$$

and the last term $\approx\Phi\left\{-1.96-\frac{0.05\sqrt{n}}{\sqrt{.95\times.05+.9\times.1}}\right\}\to 0$

so we require $\Phi\left\{\frac{1.96\sqrt{2\times.9\times.1}-0.05\sqrt{n}}{\sqrt{.95\times.05+.9\times.1}}\right\}\approx 0.1$

i.e. $n\approx\frac{(.95\times.05+.9\times.1)}{.05^2}\left\{\Phi^{-1}(0.1)-1.96\sqrt{\frac{2\times.9\times.1}{.95\times.05+.9\times.1}}\right\}^2$

i.e. need around 580 patients in each ‘arm’ of the trial (1,160 in total) or more if drop out rate known. Could inflate these by 20% to allow for losses.

General formula:

$$n\approx\frac{\theta_2(1-\theta_2)+\theta_1(1-\theta_1)}{(\theta_2-\theta_1)^2}\left\{\Phi^{-1}(\beta)+\Phi^{-1}(\alpha/2)\right\}^2$$

**{N.B. both $\Phi^{-1}(\beta)$ and $\Phi^{-1}(\alpha/2)<0$}**

$\theta_1$ and $\theta_2$ are the hypothetical percentage successes on the two treatments that might be achieved if each were given to a large population of patients. They reflect the realistic expectations of goals which one wishes to aim for when planning the trial and do not relate directly to the eventual results.

$\alpha$ is the probability of saying that there is a ‘significant difference’ when the treatments are really equally effective

(i.e $\alpha$ represents the risk of a **false positive** result)

$\beta$ is the probability of not detecting a significant difference when there really is a difference of magnitude $\theta_1 - \theta_2$ (**false negative**).

**Notes:–**

1. Approximation requires $\dfrac{\sqrt{2\theta_1(1-\theta_1)}}{\sqrt{\theta_2(1-\theta_2)+\theta_1(1-\theta_1)}} \approx 1$

   which here = 1.14, so reasonable, — otherwise need to use more complex methods.

2. Machin & Campbell (Blackwell, 1997) provide tables for various $\theta_1$, $\theta_2$, $\alpha$ and $\beta$. There are also computer programmes available.

3. If we can really justify a 1-sided test (e.g. from a pilot study) then put $\Phi^{-1}(\alpha/2) \rightarrow \Phi^{-1}(\alpha)$. 1–sided testing reduces the required sample size.

4. For given $\alpha$ and $\beta$, n depends mainly on $(\theta_2 - \theta_1)^2$ (& is roughly inversely proportional) which means that for fixed type I and type II errors if one **halves** the difference in response rates requiring detection one needs a **fourfold** increase in trial size.

5. Freiman *et al* (1978) *New England Journal of Medicine* reviewed 71 binomial trials which reported no statistical significance. They found that 63% of them had power < 70% for detecting a 50% difference in success rates. (??unethical to spend money on such trials?? [Pocock])

6. N depends very much on the choice of type II error such that an increase in power from 0.5 to 0.95 requires about 3 times the number of patients.

7. In practice, the determination of trial size does not usually take account of patient factors which might influence predicted outcome.

## 4.3 Quantitative Data

(i) Quantitative response — standard has mean $\mu_1$ and new treatment has mean $\mu_2$.

(ii) Two-sample t-test, but assume n large, so use Normal approximation: $X_1 \sim N(\mu_1, \sigma^2/n)$ and $X_2 \sim N(\mu_2,\sigma^2/n)$ assume equal sample sizes n and equal ***known*** variance $\sigma^2$.

The test works well in practice provided the variances are not very different.

(iii) Assume $\mu_1$ known

(iv) Want to detect a 'new' mean of size $\mu_2$, (or $\delta = \mu_2 - \mu_1$ the difference in mean response that it is important to detect).

(v) Power at $\mu_2$ is 1-$\beta$, i.e. $\gamma(\mu_2)$= 1-$\beta$, the degree of certainty to detect such a difference exists.

Test statistic under $H_0$: $\mu_1=\mu_2$ is $T=\dfrac{\overline{X}_2-\overline{X}_1-0}{\sqrt{2\sigma^2/n}} \sim N(0,1)$

2-sided $\alpha$ test rejects $H_0$ if $\left|\dfrac{\overline{x}_2-\overline{x}_1}{\sqrt{2\sigma^2/n}}\right| > -\Phi^{-1}(\alpha/2)$

Power function if new mean = $\mu_2$ is

$$\gamma(\mu_2) = 1 - P\left\{\left|\frac{\overline{X}_2-\overline{X}_1-0}{\sqrt{2\sigma^2/n}}\right| \le -\Phi^{-1}(\alpha/2) \;\middle|\; (\overline{X}_2-\overline{X}_1) \sim N(\mu_2-\mu_{\cdot}),{}^{2\sigma^2}/_{n})\right\}$$

$$= 1 - \left[\Phi\left\{-\Phi^{-1}(\alpha/2) - \frac{\mu_2-\mu_1}{\sqrt{2\sigma^2/n}}\right\} - \Phi\left\{+\Phi^{-1}(\alpha/2) - \frac{\mu_2-\mu_1}{\sqrt{2\sigma^2/n}}\right\}\right]$$

and we require $\gamma(\mu_2)$=1–$\beta$, i.e. set

$$\beta = \Phi\left\{-\Phi^{-1}(\alpha/2) - (\mu_2-\mu_1)\sqrt{n/2\sigma^2}\right\} - \Phi\left\{+\Phi^{-1}(\alpha/2) - (\mu_2-\mu_1)\sqrt{n/2\sigma^2}\right\}$$

As before, 2nd term → 0 as n → ∞

so we need $\Phi^{-1}(\beta) \approx -\Phi^{-1}(\alpha/2) - (\mu_2-\mu_1)\sqrt{n/2\sigma^2}$

or $n = \dfrac{2\sigma^2}{(\mu_2-\mu_1)^2}\left\{\Phi^{-1}(\beta)+\Phi^{-1}(\alpha/2)\right\}^2$

**<u>Notes:–</u>**

1. All comments in binomial case apply here also.

2. Need to know the variance $\sigma^2$ which is difficult in practice:– may be able to look at similar earlier studies, may be able to run a small pilot study, may be able to say what the likely maximum and minimum possible responses under standard treatment could be and so calculate the likely maximum possible range and then get an approximate value for $\sigma$ as one quarter of the range.

## 4.4 One-Sample Tests

The two formula given above apply to two-sample tests for proportions (§4.2) and means (§4.3). It is straightforward to derive similar formula for the corresponding one-sample tests.

In the case of a one sample test, the required sample size to achieve a power of (1– $\beta$) when using a size $\alpha$ test of detecting a change from a proportion $\theta_0$ to $\theta$ is given by

$$n=\frac{\left\{\Phi^{-1}(\beta)\sqrt{\theta(1-\theta)}+\Phi^{-1}(\tfrac{\alpha}{2})\sqrt{\theta_0(1-\theta_0)}\right\}^2}{(\theta-\theta_0)^2}$$

In the case of a one sample test on means, the required sample size to achieve a power of (1– $\beta$) when using a size $\alpha$ test of detecting a change from a proportion $\mu_0$ to $\mu$ is given by

$$n=\frac{\sigma^2}{(\mu_0-\mu)^2}\left\{\Phi^{-1}(\beta)+\Phi^{-1}(\tfrac{\alpha}{2})\right\}^2$$

The prime use of this formula would be in a paired t-test with $\mu_0$=0.

## 4.5 Practical problems

1. If recruitment rate of patients is low, it may take a long time to complete trial. This may be unacceptable and may lead to loss of interest. We could

   a) increase $\delta$

   b) relax $\alpha$ and $\beta$

   (and accept that small differences may be missed)

   c) think of using a multicentre trial (see later)

2. Allow for dropouts, missing data, etc.
   e.g. inflate required numbers by 20% to allow for losses

3. Statistical procedures must be as efficient as possible
   — consider more complex designs.

## 4.6 Computer Implementation

R, S-PLUS and MINITAB provide extensive facilities for power and sample size calculations and these are easily found under the Statistics and Stat menus under Power and Sample Size in the last two packages. SPSS does not currently provide any such facilities (i.e. up to version 16). Note that the formulae given above are approximations and so results may differ from those returned by computer packages, perhaps by as much as 10% in some cases. Further, S-PLUS and MINITAB use different approximations and continuity corrections. There are many commercial packages available, perhaps the industry standard is nQuery Advisor which has extensive facilities for more complex problems (analysis of variance, regression etc).

The course web page provides a link to small DOS program, POWER.EXE which has good facilities and this can be downloaded from the page. There are also links to other free sources on the web (and a Google search on power sample size will find millions of references). If you use these free programs you should remember how much you have paid for them.

### 4.6.1 Implementation in R

In **R** the functions `power.t.test()`, `power.prop.test` and `power.anova.test()` provide the basic calculations needed for finding any one from the remaining two of power, sample size and CRD (referred to as “delta” in R) from the other two in the commonly used statistical tests of means, proportions and one-way analysis of variance. The HELP system provides full details and extensive examples. `power.t.test()` can handle both two-sample and one-sample tests, the former is the default and the latter requires `type="one.sample"` in the call to it. `power.prop.test()`***only*** provides facilities for two-sample tests. For one-sample the programme `power.exe` (available from the course web page) is available.

### 4.6.1.1 Example: test of two proportions

Suppose it is wished to determine the sample size required to detect a change in proportions from 0.9 to 0.95 in a two sample test using a significance level of 0.05 with a power of 0.9 (or 90%).

```
> power.prop.test(p1=0.9,p2=0.95,power=0.9,sig.level=0.05)
     Two-sample comparison of proportions power calculation
              n = 581.082
             p1 = 0.9
             p2 = 0.95
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
 NOTE: n is number in *each* group
```

Thus a total sample size of about 1162 is needed, in close agreement with that determined by the approximate formula in §5.2.

### 4.6.1.2 Example: t-test of two means

What clinically relevant difference can be detected with a two sample t-test using a significance level of 0.05 with power 0.8 (or 80%) and a total sample size of 150 when the standard deviation is 3.6?

```
> power.t.test(n=75,sd=3.6,power=0.8,sig.level=0.05)
     Two-sample t test power calculation
              n = 75
          delta = 1.657746
             sd = 3.6
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
 NOTE: n is number in *each* group
```

## 4.7 Summary and Conclusions

Sample size calculation is ethically important since

- Samples which are too small may have little chance of producing a conclusion, so exposing patients to risk with no outcome
- Samples which are needlessly too large may expose more subjects than necessary to a treatment later found to be inferior

For sample size calculation we need to know

- outcome measure
- method of analysis (including desired significance levels)
- clinical relevant difference
- power
- results on standard treatment (including likely variability)

For practical implementation we need to know the maximum achievable sample size. This could be limited by

- Recruitment rate and time when analysis of results must be performed
- Total size of target population (number of subjects with the condition which is to be the subject of the clinical trial)
- Available budget

In cases where the maximum sample size is limited it is more useful to calculate a table of clinically relevant differences that can be detected with a range of powers using the available sample size.

Sample size facilities in **R** in the automatically loaded `stats` package are provided by the three functions `power.t.test()`, `power.prop.test()` and `power.anova.test()`. The first handles one and two sample t-tests for equality of means, the second handles two-sample tests on binomial proportions (but ***not*** one-sample tests) and the third simple one-way analysis of variance. The first two will calculate any of sample size, power, clinically relevant difference and significance level given values for the other three. The third will calculate the number of groups, the [common] size of each group, the within groups variance, the between groups variance, power and sample size given values for the other five.

Programme `power.exe` (available from the course web pages) will calculate

- one and two-sample t-tests (including paired t-test)
- one ***and*** two-sample tests on binomial proportion
- test on single correlation coefficient
- one sample Mann-Whitney U-test
- Mcnemar’s test
- multiple comparisons using 2-sample t-tests
- cross-over trial comparisons
- log rank test (in survival)

Facilities are available in a variety of freeware and commercial software for many more complex analyses (e.g. regression models) though in many practical cases substantial simplification of the intended analysis is required and so calculations can only be used as a guide.

# 5. Multiplicity and interim analysis

## 5.1 Introduction

This section outlines some of the practical problems that arise when several statistical hypothesis tests are performed on the same set of data. This situation arises in many apparently quite different circumstances when analyzing data from clinical trials but the common danger is that the risk of *false positive* results can be much higher than intended. The particular danger is when the *most statistically significant* result is selected from amongst the rest for particular attention, perhaps quite unintentionally.

The most common situations where problems of multiplicity (or multiple testing) arise are encountered are

- multiple endpoints
- subgroup analyses
- interim analyses
- repeated measures

The remedies for these problems include adjusting nominal significance levels to allow for the multiplicity (e.g. *Bonferroni adjustments* or more complex methods in interim analyses), use of special tests (e.g. *Tukey's test for multiple comparisons* or *Dunnett's Test for multiple comparisons with a control*) or use of more sophisticated statistical techniques (e.g. Analysis of Variance or Multivariate Analysis).

We begin with a brief example (constructed artificially but not far from reality).

### 5.1.1 Example: Signs of the Zodiac

(Effect of new dietary control regime.)

**Data:** 250 subjects chosen 'randomly'. Weighed at start of week and again at end of week. Data in kg.

**Results:**

```
                     N      Mean      StDev    SE Mean
Weight before      250    58.435     12.628     0.799
Weight after       250    58.309     12.636     0.799
Difference         250     0.126      1.081     0.068
```

So, average weight loss is 0.13kg (≈1/4 pound)

Confidence interval for mean weight loss is (–0.009, 0.260)kg.

Paired t-test for weight loss gives a t-statistic of 1.84, giving a p-value of 0.067 (using a two-sided test). (t=0.126/0.068)

**Not quite significant at the 5% level !**

Can anything be done to 'squeeze' a significant result out of this expensive study (we've been told we cannot change our mind and use a one-sided test instead!) ?????

— luckily, the birth dates are available. Perhaps the success of the diet depends upon the personality and determination of the subject. So, look at subgroups of the data by their sign of the Zodiac:–

## Mean weight loss by sign of the Zodiac

| Zodiac sign | n | mean weight loss | standard error of mean | t | p-value | |
|---|---|---|---|---|---|---|
| **Aquarius** | 26 | 0.313 | 0.217 | 1.44 | 0.161 | |
| **Aries** | 15 | **0.543** | **0.205** | **2.65** | **0.019** | ★★ |
| **Cancer** | 21 | 0.271 | 0.249 | 1.09 | 0.289 | |
| **Capricorn** | 27 | -0.191 | 0.222 | -0.86 | 0.397 | |
| **Gemini** | 18 | 0.068 | 0.266 | 0.26 | 0.801 | |
| **Leo** | 22 | 0.194 | 0.234 | 0.83 | 0.416 | |
| **Libra** | 26 | 0.108 | 0.217 | 0.50 | 0.623 | |
| **Pisces** | 19 | 0.362 | 0.232 | 1.56 | 0.136 | |
| **Sagittarius** | 12 | 0.403 | 0.294 | 1.37 | 0.197 | |
| **Scorpio** | 20 | 0.030 | 0.274 | 0.11 | 0.248 | |
| **Taurus** | 22 | **-0.315** | **0.183** | **-1.72** | **0.099** | **?** |
| **Virgo** | 22 | 0.044 | 0.238 | 0.18 | 0.955 | |

**Conclusions:** those born under the sign of Aries are particularly suited to this new dietary control. It is well known that Arieans have the strength of character and determination to pursue a strict diet and stick to it. On the other hand, there seems to be some suggestion that those under the sign of Taurus have actually put on weight. Again, not really surprising when one considers the typical characteristics of Taurus…………… . (& if we also used a 1-sided p-value……… .)

**Comment:** This is nonsense**!** The fault arises in that ***the most significant result*** was selected for attention without making any allowance for that selection. The subgroups were considered after the first test had proved inconclusive, not before the experiment had been started so the hypothesis that Aireans are good dieters was only suggested by the data and the fact that it gave an apparently significant result. This is almost certainly a

***false positive result.***

**Note:** The data for weight before and weight after were artificially generated as two samples from a Normal distribution with mean 58.5 and variance 12.5, i.e. there should be no significant difference between the mean weights before and after (as indeed there is not). Birth signs were randomly chosen with equal probability. Two sets of data had to be tried before finding this feature of at least one Zodiac sign providing a false positive.

This example will be returned to later, including ways of analysing the data more honestly.

## 5.2 Multiplicity

### 5.2.1 Fundamentals

In clinical trials a large amount of information accumulates quickly and it is tempting to analyse many different responses: i.e. to consider multiple end points or perform many hypothesis tests on different combinations of subgroups of subjects.

**Be careful!**

All statistical tests run the risk of making mistakes and declaring that a real difference exists when in fact the observed difference is due to natural chance variation. However, this risk is controlled ***for each individual single test*** and that is precisely what is meant by the *significance level* of the test or the *p-value*. The p-value is the more precise calculation of the risk of a false positive result and is more commonly quoted in current literature. The significance level is usually the broader range that the p-value falls or does not fall in, e.g. ‘not significant at the 5% level’ means that the p-value exceeds 0.05 (& may in fact be much larger than 0.05 or possibly only slightly greater).

However, it is difficult to control the overall risk of declaring *at least one false positive somewhere* if many separate significance tests are performed. If each test is operated at a separate significance level of 5% then we have a 95% chance of not making a mistake on the first test, a 95%×95% (= 90.25%) of avoiding a mistake on either of the first two and so nearly a 10% risk of one or other (or both) of the first two tests resulting in a false positive.

If we perform 10 (independent) tests at the 5% level, then

Prob [reject $H_0$ in at least one test when $H_0$ is true in all cases] =

$$1 - (1- 0.05)^{10} = 0.4$$

i.e. a 40% chance of declaring a difference when none exists!!!!

Perhaps a more familiar situation is the calculation of ***Normal Ranges*** in clinicochemcal tests. A ‘normal person’ has been defined as one who has not been sufficiently investigated. A *normal range* comprise 95% of the values. If 100 normal persons are evaluated by a clinical test then only 95 of them will be declared normal. If they are then subjected to another independent test then only 90 of them will remain as being considered normal. After another 8 tests there will be only 60 normals left.

**Aside:** A complementary problem is that of ***false negatives,*** i.e. failing to detect a difference when one really exists. Clearly the risk diminishes as more and more tests are performed but at the greatly increased risk of more false positives. (If you buy more Lotto tickets you are more likely to win, but at increasing expense). These problems are more complex and are not considered here, nor are they commonly considered in the medical statistical literature.

### 5.2.2 Bonferroni Corrections

A simple but very conservative remedy to control the risk of making a false positive is to lower the nominal significance level of the individual tests so that when you calculate the overall final risk after performing k tests it turns out to be closer to your intended level, typically 5%. This is known as a *Bonferroni correction.* The simplest form of the rule is that if you want an overall level of $\alpha$ and you perform k (independent) significance tests then each should be run at a nominal $\alpha$/k level of significance.

**Examples:**

**(a)** 5 separate tests will be performed, so to achieve an overall 5% level of significance a result should only be declared if any test is nominally significant at the 5%/5=1% significance level.

**(b)** 25 tests are to be performed, an overall level of 1% is intended, so each should be run at a nominal level of 1/25=0.04%, i.e. a result should not be claimed unless p<0.0004 in any one of them.

**(c)** 12 tests have been performed and the smallest p-value is 0.019. What is the overall level of significance? The Bonferroni method suggests that it is safe to claim only an overall level of $12\times0.019 = 0.228$. Note that this is the situation in the Signs of the Zodiac example above. This suggests we have no worthwhile evidence of any birth sign being particularly suited to dieting. (We will return later to this example).

**Note:** Clearly, if a large number of tests is to be performed the Bonferroni correction will demand a totally unrealistically small p-value. This is because the Bonferroni method is very conservative — it over-corrects and in part this is because a simple but only roughly approximate formula has been used.

We can make a more exact calculation which says that to achieve a desired *overall* level of $\alpha$ when performing k tests you should use a nominal level of $\varepsilon$ where $\alpha = 1 - (1-\varepsilon)^k$, i.e. only declare a result significant at level $\alpha$ if $p < \varepsilon$, where $\varepsilon$ is given by the formula above. It may not appear very easy to calculate the level from this formula and usually it is not worthwhile since it would not really cure the problem of it being over conservative and usually there are better ways of overcoming the problem of multiplicity, by concentrating on the more important objectives of the trial or using a more sophisticated analysis.

**Aside:** an approximately solution to the formula above is $\varepsilon = \alpha/k$ which is the derivation of the simple Bonferroni correction.

The exact solution is $\varepsilon = 1 - \exp\{{}^1\!/_k \log(1 - \alpha)\}$.

### 5.2.3 Multiple End-points

The most common situation where problems of multiple testing arise is when many different outcome measures are used to assess the result of therapy. It is rare that only a single measure is used ('once you have got hold of the subject then measure everything in sight'). For example, it is routine to record pulse rate, systolic and diastolic blood pressure, perhaps sitting, standing and supine before and after exercise in hypertensive studies. However, separate significance tests on each separate end-point comparison increases the chance of some false positives.

**Remedies:**

- Bonferroni correction
- choose primary outcome measure
- multivariate analysis

Applying Bonferroni corrections is unduly conservative, i.e. it means that you are less likely to be able to declare a real difference exists even if there is one. The reason for this is that the results from multiple outcome measures are likely to be *highly correlated*. If the drug is successful as judged by standing systolic blood pressure it is quite likely that the sitting systolic blood pressure would provide similar evidence. If you had not measured the other outcomes and so been forced to use a Bonferroni adjustment in multiplying all your p-values by the number of tests and had instead stayed with just the single measure you might have had an interesting result. This would be particularly frustrating if you had considered 20 highly correlated measures, each providing a nominal p-value of around 0.01 and Bonferroni told you that you could only claim an overall p-value of 0.2.

The recommended remedy is to concentrate on a *primary outcome measure* with perhaps a few (two or three) secondary measures which you consider as well (perhaps making an informal Bonferroni correction). Of course it is essential that these are decided in advance of the trial and ***this is stated in the protocol.*** The choice can be based on medical expertise or from initial results from a pilot study if the trial is a novel situation. This does not preclude recording all measures that you wish but care must be taken in reporting analyses on these — this is particularly true of clinicochemcial laboratory results (and especially when they are recorded as within or without 'Normal Ranges', see above). Of course these should be scrutinized and any causes for concern reported.

The ideal statistical remedy is to use a multivariate technique though this may require seeking more specialist or professional statistical assistance. Multivariate techniques will make proper allowance in the analysis for correlated observations (e.g. sitting and standing systolic blood pressure). There are multivariate equivalents of routine univariate statistical analyses such as Student's t-test (it is Hotelling's $T^2$-test), Analysis of Variance or ANOVA (it is Multivariate Analysis of Variance or MANOVA, with Wilks' test or the Lawley-Hotelling test).

The advantage of multivariate analysis is that it will handle all measurements simultaneously and return a single p-value assessing the evidence for departure from the null hypothesis, e.g. that there is a difference between the two treatment groups as revealed by the battery of measures. This advantage is balanced by the potential difficulty of interpreting the nature of the difference detected. It may be that all outcome measures 'are better' in one group in which case common sense prevails. Practical experience reveals this is often not so simple and experience is needed in interpretation. This is in part the reason that they are perhaps not so widely used in clinical trials. Further, it is not so easy to define criteria of effectiveness in advance for inclusion in a protocol. Many of these multivariate statistical procedures are now included in widely available statistical packages but advice must be to use them with caution unless experienced help is to hand.

### 5.2.4 Cautionary Examples

Andersen (1990) reports several examples of ignoring the problems of multiplicity. First, (ref: Br J Clin Pharmacol [Suppl.], 1983, **16**: 103) a study of the effect of midazolan on sleep in insomniac patients presented a table of 2×9 tests of significance on measures of platform balance (seconds off balance) made at various times. The case of measuring the same outcome at successive times is a common one which requires a particular form of multivariate analysis termed ***repeated measures analysis***.

Next, (ref: Basic Clin Med 1981, **15**: 445) a report of a new compound to treat rheumatoid arthritis evaluated in a double-blind controlled clinical trial, indomethacin being the control treatment. Andersen reports that there were several criteria for effect (i.e. end-points), repeated at various timepoints and various subdivisions. A total of 850 pairwise comparisons were made (t-tests and Fisher's exact test in 2×2 contingency tables) and 48 of these gave p-values < 0.05. If there were no difference in the treatment groups and 850 tests were made then one might expect that 5% of these would shew 'significant' results. 5% of 850 = 850/20 = 42.5 so finding 48 is not very impressive.

Andersen quotes *The Lancet* (1984, **ii**: 1457) in relation to measuring everything that you can think of (or 'casting your net widely') as saying "Moreover, submitting a larger number of factors to statistical examination not only improves your chances of a positive result but also enhances your reputation for diligence".

## 5.3 Subgroup analyses

### 5.3.1 Fundamentals

Problems of multiplicity arise when separate comparisons are made within each of several subgroups of the subjects, for example when the sample of patients is subdivided on baseline factors, e.g. on gender and age for example resulting in four subgroups: (i) M>50; (ii) F>50; (iii) M≤50 & (iv) F≤50. Just as with multiple end-points, the chance of picking up an effect when none exists increases with the number of subdivisions.

Often subgroups are quite naturally considered and there are good *a priori* reasons for investigating them. If so, then this would of course be recorded in the protocol. If the subgroups are only investigated when an overall analysis gives a non-significant result and so subgroups are dredged to retrieve a significant result (as in the Zodiac example) then extreme care is needed to avoid charges of dishonesty. A safe procedure is only to use [post-hoc] subgroup analyses to suggest future hypotheses for testing in a later study.

**Remedy:**

- ♦ Bonferroni adjustments
- ♦ Analysis of Variance
- ♦ Follow-up tests for multiple comparisons

Bonferroni adjustments can be used but suffer from the same element of conservatism as in other cases but not so acutely since typically tests on separate subgroups are independent (unlike tests on multiple end-points).

The recommended routine remedy is to perform an Analysis of Variance (ANOVA) to investigate differences between the subgroups and then follow up the result of this (if a significant result is detected) to determine which subgroups are ‘interesting’. A one-way analysis of variance can be thought of as a generalisation to several samples of a two-sample t-test to test for the differences between several subgroups. The test examines the null hypothesis that all subgroups have the same mean against the alternative that at least one of them is different from the rest. The rationale for performing this as a preliminary is that if you think that the effect (e.g. a treatment difference) may only be exhibited in one of several subgroups then it means that one (or more) of the subgroups is different from the rest and so it makes sense to examine the statistical evidence for this. Follow-up tests can then be used to identify which one is of interest. There are many possible follow-up tests which are designed to examine slightly different situations. Examples are Tukey’s multiple range test which examines whether the two most different means are ‘significantly different’, Dunnett’s test which examines whether any particular group mean is ‘significantly different’ from a control group, the Neuman-Keuls test which looks to see which pairs of treatments are different and there are many others which may be found in commonly used statistical packages.

## 5.3.2 Example: Zodiac (Con$^{t.}$)

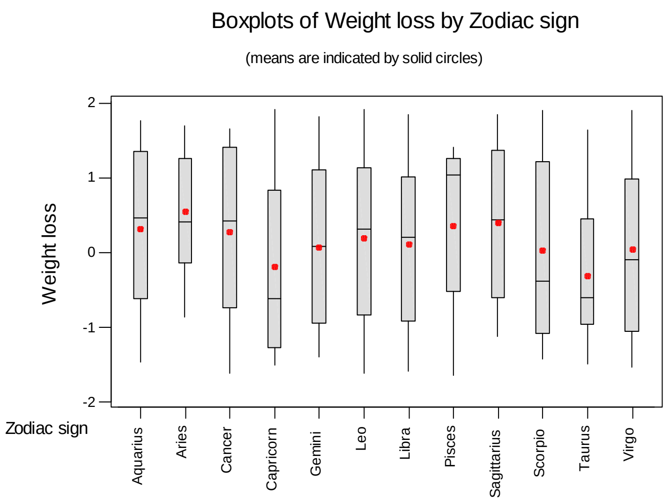
returning yet gain to the signs of the Zodiac example the appropriate analysis when the subjects are classified by Zodiac sign is to perform a one-way analysis of variance of the weight losses with the Zodiac sign as the classification variable. the analysis presented here is performed in Minitab but other packages would (should) give identical results:

**One-way ANOVA: Weight loss versus Zodiac sign**

```
Analysis of Variance for Weight loss
Source     DF        SS        MS        F        P
Zodiac s   11     13.44      1.22     1.05    0.405
Error     238    277.49      1.17
Total     249    290.93
                                       Individual 95% CIs For Mean
                                          Based on Pooled StDev
                                     -0.60      0.00      0.60      1.20
Level         N       Mean     StDev  ---+---------+---------+---------+---
Aquarius     26     0.313     1.106             (------*------)
Aries        15     0.543     0.794               (--------*--------)
Cancer       21     0.271     1.140            (-------*------)
Capricorn    27    -0.191     1.155     (------*------)
Gemini       18     0.068     1.128        (-------*-------)
Leo          22     0.194     1.096           (------*-------)
Libra        26     0.108     1.105          (------*------)
Pisces       19     0.362     1.010             (-------*-------)
Sagittarius  12     0.403     1.018           (----------*---------)
Scorpio      20     0.030     1.226        (-------*------)
Taurus       22    -0.315     0.860  (-------*------)
Virgo        22     0.044     1.117        (-------*------)
                                   ---+---------+---------+---------+---
Pooled StDev =    1.080            -0.60      0.00      0.60      1.20
```

This shews that the overall p-value for testing for a difference between the means of the twelve groups is 0. 405 >> 0.05 (i.e. non-significant).

The sketch confidence intervals for the means give an impression that the interval for the mean weight loss for Aries just about excludes zero but this makes no allowance for the fact that this is the most extreme of twelve independent intervals. The box pot on the next page gives little indication that any mean is different from zero:



Here the grey boxes indicate inter-quartile ranges (i.e. the 'middle half').

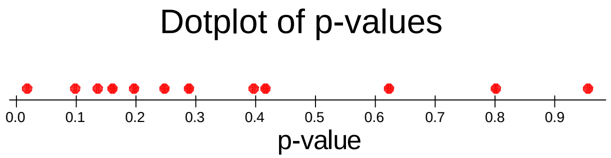
At this stage one would stop since there is no evidence of any difference in mean weight loss between the twelve groups but for illustration if we arbitrarily take the final sign (Virgo) as the 'control' and use Dunnett's test to compare each of the others with this then we obtain

```
Dunnett's comparisons with a control
    Family error rate = 0.0500        Individual error rate = 0.00599
Critical value = 2.77:   Control = level (Virgo) of Zodiac sign:
Intervals for treatment mean minus control mean

Level          Lower  Center  Upper   ------+---------+---------+---------+-
Aquarius      -0.598   0.269  1.136          (---------*----------)
Aries         -0.503   0.500  1.502           (-----------*------------)
Cancer        -0.686   0.227  1.141        (-----------*----------)
Capricorn     -1.095  -0.235  0.625   (----------*----------)
Gemini        -0.927   0.024  0.976     (-----------*-----------)
Leo           -0.753   0.150  1.053        (----------*----------)
Libra         -0.803   0.064  0.931       (----------*----------)
Pisces        -0.620   0.318  1.256         (-----------*-----------)
Sagittarius   -0.716   0.359  1.433        (------------*-------------)
Scorpio       -0.939  -0.014  0.911     (-----------*----------)
Taurus        -1.261  -0.359  0.544 (-----------*----------)
                                     ------+---------+---------+---------+-
                                         -0.80      0.00      0.80      1.60
```

This gives confidence intervals for the difference of each mean from that of the Virgo group, making proper allowance for the multiplicity and it is seen that all of these comfortably include zero so indicating that there is no evidence of any difference when due allowance is made for the multiple comparisons.

Another useful technique in this situation is to look at the twelve p-values associated with the twelve separate tests. If there were any underlying evidence that some groups were shewing an effect then some of them would be clustered towards the lower end of the scale from 0.0 to 1.0 (the values are given in the table on P5).



This shews that the values are reasonably evenly spread over the range from 0.0 to 1.0 and in particular that the lowest one is not extreme from the rest.

## 6.3.3 More Cautionary Examples

First, a report of an actual clinical double-blind study where two treatments were compared and there was an extra unusual element of blinding in that in fact the two treatments were actually identical, see Lee, McNear et al (1980), *Circulation.*

1073 patients with coronary heart disease were randomized into group 1 and group 2, baseline factors were reasonably balanced. The response was survival time and on initial analysis the overall differences between treatment groups non-significant.

Then subgroup analyses were performed: 6 groups were identified on the basis of 2 baseline factors (left ventricular contraction pattern:- normal/abnormal; number diseased vessels 1/2/3). A significant difference in survival times was found in one of the groups (abnormal/3, $\chi^2$=5.4, p<0.023) and could be justified scientifically. Sample sizes were quite large:–

$$n=397: \qquad n_1=194, \qquad n_2=203$$

In fact, all patients were treated in the **SAME** way — the ‘treatment’ corresponded to the random allocation into 2 groups. Thus a ***false positive effect*** had been discovered.

## 5.4 Interim analyses

### 5.4.1 Fundamentals

It may be desirable to analyse the data from a trial periodically as it becomes available and again problems of multiple testing arise.

Here the remedies are rather different (and considerably more complex) since not only are the sequence of tests not independent but successive tests are based on accumulating data, i.e. the data from the first period test are pooled into that collected subsequently and re-analyzed with the newly obtained values.

The main objectives of this periodic checking are:–

- To check protocol compliance, e.g. compliance rate may be very low. Check that investigators are following the trial protocol and quick inspection of each patient's results provides an immediate awareness of any deviations from intended procedure. If early results indicate some difficulties in the compliance it may be necessary to make alterations in the protocol.
- To pick up bad side effects so that quick action can be taken and warn investigators to look out for such events in future patients.

- Feedback:– helps maintain interest in trial and satisfy curiosity amongst investigators. Basic pre-treatment information such as numbers of patients should be available. Overall data on patient response and follow up for all treatments combined can provide a useful idea of how the trial is proceeding.
- Detect large treatment effects quickly so one can stop or modify trial.

The primary reason for monitoring trial data for treatment differences is the ethical concern to avoid any patient in the trial receiving a treatment known to be inferior. In addition, one wishes to be efficient in the sense of avoiding unnecessary continuation once the main treatment differences are reasonably obvious.

However, multiplicity problems exist here too. We have repeated significance tests although not independent — so the overall significance level will be much bigger than the nominal level of $\alpha$ used in each test.

### 5.4.2 Remedy:

To incorporate such interim analyses we must:–

- build them into the protocol (e.g. a group sequential design)
- reduce the nominal significance level of each test, so overall level is required $\alpha$

However, if we use the standard Bonferroni adjustment then we obtain very conservative procedures for exactly the same reasons as detailed in earlier sections. Instead we need refined calculations for the appropriate nominal p-values to use at each step to achieve a desired overall significance level. These calculations are different from those given earlier since there the tests were assumed entirely independent; here they assume that the data used for the first test is included in that for the second, both sets in that for the third etc. (i.e. accumulating data) — the exact calculations are complicated. The full details are given in Pocock (1983) and summarized from there in the tables below:–

**Repeated significance tests on accumulating data**

| Number of repeated tests at the 5% level | overall significance level |
|---|---|
| 1 | 0.05 |
| 2 | 0.08 |
| 3 | 0.11 |
| 4 | 0.13 |
| 5 | 0.14 |
| 10 | 0.19 |
| 20 | 0.25 |
| 50 | 0.32 |
| 100 | 0.37 |
| 1000 | 0.53 |
| ∞ | 1.0 |

**Nominal significance levels required for repeated two-sided significance testing for various α**

| N | α=0.05 | α=0.01 |
|---|---|---|
| 2 | 0.029 | 0.0056 |
| 3 | 0.022 | 0.0041 |
| 4 | 0.018 | 0.0033 |
| 5 | 0.016 | 0.0028 |
| 10 | 0.0106 | 0.0018 |
| 15 | 0.0086 | 0.0015 |
| 20 | 0.0075 | 0.0013 |

Here N is the maximum number of interim analyses to be performed and this is decided in advance (and included in the protocol of course).

## 5.4.3 Yet More Cautionary Examples

First an example quoted by Pocock (1983, p150). This is a study to compare of drug combinations CP and CVP in non-Hodgkins lymphoma. The measure was occurrence or not of tumour shrinkage. The trial was over 2 years and likely to involve about 120 patients. Five interim analyses planned, roughly after every 25th result. The table below gives numbers of ‘successes’ and ***nominal*** p-values using a $\chi^2$ test at each stage.

| | **response rates** | | |
|---|---|---|---|
| **Analysis** | **CP** | **CVP** | **statistic & p-value** |
| **1** | 3/14 | 5/11 | 1.63 (p>0.20) |
| **2** | 11/27 | 13/24 | 0.92 (p>0.30) |
| **3** | 18/40 | 17/36 | 0.04 (p>0.80) |
| **4** | 18/54 | 24/48 | 3.25 (0.05<p<0.1) |
| **5** | 23/67 | 31/59 | 4.25 (0.025<p<0.05) |

**Conclusion**: Not significant at end of trial (overall p>0.05) since p>0.016, the required nominal value for 5 repeat tests (see table above).

### 5.4.3.1 Notes:–

- If there had been **NO** interim analyses and only the final results available then the conclusion would have been **different** and CVP declared significantly better at the 5% level.
- In the early stages of any trial the response rates can vary a lot and one needs to avoid any over reaction to such early results on small numbers of patients. For instance, here the first 3 responses occurred on CVP but by the time of the first analysis the situation had settled down and the $\chi^2$ test showed no significant difference. By the fourth analysis, the results began to look interesting but still there was insufficient evidence to stop the trial. On the final analysis, when the trial was finished anyway, the $\chi^2$ test gave p=0.04 which is not statistically significant, being greater than the required nominal level of 0.016 for N=5 analyses.

A totally negative interpretation would not be appropriate from these data alone. One could infer that the superiority of the CVP treatment is interesting but not conclusive.

Next, an example quoted by Andersen (1990), (ref: Br J Surg, (1974), **61**: 177). “A randomized trial of Trasylol in the treatment of acute pancreatitis was evaluated statistically when 49 patients had been treated. No statistically significant difference was evident between the two groups, but a trend did emerge in favour of one group. The trial was therefore continued. When altogether 100 cases had been treated, the data were analyzed again. There was now a significant difference ($\chi^2$ = 4.675, d.f. = 1, p< 0.05) and the trial was published.”

In fact the p-value is 0.031and even if only two interim analyses (including the final one) had been planned this is greater than the necessary 0.029 to claim 5% significance.

Continuing to collect data until a significant result is obtained is clearly dishonest — eventually an apparently significant result will be obtained.

### 5.4.3.2 Further Notes:–

- One decides in advance what is expected as the maximum number of interim analyses and accordingly makes the nominal significance level smaller. e.g. with at most 10 analyses and overall type I error = 0.05 one uses p<0.0106 as the stopping rule at each analysis for a treatment difference. One should also consider whether an overall type I error $\alpha$=0.05 is sufficiently small when considering a stopping rule. There are 2 situations where $\alpha$=0.01 may be more appropriate:

  i) if a trial is unique in that its findings are unlikely to be replicated in future research studies

  ii) if there is more than one patient outcome used in interim analyses and stopping rule is applied to each outcome. However, one possibility would be to have one principal outcome with a stopping rule having $\alpha$=0.05 and have lesser outcomes with $\alpha$=0.01. It has been suggested that a very stringent stopping criterion, say p<0.001, should be used, on the basis that no matter how often one performs interim analyses the overall type error will remain reasonably small. It also means that the final analysis, if the trial is not stopped early, can be interpreted using standard significance tests without any serious need to allow for earlier repeated testing.

- See Pocock (1983) for more detail.

## 5.5 Repeated Measures

### 5.5.1 Fundamentals

Repeated measures arise when the same feature on a patient is measured at several time points, e.g. blood concentration of some metabolite at baseline and then at intervals of 1, 3, 6, 12 and 24 hours after ingestion of a drug. If, for example, there are two groups of subjects (e.g. two treatment groups) it is tempting to use two-sample t-tests on the measures at each time point in sequence. Of course this is incorrect unless adjustments are made. However, diagrams which shew mean values of the two treatment groups plotted against time and which ***shew error bars for each mean*** invite the eye to do exactly that and this must be resisted.

**Remedies:**

- Bonferroni adjustments
- Multivariate analysis for repeated measures
- Construction of summary measures.

No essentially new comments apply to this situation and indeed some examples discussed earlier include a repeated measure element. Bonferroni adjustments are very conservative since the tests will be highly correlated (as with multiple end-points).

Multivariate analysis of repeated measures can take advantage of the fact that the observations are obtained in a sequence and it may be possible to model the correlation structure.

There are special techniques which do this and specialist or professional advice should be sought. Some so-called ‘repeated measures analyses’ in some statistical packages are in fact quite spurious.

Calculation of summary measures includes calculating quantities such as ‘area under the curve’ (AUC) which may have an interpretation as reflecting bioavailability, another is concentrating on change from baseline. As always, the form of the analysis should be fixed before collection of the data.

## 5.6 Miscellany

### 5.6.1 Regrouping

The example below illustrates the dangers of post-hoc recombining subgroups, perhaps a complementary problem to that of post-hoc dividing into subgroups. The example is taken from Pocock (1983) who quotes Hjalmarson *et* a*l* (1981), The Lancet, **ii**: 823. The table gives the numbers of deaths or survivals in 90 days after acute myocardial infarction with the subgroup for age-group 65-69 combined first with the older subgroup and then with the younger one. For this subgroup the death rates on placebo and metoprolol were 25/174 (14.4%) and 17/165 (6.7%) respectively.

| | placebo | metoprolol | |
|---|---|---|---|
| deaths | 62/697 (8.9%) | 40/698 (5.7%) | p<0.02 |
| age 40–64 | 26/453 (5.7%) | 21/464 (4.5%) | p>0.2 |
| age 65–74 | 36/244 (14.8%)<br>**Metoprolol better for elderly?** | 19/234 (8.1%) | p=0.03 |
| age 40–69 | 51/627 (8.1%) | 32/629 (5.1%) | p=0.04 |
| age 70–74 | 11/70 (15.7%)<br>**Metoprolol better for younger?** | 8/69 (11.6%) | p>0.2 |

As well as the dangers of multiple testing, this example illustrates the dangers of post-hoc re-grouping, subgroups should be defined on clinical grounds before the data are collected.

Some subgroup effects could be real of course. However, we should only use subgroup analyses to generate future hypotheses.

### 5.6.2 Multiple Regression

A further situation where multiplicity problems arise in a well-disguised form and which is often ignored is in large regression analyses involving many explanatory variables. This applies whether the model is ordinary regression with a quantitative response or whether it is a logistic regression for success/failure data or even a Cox proportional hazards regression for survival data.

When analysing the results of estimating such models it is usual to look at estimates of the individual coefficients in relation to their standard errors, declare the result 'significant' at the 5% level if the ratio is more than 1.96 (or 2) in magnitude and conclude that the corresponding variable 'is important' in affecting the response. It is customary for problems of multiplicity to be ignored on the grounds that although there are several or even many separate (non-independent) t-tests involved, each of the variable si of interest in its own right and that is why it was included in the analysis.

However, there are situations where the regression analysis is more of a fishing expedition and it is more a case of 'lets plug everything in and see what comes out', effectively selecting the most significant result for attention.

If this is the case then an honest analysis would have to include this feature and make an appropriate correction, such as a Bonferroni one. Beware especially of including an ***interaction*** in a model when there is no *a priori* reason to expect it.

#### 5.6.2.1 Example: shaving & risk of stroke

In the Autumn of 2003 it was reported widely in the media that men who did not shave regularly were '70% more likely to suffer a stroke and 30% more likely to suffer heart disease, according a study at the University of Bristol'. This is an eye-catching item and so was easily accepted as true.

It is likely that these conclusions were based on a logistic regression model, looking at the probability of suffering a stroke, or on some similar regression model. However, it is of importance to know whether firstly there was any *a priori* medical hypothesis that suggested that diligence in shaving was a feature to be investigated and secondly how many other variables were included in the study. The exact reference for this study is Shaving, Coronary Heart Disease, and Stroke: The Caerphilly Study Ebrahim et al. *Am. J. Epidemiol.*2003; 157: 234-238, see http://aje.oxfordjournals.org/cgi/content/full/157/3/234 , and you are invited to read this article critically.

## 5.7 Summary and Conclusions

Multiplicity can arise in

- testing several different responses
- subgroup analyses
- interim analyses
- repeated measures
- &c.

The effect of multiplicity is to increase the **overall risk of a false positive (i.e. the overall significance level).**

Problems of multiplicity can be overcome by

- Bonferroni corrections to nominal significance levels
- Other adjustments to nominal significance levels in special cases, e.g. for accumulating data in interim analyses where adjusting for multiplicity can have counter-intuitive effects.
- more sophisticated analyses, e.g. ANOVA or multivariate methods.

Bonferroni adjustments are typically very conservative because in many situations the tests are highly correlated (especially with multiple end-points and repeated measures).

Conservative means 'safe' — i.e. you preserve your scientific reputation by avoiding making mistake but at the expense of failing to discover something scientifically interesting.
A final comment is to remember that

*"If you torture the data often enough it will eventually confess"*

# 6. Crossover Trials

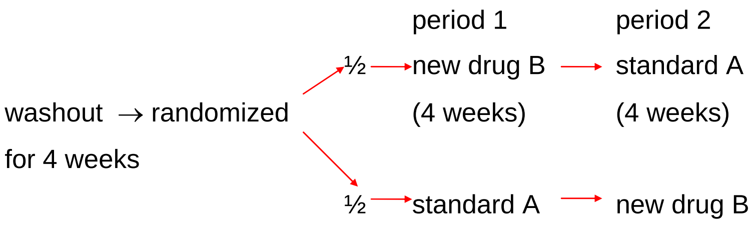
## 6.1 Introduction

Where it is possible for patients to receive both treatments under comparison, crossover trials may well be more efficient (i.e. need <u>fewer patients</u>) than a parallel group study.

Recall idea from section 2.: by acting as his/her own control, the effect of large differences <u>between patients</u> can be lessened by looking at <u>within patient</u> comparisons.

Example 6.1 (Pocock, p112)

Hypertension trial:



period 1 period 2

washout → randomized
for 4 weeks

½ → new drug B → standard A
(4 weeks) (4 weeks)

½ → standard A → new drug B

Response is systolic blood pressure at end of 5 minute exercise test.

B → A: 55 patients, A → B: 54 patients.

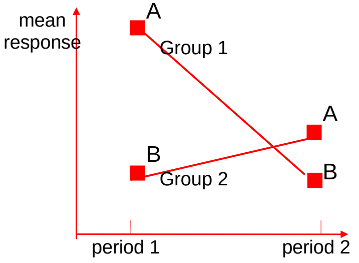
| | | |
|---|---|---|
| Possible effect: | treatment effects | $\tau$ |
| | period effect | $\pi$ |
| | carryover effect | $\lambda$ |

## 6.2 Illustration of different types of effects

Note: assuming that 'low' is good throughout

a) <u>Carryover effect</u>

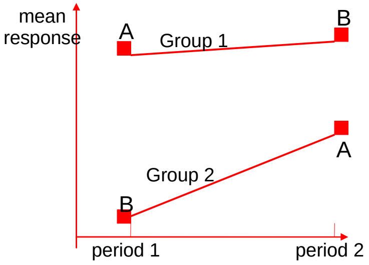
(i)



possible explanation:
beneficial effect of B carries over into period 2

<u>Carryover effect</u>

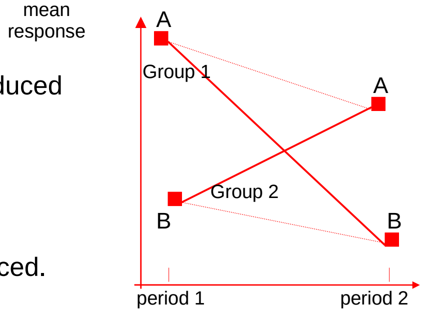
(ii)



<u>Direction</u> of treatment effect different for different periods caused by carryover.

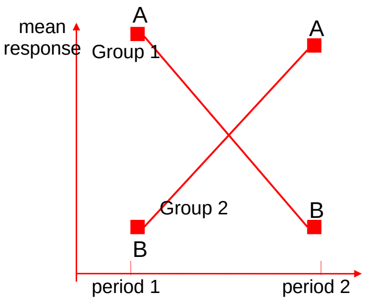
(ii) is more serious, (i) is unlikely to be detected because of low power.

b) Period effect



response in period 2 reduced for both treatments, i.e. patients generally improve so period 2 values on average reduced.

c) treatment effect



B better than A

## 6.3 Model

| | period 1 | | period 2 | |
|---|---|---|---|---|
| group 1 | A | $Y_{11k}$ | B | $Y_{12k}$ |
| group 2 | B | $Y_{21k}$ | A | $Y_{22k}$ |

response $Y_{ijk}$ for

group i (order); i=1,2

period j; j=1,2

patient k; k=1,2,...,$n_i$ . ($n_1$=$n_2$ in balanced case)

Effects

$\mu$ — overall mean

$\tau_A$, $\tau_B$ — treatment effects

$\pi_1$, $\pi_2$ — period effects

$\lambda_A$, $\lambda_B$ carryover effects (treatment x period interaction)

$\alpha_k$ — random patient effect ~ N(0,$\phi^2$) (between patients)

$\varepsilon_{ijk}$— random errors ~ N(0, $\sigma^2$) (independently)

Identifiability

$$\boxed{\begin{array}{c} \tau_A + \tau_B = 0 \\ \pi_1 + \pi_2 = 0 \end{array}}$$

Model

|  | period 1 | period 2 |
|---|---|---|
| group 1 | $\mu+\alpha_k+\tau_A+\pi_1+\varepsilon_{11k}$ | $\mu+\alpha_k+\tau_B+\pi_2+\lambda_A+\varepsilon_{12k}$ |
| group 2 | $\mu+\alpha_k+\tau_B+\pi_1+\varepsilon_{21k}$ | $\mu+\alpha_k+\tau_A+\pi_2+\lambda_B+\varepsilon_{22k}$ |

If we take expected values, $\alpha_k$ and $\varepsilon_{ijk}$ disappear.

$Y_{ijk} = \mu+\alpha_k+\tau+\pi+\lambda+\varepsilon_{ijk}$

$E(Y_{11k}) = \mu+\tau_A+\pi_1$

$E(Y_{12k}) = \mu+\tau_B+\pi_2+\lambda_A$

To isolate $\tau$, $\pi$ and $\lambda$ effects we consider sums and differences of the $Y_{ijk}$'s.

### 6.3.1. Carryover effect

Compute $T_{ik} = ½(Y_{i1k} + Y_{i2k})$ i.e. the average of the 2 values for patient k.

Then $T_{1k} \sim N(\mu+½\lambda_A, \phi^2+½\sigma^2)$ and $T_{2k} \sim N(\mu+½\lambda_B, \phi^2+½\sigma^2)$

If $\lambda_A = \lambda_B$ i.e. no (differential) carryover, $T_{1k}$ and $T_{2k}$ have identical Normal distributions.

Thus we can test for equality of means of group 1 and group 2 using a 2-sample t-test to establish whether

$H_0$: $\lambda_A = 0 = \lambda_B$ is plausible.

$$\text{i.e. use} \frac{\bar{T}_1 - \bar{T}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_r$$

where $s_1^2$ is the sample variance of the $T_{1k}$ so $\hat{\text{var}}(\bar{T}_1) = \frac{s_1^2}{n_1}$, etc. and we take [conservatively] $r=\min(n_1, n_2)$ or use a more sophisticated formula.

[Note that our model does specify equal variances and so we could use the 'pooled variance version' of the t-test

$$\frac{\bar{T}_1-\bar{T}_2}{\sqrt{\hat{\text{var}}(\bar{T}_1-\bar{T}_2)}} \sim t_{n_1+n_2-2}$$

where $\hat{\text{var}}(\bar{T}_1-\bar{T}_2)=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ but it should make little difference in practice.

Ex 6.1 (continued)

| | B → A | A → B |
|---|---|---|
| $n_i$ | 55 | 54 |
| $\bar{T}_i$ | 176.28 | 180.17 |
| $s_i$ | 26.56 | 26.27 |

so $t=\frac{180.17-176.28}{\sqrt{\frac{26.27^2}{54}+\frac{26.56^2}{55}}}=0.769$ which is clearly non-significant when compared with $t_{54}$ and so the data provide no evidence of a carry-over effect.

NB 'pooled' 2-sample $t=\frac{180.17-176.28}{\sqrt{\frac{54\times26.27^2+53\times26.56^2}{107}\times\left(\frac{1}{55}+\frac{1}{54}\right)}}=0.769$

(little difference because the variances are almost equal anyway)

### 6.3.1.1 Notes

- Test for carryover typically has low power since it involves **between** patient comparisons.

- If there is a significant carryover effect (i.e. treatment x period interaction) then it is NOT SENSIBLE to test for period and treatment separately, so
  a) plot out means and inspect
  b) just use first period results and
  compare A and B as a parallel group study.

- If just first period results are used then the treatment comparison is **between** patients (so also of low power).

- If there is a carryover then it means that the results of the second period are 'contaminated' and give no useful information on treatment comparisons — the trial should have been designed with a longer washout period.

- **NB** we used the *average* of the two values for each patient (i.e. from period 1 and period 2) in describing the carryover test since then the model indicates this has a mean of $\mu$ when there is no carryover. The value of the t-statistic would be ***exactly*** the same if we used just the sum of the two period values — this is easier (avoids dividing by 2!) and this will be the procedure in later examples.

### 6.3.2 Treatment & period effects

Consider $D_{ik} = Y_{i1k} – Y_{i2k}$ i.e. within subject differences.

Then $D_{1k} \sim N((\tau_A\text{-}\tau_B)+(\pi_1\text{-}\pi_2), 2\sigma^2)$ group 1 and

$D_{2k} \sim N((\tau_B\text{-}\tau_A)+(\pi_1\text{-}\pi_2), 2\sigma^2)$ group 2

#### 6.3.2.1 Treatment test

$H_0$: $\tau_A = 0 = \tau_B$

If this is true, then $D_{1k}$ and $D_{2k}$ have identical distributions so we can test $H_0$ by a t-test for equality of means as before.

$$\frac{\bar{D}_1 - \bar{D}_2}{\sqrt{\frac{s^2_{D1}}{n_1} + \frac{s^2_{D2}}{n_2}}} \sim t_r$$

where now $s^2_{D1}$ is the sample variance of the ***differences*** $D_{1k}$.

Notice that $\bar{D}_1$ is the difference between *period 1* and *period 2* results averaged over those in group 1 and $\bar{D}_2$ is the difference between *period 1* and *period 2* results averaged over those in group 2. Thus this test can be regarded as a *two-sample t-test on period 1 – period 2 differences between the two groups of subjects.*

Ex 6.1 (continued again)

| | B → A | A → B |
|---|---|---|
| $n_i$ | 55 | 54 |
| $\bar{D}_i$ | 5.04 | –2.81 |
| $s_i$ | 15.32 | 19.52 |

We have $t = \dfrac{5.04 - (-2.81)}{\sqrt{\frac{15.32^2}{55} + \frac{19.52^2}{54}}} = 2.33$

so p=0.024 when compared with $t_{54}$ — significant evidence of treatment effects.

[The pooled t-statistic is $t = \dfrac{5.04 - (-2.81)}{\sqrt{\frac{54\times15.32^2+53\times19.52^2}{107} \times \left(\frac{1}{55} + \frac{1}{54}\right)}} = 2.34$ with a p-value of 0.021 when compared with $t_{107}$ (i.e. no material or practical difference)]

### 6.3.2.2 Period test

$$H_o: \pi_1 = 0 = \pi_2$$

If $H_0$ is true then $D_{1k}$ and $-D_{2k}$ will have identical distributions and so the test will be based on

$$\frac{\overline{D}_1 - (-\overline{D}_2)}{\sqrt{\frac{s_{D1}^2}{n_1} + \frac{s_{D2}^2}{n_2}}} \sim t_r$$

**NB** it is + in the numerator (not –) since it is still a 2-sample t-test of 2 sets of numbers the $\{(Y_{11k} - Y_{12k});\ k=1,\ldots,n_1\}$ from group 1 and the $\{(Y_{21k} - Y_{22k});\ k=1,\ldots,n_2\}$ from group 2.

Notice that $\overline{D}_1$ is the difference between *Treatment A* and *Treatment B* results averaged over those in group 1 and $(-\overline{D}_2)$ is the difference between *Treatment A* and *Treatment B* results averaged over those in group 2. Thus this test can be regarded as a *two-sample t-test on Treatment A – Treatment B differences between the two groups of subjects.*

Ex 7.1 (continued yet again)

We have $t = \dfrac{5.04 - (+2.81)}{3.365} = 0.66$

so no significant evidence of a period effect.

[Same conclusion from the pooled test]

## 6.4★ Analysis with Linear Models

### 6.4.0★ Introduction

The analyses presented above using carefully chosen t-tests provide an illustration of the careful use of an underlying model in selecting appropriate tests to examine hypotheses of interest. However, to extend the ideas to more complicated cross-over trails with more treatments and periods it is necessary to use a more refined analysis with linear models. The basic model for a multi-period multi-treatment trial for the response of patient k to treatment i in period j is:

$$Y_{ijk} = \mu + \tau_i + \pi_j + \lambda_{ij} + \alpha_k + \varepsilon_{ijk}$$

where $\varepsilon_{ijk} \sim N(0, \sigma^2)$, $\alpha_k \sim N(0, \phi^2)$, $\Sigma\alpha_i = \Sigma\tau_j = \Sigma\Sigma\lambda_{ij} = 0$ and where $\lambda_{ij}$ denotes the carryover effect which mathematically is identical to an interaction between the factors treatment and period. Note that this model is slightly different from that given in §7.3 where the suffix i was used to indicate which group a patient belonged to and here it denotes the treatment received. The essence of a cross-over trial is that not all combinations of i, j and k are tested. For example in a trial with two periods and two treatments only about half of the patients will receive treatment 1 in period 1 and for others the combination i = j = 1 will not be used. Since the patient effect $\alpha_k$ is specified as a random variable this is strictly a *random effects model* which is a topic covered in the second semester in MAS473/6003 so we present first an approximate analysis with a *fixed effects model* which alters the assumption that the $\alpha_k$ are random variables and instead have the identifiability constraint $\Sigma\alpha_k = 0$.

### 6.4.1* Fixed effects analysis

The data structure presumed is that the dataframe consists of variable response with factors treatment, period and patient. Dataframes provided in the example data sets with this course are generally not in this form. Typically, in the example data sets the responses in the two periods are given as separate variables so each record consists of responses to one subject, which is convenient for performing the two sample t-tests described in earlier sections and these will require some manipulation.

The R analysis is then provided by:

```
> crossfixed<-
lm(result ~ period + treatment + patient +
   treatment:period)
> anova(crossfixed)
```

This will give an analysis of variance with entries for testing with F-tests differences between periods, treatments and the carryover (i.e. treatment×period interaction). The p-values will be almost the same as those from the separate t-tests and will be identical if non-default pooled variance t-tests are used by including `var.equal = TRUE` in the `t.test(.)` command.

Strictly speaking it has been presumed here that the numbers of subjects allocated to the various groups receiving treatments in the various orders have ensured that the factors period and treatment are orthogonal (e.g. equal number to two groups in a 2 periods 2 treatments trial). If this is not the case then the above analysis of variance will give a ‘periods ignoring treatments’ sum of squares and a ‘treatments adjusted for periods’ sum of squares. This aspect of the analysis may be discussed more fully in a course on random and mixed effects linear models.

### 6.4.2* Random effects analysis

The same data structure is used and here the library `nlme` for random effects analysis is required and a random effects linear model is fitted with `lme(.)`

The R analysis is then provided by:

```
> library(nlme)
> crossrandom<-
   lme(result ~ period + treatment
    + treatment:period, random = ~ 1|patient)
> anova(crossrandom)
```

The analysis of variance table will usually be very similar to that provided by the fixed effects model except that the standard errors of estimated parameters will be a little larger (to allow for the additional randomness introduced by regarding the patients as randomly selected from a broader population) and consequently the p-values associated with the various fixed effects of treatment, period and interaction will be a little larger (i.e. less significant).

### 6.4.3* Deferment of example

An example is not provided here but analyses using the two forms of model will be given on the hours sleep data used in Q2 on Task Sheet 4.

## 6.5 Notes

- If there is a substantial period effect, then it may be difficult to interpret any overall treatment difference within patients, since the observed treatment difference in any patient depends so much on which treatment was given first.

- Some authors (e.g. Senn, 2002) strongly disagree with the advisability of performing carryover tests. In part, the argument is based upon the difficulty introduced by a two-stage analysis, i.e. where the result of the first stage (a test for carryover) determines the form of the analysis for the second stage (i.e. whether data from both periods or just the first is used). This causes severe inferential problems since strictly the second stage is conditional upon the outcome of the first. In practice, most pharmaceutical companies rely upon medical considerations to eliminate the possibility of any carryover of treatments. In any case, the test for carryover typically has low power needs to be supplemented by medical knowledge — i.e. need expert opinion that either the two treatments cannot interact or that the washout period is sufficient, cannot rely purely on statistical evidence.

- We can obtain confidence intervals for treatment differences since $\tfrac{1}{2}(\bar{D}_1 - \bar{D}_2) \sim N(\tau_A - \tau_B, \tfrac{1}{2}\sigma^2(n_1^{-1} + n_2^{-1}))$ and estimate $\sigma^2$ with a pooled variance estimate or else say that the standard error of $\tfrac{1}{2}(\bar{D}_1 - \bar{D}_2)$ is $\sqrt{\tfrac{1}{4}\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$ and use the approximate formula for [say] a 95% CI of $\tfrac{1}{2}(\bar{D}_1 - \bar{D}_2) \pm 2\times\text{s.e.}\{\tfrac{1}{2}(\bar{D}_1 - \bar{D}_2)\}$ (2 rather than 1.96 is adequate given the approximations made anyway in assuming normality etc).

- If it is unsafe to assume normality the various two-sample t-tests above can be replaced by non-parametric equivalents, e.g. a Wilcoxon-Mann-Whitney test.
  The simpler non-parametric test, a sign test, is essentially identical to the case of binary responses considered in §7.4 below.

- Sample size & efficiency of crossover trials:–
  it can be shown that the number of patients required in a crossover trial is $N = n(1–\rho)$ where n= number required in *each* arm of a parallel group study and $\rho$= correlation between the2 measurements on each patient (assuming no carryover effect). Since $\rho > 0$ usually, need fewer patients in a crossover than in a parallel group study. Sample size calculation facilities for cross-over trials are available in `power.exe` .

- Can be extended to > 2 treatments and periods, usually when intervals between treatments can be very short.

  e.g.

  |   | period |   |
  |---|---|---|
  | 1 | 2 | 3 |
  | A | B | B |
  | B | A | A |

  |   |   |   |
  |---|---|---|
  | A | B | C |
  | C | A | B |
  | B | C | A |

- In trials involving several treatments it is unrealistic to consider all possible orderings and so need ideas of *incomplete block designs* [balanced or partially balanced] to consider a balanced subset of orderings. (See MAS370 or MAS6011 second semester).

- Crossover trials are most suitable for short acting treatments where carryover effect is not likely, but usually not curative so baseline is similar in period 2.

## 6.6 Binary Responses

The analysis of binary responses introduces some new features but is essentially identical in logic to that of continuous responses considered above. The key idea is to consider *within subject* comparisons as before. This is achieved by considering whether the difference between the responses to the two treatments for the same subject indicates treatment A is ‘better’ or ‘worse’ than treatment B. If the responses on the two treatments are identical then that subject provides essentially no information on treatment differences.

### 6.6.1 Example: (Senn, 2002)

A two-period double blind crossover trial of 12μg formoterol solution compared with 200μg salbutamol solution administered to 24 children with exercise induced athsma. Response is coded as + and – corresponding to ‘good’ and ‘not good’ based upon the investigators overall assessment. Subjects were randomised to one of two groups: group 1 received the treatments in the order formoterol → salbutamol; group 2 in the order salbutamol → formoterol.

The results are given below:

| group | subject | formoterol | salbutamol | preference |
|---|---|---|---|---|
| group 1<br>for →sal | 1 | + | + | — |
| | 2 | – | – | — |
| | 3 | + | – | f |
| | 4 | + | – | f |
| | 5 | + | + | — |
| | 6 | + | – | f |
| | 7 | + | – | f |
| | 8 | + | – | f |
| | 9 | + | – | f |
| | 10 | + | – | f |
| | 11 | + | – | f |
| | 12 | + | – | f |
| group2<br>sal → for | 13 | + | – | f |
| | 14 | + | – | f |
| | 15 | + | + | — |
| | 16 | + | + | — |
| | 17 | + | + | — |
| | 18 | + | + | — |
| | 19 | + | + | — |
| | 20 | – | + | s |
| | 21 | + | – | f |
| | 22 | + | – | f |
| | 23 | + | – | f |
| | 24 | + | – | f |

To test for a difference between treatments we test whether the proportion of subjects preferring the ***first period treatment*** is associated with which order the treatments are given in, (c.f. performing a two sample t-test on the period 1 – period 2 responses). This test is sometimes known as the Mainland-Gart Test:

| | preference | | |
|---|---|---|---|
| sequence | first period | second period | total |
| for → sal | 9 | 0 | 9 |
| sal → for | 1 | 6 | 7 |
| total | 10 | 6 | 16 |

The value of the Pearson chi-squared test statistic is

$$(9\times6 – 1\times0)^2\times16/[10\times6\times7\times9] = 12.34$$

which is clearly significant at a level <0.001 and so the data provide strong evidence of superiority of the treatment by formoterol.

To test for a period effect we similarly test whether the proportion of subjects preferring treatment A is associated with the order in which the treatments are given:

| | preference | | |
|---|---|---|---|
| sequence | formoterol | salbutamol | total |
| for → sal | 9 | 0 | 9 |
| sal → for | 6 | 1 | 7 |
| total | 15 | 1 | 16 |

Now the test statistic is $(9\times1 - 6\times0)^2\times16/[15\times1\times7\times9] = 1.37$ and we conclude that there is no evidence of a period effect.

## 6.7 Summary and Conclusions

Possible effects that must be tested in a two-treatment two-period crossover trial (whether continuous or binary outcomes) are:

- **carryover**:– test by two-sample test on average response over both periods
- **treatment**:– test by two-sample test on differences of *period I – period II* results between the two groups of subjects
- **period**:– test by two-sample test on differences of *treatment A – treatment B* results between the two groups of subjects.

If carryover (i.e. treatment×period interaction) is present then use only results from period I, in which case treatment comparisons are **between** subjects. A full crossover analysis gives a **within** subject comparison.

- Use of a preliminary test for carryover is not recommended by some authorities and it is preferable to rely upon medical considerations to eliminate the possibility of a carryover.
- If normality is assumed then the tests can be performed with two sample t-tests. These can be replaced with non-parametric equivalents such as a Wilcoxon-Mann-Whitney test.
- binary responses can be analyzed with a Mainland-Gart test which considers only those subjects exhibiting different responses to the treatments.

# 7. Binary Response Data

## 7.1 Background

Responses are often measured on a binary or categorical scale. Here we only look a the binary case, so we can represent the response of the $i^{th}$ patient by $y_i = 1$ (success) or $y_i = 0$ (failure). We can use standard Pearson $\chi^2$ or Mantel-Haenszel tests but not all cross-classified tables are appropriate for application of these hypotheses tests of independence of classification or homogeneity. In some cases it is appropriate to consider different statistics calculated from the table to reflect on the key question of interest there are further techniques for special designs (e.g. paired observations) or observational studies or if we have additional data, e.g. on covariates (such as different centres).

## 7.2. Combining trials and the Mantel-Haenszel Test

We may have results from several trials or centres. How should we combine them?

e.g. For a binary response of treatment *vs* placebo

e.g. trial j (for j=1,2,.....,N)

| | Successes | Failures | |
|---|---|---|---|
| Treatments | $Y_{1j}$ | $n_{1j}–Y_{1j}$ | $n_{1j}$ |
| Placebo | $Y_{2j}$ | $n_{2j}–Y_{2j}$ | $n_{2j}$ |
| | $t_j$ | $n_j–t_j$ | $n_j$ |

It can be dangerous to collapse these N 2×2 separate tables into 1 single 2×2 table:

centre 1

| | S | F | |
|---|---|---|---|
| trt | 30 | 70 | 30%S |
| plac | 120 | 180 | 40%S |
| | 150 | 250 | |

centre 2

| | S | F | |
|---|---|---|---|
| trt | 210 | 90 | 70%S |
| plac | 80 | 20 | 80%S |
| | 290 | 110 | |

looks like **placebo** better?

($\chi^2$ = 3.2, n.s.)

looks like **placebo** better?

($\chi^2$ = 3.76, n.s.)

but if we collapse the two tables into one:

**centre 1 & 2**

| | S | F | |
|---|---|---|---|
| trt | 240 | 160 | 60%S |
| plac | 200 | 200 | 50%S |
| | 440 | 360 | |

It looks like the **treatment** is better; ($\chi^2$=8.08, highly significant)

This is known as **Simpson's Paradox** — it is misleading to look at margins of higher dimensional arrays, especially when there are imbalances in treatment numbers.

The root cause of the paradox here is that the overall success rates in the two centres is markedly different (30–40% in centre 1 but 70–80% in centre 2) so it is misleading to ignore the ***centre differences*** and add the results together from them.

## 7.3 Mantel-Haenszel Test

One way of combining data from such trials is using the Mantel-Haenszel test (***but this does not necessarily overcome Simpson's Paradox — it only avoids differences BETWEEN trials and assesses evidence WITHIN trials)***.

Consider a single 2×2 table:

| | Successes | Failures | |
|---|---|---|---|
| Treatments | $\mathbf{Y_1}$ | $n_1 – Y_1$ | $\mathbf{n_1}$ |
| Placebo | $Y_2$ | $n_2 – Y_2$ | $n_2$ |
| | **t** | n–t | n |

and assume $Y_i \sim B(n_i,\theta_i)$ ; i=1,2

interested in $H_0$: $\theta_1 = \theta_2$

Fisher's exact test considers

$P(y_1,y_2|y_1+y_2=t)$ i.e. conditions on the total number of successes

If $\theta_1 = \theta_2$ then $P(y_1,y_2|y_1+y_2=t) = \dfrac{\binom{n_1}{y_1}\binom{n_2}{t-y_1}}{\binom{n}{t}}$

(i.e. a hypergeometric probability)

$\Rightarrow E(Y_1)=n_1t/n$ and $V(Y_1)=n_1n_2t(n\text{-}t)/n^2(n\text{-}1)$

So, if we have large margins, a means of analysis is to say that

$$T_{MH} = [Y_1\text{-}E(Y_1)]^2/V(Y_1) \sim\chi_1^2 \text{ under } H_0$$

If $T_{MH} > \chi_1{}^2{}_{;1\text{-}\alpha}$ then $p < \alpha$ and there is a significant treatment difference.

### 7.3.1 Comments

1. Asymptotically equivalent to usual $\chi^2$ test.

2. Known as the ***Mantel-Haenszel*** or [very misleadingly as a *Randomization test*].

3. Does not matter whether you use $Y_1$, $Y_2$, $n–Y_1$ or $n–Y_2$.

4. The extension to several tables is simple. We use $W=\Sigma Y_{1j}$ and under $H_0$: $\theta_1 = \theta_2$ in each table, i.e. $\theta_{1j}=\theta_{2j}$, i.e. response ratio equal within each study we have $E(W)= \Sigma E(Y_{1j})$ and $V(W)=\Sigma V(Y_{1j})$ and $[W\text{-}E(W)]^2/V(W) \sim\chi_1^2$ under $H_0$ again.

5. This test is most appropriate when treatment differences are consistent across tables (we can test this but it is easier in a logistic regression framework — see later) — the test pools evidence from ***within*** the different trials whilst avoiding differences ***between*** trials.

### 7.3.2 Possible limitations of M-H test

- Randomness dubious
- reporting bias
- not clear that $\theta_l$ is the same for all trials.

### 7.3.3 Relative merits of M-H & Logistic Regression approaches

The Mantel-Haenszel test is simpler if one has just 2 qualitative prognostic factors to adjust for and wishes only to assess significance, not magnitude, of a treatment difference. The logistic approach (see below) is more general and can include other covariates, further, it can test whether treatment differences are consistent across tables. The M-H test is not very appropriate for assessing effects if tables are inhomogeneous, i.e. if treatment differences are inconsistent across tables, and must be used with care if success rates differ markedly (i.e. leading to Simpson's Paradox).

### 7.3.4 Example: pooling trials

A research worker in a skin clinic believes that the severity of eczema in early adulthood may depend on breast or bottle feeding in infanthood and that bottle fed babies are more likely to suffer more severely in adulthood. Sufferers of eczema may be classified as 'severe' or 'mild' cases. The research worker finds that in a random sample of 20 cases in his clinic who were bottle fed, 16 were 'severe' whilst for 20 breast fed cases only 10 were 'severe'. How do you assess the research workers belief?

In a search through the recent medical literature he finds the results, shown below, of two more extensive studies which have been carried out to investigate the same question. Assess the research worker's belief in the light of the evidence from these studies.

| | Bottle fed | | Breast fed | |
|---|---|---|---|---|
| study | severe | mild | severe | mild |
| 2 | 34 | 16 | 30 | 20 |
| 3 | 80 | 34 | 48 | 50 |

Analysis

Study 1

| | Severe | Mild | |
|---|---|---|---|
| Bottle | 16 | 4 | 20 |
| Breast | 10 | 10 | 20 |
| | 26 | 14 | 40 |

$Y_1$ =number of response 'severe' on bottle fed.

Under $H_0$ response ratios equal:

$E(Y_1) = 20\text{x}26/40 = 13$

$V(Y_1) = 20\text{x}20\text{x}26\text{x}14/40\text{x}40\text{x}39 = 2.333$

So Mantel-Haenszel test statistic is

$(16\text{-}13)^2/2.333 = 3.86 > \chi_1{}^2{}_{;0.95} = 3.84$

and so is just significant at 5% level, i.e. more severe cases on bottle feed

Study 2

| | Severe | Mild | |
|---|---|---|---|
| Bottle | 34 | 16 | 50 |
| Breast | 30 | 20 | 50 |
| | 64 | 36 | 100 |

$E(Y_2) = 50\text{x}64/100 = 32$

$V(Y_2) = 5.8182$

M-H test statistic =0.687, $p > 0.05$, n.s.

Study 3

| | Severe | Mild | |
|---|---|---|---|
| Bottle | 80 | 34 | 114 |
| Breast | 48 | 50 | 98 |
| | 128 | 84 | 212 |

$E(Y_3) = 68.83$, $V(Y_3) = 12.6668$,

M-H test statistic = 9.850, $p < 0.005$

Combining all 3 studies

Use W = $Y_1+Y_2+Y_3$ .

Under $H_0$: response ratios equal,

W=130, E(W)=113.83, V(W)=20.8183 so

M-H test statistic = 12.56, p < 0.0005, highly significant

**Caution**: the response ratios in the three studies differ quite a lot

(80%, 68% and 70% in studies 1, 2 and 3)

For interest, combining all 3 tables gives:

| | Severe | Mild | |
|---|---|---|---|
| Bottle | 130 | 54 | 184 |
| Breast | 88 | 80 | 168 |
| | 218 | 134 | 352 |

giving an Pearson $\chi^2$–statistic of 12.435, p < 0.0005. It might also be noted that the M-H statistic calculated from this table is slightly different, 12.400. These small differences are inconsequential in this case. The combined M-H statistic tests for association *within strata*, i.e. within studies, and so avoids differences *between strata*, thus avoiding Simpson's paradox (rather than overcoming it).

**Note:** We could also calculate the ordinary Pearson chi-squared values for each of these tables; the results are very close to (actually slightly greater than) the Mantel-Haenszel values since the numbers are large.

### 7.3.5 Example of Mantel-Haenszel Test in R

The function for performing a Mantel-Haenszel test in **R** is `mantelhaen.test()`. The Help system gives full details and examples.

The data are from the example 8.1 in §8.3.4 on page 135

The first example shews how to set up **R** to run a MH test on just one table by creating a factor z which has just one level.

```
> x<-factor(rep(c(1,2),c(20,20)),labels=c("bottle","breast"))
> y<-factor(rep(c(1,2,1,2),c(16,4,10,10)),labels=c("severe","mild"))
> z<-factor(rep(1,40),labels="study 1")
> table(x,y,z)

, , study 1
       severe mild
bottle     16    4
breast     10   10
> mantelhaen.test(x,y,z,correct=F)

Mantel-Haenszel chi-square test without continuity correction

data:  x and y and z
Mantel-Haenszel chi-square = 3.8571, df = 1
, p-value = 0.0495

>
```

The second example shews how to calculate the MH statistic for all three tables combined.

```
> x<-factor(rep(c(1,2,1,2,1,2),c(20,20,50,50,114,98)),
+ labels=c("bottle","breast"))
> y<-factor(rep(c(1,2,1,2,1,2,1,2,1,2,1,2),
+ c(16,4,10,10,34,16,30,20,80,34,48,50)),
+ labels=c("severe","mild"))
> z<-factor(rep(c(1,2,3),c(40,100,212)),
+ labels=c("study 1" ,"study 2","study 3"))
> table(x,y,z)

, , study 1
       severe mild
bottle     16    4
breast     10   10

, , study 2
       severe mild
bottle     34   16
breast     30   20

, , study 3
       severe mild
bottle     80   34
breast     48   50
> mantelhaen.test(x,y,z,correct=F)

     Mantel-Haenszel chi-square test wit
hout continuity correction

data:  x and y and z
Mantel-Haenszel chi-square = 12.5593, df =
1, p-value = 0.0004

>
```

## 7.4 Observational Studies

### 7.4.1 Introduction

In epidemiological studies where it is not possible to control treatments or other factors administered to subjects inferences have to be based on observing characteristics and other events on subjects. For example, to investigate the effect of smoking on health (e.g. heart disease) *cases* of subjects with heart disease might be collected. These would be compared with *controls* who do not exhibit such symptoms but are otherwise similar to the cases in general respects (e.g. age, weight etc.) and the incidence of smoking in the two groups would be compared. This is an example of a ***retrospective study.*** A different form of observational study is a ***prospective study*** where a cohort of subjects who are known to have been exposed to some risk factor (e.g. a very premature birth) and are followed up through a period. They are then observed at some later date and the incidence of a condition (e.g. school achievement very far below average) is assessed. In such studies the numbers of observations is typically very large since the incidence of the condition is often rare. It would be possible to use a chi-squared or a Mantel-Haenszel test for comparing the proportions but this would not be informative, either because with such large numbers of subjects the statistical test is very powerful and so return a highly significant result without saying anything about the magnitude of the effect or because the incidence is so rare that expected numbers in some cells are unduly low. Instead such observational studies are more traditionally analysed by estimating quantities that are of direct interpretability (odds ratios and relative risks) and they are assessed by calculating confidence intervals for their true values using formulae giving approximations to their standard errors.

### 7.4.2 Prospective Studies — Relative Risks

Prospective studies follow a group of subjects with different characteristics to see if an outcome of interest occurs. These would be used where the characteristic is not a 'treatment' that can be administered to a randomly selected group of subjects but some 'risk factor' such as very low birth weight or more than one month premature birth or blood group. The outcome may be some feature which occurs at some time later. The analysis would be based on calculating the risks of developing the feature for the different groups and, in the case of two outcomes (positive and negative say) and two groups (exposed and non-exposed say) calculating the relative risks.

| | Outcome | | |
|---|---|---|---|
| | Positive | Negative | Total |
| Exposed | a | b | a+b |
| Non-exposed | c | d | c+d |

The risk of a positive outcome for the exposed group is a/(a+b) and for the non-exposed group it is c/(c+d). The ***relative risk*** is the ratio of these two

$$\mathrm{RR}=\frac{a/(a+b)}{c/(c+d)}=\frac{a(c+d)}{c(a+b)}$$

and we compare this with the value 1 (the RR if there is no difference in risks for the two groups) by using its standard error.

The formula for the standard error of $\log_e(\mathrm{RR})$ is

$$\mathrm{S.E.}\{\log_e(\mathrm{RR})\}=\sqrt{\tfrac{1}{a}-\tfrac{1}{a+b}+\tfrac{1}{c}-\tfrac{1}{c+d}}$$

### 7.4.2.1 Example

The data are taken from a study of 'small-for-date' babaies who were classifie as having symmetric or asymmetric growth retardation in relation to their Apgar score.

| | Apgar < 7 | | |
|---|---|---|---|
| | Yes | No | Total |
| Symmetric | 2 | 14 | 16 |
| Asymmetric | 33 | 58 | 91 |

The calculations give $RR=0.3447$, $\log_e(RR) = -1.0651$, $s.e.(\log_e(RR)) = 0.6759$.

A 90% CI for $\log_e(RR)$ is $-1.0651 \pm 1.645\times 0.6759 =$ $(-2.1769,\ 0.0467)$and taking exponentials of this gives a 90% CI for the RR as (0.11, 1.05). Since this interval contains 1 there is no evidence at the 10% level of a difference in risk of a low Apgar score between the two groups.

### 7.4.3 Retrospective Studies — Odds Ratios

Retrospective studies identify a collection of ***cases*** (e.g. with a disease) and compare these with respect to exposure to a risk factor with a group of ***controls*** (without the disease). The selection of the subjects is based on the outcome and not the characteristic defining the group as with prospective studies.

| | Cases | Controls |
|---|---|---|
| Exposed | a | b |
| Non-exposed | c | d |
| Total | a+c | b+d |

It is not sensible to calculate the risk of 'being a *case'* (a/(a+b)) since this can apparently be made any value just by selecting more or fewer controls which would increase or decrease b but not any other value.

Instead it is sensible to look at the ***odds*** of exposure for the cases and for the controls and look at the ratio between these. If exposure is not a risk factor for being a case then this ***odds ratio*** will be close to 1. As before there is a simple formula for the standard error of the $\log_e$ of the odds ratio

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

and

$$S.E.\{\log_e(OR)\} = \sqrt{\tfrac{1}{a} + \tfrac{1}{b} + \tfrac{1}{c} + \tfrac{1}{d}}$$

#### 7.4.3.1 Example

The following gives the results of a case-control study of erosion of dental enamel in relation to amount of swimming in a chlorinated pool.

| | Enamel erosion | |
|---|---|---|
| Swimming per week | Yes | No |
| ≥ 6 hours | 32 | 118 |
| < 6 hours | 17 | 127 |

The calculations give OR=2.0259, s.e.($\log_e$(RR))=0.3262 and so a 95% for the log odds ratio is (0.0666, 1.3454) and the confidence interval for the odds ratio itself is thus (1.0689, 3.8397) which excludes the value 1 and so provides evidence at the 5% level of a raised risk of dental erosion in those swimming more than 6 hours a week.

## 7.5 Matched pairs

### 7.5.1 Introduction

In the comparison of two treatments A & B, suppose each patient receives both treatments (in random order), e.g. a crossover or matched-pair trial. We then observe pairs:

$(y_{i1}, y_{i2})$
↗ ↖
response to A  response to B

of the form (0, 0), (0, 1), (0, 1), (1, 1), (1, 0), (1, 1), ........
e.g. Rheumatoid arthritis study, two treatments A & B.
Response caused? 1=yes, 0=no
Could present results as:

response

| | | yes | no | |
|---|---|---|---|---|
| treatment | A | 11 | 37 | 48 |
| | B | 20 | 28 | 48 |

and then it is tempting to analyse this as an ordinary 2×2 table with a $\chi^2$-test.

This ☠ **INVALID** ☠ since it ignores the double use of each patient (there are only 48 independent subjects in the table not 96).

A more useful summary is

| | | B | | |
|---|---|---|---|---|
| | | yes | no | |
| A | yes | 8 | **3** | 11 |
| | no | **12** | 25 | 37 |
| | | 20 | 28 | 48 |

A suitable test for what is really of interest (treatment difference) — not 'no association') is:

### 7.5.2 McNemar's Test

Ignore (1,1) and (0,0), use the unlike pairs only. If no treatment differences exist, then the proportions of (1,0)'s (say) out of the total number of (1,0)'s and (0,1)'s should be consistent with binomial variation with p=½.

<u>In example</u>

There are 3 (1,0)'s out of a total of 15 unlike pairs.

i.e. significance probability = $2\times\sum_{x=0}^{3}\binom{15}{x}\left(\tfrac{1}{2}\right)^{15}$ =0.035 which is significant at the 5% level.

For larger n use the Normal approximation

$$\frac{(n_{10}-n_{01})^2}{n_{10}+n_{01}} \sim \chi_1^2$$

**Note:** We have not used the data from subjects where the responses were the same, i.e. subjects for whom both treatments produced successes or both failures. This is sensible since these subjects provide no evidence on *treatment differences*, even though intuitively the results from these subjects might suggest that the two treatments are equivalent.

## 7.6 Logistic Modelling

### 7.6.1 Introduction

(for more details of logistic models see PAS372 or PAS6003)

Logistic modelling has become a very popular way of handling binary data and the analyses can be handled in most standard statistical packages.

In the clinical trials context define:

For patient i, outcome = $Y_i$ = 0 (failure) or 1 (success).

treatment $x_i$ = 0 (placebo) or 1 (treatment)

Then an alternative parameterization of the 2×2 set up is

$$P[Y_i=1]=\frac{e^{\beta_0+\beta_1x_i}}{1+e^{\beta_0+\beta_1x_i}} \quad \text{and} \quad P[Y_i=0]=1-P[Y_i=1]=\frac{1}{1+e^{\beta_0+\beta_1x_i}}$$

i.e. on placebo

$$P[Y_i=1]=\frac{e^{\beta_0}}{1+e^{\beta_{0i}}}$$

and on treatment $P[Y_i=1]=\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$

We can see that $\ln\left\{\dfrac{P[Y_i=1]}{P[Y_i=0]}\right\}=\beta_0+\beta_1x_i$

The model extends naturally to include other <u>prognostic factors</u> or <u>covariates</u>: $\ln\left\{\dfrac{P[Y_i=1]}{P[Y_i=0]}\right\}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+.....+\beta_px_{ip}$

$$=\beta_0+\underline{\beta}'\,\underline{x}_i$$

where the $x_{ij}$ can be continuous or discrete or dummy.

$$\frac{P[Y_i=1\,|\,\underline{x}_i]}{P[Y_i=0\,|\,\underline{x}_i]}=\exp\{\beta_0+\underline{\beta}'\,\underline{x}_i\}$$

In this case $P(Y_i=1)$ = P(success) = $\dfrac{e^{\beta_0+\underline{\beta}'\underline{x}}}{1+e^{\beta_0+\underline{\beta}'\underline{x}}}=\theta_i$

and $\ln\left\{\dfrac{P[Y_i=1]}{P[Y_i=0]}\right\}=\ln\left(\dfrac{\theta_i}{1-\theta_i}\right)=\beta_0+\underline{\beta}'\,\underline{x}_i$

### 7.6.2 Interpretation

For comparative trials

$$\ln\left\{\frac{P[Y_i=1]}{P[Y_i=0]}\right\}=\beta_0+\ \mathbf{0}\ +\beta_2x_{i2}+\beta_3x_{i3}+.....+\beta_px_{ip}\quad \text{if } x_{i1}=0,$$

i.e. on placebo

$$\ln\left\{\frac{P[Y_i=1]}{P[Y_i=0]}\right\}=\beta_0+\ \boldsymbol{\beta_1}+\beta_2x_{i2}+\beta_3x_{i3}+.....+\beta_px_{ip}\quad \text{if } x_{i1}=1,$$

i.e. on treatment

so if $\beta_1>0$, odds in favour of success are greater in treatment group

and if $\beta_1<0$, odds in favour of success are greater in placebo group

Similar interpretations for other factors:

$\beta_j > 0 \Rightarrow$ P(success) ↗ as $x_j$ ↗ and P(success) ↘ as $x_j$ ↘

$\beta_J < 0 \Rightarrow$ P(success) ↘ as $x_j$ ↗ and P(success)↗ as $x_j$ ↘.

### 7.6.3 Inference

$\beta_0$ and $\underline{\beta}$ are estimated by Maximum Likelihood:

$$L(\beta_0,\underline{\beta})=\prod_{i=1}^{n}\theta_i^{y_i}(1-\theta_i)^{1-y_i}\ ;$$

$$\ln L(\beta_0,\underline{\beta})=\Sigma y_i\ln\{\theta_i/(1-\theta_i)\}+\Sigma\ln(1-\theta_i)$$

$$\ln L(\beta_0,\underline{\beta}) = \ell(\beta_0,\underline{\beta}) = \Sigma y_i(\beta_0+\underline{\beta}'\,\underline{x}_i) - \Sigma\ln[1+\ \exp(\beta_0+\underline{\beta}'\underline{x}_i)]$$

Standard iterative methods (e.g. Newton-Raphson)

give m.l.e.’s $\beta_0,\underline{\beta}$

$$\frac{\partial\ell}{\partial\beta_0}=\sum_1^n(y_i-\theta_i)\ ;\qquad \frac{\partial\ell}{\partial\beta_i}=\sum_1^n x_i(y_i-\theta_i)$$

Estimated standard errors of these estimates can be obtained from the diagonal of the estimated variance matrix

$$\hat{\text{var}}\begin{pmatrix}\hat\beta_0\\ \hat{\underline{\beta}}\end{pmatrix}\approx\left\{-E\left[\frac{\partial^2\ell}{\partial(\beta_0,\underline{\beta})}\right]\right\}^{-1}_{@\hat\beta_0,\hat{\underline{\beta}}}$$

R or MINITAB or SAS or SPSS or S-PLUS will fit the model and give estimates and standard errors. We can test significance in terms of:–

a) partial z-test

$H_0$: $\beta_j = 0$

test compares $\dfrac{\hat{\beta}_j}{\sqrt{\mathrm{var}(\hat{\beta}_j)}}$ with N(0,1) %-points

(ignore strict need for t-test)

b) likelihood ratio

compare $2|\ell_{\text{full model}} - \ell_{\text{reduced model with } \beta=0}|$ with $\chi_1^2$

where $\ell$ is the maximized log likelihood (or ***deviance***)

### 7.6.4 Example (Pocock p.219)

A trial to assess the effect of the treatment *clofibrate* on ischaemic heart disease (IHD). Subjects were men with high cholesterol, randomized into placebo and treatment groups.

Prognostic factors (i.e. factors which also affect risk of IHD and which can be identified in advance) were:

age; smoking; father's 'history'; systolic BP; cholesterol

Response: $Y_i$ : 'success' (!!) = patient subsequently suffers IHD

Each patient has a certain probability $p_i$ of achieving a response. $p_i$ is the probability of getting IHD. Define the following multiple logistic model for how $p_i$ depends on the prognostic variables:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \ln\left\{\frac{P[\text{suffers IHD}]}{P[\text{does not}]}\right\} = \beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+.....+\beta_6x_{i6}$$

where $\beta_0,....,\beta_6$ are numerical constants called logistic coefficients. This is sometimes written $\mathrm{logit}(p_i) = \beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+.....+\beta_6x_{i6}$.

$x_1$=0 (placebo), 1 (clofibrate)

$x_2$=ln(age)

$x_3$=0 (non-smoker),1 (smoker)

$x_4$=0 (father alive), 1 (dead)

$x_5$=systolic BP in mm Hg

$x_6$=cholesterol in mg/dl

Apply maximum likelihood to estimate values of $\beta_l$ (l=0,1,...6):

| | Numerical variable | logistic coef | |
|---|---|---|---|
| factor | $x_j$ | $\beta_j$ | z-value |
| 1:treatment | 0=placebo,1=treatment | –0.32 | –2.9 |
| 2:age | ln(age) | 3.0 | 6.3 |
| 3:smoking | 0=non-smok, 1=smoker | 0.83 | 6.8 |
| 4:father's hist | 0=alive, 1=dead | 0.64 | 3.6 |
| 5:systolic BP | Systolic BP in mm Hg | 0.011 | 3.7 |
| 6:cholesterol | Cholesterol in mg/dl | 0.0095 | 5.6 |
| | constant term $\beta_0 = –19.60$ | | |

$\Phi^{-1}(.005) = z_{.005} = –2.58$ $\qquad$ $z_{.025} = –1.96$

(1% level) $\qquad$ (5% level)

Treatment: significant, $p < 0.01$; $\beta_1 < 0$;

Probability of IHD is smaller on treatment than on placebo

Prognostic factors: all five significant ($p < 0.01$); all have positive m.l.e.'s, ∴ probability of IHD increases with age, smoking, 'poorer heredity', high blood pressure, high cholesterol.

Another useful way of describing the importance of each factor is to look at ***odds ratios***. The odds ratio is approximately equal to the relative risk if the probability of the event is small and consequently the term *relative risk* is often [technically mistakenly] used in this context.

e.g. the odds ratio of getting IHD on clofibrate compared with placebo is the ratio of odds:

$$\frac{\dfrac{P[Y=1\,|\,x_1=1]}{P[Y=0\,|\,x_1=1]}}{\dfrac{P[Y=1\,|\,x_1=0]}{P[Y=0\,|\,x_1=0]}}$$

$$= \exp\{\beta_1\}$$

The estimated odds ratio is $e^{–0.32} = 0.73 < 1$

i.e. odds of getting IHD are 27% lower on clofibrate after allowing for the other prognostic factors.

The standard error of $\beta_1$ is 0.11 (= –0.32/–2.9, but actually obtained direct from diagonal of information matrix [not given here]). So approximate 95% confidence limits for $\beta_1$ are $–0.32 \pm 2\text{x}0.11 = –0.10$ and $–0.54$. Hence $\exp\{\beta_1\}$ has 95% confidence limits $e^{–0.1}$ and $e^{–0.54}$ = 0.90 and 0.58 so that 95% confidence limits for the reduction due to clofibrate in odds of getting IHD are 10% and 42%.

Similar calculations for smoking show 95% limits for the increase in odds of getting IHD for smokers are 80% and 193%.

### 7.6.5 Interactions

Interaction terms would be handled by creating a new variable as the product of the treatment and the covariate values. In the example above the treatment is coded as 0 for placebo and 1 for clofibrate, so the value of this interaction term would be 0 for all subjects receiving placebo and the same as the covariate for those on clofibrate. In the example above Treatment is variable $x_1$ and $\log_e$(age) is variable $x_2$ and there are six variables in all. We create a new variable $x_7 = x_1 \times x_2$ and then our model is

$$\text{logit}(p_i) = \beta_0+\beta_2x_{i2}+\beta_3x_{i3}+.....+\beta_6x_{i6} \text{ for placebo, and}$$

$$\text{logit}(p_i) = \beta_0+\beta_1x_{i1}+(\beta_2+\beta_7)x_{i2}+\beta_3x_{i3}+.....+\beta_6x_{i6} \text{ for clofibrate}$$

and $\beta_7$ reflects the interaction effect, (note that $x_7$ is identical to $x_2$ for those on clofibrate but 0 for those on placebo).

Exactly the same method is appropriate for handling interactions between two continuous covariates and between two 2-level factors. Interactions involving a k-level factor can only be handled by converting the factor into k–1 dummy binary variables. In this case the interaction term has k–1 degrees of freedom if it is a k-level factor×covariate interaction or (k–1)(j–1) degrees of freedom for an interaction between a k-level and a j-level factor. This also means that the separate parts of the chi-squared statistic must be combined before assessing significance. However, be cautious in including interactions if there is no *a priori* reason to expect them since problems of multiplicity can arise, especially if the number of levels of a factor is large.

### 7.6.6 Combining Trials

Within the context of combining trials we might keep $\beta_1$ the same in each trial, but allow $\beta_0$ to vary to reflect possible differences in trial j conditions:

i.e.
$$\ln\left\{\frac{P[Y_{ij} = 1]}{P[Y_{ij} = 0]}\right\} = \beta_j+\beta_1x_{ij}$$

e.g. 3 clinics

$$=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}$$

where the last two terms are the clinic coding $x_{i2}$ and $x_{i3}$ are *dummy variables*, i.e.

$(x_{i2}, x_{i3}) =$ (0,0) for clinic 1
(1,0) for clinic 2
(0,1) for clinic 3

which gives $\beta_0+\beta_1x_{i1}$ for clinic 1
$(\beta_0+\beta_2)+\beta_1x_{i1}$ for clinic 2
$(\beta_0+\beta_3)+\beta_1x_{i1}$ for clinic 3

## 7.7 Summary and Conclusions

- ♦ Combining trials can give paradoxical results if response rates and sample sizes are very different in the trials (Simpson's Paradox)
- ♦ Simpson's paradox can be resolved by more sophisticated modelling allowing for a separate 'trial effect'
- ♦ The Mantel-Haenszel test provides an alternative way of analysing 2×2 tables which makes it easier to combine results from different trial but which does not overcome Simpson's Paradox but avoids it.
- ♦ Care needs to be taken in analysing matched pairs binary responses. McNemar's test uses only the information from **unlike pairs**
- ♦ Logistic Regression allows the **log-odds** to be modelled as a linear model in the covariates.
- ♦ Logistic models can be implemented in most standard statistical packages
- ♦ Logistic models allow **relative risks** to be estimated (including confidence intervals).
- ♦ Positive coefficients in a logistic model indicate that the factor increases the risk of the 'success'