MEDICAL STATISTICS

Part 1: Clinical Trials

MAS361/MAS61002

School of Mathematics and Statistics
The University of Sheffield

2022–2023

Contents

Chapter 1

Introduction

1.1 Definition of a Clinical Trial

“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.” (Pocock, 1983)
NB:

- Planned experiment – not an observational study and is pre-planned/designed
- On patients – not a study on healthy volunteers
- Inferential procedure – want to use results from limited sample of patients to identify best treatment for the general future population of patients.

1.2 Historical Background

(see e.g. Pocock Ch. 2, Matthews Ch. 1)
Prior to 1950s medicine developed in a haphazard way. Some advances were made (chiefly in communicable diseases) perhaps because the improvements could not be masked by poor procedure. Medical literature emphasized individual case studies and treatment was copied; unscientific and inefficient.

There were occasional exceptions.

For example:
In 1537: **Treatment of battle wounds**:
Standard Treatment: Boiling Oil
versus
New Treatment: Egg Yolk + Turpentine + Oil of Roses

appreciation only ↓

New treatment found to be better.

In 1741: **Treatment of Scurvy, HMS Edinburgh**:
Two patients allocated to each of 6 treatments:

1. cider
2. elixi vitriol
3. vinegar
4. nutmeg
5. sea water
6. oranges and lemons

Treatment 6 produced "the most sudden and visible good effects."

Incorporation of statistical techniques is more recent. Four milestones were:

1. Medical Research Council (MRC) **Streptomycin trial for tuberculosis** (1948) was first to use a randomized control.

2. MRC **cancer trials** (with statistician Austin Bradford-Hill) first recognizably modern sequence – laid down the (now) standard procedure.

3. An early study of the **validity of controlled and uncontrolled trials** (Foulds, 1958) examined reports of psychiatric clinical trials and found:
   in 52 uncontrolled trials, treatment was declared 'successful' in 43 cases (83%)

in 20 controlled trials, treatment was 'successful' in only 5 cases (25%). This suspicious difference added weight to the call for controlled trials (and may also be evidence of publication bias – see below).

4. **Salk Polio Vaccine** – An (accidental) comparison between a randomized controlled double-blind clinical trial and a non-randomized open trial; demonstrating superiority of the former and establishing it as standard practice (see below).

1.2.1 Field Trial of Salk Polio Vaccine

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

Certain areas of the US, 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:

- only volunteers could be used – these tended to be from wealthier/better educated background (i.e. volunteer bias)
- 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 randomized double-blind trial simultaneously. Every child received an injection but half these did not contain vaccine and child/parent/evaluating physician did not know which.

Results (excluding some dropouts)

Study	Group	No.	Cases	Rate per 100 000
Observed control	Vaccinated 2nd grade	221 998	38	17
	1st and 3rd grade control	725 173	330	46
	Unvaccinated 2nd grade	123 605	43	35
Randomized control	Active vacc	200 745	33	16
	Placebo	201 229	115	57
	Not vaccinated	338 778	121	36

Results from second part conclusive:

- incidence in vaccine group reduced to 1/3 placebo level
  Not clear from this data, but also found

- paralysis from those getting polio 70% less
- 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* suitable control) was difficult. This provides an example of **volunteer bias**. A similar difference is noted between placebo and not vaccinated groups in the RCT, but here we have the placebo group to form a suitable control.

appreciation
only ends ↑

Thus, the Salk 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 randomized 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 around 1900.

1.3 Role of Clinical Trials in Drug Development

Typically a new treatment develops through a research programme (at a pharmaceutical company) who test MANY different manufactured/synthesized compounds at vast cost. Approximately 5 in 10 000 of those synthesized get to a clinical trial stage (via chemical analysis, preliminary animal testing, pre-clinical screening, etc.). Of these, 1 might reach marketing. The 4 stages of a clinical trial programme (after the pre-clinical stage) are

- **Phase I trials**. Clinical pharmacology and toxicity tests, 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 and investigate metabolism of drugs. n=10 - 50.

- **Phase II trials**. Initial clinical investigation for treatment effect. Concerned with safety and 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 including long-term studies of side effects, morbidity and mortality. n= as many as possible.

In this course, we are mainly concerned with Phase III trials.

1.3.1 Further Notes

appreciation only ↓

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. Usually the 'standard' drug is already on the market and the company want the new drug to be at least equally good so as to get a share of the market (this leads to a distinction between **superiority** and **non-inferiority** trials; see Further Clinical Trials course).

Almost all Phase III trials now are **randomized controlled (comparative) studies**, or **RCTs**, i.e. they involve comparison of effects in a group receiving new drug and a group receiving the standard drug.

To avoid bias (subconscious or otherwise), patients must be assigned to groups **at random**. For example, bias might occur if we give very ill people the new drug since there is no chance of standard drug working; typically we would then see 'poor' results from the new drug. Alternatively we might see bias in the opposite direction, because there is more chance of the very ill 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 nature 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. using 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. appreciation only ends ↑

1.4 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).

1.4.1 Nocebo Effect

appreciation only ↓

Originally 'placebo effect' was taken to refer to both pleasant and harmful effects of a treatment believed to be inert but sometimes this is reserved just for pleasant effects and the term **nocebo effect** used to refer to a harmful effect (placebo and nocebo are the Latin for 'I will please' and 'I will harm', respectively). There are anecdotal reports of nocebo effects being surprisingly extreme, such as the case of an attempted suicide with placebo pills during a clinical trial which was only averted by emergency medical intervention, see Reeves et al., (2007), General Hospital Psychiatry, 29, 275-277. appreciation only ends ↑

1.5 Blindness of Trials

Using placebos allows the opportunity to make a trial blind– i.e. the patient or the doctor does not know which treatment was received. This avoids bias from patient or evaluator attitudes.
Single blind – either patient or evaluator blind.
Double blind – both patient and evaluator blind.

Clearly, blind trials are not always possible (e.g. cannot compare a drug treatment with a surgical treatment), but good practice is that double blind trials should be used where possible. In organizing such a trial there is a coded list which records each patient's treatment. This is held by a co-ordinator and only broken at analysis (or in emergency).

1.6 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 so. In the EBM framework statistical methodology has a key 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 the drug has no effect, nor that it is effective. The statistical test can only add to the overall evidence.

1.6.1 The Bradford-Hill Criteria

To help answer the **specific question of causality**, Austin 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 other 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 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.

1.7 Ethical Considerations

Unsurprisingly, since they are experiments on humans, great concern is given to the ethical conduct of clinical trials and the area is heavily regulated. All general ethical considerations apply here also, but there are some special points to be noted.

1.7.1 Medical Ethics

Specified in **Declaration of Helsinki** (1964+amendments) consisting of 32 paragraphs, see http://www.wma.net/

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

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, asking "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.

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. Thus it may 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. It is also unethical to perform a trial which has little prospect of reaching any conclusion, e.g. because of insufficient numbers of subjects (or some other aspect of poor design). These considerations often lead to **sample size calculations**; see later.

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 (in each hospital, city or region) monitor and ‘licence’ all clinical trials.

1.7.2 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 Medicine, . . .).

Because of the huge amounts of money involved, concerns were mounting that the usual standards of professional academic research did not necessarily apply in the medical area. In particular, there was concern at articles where declared authors had

- not participated in design of study
- 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) had 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

Beware also of publication bias where authors only publish if results say that new drug is better. Studies which are inconclusive or reveal preference for the standard treatment are not written up. Efforts to minimize this centre round monitoring the registers of trials, access to trial data, etc, but commercial interests inhibit full and rapid disclosures.

1.8 Summary and 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 statistician's responsibility

Chapter 2

Trial Designs

2.1 Why Design Trials?

To ensure that treatment comparisons are made fairly and efficiently we should design our study carefully. This helps meet our ethical obligations. It is also good scientific practice and should be cost-efficient too. The key technical issue is whether comparions are made 'between' or 'within' patients. If comparisons are made using responses from different patients, then there is probably no concern about lack of independence, but we will have to contend with potentially substantial inherent differences from person to person. To make comparisons within patients, we must have two or more observations from each person. These can, in some way, act as their own baseline, so we may be able to construct very powerful tests that are able to pick up subtle treatment effects. However, construction of these tests is much more technically challenging because the multiple observations from each person will typically not be independent.

2.2 Parallel Group Designs

These compare k treatments by dividing patients, at random, into k groups – the n_i patients in group i receive treatment i. Thus each patient receives just one treatment. Comparisons are made between patients. Representing patients with an 'X ':

Group				
1	2	3	$\ldots$	k
X	X	X	$\ldots$	X
X	X	X	$\ldots$	X
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
X	$\vdots$	$\vdots$	$\ldots$	$\vdots$
	X	$\vdots$	$\ldots$	X
		X		

Usually the groups are of the same size, but this is not necessary. This is attractive as it means more weight can be given to comparisons of particular interest and that dropouts do not cause great difficulty.

2.3 In Series Designs

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. Thus we consider 'in series' designs, where each patient receives all k treatments in the same order and comparisons can be made within patients. Here an individual's different treatment applications are linked by arrows.

Patient	Treatment							
	1	$\longrightarrow$	2	$\longrightarrow$	3	$\ldots$	$\longrightarrow$	k
1	X	$\longrightarrow$	X	$\longrightarrow$	X	$\ldots$	$\longrightarrow$	X
2	X	$\longrightarrow$	X	$\longrightarrow$	X	$\ldots$	$\longrightarrow$	X
$\vdots$	$\vdots$		$\vdots$		$\vdots$			$\vdots$
n	X	$\longrightarrow$	X	$\longrightarrow$	X	$\ldots$	$\longrightarrow$	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. A similar (reversed) effect can be occur for patients with a progressive disease. The key is that difficulties occur if the underlying disease is not stable.

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/modify 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

2.3.1 Crossover Designs

Problems with ‘period’, ‘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 necessary to use a between-patient (parallel group) analysis of the results for period 1 only, though, not having been designed for this, power may be low.

Example 2.1 Stage-fright trial (Pocock, p112)

Evaluated the effect of the drug oxprenolol on stage-fright in musicians. Trial was double blind in that neither the musician nor the assessor knew the order of treatment. Each musician assessed on each day for nervousness and performance quality.

N = 24 musicians split at random into two groups, each following one ‘arm’ of the trial.

No.	day 1		day 2
12	oxp	$\longrightarrow$	placebo
12	placebo	$\longrightarrow$	oxp

NB More typically design arms are

washout → treatment → washout → treatment

where **washout** is a period with no treatment at all.

Example 2.1 Plaque removal of mouthwashes

Here we have three treatments
A: water
B: brand X
C: brand Y

Patient	Order		
	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.

NB If it is not possible for each patient to have each treatment, use a balanced incomplete block design.

2.4 Factorial Designs

In some situations, it may be possible to investigate the effect of two or more treatments by allowing patients to receive combinations of treatments (or factors). For example, in the 2×2 case, each patient takes two 'drugs', with each drug in either active or placebo form and so the possible combinations are:

active A, active B
active A, placebo B
active B, placebo A
placebo A, placebo B

often written as:

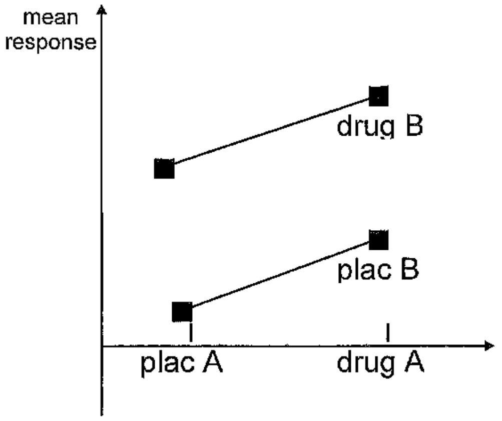
AB
A
B
placebo

Suppose we had 40 patients and allocated 10 at random to each combination, then overall 20 have had (active) 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 $\simeq$ 40).

This factorial design might lead to more efficient comparisons, because of 'larger' numbers in each comparison. For example, the effect of B can be calculated by looking at **both** the difference between B and placebo **and** that between AB and A.

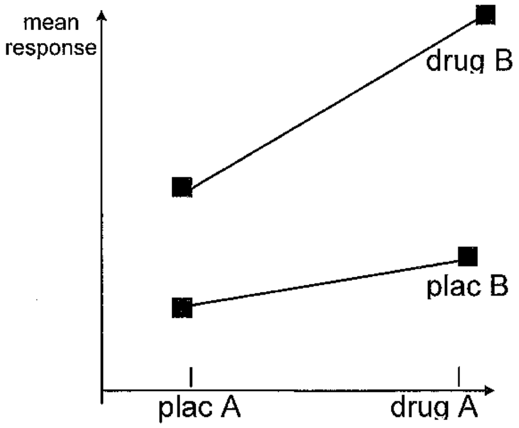
Obviously such designs not always applicable because of problems with interactions of drugs, but these might themselves be of interest.

Types of interaction



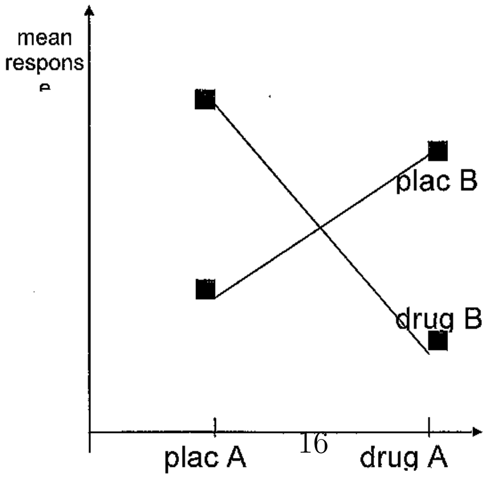
lines parallel ⇒ <u>no interaction</u>

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>

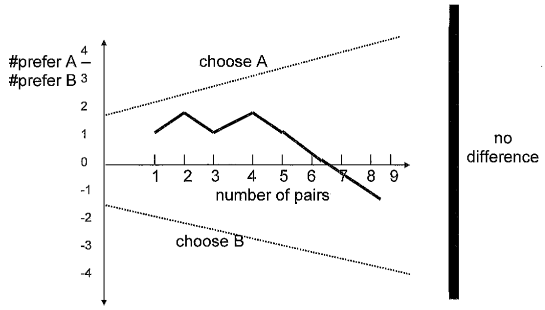
A increases response when given alone, but decreases response when in combination with B

2.5 Sequential Designs

In its simplest form, patients are entered into the trial in matched pairs, one receives A, the other B (allocated at random). Test for treatment superiority after results from **each pair** are known. E.g. simple preference data (i.e. examining the responses of a matched pair, we record whether the pair ‘think’ A or B is better), for example

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

We then plot the difference in the number of A and B preferences against the number of pairs so far. To reach a decision as to the superiority of either treatment (or whether the trial is inconclusive) we need ‘boundary stopping rules’.



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
5. Calculation of boundaries is highly complex. With paired success/failure data (taking A as preferable as a 'success') the underlying test is based on a binomial calculation but for individual patients with a quantitative response it is based on a t-test calculation with adjustments made for multiple testing and interim analyses on accumulating data, topics which are discussed further in Chapter 7.

2.6 Other Designs

We have covered here only the simplest designs for clinical trials. We will look at the analysis of parallel group and crossover studies later and you may see more on sequential designs if you take the Further Clinical Trals course. You will also see the special structure of survival analysis studies in Part 2. However, in practice, studies are often more complex. It may be possible to combine elements seen here. For instance

Treatment 1	Treatment 2	Treatment 3
Before After	Before After	Before After
$X \longleftrightarrow X$	$X \longleftrightarrow X$	$X \longleftrightarrow X$
$X \longleftrightarrow X$	$X \longleftrightarrow X$	$X \longleftrightarrow X$
$X \longleftrightarrow X$	$X \longleftrightarrow X$	$X \longleftrightarrow X$
$\vdots$	$\vdots$	$\vdots$
$X \longleftrightarrow X$	$X \longleftrightarrow X$	$X \longleftrightarrow X$

However, there are many other special situations arising in medical statistics that we will not cover, e.g. first time in man studies, more elaborate repeated measures (including continuous trace) responses, dose-response studies, pharmacokinetic models.

2.7 Summary and Conclusions

- Efficient design of clinical trials is a crucial ethical element contributed by statistical theory and practice

- 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, issues over interactions (quantitative or qualitative), comparisons are between patients but more available than in parallel designs
- Sequential designs – suitable for rapidly evaluated outcomes, minimizes numbers of subjects when clear differences between treatments

Chapter 3

Protocols and Protocol Deviations

3.1 Protocol

The protocol for any trial is a written document containing **all** details of trial conduct. It is needed to gain permission to conduct any trial. It should contain items on the purpose, design and conduct of the study (see Pocock, table 3.1).

Typical elements:

Purpose	motivation	
	aims	
Design & conduct	patient selection criteria	inclusion/exclusion criteria
	treatment schedule	
	number of patients (and why)	
	assignment of patients	trial design
		randomization
	evaluation of response	baseline measure
		principal response
		subsidiary criteria
	admin	‘informed consent’ form
		monitoring/record forms
		records & responsible persons
	techniques for analysis	

3.2 Protocol Deviations

Things always go wrong! A protocol deviation occurs when a patient departs from the defined experimental procedure (e.g. does not meet the inclusion/exclusion criteria (e.g. too young), takes 2 tablets instead of 1, forgets to take medicine, takes additional other medicine,.....).

All protocol violations and major deviations should be recorded as they occur. They should be noted in the report and in the analysis. Our aim in the analysis is to minimize bias in the treatment comparison of interest, i.e. to ensure treatment comparisons are not affected by factors other than treatment differences. Two possibilities are '**per protocol**' and '**intention to treat**' analyses.

In per protocol (or 'on treatment') analysis we consider only those patients who completed the trial in accordance with the protocol. Those patients who deviate from the protocol are excluded. This analysis is simple, but it has important deficiencies. Firstly, if the numbers of patients is reduced there is a loss of power. Secondly, the randomization is compromised (i.e. no longer completely valid) and potential biases appear because failure of patient to follow protocol may be related to treatment (e.g. did patient forget to take enough pills because the drug was very strong, or choose not to do so because they felt it ineffective or that it had unpleasant side effects?).

In intention to treat' (or 'pragmatic') analysis, we include all patients as originally randomized and assigned to treatments. In fact, we often modify this slightly to remove clear mistakes (e.g. patient too young or too old) which should have been picked up at the entry/randomization stage. We then speak of only analyzing data from 'eligible' patients, with those who violated inclusion criteria being excluded. This analysis rarely excludes many patients, so is rarely subject to critical loss of power and the randomization is not compromised, but patients who deviated from the protocol may give very odd responses and this may distort the results since the groups being compared are not homogeneous. Withdrawals cause particular problems since we will not have (complete) responses for them. In a 'count' experiment, it may be easy to assign the supposed response of withdrawals to a sensible category (often 'worst case scenario'), but this can be more difficult in the case of continuous responses (possibilities include imputation, value at intermediate timepoint,...).

In practice, the intention to treat analysis is preferred because of its lower risk of bias, but a per protocol analysis is often reported as well (ideally the two will not give substantially different results).

Example 3.1 MRC (1966) study of surgery v radiotherapy for operable lung cancer

appreciation only ↓

In the group assigned to receive surgery, a certain proportion were found to have tumours which could not be removed (i.e. they were not operable and so should not have been included in the trial as they did not meet the inclusion criteria). In the radiotherapy group, there was no opportunity to detect similar patients (so there may or may not have been patients who did not meet this inclusion criterion). The only fair comparison is between the groups as randomized, even though not all in the surgery group received treatment. If the inoperable cases (likely to have a poorer expected outcome) were removed before analysis, the remainder in the surgery group would have different (and probably lower risk) characteristics from the group as a whole.

Example 3.2 Salk polio vaccine

Note that the data presented in Section 1.2.1 on the field trial of the Salk polio vaccine, the non-randomized part of the study can be subjected to an intention to treat analysis. It was intended that all 2nd grade children would be vaccinated but some of them (in fact more than 35% of them) refused the vaccine. If the treatment is regarded as offering the vaccination and inoculating those who accept (rather than giving the vaccination itself) then the rate for all 2nd grade children could be compared to that for the observed controls.

appreciation only ends ↑

Example 3.3 Antidepressant trial (Pocock, p182)

Randomized double-blind trial compared
low and high doses of new antidepressant with a control treatment.

50 patients entered the trial but 15 withdrew because of possible side effects. Results were originally recorded on a 3-level effectiveness scale, but we have simplified this here to ‘very effective’ or ‘not very effective’.

Results:

clinical assessment	low dose	high dose	control	TOTAL
very effective	2	8	6	
not	7	4	8	
total assessed	9	12	14	35
withdrawn	6	8	1	15
total randomized	15	20	15	50

NB It looks as if withdrawals are not random and might indeed be due to some other reason, such as side effects, as different proportions withdrew in each case.

Analysis 1: per protocol (i.e. only those assessed)
Analyzing informally by just examining percentage very effective:

	low	high	control
% very effective	22%	67%	43%

Conclusion: high dose produced the highest proportion of 'very effective' assessments.

Analysis 2: Intention to treat (i.e. including patient withdrawals)
BUT regarding all withdrawals as 'ineffective', i.e. **worst case scenario**

	low	high	control
% very effective	13%	40%	40%

Conclusion: no difference between high dose and control.

Thus the conclusions from the trial are reversed once withdrawals are taken into account.

3.3 Summary and Conclusions

- Protocols specify all aspects of a clinical trial, including trial purpose, patient selection criteria, methods of design and analysis, including randomization, numbers of subjects, techniques for analysis, informed consent form
- Protocol deviations
  intention to treat analysis (default) may lose power for comparison since subjects in treatment groups may not be homogeneous
  per protocol analysis may lead to bias since randomization is compromised, may also lose power by reducing numbers of subjects.

Chapter 4

Randomization

4.1 Why Randomize?

Randomization of patients to treatment groups is necessary

- to safeguard against selection bias
- to try and avoid accidental bias
- to provide the basis for statistical tests

4.1.1 Historical/database controls

One way to avoid the need for choosing which treatment to assign to each patient would be to put all current patients on the new treatment and compare the results with records of previous patients on the 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 and reduce numbers of new patients needed for a trial.

However, this approach suffers from two major problems

1. The patient population may have changed as there are no formal inclusion/exclusion criteria for the historical patients.
2. Ancillary care may improve with time, giving a tendency for performance of the new drug to be exaggerated.

Database controls (controls selected from (current) databases of patients held for any of a variety of reasons) suffer from similar problems. Essentially we cannot say whether any improvement in patients is due to the drug or to the act of being treated (the placebo effect). If recruiting sufficient patients for a trial is problematic, 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.

4.2 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 philosophically reasonable but clumsy, time consuming and **unprofessional** in front of patients! Thus people use tables of random numbers, or generate random numbers in a statistical computer package, instead.

To avoid bias in assigning patients to treatment groups, we need to assign them at random. We create a randomization list in advance, so that when an eligible patient arrives they can be assigned to a treatment according to the next number on the list. Randomization lists can be made as long as desired and one should make it long enough to complete the whole trial. In fact, in double blind trials, the randomization list is produced centrally and then numbered packs are assembled containing the treatment assigned, but otherwise identical. Each patient receives the next numbered pack when entering the trial and neither the doctor nor the patient knows what treatment the pack contains. The randomization code is 'broken' only at the end of the trial, when the analysis starts. Even then, the statistician may not be told which of A and B is the placebo and which the active treatment.

Throughout this chapter we will use the following random digits in our examples. They are taken from Neave, table 7.1, starting at row 26, column 1.

3 0 4 5 8 4 9 2 0 7 6 2 3 5 8 4 1 5 3 2 . . . .

Example 4.1 12 patients to be assigned at random to 2 treatments, A and B. We might decide

0 to 4 ⟶ A
5 to 9 ⟶ B

and so obtain

A A A B B A B A A B B A

Example 4.2 With 3 treatments A, B, C
Decide

1 to 3 ⟶ A
4 to 6 ⟶ B
7 to 9 ⟶ C
0 ⟶ ignore

and obtain

A B B C B C A C B A A B

Disadvantage
May lack balance, especially in small trials.
E.g. in Ex 4.1 7A's, 5B's and in Ex 4.2, 4A's, 5B's, 3C's

Advantages

1. Each treatment is completely unpredictable
2. Probability theory guarantees that, in the long run, the numbers of patients on each treatment will not be substantially different.

4.3 Restricted Randomization

4.3.1 Blocking

Block randomization ensures equal treatment numbers (balance) 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 (not just to a single patient).

In Ex 4.1 (12 patients, 2 treatments A & B)
Let

0 to 4 ⟶ AB

and

5 to 9 ⟶ BA,

Then our standard sequence of digits implies

AB AB AB BA BA AB

NB Gaps here are just for convenience in reading.

In Ex 4.2 (3 treatments A, B & C)

1	⟶	ABC
2	⟶	ACB
3	⟶	BAC
4	⟶	BCA
5	⟶	CAB
6	⟶	CBA
7,8,9,0	⟶	ignore

Gives 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.

Example 4.3

1	⟶	AABB
2	⟶	ABAB
3	⟶	ABBA
4	⟶	BBAA
5	⟶	BABA
6	⟶	BAAB
7,8,9,0	⟶	ignore

Gives ABBA BBAA BABA

Problem Towards 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.

4.3.2 Unequal Allocation

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

E.g.
A – Standard
B – New

May decide we need most information on B to get more accurate estimates of the B effect; A variation is probably known reasonably well already if it is the standard.

Suppose we decide on a fixed ratio of 1:2, then again we need blocking.

Identify all the 3!/(1!2!) possible orderings of ABB and assign to digits:

1 to 3	⟶	ABB
4 to 6	⟶	BAB
7 to 9	⟶	BBA
0	⟶	ignore

gives ABB BAB BAB BBA.

A trial without 'stratification' (i.e. having 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.

4.3.3 Stratified Randomization (Random Permuted Blocks Within Strata)

It is desirable that treatment groups should be as similar as possible in regard of patient characteristics which might affect the response.

For example, relevant patient factors might be

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. For example, 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. This would affect the credibility of any treatment comparisons and doubt might be cast on whether the randomization had been done correctly. However we could avoid it by using a **stratified randomization scheme**.

Here we prepare a separate randomization list for each stratum.

Example 4.4 Considering age and sex only, the 4 possible strata are
<50, M ≥ 50, M <50, F ≥ 50, F

We produce separate lists:

$<$50, M	A B B A	B B A A	...
$\geq$50, M	B A B A	B A A B	...
$<$50, F	A B A B	B A A B	...
$\geq$50, F	A B A B	A B B A	...

So, as a new patient enters the trial, the treatment assigned is the next specified on the list corresponding to their age and sex.

4.3.4 Minimization

If there are many factors, stratification may be practically unmanageable. 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 loses 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, as shown in the example below.

Example 4.5 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' ($<$50/$\geq$50)
'disease free-time' ($<$2/$\geq$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 $\geq$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 is given in the table below). We now calculate a 'score':

	Factor	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.

Therefore, put new patient on A, because this will balance up the scores.

NB 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 updated record of treatment assignments by patient factors. Computer software is available to help with this (see below).

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. Nonetheless, it may be useful to introduce some randomness into the minimization procedure. This can be done by assigning the treatment of choice (i.e. the one with smaller sum of marginal totals or 'score') with probability p where 0.5<p<1 (e.g. p= 0.75 might be a suitable choice).

Hence, before the trial starts, one could prepare two randomization lists. The first is a simple randomization list where A and B occur equally often, for use only when the two treatments have equal scores. The second is a list in which the treatment with the smaller score occurs with probability 0.75 while the other treatment occurs with probability 0.25. Using a table of random numbers this is prepared by assigning S (=Smaller Score) for digits 1 to 6 and L (=Larger Score) for digits 7 or 8 (ignoring 9 and 0).

Terminology

Note that some authors use the term ‘Adaptive Randomization’ as a synonym for ‘Minimization’, but that term is best reserved for situations where the **outcomes** of the treatment are available before the next subject is randomized and the randomization scheme is adapted to incorporate information from (results from) the earlier subjects; we do not cover this methodology in this course.

4.4 Randomization Software

A directory of randomization 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.

4.5 Summary and Conclusions

Randomization

- protects against accidental and selection bias
- provides a basis for statistical tests

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
- may not reflect change in ancillary care
- are unable to allow for placebo effect.

Chapter 5

Basic Trial Analysis

5.1 Basics of Hypothesis Testing

Before considering the basic analysis of the common experimental designs outlined in Chapter 2, we recap on the basics of conducting and reporting statistical hypothesis tests, mainly via two common examples. You should note that although Bayesian anlyses are commonplace in some areas of Medical Statistics (e.g. Health Economics), they are not yet fully established in the area of Clinical Trials; thus we restrict our attention to the classical framework and, in essence, to the use of hypothesis tests.

A hypothesis test is a method for deciding between two hypotheses: a **null hypothesis** which will be adopted unless there is significant evidence from the available data that the **alternative hypothesis** is, in fact, more plausible. You will have studied hypothesis testing before, but key points to recall are:

- the treatment of the hypotheses is asymmetric, so that the null and alternative must be correctly identified. The null is the hypothesis of 'no difference', 'status quo', 'one we don't want to be true',...

- A suitable **test statistic** must be identified. This can be done by formal methods (Rao-Blackwell Theory, GLRT, etc) but usually agrees with common sense. A statistic must highlight differences between the hypotheses and must increase as this difference increases. It must also have a known (or calculable) **sampling distribution** under the assumption that the null hypothesis is true (thus initial 'guesses' at test statistics are often modified, e.g. by scaling according to anticipated variability, to give a 'nice' distribution).

- A hypothesis test is conducted by calculating the probability of observing a test statistic as extreme as, or more extreme than, the value observed in the sample data. This is calculated from its sampling distribution, i.e. under the assumption that the null hypothesis is true.

- The test result is reported in terms of this **significance probability** or ***p*-value** and its interpretation on a conventional scale. Suitable adjectives for describing the weight of evidence against the null hypothesis are

	p > 0.1	No evidence against H_0/Data consistent with H_0
0.05 <	p <0.1	Weak evidence against H_0
0.01 <	p <0.05	Some evidence against H_0
0.001 <	p <0.01	Strong evidence against H_0
<	p <0.001	Very strong evidence against H_0

- If a significant effect is found, its size should be quantified in some way; typically by giving a confidence interval.

- The phrasing of the conclusion is important and prone to error. Always explain technical conclusion in words (so never talk about rejecting H_0, etc). Take care over phrasing of conclusions, even if the wording then seems formulaic. Consider the following example:

  - DO NOT SAY
    p< 0.05, reject H_0
  - TRY
    There is some evidence ($p = 0.045$) to reject the hypothesis that students weigh the same, on average, before and after a semester in hall. It appears that students tend to weigh less afterwards, by an average of 3.2lbs, 95% CI (0.09,6.31)lbs.
  - DO NOT CONTRACT THIS IMPLYING 1-SIDED TEST
    There is some evidence ($p = 0.045$) that students weigh less after a semester in hall.

- It is important to take care over what can be concluded from a significant *p*-value. Even if there is substantial evidence in favour of the alternative, it does not mean that it is definitely correct. Further, an effect can be statistically significant without being practically important. For example, it may be detected that mean systolic blood pressure in two groups differs, but an estimate of that difference might be 0.2mmHg and since blood pressure is often only recorded to the nearest 5mmHg, this makes little practical difference.

- It is also important to take care over what can be concluded from a non-significant p-value. Insufficient evidence to reject the null hypothesis does not mean that it is actually true. It may be true, or it may simply be that we have not found enough evidence to detect its falsity (perhaps because it, randomly, happened to exhibit characteristics similar to the null, but perhaps because our sample was too small and we have simply not looked 'properly').

- It is important to check validity of the assumptions underlying applicaibility of any test. Sometimes alternative tests, making weaker assumptions, are available if the first choice test is inappropriate.

Although we have considered several possible designs for Clinical Trials, and hinted at many more, it is true that many of their analyses come down to (combinations of) two standard situations:

- comparing 'typical levels' in two or more groups

- assessing whether counts follow an 'expected' pattern.

Classical tests for these situations are **t-test** and χ^2 tests. We review these via examples in the following sections.

Note: In practice, Clinical Trials do not come down to single, simple questions of this type, and will typically be analyzed using a range of tests and regression models, but these simple problems do lie at the conceptual core of the methods. We give an example of multiple regression modelling to illustrate the contribution of basic tests to more complex approaches.

5.2 Example: Two-Sample *t*-Test

Here we review the two-sample t-test, in its 'two-sided, separate variance with simplified degrees of freedom' form (see Sections 5.5.1 and 5.5.2). If we wish to compare levels in more than two groups, this generalizes to an F-test (ANOVA).

If we wish to the level in one group with a standard value, we will require instead a one-samplet-test. This will also be suitable if we wish to compare levels in two **matched** groups.

Example 5.1 Birthweights and Smoking
Question: As part of a study of birth weight of babies and parental smoking habits the following data (in kg) were collected on 16 babies whose mothers smoked more than 20 cigarettes a day and 64 babies whose mothers were non-smokers.

	n	$\bar{x}$	s^2
smoking mother	16	3.54	0.151
non-smoking mother	64	3.91	0.164

Do the data provide evidence that the mean birth weight of babies born to smoking mothers is lower than that of babies born to non-smoking mothers?

What assumptions do you make to have a valid test?

Solution:
We have two independent random samples, and population variances are unknown $\Rightarrow$ use 2 sample t-test.

The first trap to avoid is conducting a one-sided test (as suggested by the wording of the question). See Section 5.4.1 for the reasoning here. In fact, we wish to test $H_0 : \mu_X = \mu_Y$ versus $H_1 : \mu_X \neq \mu_Y$, where μ_X and μ_Y denote the mean birth weights in the populations of smoking and non-smoking mothers respectively.

Under H_0, and some assumptions – see later, $(\overline{X} - \overline{Y})/(S_X^2/16 + S_Y^2/64)^{1/2} \sim t_\nu$, approx, where ν is given by $min(16, 64) = 16$. The observed value of the test statistic is

$$t_{obs} = |\frac{\overline{x} - \overline{y}}{(s_X^2/16 + s_Y^2/64)^{1/2}}| = |\frac{3.54 - 3.91}{(0.151/16 + 0.164/64)^{1/2}}| = |\frac{-0.37}{\sqrt{0.012}}| = 3.378.$$

The approx p-value of this two-sided test is $P(|t_{16}| > 3.378)$. From tables (or from R):

q	0.995	0.999
$t_{16,q}$	2.921	3.686

Remembering to double tail probabilities to allow for our two-sided test (or the modulus in the test statistic), we see that therefore, $0.002 < p$-value < 0.01, and there is “strong” evidence against H_0.

A 95% C.I. for $\mu_X - \mu_Y$ is $-0.37 \pm 2.12 \times \sqrt{0.012} = (-0.60, -0.14)$kg.

Report result in context of problem:
“The sample data provide strong evidence, $0.002 <$ p-value<0.01, that the average birth weight of babies with smoking mothers is different from that for non-smoking mothers. It seems that typically babies of smoking mothers weigh less and the sample results in a 95% confidence interval for the difference in average birth weights of (0.14,0.60)kg.”

Assumptions are that birth weights in both populations are normally distributed and that we have independent random samples.

5.3 Example: Chi-Squared Test of Independence

χ^2 tests occur in many situations where counts are recorded. This is because they are suitable for testing whether observed counts agree reasonably with ‘expected’ counts. The hypothesis specifying what is expected may be of several types:

- that two groups have similar ‘typical’ levels, i.e. like a discrete version of the t-test (see Chapter 6)
- that two classifying factors act independently
- that some underlying theory is true (e.g. a simple model for dominant and recessive genetic inheritance might predict phenotypes in a 3:1 ratio).

Essentially they occur because we often assume basic quantities are Normally distributed and then look at (sums of) squared quantities, which means these sums should have χ^2 distributions.

In this section we look at the test in the context of assessing independence (but see below for an alternative phrasing).

Example 5.2 Blood Groups and Social Class
Question: A survey reported in *Nature* in 1983 found that 2648 native Yorkshire

blood donors could be classified according to blood group and social class as follows:

		Blood group	
		A	not A
Social	I–II	257	297
Class	III–V	866	1228

Carry out a χ^2 test to determine whether there is an association between blood group and social class.

Solution:

Expected numbers under H_0, e_{ij} say, are:

234.9	319.1
888.1	1205.9

and column percentages are:

22.9	19.5
77.1	80.5

Recall that e_{ij} is given by

$$e_{ij} = \frac{\text{row } i \text{ total * column } j \text{ total}}{\text{overall total}}.$$

From the e_{ij} and the observed data, o_{ij}, or directly using R,

$$X^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 4.56,$$

and comparing with χ^2_1 gives a p-value of 0.033.

Conclusion: there's evidence ($p = 0.03$) of association between blood group and social class in the population from which the blood donors were drawn. Those with blood group A are more often of social classes I-II than those of other blood groups (22.9% vs 19.5%).

Note that this question might equally have been phrased as ‘test the hypothesis that there is no difference in the proportion of donors with blood group A in classes I-II and III-V’ and approached through a Binomial test. After approximation, the results would be identical (apart from rounding error), since tests are equivalent.

5.4 Example: Multiple Regression

Typically in clinical trials there will be many factors which affect a patient's response to a treatment. A simple experiment would just construct two 'similar' groups, apply the treatment to one and compare the group responses. This relies on random allocation to keep the baseline for the comparison reasonably constant, allowing any significant difference to be attributed to a treatment effect. A much more flexible approach is to record the value of many potentially influential covariates (often called 'prognostic factors') and use regression to adjust for these, so obtaining a 'purer' test of treatment effect. In this form regression is sometimes called 'analysis of covariance'. Here we illustrate how the one-sample *t*-test, in its 'two-sided form, is used to test for necessity of a various terms, most particularly a treatment effect, in a regression model.

Example 5.3 Infant Calcium Levels (Pocock, 1983)
Study: A study investigating whether treating pregnant mothers with calcium affected infant calcium levels wanted also to adjust for various prognostic factors:

Factor	**Coding**
Treatment	Control=0, vit D=1
Type of feed	Artificial=0, breast=1
Sex of infant	Male=0, female=1
Maternal age	Age in years
Total parity	Parity (with $\geq$3 set=3)
Social class	Classes I to V scored 1 to 5, unmarried mothers=3
Marital status	Married=0, unmarried=1
Birth weight	Weight in Kg
Gestation period	Gestation in weeks (with $\leq$37 set=37)
Special care unit (SCU)	Not in SCU=0, in SCU=1
Pre-eclamptic toxaemia (PET)	No PET=0, PET=1

A multiple regression approach was used. The full model output is shown below. One sample *t*-tests are used to establish the significance level of individual terms (details not shown, part of standard package output).

Factor	**Regression coefficient**	**Standard error**	**Significance**
Constant	7.488		
Treatment	+0.354	0.103	$p < 0.001$
Type of feed	+0.717	0.115	$p < 0.001$
Sex of infant	+0.256	0.100	$p = 0.01$
Maternal age	-0.225	0.270	
Total parity	-0.014	0.058	
Social class	-0.067	0.054	
Marital status	-0.025	0.192	
Birth weight	+0.070	0.120	
Gestation period	+0.053	0.047	
Special care unit (SCU)	-0.254	0.170	
Pre-eclamptic toxaemia (PET)	-0.425	0.470	

This full model suggests that the first three factors might be important, but formal model selection procedures should be followed to establish a suitable reduced model (NB because data in clinical trials are likely to be unbalanced, significance may depend on the order in which terms are added and which other terms are present). A suitable reduced model is shown below (note the values of the coefficients for the retained terms have changed slightly).

Factor	**Regression coefficient**	**Standard error**	**Significance**
Constant	8.686		
Treatment	+0.336	0.101	$p < 0.001$
Type of feed	+0.771	0.111	$p < 0.001$
Sex of infant	+0.254	0.098	$p = 0.01$

5.5 Additional Comments on Testing Procedure

We make two further general comments on the use of statistical tests in Medical Statistics. The first is on the question of whether to use one- or two-sided tests; the other is when considering use of a t-test whether to use the separate or pooled version and about testing for equality of variance first.

5.5.1 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 clinicians who often say things like: "I know this drug can only possibly **lower** mean systolic blood pressure, so I **must** use a one-sided test of $H_0 : \ \mu = \mu_0 \ v \ H_A : \ \mu < \mu_0$ 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 'show' 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 statistical theory would declare that you should ignore this evidence and so fail to detect that the drug is in fact harmful.

appreciation only ↓

A more difficult case is if there is a suspicion that a supplier is adulterating milk with water. This would result in a lower freezing point for the milk. If we propose to test the suspicions by testing the freezing point of several samples of milk, should we use a one- or two-sided test? If our narrow interest is only in determining whether the milk is being adulterated with water, we should indeed use a one-sided version, but we must realize that this would exclude chances of detecting other additions, some of which might raise the freezing point.

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 in advance (with certified documentary evidence)).

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 show 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 course Further Clinical Trials. appreciation only ends ↑

5.5.2 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, sample means $\bar{x_1}$ and $\bar{x_2}$ and sample variances S_1^2 and S_2^2 of the two samples, then the two versions of a 2-sample t-test are (NB do not forget the modulus sign):

1. **Separate variance**

$$t_r = |\frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}|$$

   Here 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.

2. **Pooled variance**

$$t_r = |\frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}(\frac{1}{n_1} + \frac{1}{n_2})}}|$$

   where $r = n_1 + n_2 - 2$. This version assumes that the variances of the two samples are equal. It 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 Chapter 6) where a pooled variance is used for entirely pragmatic reasons and because many approximations are necessary to obtain any answer at all and this one is not as 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.

5.5.3 Testing 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, to 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, though 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 such an approach are

1. tests of equality of variance are very low powered (unless you have large amounts of data)
2. as usual, a non-significant result does not mean that the variances are truly equal, only that the evidence for them being different is weak
3. also, 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 (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'.

5.6 Possible Analyses for Specific Designs

NB The suggestions given here assume the response is continuous. We say more about analysis of trials with binary outcomes in Chapter 11.

5.6.1 Parallel Group Designs

Data type	2 groups	>2 groups
Normal	t-test	1-way ANOVA
Non-parametric	Mann-Whitney	Kruskal-Wallis

5.6.2 In Series Designs

Data type	2 groups	>2 groups
Normal	Paired t-test (on diffs)	2-way ANOVA
Non-parametric	Wilcoxon signed rank test	Friedman's test

Crossover Design

Once again, as an in series design, the test for the treatment effect is a paired t-test (one-sample t-test) on the differences between the responses under the two treatments (or non-parametric 'equivalent'). However, one might also wish to test for period and carryover effects; see Chapter 8.

5.6.3 Factorial Designs

An appropriate (multi-way) ANOVA (or non-parametric 'equivalent') is needed to take account of the particular combinations applied and their potential interactions.

5.6.4 Sequential Designs

An outline of the analysis was given when the design was introduced. Some more pertinent comments are given in Chapter 7.

5.7 Summary and Conclusions

- Recall previous instruction in use of hypothesis tests, especially care over quantifying and phrasing of 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–treatment effect is tested using suitable two- (or more) group test

- In series designs – treatment effect is tested using suitable paired test

- Crossover designs – treatment effect is tested using suitable paired test, but also need to consider other potential effects.

- Factorial designs – analysis via suitable ANOVA, allowing for particular treatment combinations applied and potential interactions between them

- Sequential designs – specialist analysis allowing for accruing evidence.

- Regression is often used to adjust for prognostic factors.

Chapter 6

Size of the Trial

6.1 Introduction

One of the most common tasks of a medical statistician is to perform **sample size calculations**.

Essentially we wish to answer the following question:
"What sample sizes are required to have a good chance of detecting clinically relevant differences if they exist?"

In order to do so, the following specifications are required:

1. Trial type
2. Main outcome measure (e.g. μ_A, μ_B estimated by $\bar{X_A}, \bar{X_B}$)
3. Method of analysis (e.g. two-sample t-test)
4. Result given on standard treatment (or pilot results)
5. How small a difference is it important to detect? i.e. the **clinically relevant difference, CRD** (e.g. $\delta = \mu_A - \mu_B$)
6. Degree of certainty with which we wish to detect it, (**power**, 1-β).

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; α represents risk (probability) of false positive result
  Type II: false negative; treatments different but result non-significant; β represents risk of false negative result.

Essentially what we do is to identify the value of the sample size for which the test would just indicate borderline significance (we may need to make assumptions in order to make the algebra viable).

6.2 Binary Data

NB This section gives an alternative formulation of the type of problem we tackled using a χ^2 test in Section 5.3.

We will work through this case via a numerical example. In Section 6.3 we work the continuous response case algebraically.

Our specifications are as follows.

1. **Trial type**
   Parallel group; 2 treatments; we will look at the (usual) case of requiring equal numbers on the two treatments.

2. **Outcome**
   Count numbers of 'Successes' and 'Failures'

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

3. **Test method**
   Model: $X_1 \sim Bi(n, \theta_1)$ and $X_2 \sim Bi(n, \theta_2)$ where X_1 and X_2 are the numbers of successes on standard and new treatments.
   Hypotheses: $H_0 : \ \theta_1 = \theta_2 \ v \ H_1 : \ \theta_1 \neq \theta_2$
   (i.e. a 2-sided test of proportions).
   Approximation: Take Normal approximation to binomial
   $X_1 \sim N(n\theta_1, n\theta_1(1-\theta_1))$ and $X_2 \sim N(n\theta_2, n\theta_2(1-\theta_2))$
   Requirement: take α = P[type I error] = level of test = 5%.

4. **Standard results**
   Suppose standard treatment gives 90% success.

5. **Clinically relevant difference (CRD)**
   It is of clinical interest if the new treatment gives 95% success (or better),
   i.e. $\theta_1 \;=\; 0.9$
   and $\theta_2 \;=\; 0.95$
   i.e. a 5% point improvement. NB it is simply coincidence that this improvement is the same as the level of the test in this example.

6. **Power requirement**
   Suppose we want to be 90% sure of detecting an improvement of 5%,
   i.e. $\gamma(0.95) \;=\; 0.9$,
   where $1-\beta \;=\; \gamma$ is the power of the test.

We now follow through the development of the test statistic.

We have (using our earlier Normality approximations)

$$\frac{X_2}{n} - \frac{X_1}{n} \;\sim\; N(\theta_2 - \theta_1, [\theta_2(1-\theta_2) + \theta_1(1-\theta_1)]/n) \tag{6.1}$$

since

$$\begin{aligned} var(\frac{X_2}{n} - \frac{X_1}{n}) &= var(\frac{X_2}{n}) + var(\frac{X_1}{n}) \\ &= \frac{\theta_2(1-\theta_2)}{n} + \frac{\theta_1(1-\theta_1)}{n}. \end{aligned}$$

The test statistic is therefore

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

and (remembering $\theta_1 = \theta_2 = 0.9$ under H_0) we will reject H_0 at the 5% level if

$$|\frac{X_2}{n} - \frac{X_1}{n}| > 1.96\sqrt{\frac{2 \times 0.9 \times 0.1}{n}}.$$

Recall that the power function of the test is

$$\begin{aligned} \gamma(\theta_2) &= P[\text{reject } H_0|\text{alternative parameter } \theta_2] \\ &= P[|\frac{X_2}{n} - \frac{X_1}{n}| > 1.96\sqrt{\frac{2 \times 0.9 \times 0.1}{n}}|\theta_1 = 0.9, \theta_2] \end{aligned}$$

and we have the specific requirement that $\gamma(0.95) = 0.9$.

Now we can identify the distribution (6.1) in the case θ_1=0.9 and θ_2=0.95, and hence evaluate the power for this θ_2 value as

$$\begin{aligned}\gamma(0.95) &= 1 - P[|\frac{X_2}{n} - \frac{X_1}{n}| \leq 1.96\sqrt{\frac{2 \times 0.9 \times 0.1}{n}}|\theta_1 = 0.9, \theta_2 = 0.95] \\ &= 1 - (\Phi[\frac{1.96\sqrt{\frac{2\times0.9\times0.1}{n}} - 0.05}{\sqrt{\frac{0.95\times0.05}{n} + \frac{0.9\times0.1}{n}}}] - \Phi[\frac{-1.96\sqrt{\frac{2\times0.9\times0.1}{n}} - 0.05}{\sqrt{\frac{0.95\times0.05}{n} + \frac{0.9\times0.1}{n}}}])\end{aligned}$$

and the last term can be approximated by

$$\Phi[-1.96 - \frac{0.05\sqrt{n}}{\sqrt{0.95 \times 0.05 + 0.9 \times 0.1}}] \longrightarrow 0$$

so we require

$$\Phi[\frac{1.96\sqrt{2 \times 0.9 \times 0.1} - 0.05\sqrt{n}}{\sqrt{0.95 \times 0.05 + 0.9 \times 0.1}}] \simeq 0.1$$

i.e.

$$n \simeq \frac{0.95 \times 0.05 + 0.9 \times 0.1}{0.05^2}[\Phi^{-1}(0.1) - 1.96\sqrt{\frac{2 \times 0.9 \times 0.1}{0.95 \times 0.05 + 0.9 \times 0.1}}]^2$$

i.e. this approximate formula suggests we need around 680 patients in each ‘arm’ of the trial (1 360 in total), or more if we anticipate drop outs (see Section 6.5).

General formula
In the general case we have, making the further approximation given in Note 3 below,

$$n \simeq \frac{\theta_2(1-\theta_2) + \theta_1(1-\theta_1)}{(\theta_2 - \theta_1)^2}[\Phi^{-1}(\beta) + \Phi^{-1}(\alpha/2)]^2$$

Notes

1. Both $\Phi^{-1}(\beta)$ and $\Phi^{-1}(\alpha/2)$ will be negative for sensible α and β.

2. θ_1 and θ_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.

3. The approximation at the final stage requires, takes the multiplier of $\Phi^{-1}(\alpha/2)$ as 1, i.e.
$$\frac{\sqrt{2\theta_1(1-\theta_1)}}{\sqrt{\theta_2(1-\theta_2)+\theta_1(1-\theta_1)}} = 1$$
In the example its actual value is 1.14, so reasonable; otherwise need to use more complex methods.

4. Machin & Campbell (Blackwell, 1997) provide tables for various θ_1, θ_2, α and β. There are also computer programmes available.

5. If we can really justify a 1-sided test, then replace $\Phi^{-1}(\alpha/2)$ by $\Phi^{-1}(\alpha)$. One-sided testing reduces the required sample size.

6. For given α and β, n depends mainly on $(\theta_2 - \theta_1)^2$ (and 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.

7. 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. Pocock suggests it is unethical to spend money on such trials.

8. 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 three times the number of patients.

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

6.3 Continuous (Quantitative) Data

Considering the same specifications as in the binary case, here we have

1. Parallel group design.

2. Quantitative outcome; standard treatment has mean μ_1, new has μ_2.

3. Analyze with two-sample t-test, **but** assume n large, so use Normal approximation and also assume that the groups have equal, known variance and consider the case of equal sample sizes. This approximation should be reasonable as long as the variances are not too diferent.

4. Assume the mean on the standard treatment, μ_1, is known.

5. Want to detect that the new mean is μ_2, or equivalently that the change is $\delta \;=\; \mu_2 - \mu_1$.

6. Degree of certainty with which we wish to detect this change is 1-β, i.e. $\gamma(\mu_2) = 1 - \beta$.

Test statistic under $H_0: \;\mu_1 = \mu_2$ is

$$T = \frac{(\bar{X_2} - \bar{X_1}) - 0}{\sqrt{\frac{2\sigma^2}{n}}} \simeq N(0,1).$$

The 2-sided, level α test rejects H_0 if

$$|\frac{(\bar{X_2} - \bar{X_1}) - 0}{\sqrt{\frac{2\sigma^2}{n}}}| \;>\; -\Phi^{-1}(\alpha/2).$$

The power function if the new mean $= \mu_2$ is therefore

$$\begin{aligned}\gamma(\mu_2) &= 1 - P[|\frac{(\bar{X_2} - \bar{X_1}) - 0}{\sqrt{\frac{2\sigma^2}{n}}}| \;\leq\; -\Phi^{-1}(\alpha/2)|\bar{X_2} - \bar{X_1} \sim N(\mu_2 - \mu_1, 2\sigma^2/n)] \\ &= 1 - [\Phi(-\Phi^{-1}(\alpha/2) - \frac{\mu_2 - \mu_1}{\sqrt{2\sigma^2/n}}) - \Phi(+\Phi^{-1}(\alpha/2) - \frac{\mu_2 - \mu_1}{\sqrt{2\sigma^2/n}})]\end{aligned}$$

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

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

As before, the second term $\longrightarrow 0$ as $n \longrightarrow \infty$, so we need

$$\Phi^{-1}(\beta) \simeq -\Phi^{-1}(\alpha/2) - (\mu_2 - \mu_1)\sqrt{n/2\sigma^2}$$

or

$$n \simeq \frac{2\sigma^2}{(\mu_2 - \mu_1)^2}[\Phi^{-1}(\beta) + \Phi^{-1}(\alpha/2)]^2.$$

Notes

1. All comments in binomial case apply here also.

2. Need to know the variance σ^2 which is difficult in practice. Techniques which can help determine a reasonable guess at a value for it are:
    - look at similar earlier studies,
    - to run a small pilot study,
    - to say what the likely maximum and minimum possible responses under the standard treatment could be and so calculate the likely maximum possible range and then get an approximate value for σ as one quarter (or one sixth) of the range. Here the rationale is the recognition that for Normal data an approximate 95% confidence interval is $\mu \pm 2\sigma$ so the difference between the maximum and minimum is roughly 4σ.

6.4 One-Sample Tests

The two formulae given above apply to two-sample tests for proportions and means. It is straightforward to derive similar formulae for the corresponding one-sample tests. Obviosuly then you do not need to double the calculated size for a second group. In the case of a one-sample test, the required sample size to achieve a power of $1-\beta$ when using a size α test of detecting a change from a proportion θ_0 to θ is given by

$$n \simeq \frac{[\Phi^{-1}(\beta)\sqrt{\theta(1-\theta)} + \Phi^{-1}(\alpha/2)\sqrt{\theta_0(1-\theta_0)}]^2}{(\theta-\theta_0)^2}$$

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 α test of detecting a change from a mean μ_0 to μ is given by

$$n \simeq \frac{\sigma^2}{(\mu-\mu_0)^2}[\Phi^{-1}(\beta) + \Phi^{-1}(\alpha/2)]^2.$$

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

6.5 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

- increase δ
- relax α and β and accept that small differences may be missed
- think of using a multicentre trial (see later).

2. We must allow for dropouts, missing data, etc. For example, we might inflate required numbers by 20% to allow for such losses.
3. Statistical procedures must be as efficient as possible; we should consider more complex designs.

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

6.6 Computer Implementation

R, S-PLUS and MINITAB provide extensive facilities for power and sample size calculations. The approximations used in the packages (and in the working above) vary and the resulting sample sizes can differ by as much as 10%. 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 a small DOS program, power.exe, which 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)

However, it may not run on 64-bit machines.

There are also links to other free sources on the web. The industry standard is probably nQuery Adviser, but this is quite costly.

6.6.1 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 three of power, sample size, significance level and CRD (referred to as 'delta' in R) 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.

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 1164 is needed, in close agreement with that determined by the approximate formula in Section 6.2.

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
```

6.7 Summary and Conclusions

1. 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
2. For sample size calculations we need to know
   - outcome measure
   - method of analysis (including desired significance levels)
   - clinically relevant difference
   - power
   - results on standard treatment (including likely variability)
3. 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

4. Sample size facilities in R 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.

   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.

Chapter 7

Multiplicity Problems

7.1 Introduction

This chapter 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. A particular danger is when the most statistically significant result is selected from amongst the rest for particular attention, perhaps quite unintentionally.

Similar problems can, of course, arise in many other areas, but they seem particularly acute in clinical trials since experiments on human subjects are 'expensive' and so there is great pressure to measure as much as possible and to extract every bit of information possible from the results.

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

- multiple endpoints
- subgroup analyses
- interim analyses
- multiple regression
- 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 (including repeated measures analysis)).

We examine each of these situations in turn, explaining the issues involved and how they might be remedied. Examples are given, artificial or from the literature. Several are taken from the book by Andersen (1990), who discusses the topic in a fair amount of detail. Fortunately, many of the examples from the literature are relatively old, since the standard of statistical reviewing for the main medical journals has improved vastly in recent years, but care should still be taken to avoid the pitfalls exemplified here.

7.1.1 Example: Signs of the Zodiac (Effect of new dietary control regime)

We illustrate the issues with a brief example (constructed artificially but not far from reality).

Data: 250 subjects chosen 'randomly'. Weighed at the start of the week and again at the 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.
95% confidence interval for mean weight loss is (-0.009, 0.260)kg.
Paired t-test for weight loss gives a t-statistic of t=0.126/0.068=1.84, giving a p-value of 0.067 (using a two-sided test).

Not quite significant at the 5% level!

Can anything be done to 'squeeze' a significant result out of this expensive study (other than a 1-sided test that we already know is not acceptable!)? 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	SE(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 ...

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 analyzing the data more honestly.

7.1.2 Fundamental Issues

In clinical trials a large amount of information accumulates quickly and it is tempting to analyze many different responses, i.e. to consider multiple end points or perform many hypothesis tests on different combinations of subgroups of subjects. We must 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. This risk is controlled for each individual test, since that is precisely what is meant by the significance level of the test or the p-value. Recall that the p-value is the precise calculation of the risk of a false positive result and is more commonly quoted in current literature. The significance level is 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 (but may either be much larger than 0.05 or only slightly greater). The level of the test is set when designing the trial.

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 $95\% \times 95\% = 90.25\%$ chance of avoiding a mistake on both 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
P[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.

Essentially the same situation arises in the use of **normal ranges** in clinico-chemical tests or the search for a **typical individual** A normal range comprises (the central) 95% of the values for that chemical. If 100 normal people 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. Using the same reasoning as above, after another 8 tests there will be only around 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 decreases as more and more tests are performed but at the greatly increased risk of more false positives. These problems are more complex and are not considered here, nor are they commonly considered in the medical statistical literature.

7.1.3 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 α and you perform k (independent) significance tests, then each should be run at a nominal α/k level of significance. This value is an approximate solution for ε to the equation

$$\alpha = 1 - (1 - \varepsilon)^k. \tag{7.1}$$

That is, only declare the result significant at level α if, in fact, $p < \varepsilon = \alpha/k$.

This solution is (usually) approximate in two senses. Firstly, equation (7.1) can instead be solved exactly as $\varepsilon = 1 - exp[1/klog(1 - \alpha)]$. This refinement is rarely employed in practice.

More importantly, the tests in question are often not independent. In this case, the overall error risk will depend on their degree of correlation and cannot be simply determined.

The simple $\varepsilon = \alpha/k$ version of the Bonferroni correction is very conservative; it over corrects. Conservative means 'safe'; i.e. you preserve your scientific reputation by avoiding making mistakes, but at the expense of failing to discover something scientifically interesting. Often it will demand a totally unrealistically small p-value and in practice there are better ways of overcoming the problem of multiplicity, by limiting the number of tests by concentrating on the more important objectives of the trial or using a more sophisticated analysis.

7.1.4 Examples

Example Suppose 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.

Example 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.

Example 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 $12 \times 0.019 = 0.228$. 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.

7.2 Multiple Endpoints

7.2.1 The issues

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. For example, in hypertensive studies it is routine to record pulse rate, systolic and diastolic blood pressure, perhaps sitting, standing and supine, before and after exercise. However, separate significance tests on each separate end point comparison increases the chance of some false positives.

7.2.2 Solutions

- 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 solution 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. Of course these should be scrutinized and any causes for concern reported.

The ideal statistical solution 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 (namely Hotelling's T^2 test), Analysis of Variance or ANOVA (namely 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.

7.2.3 Examples

appreciation only ↓

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 (see below).

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. endpoints), repeated at various timepoints and for 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 show 'significant' results. 5% of $850 = 42.5$, so finding 48 is not very impressive.

Finally, 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".

appreciation only ends ↑

7.3 Subgroup Analyses

7.3.1 The issues

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

$\text{M} \leq 50, \quad \text{F} \leq 50, \quad \text{M} > 50 \quad \& \text{ F} > 50.$

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.

7.3.2 Solutions

Possible ways of avoiding these issues are

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

Bonferroni adjustments can be used. They 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 a single test to answer the general question 'Are the subgroups homogeneous in their responses, or do they differ?'. A suitable Analysis of Variance (ANOVA) is the usual approach. However, as soon as a significant result is detected, a follow-up is called for to determine **which** subgroups are 'interesting'. A one-way analysis of variance can be thought of as a generalization 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,; and the Neuman-Keuls test which looks to see which pairs of treatments are different. However there are many others which may be found in commonly used statistical packages.

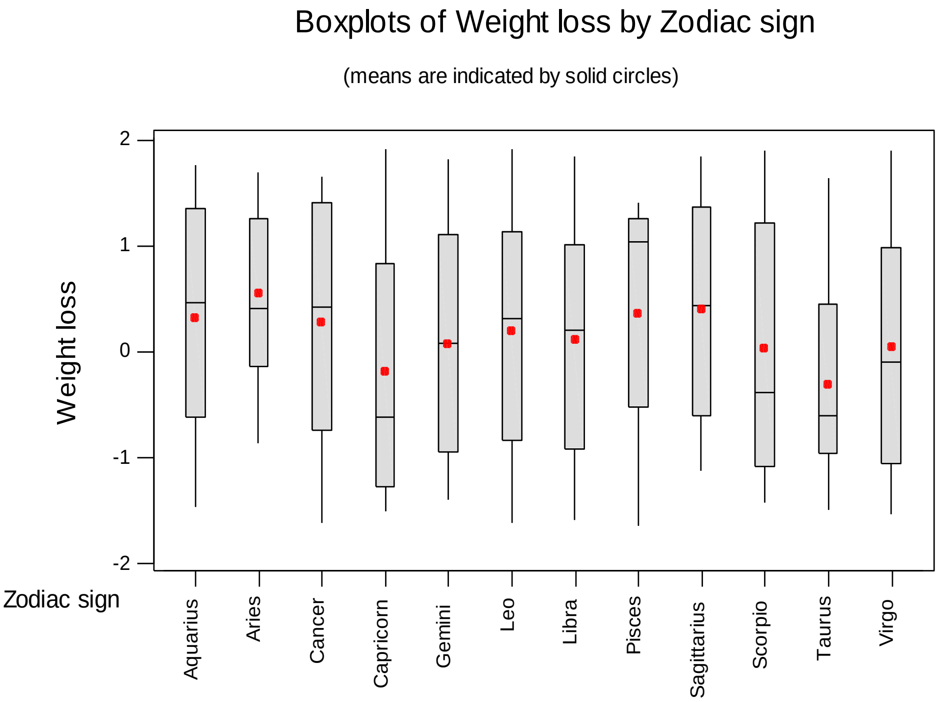
7.3.3 Examples

Zodiac (cont.) 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 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 shows 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 plot below gives little indication that any mean is different from zero.



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 showing an effect then some of them would be clustered towards the lower end of the scale from 0.0 to 1.0. However, examination shows 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.

Identical treatment study, Lee, McNear et al (1980), Circulation

This paper reports an actual clinical double-blind study where two treatments were apparently compared. The 1073 patients with coronary heart disease were randomized into two groups, baseline factors were reasonably balanced, the response was survival time. However, there was an extra unusual element in that, in fact, the two treatments were actually identical and the 'treatment' corresponded just to the random allocation into two groups.

appreciation only ↓

On initial analysis, the overall differences between treatment groups was non-significant. Then subgroup analyses were performed as follows. Six groups were identified on the basis of two baseline factors (left ventricular contraction pattern: normal or abnormal; number of diseased vessels: 1, 2, or 3). A significant difference in survival times was found in one of the groups (abnormal/3, χ^2=5.4, p<0.023) and could be justified scientifically. Sample sizes were quite large: n=397, n_1=194, n_2=203.

In fact, since all patients were treated in the **same** way, a false positive effect had been discovered.

Race and anaemia Andersen (1990) reporting a study (ref: N Engl J Med 1978, 298: 647)

"A survey of racial patterns in pernicious anaemia assessed for age distributions (at presentation) in relation to sex and ethnic group ('European' origin, black patients and Latin American patients). The statistical method was Student's t test.

Blacks (p<0.001) and Latin Americans (p<0.05) were younger than 'Europeans'. However, the significant age differences were confined to the women; the three male groups did not differ significantly from each other. The black women were significantly younger than all the other groups of patients (p<0.001) except the Latin American women and black men, in whom the age difference did not attain statistical significance. Furthermore, a smaller proportion of the black women were 70 years older, and a larger proportion were 40 years or younger than all the other groups. In fact, the age distribution among the black women may be a bimodal one, with one cluster around a median age of 62 and the other around a median age of 31. The Latin American women were not significantly younger than any other group except the 'European' men (p<0.05). Within each racial category, the women tended to be younger than the men, but the differences never reached statistical significance."

It is clear that somewhere in here is evidence of interesting interactions between age, sex and race and a full three-way analysis of variance would elicit this. The p values clearly make no allowance for multiple testing and it is not clear how many were actually performed since only (almost) the significant ones were reported.

appreciation only ends ↑

7.3.4 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 al.* (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 firstly 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 11/165 (6.7%) respectively.

	placebo deaths	metoprolol deaths	
all ages	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%)	19/234 (8.1%)	p=0.03
	?Metoprolol better for elderly		
age 40-69	51/627 (8.1%)	32/629 (5.1%)	p=0.04
age 70-74	11/70 (15.7%)	8/69 (11.6%)	p>0.2
	?Metoprolol better for younger		

As well as the dangers of multiple testing, this example illustrates the dangers of *post hoc* regrouping. Subgroups should be defined on clinical grounds, before the data are collected. Some subgroup effects could be real, of course. However, again we should probably only use subgroup analyses to generate future hypotheses.

7.4 Interim Analyses

7.4.1 The issue

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

The objectives of this periodic checking may be laudable. Typical reasons are to

- check protocol compliance
  Checking 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.

- pick up bad side effects
  Quick action can then be taken and investigators warned to look out for such events in future patients.

- provide feedback
  This 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
  The ability to stop or modify the trial under such circumstances is an ethically attractive reason for interim analyses. 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. Here the remedies are rather different (and considerably more complex) since not only are the tests in the sequence 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.

7.4.2 Solution

To incorporate such interim analyses we must

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

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. The first shows the true false positive risk if we perform multiple tests.

appreciation only ↓

Repeated Significance Tests on Accumulating Data

No. Tests at 5% Level	Overall Sig. 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
∞	1.00

The second table shows how we can

appreciation only end ↑

'invert' the information above and correct the significance levels of the interim tests to maintain our overall risk. Here N is the **maximum** number of interim analyses to be performed. This is decided in advance and included in the protocol. The understanding is that if an interim test proves significant, the trial is stopped at that point.

Nominal Significance Levels Required for Repeated Two-Sided Significance Testing for Various α

N	α=0.05	α=0.01
2	0.0290	0.0056
3	0.0220	0.0041
4	0.0180	0.0033
5	0.0160	0.0028
10	0.0106	0.0018
15	0.0086	0.0015
20	0.0075	0.0013

7.4.3 Examples

Non-Hodgkins Lymphoma, from Pocock (1983, p150)
This is a study to compare two 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 were analyses planned, roughly after every 25th result. The table below gives numbers of 'successes' and nominal p-values using a χ^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).

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. Medics find this difference difficult to understand and accept (after all, the actual data are the same), but this is because they fail to appreciate the way in which statistical tests are constructed in order to limit risk of a Type I error.

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 χ^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 test gave p=0.04 which is not statistically significant, being greater than the required nominal level of 0.016 for N=5 analyses. However, a totally negative interpretation may not be appropriate from these data alone. One might infer that the superiority of the CVP treatment is interesting but not conclusive.

Pancreatitis 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 0.031, and 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 as eventually an apparently significant result will be obtained.

7.4.4 Further Notes

We have seen how one decides in advance what is expected as the maximum number of interim analyses and accordingly makes the nominal significance level smaller. However, one should also consider whether an **overall** type I error α=0.05 is sufficiently small when considering such a stopping rule.

There are two situations where α=0.01 may be more appropriate

- if a trial is unique in that its findings are unlikely to be replicated in future research studies
- if there is more than one patient outcome used in interim analyses and the stopping rule is applied to each outcome.

However, in the latter case, one possibility would be to have one principal outcome with a stopping rule having α=0.05 and have lesser outcomes with α=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 I 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.

7.5 Multiple Regression

7.5.1 The issue

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, logistic regression, or Cox proportional hazards regression (see later).

When analyzing 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 variables is 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 "let's plug everything in and see what comes out", effectively selecting the most significant result for attention.

A trap that is all too easy to fall into arises with interactions. Even with a modest number of variables, the number of possible pairwise interactions can be large: including all of them in a model "to see if any turn out to be significant" invites a false positive result which can be seriously misleading. Considering higher order interactions only exacerbates the problem.

7.5.2 Solutions

Background knowledge should be key as early as suggesting **possible** models. Interaction terms should only be included where prior knowledge indicates they could naturally arise (as should be the case for all postulated variables).

If many variables and interactions really are plausible, then an honest analysis would have to include this feature and make an appropriate correction, such as a Bonferroni one.

7.5.3 Example

appreciation only ↓

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; you are invited to read this article critically.

appreciation only ends ↑

7.6 Repeated Measures

7.6.1 The issue

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 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 show mean values of the two treatment groups plotted against time and which show error bars for each mean invite the eye to do exactly that and so this 'analysis' must be specifically resisted.

7.6.2 Solutions

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

No essentially new comments apply to this situation and indeed some examples discussed earlier include a repeated measure element.

Calculation of summary measures includes calculating quantities such as 'growth rate', 'time to peak value', 'area under the curve' (AUC) which may have an interpretation as reflecting bioavailability. Another common possibility is concentrating on change from baseline (for 'before' and 'after' studies). These methods work by reducing the multivariate responses to a single, univariate, response. In general, they provide good ways of summarizing the data and might well be employed in a preliminary, graphical display, even if more complex responses are retained for the analysis.

Bonferroni adjustments will be 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 which are introduced in Extended Linear Models. They are typically considerably more complex than the simple fixed effects linear models encountered up to level 3.

As always, the form of the analysis should be fixed before collection of the data.

7.7 Summary and Conclusions

Multiplicity can arise in

- testing several different responses
- subgroup analyses
- interim analyses
- multiple regression analyses
- repeated measures

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 (including repeated measures) methods.

Bonferroni adjustments are typically very conservative because in many situations the tests are highly correlated (especially with multiple end points and repeated measures); this is why other, more complex, methods are usually needed.

Chapter 8

Crossover Trials

8.1 Introduction

Recall from Chapter 2 that where it is possible for patients to receive both treatments under comparison, crossover trials may well be more efficient (i.e. need fewer patients) than a parallel group study. The reason is that by acting as his/her own control, the effect of large differences between patients can be lessened by looking at within-patient comparisons.

Example 8.1 (Pocock, p112) Hypertension trial with basic design:

$$\begin{tikzcd}
 & & \text{period 1} & \text{period 2} \\
 & n_1 \arrow[r] & \text{new drug A} \arrow[r] & \text{standard drug B} \\
\text{randomized} \arrow[ur] \arrow[dr] & & & \\
 & n_2 \arrow[r] & \text{standard drug B} \arrow[r] & \text{new drug A}
\end{tikzcd}$$

Response is systolic blood pressure at end of a 5 minute exercise test. There is a total of 109 patients, split
A⟶ B: 55 patients   B ⟶A: 54 patients
Suitable washouts were used before the start of the trial and before each period's active treatment.

Possible effects:

- **treatment effects** – drug A or B is superior

- **period effect** – responses may be generally higher in one period or another, eg for seasonal diseases (eg hayfever) or unstable baselines (eg disease cured)
- **carryover effect** – or **treatment** × **period interaction**; the effect of a drug may depend on when it is given

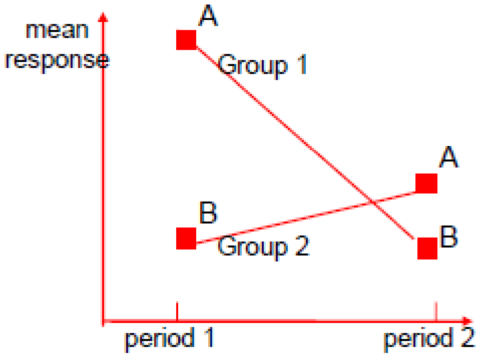
Our main interest is in detecting treatment effects. Period and carryover are generally nuisance effects, hampering our determination of whether there is a treatment effect (though occasionally they may be of interest in their own right).

8.2 Illustration of Effects

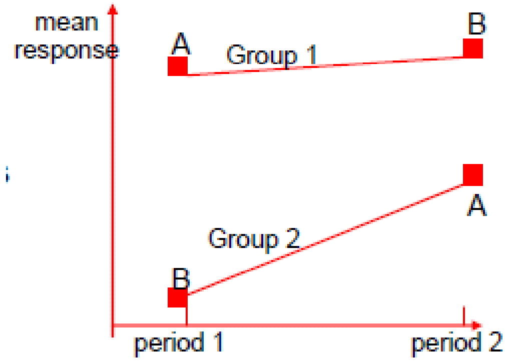
Notes:

1. Assume that 'low' is good throughout.
2. Note similarity to plot in Section 2.4, since we are looking at potential combinations of two main effects and their interaction.

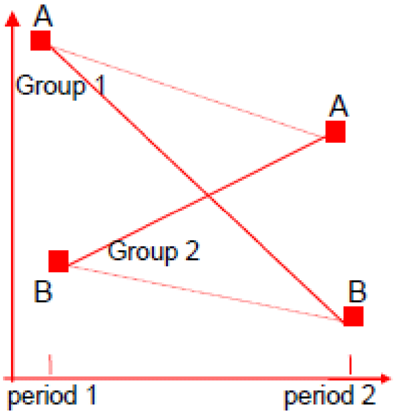
- Carryover Effect (simple)
  Possible explanation: beneficial effect of B carries over into period 2. Unlikely to be detected because of low power.



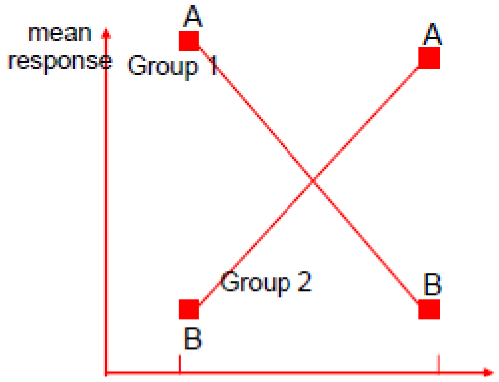
- Carryover Effect (complex)
  Direction of treatment effect different for different periods, caused by carryover. More serious than simple form above.



- Period Effect
  Response in period 2 reduced for both treatments, i.e. patients generally improve so period 2 values on average reduced.



- Treatment Effect
  Treatment effect: B better than A



8.3 Model

Response

Denote the response Y_{ijk} for

group (reflecting treatment order) i;	$i = 1, 2$
period j;	$j = 1, 2$
patient k;	$k = 1, 2, ..., n_i$. NB $n_1 = n_2$ in the balanced case.

Data structure

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

Effects

μ	– overall mean
$\tau_A,\ \tau_B$	– treatment effects
$\pi_1,\ \pi_2$	– period effects
$\lambda_A,\ \lambda_B$	– carryover effects
α_k	– **random** patient effect $\sim$ N(0,ϕ^2) (between patients)
ε_{ijk}	– random errors $\sim$ N(0,σ^2) (independently)

NB Random effects models are discussed in Extended Linear Models.

Model
We can now identify the value of Y_{ijk} for any combination of ijk as

	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	$\mu + \alpha_k + \tau_B + \pi_1 + \varepsilon_{21k}$	$\mu + \alpha_k + \tau_A + \pi_2 + \lambda_B + \varepsilon_{22k}$

Note that if we take expected values, α_k and ε_{ijk} disappear. Thus $E(Y_{ijk})$ is given by

	period 1	period 2
group 1	$\mu + \tau_A + \pi_1$	$\mu + \tau_B + \pi_2 + \lambda_A$
group 2	$\mu + \tau_B + \pi_1$	$\mu + \tau_A + \pi_2 + \lambda_B$

Note: There are several different ways of parameterizing the crossover model; here we are following Jones & Kenward (1989).

8.4 Test Procedure

To isolate τ, π and λ effects, we consider averages (or sums/totals) and differences of the Y_{ijk}s.

8.4.1 Carryover Effect

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

$$T_{1k} \sim N(\mu + (\tau_A + \tau_B + \pi_1 + \pi_2 + \lambda_A)/2, \phi^2 + \sigma^2/2)$$

and

$$T_{2k} \sim N(\mu + (\tau_A + \tau_B + \pi_1 + \pi_2 + \lambda_B)/2, \phi^2 + \sigma^2/2)$$

So, 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

$$H_0: \ \lambda_A = \lambda_B$$

is plausible. So we use

$$\frac{\bar{T}_1 - \bar{T}_2}{\sqrt{\frac{S^2_{T_1}}{n_1} + \frac{S^2_{T_2}}{n_2}}} \sim t_r$$

where $S^2_{T_1}$ is the **sample variance of the** T_{1k} so $v\hat{a}r(\bar{T}_1) = \frac{S^2_{T_1}}{n_1}$, 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

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

where

$$v\hat{a}r(\bar{T}_1 - \bar{T}_2) = \frac{(n_1-1)S^2_{T_1} + (n_2-1)S^2_{T_2}}{n_1+n_2-2}(\frac{1}{n_1} + \frac{1}{n_2}),$$

but it should make little difference in practice.

Example 8.1 (cont.)

	A ⟶ B	B ⟶ A
n_i	55	54
$\bar{t}_i$	176.28	180.17
s_{T_i}	26.56	26.27

We have

$$t = \frac{176.28 - 180.17}{\sqrt{\frac{26.56^2}{55} + \frac{26.27^2}{54}}} = -0.769$$

so clearly non-significant when compared with t_{54}, giving no significant evidence of a carryover effect. NB Don't forget to take the modulus and conduct the required 2-sided test.

Here the pooled t-statistic is

$$t = \frac{176.28 - 180.17}{\sqrt{\frac{54 \times 26.56^2 + 53 \times 26.27^2}{107}(\frac{1}{55} + \frac{1}{54})}} = -0.769$$

i.e. there is little difference because the variances are almost equal anyway, though we would now conduct the test by reference to a t_{107} (not t_{54}) distribution.

Notes

- Ideally we would like to test for $\lambda_A = 0 = \lambda_B$ rather than just $\lambda_A = \lambda_B$, i.e. no carryover at all rather than just no differential carryover. However, this is not possible if we wish to retain a period effect in our model, because the equal carryover and period effect cannot be separately identified.

- The test for carryover typically has low power since it involves between patient comparisons.

- 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.

- If there is a significant carryover effect (i.e. treatment × period interaction), then it is **not sensible** to test for period and treatment separately, so 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). Also, since the trial has been designed as a crossover, it will typically have too few patients to meet power requirements when used as a parallel group study (see Section 8.5).

- We used the average of the two values for each patient (i.e. from period 1 and period 2) in constructing the carryover test. However, the value of the t-statistic would be exactly the same if we used just the **sum** of the two period values. This is slightly easier (as it avoids dividing by 2) and will be seen in some examples.

8.4.2 Treatment and Period Effects

Testing for both of these is carried out on the **assumption that** $\lambda_A = \lambda_B$ and is based on the within subject differences

$$D_{ik} = Y_{i1k} - Y_{i2k}.$$

Note that for group 1 we have

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

while for group 2 we have

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

Treatment Test

$H_0: \ \tau_A = \tau_B$

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.

$$\frac{\bar{D}_1 - \bar{D}_2}{\sqrt{\frac{S^2_{D_1}}{n_1} + \frac{S^2_{D_2}}{n_2}}} \sim t_r$$

where now $S^2_{D_i}$ is the **sample variance of the differences** D_{ik}. 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.

Example 8.1 (cont.)

	A $\longrightarrow$ B	B $\longrightarrow$ A
n_i	55	54
$\bar{d}_i$	5.04	-2.81
s_{D_i}	15.32	19.52

We have

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

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

For comparison, the pooled t-statistic is

$$t = \frac{5.04 - (-2.81)}{\sqrt{\frac{54\times15.32^2+53\times19.52^2}{107}(\frac{1}{55}+\frac{1}{54})}} = 2.34$$

with a p-value of 0.021 when compared with t_{107}, so again there is no material or practical difference from using the different variance estimators.

Period Test

$H_0: \ \pi_1 = \pi_2$

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

$$\frac{\bar{D}_1 - (-\bar{D}_2)}{\sqrt{\frac{S^2_{D_1}}{n_1} + \frac{S^2_{D_2}}{n_2}}} \sim t_r$$

NB it is + in the numerator (not -) since it is still a 2-sample t-test of two 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 $\bar{D_1}$ is the difference between Treatment A and Treatment B results averaged over those in group 1 and (-$\bar{D_2}$) 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.

Example 8.1 (cont.) We have

$$t = \frac{5.04 - (+2.81)}{3.365} = 0.66$$

so no significant evidence of a period effect.

NB We reach the same conclusion from the pooled test.

8.5 Sample Size and Efficiency

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 ρ= correlation between the two measurements on each patient (assuming no carryover effect). Since $\rho > 0$ usually, we will typically need fewer patients in a crossover than in a parallel group study, in fact fewer than half as many.

Sample size calculation facilities for cross-over trials are available in power.exe.

8.6 Further 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. That is, we need expert opinion that either the two treatments cannot interact or that the washout period is sufficient, and cannot rely purely on statistical evidence.

- We can obtain confidence intervals for treatment differences, since

$$(\bar{D}_1 - \bar{D}_2)/2 \sim N(\tau_A - \tau_B, \sigma^2(1/n_1 + 1/n_2)/2)$$

  We estimate σ^2 with a pooled variance estimate or else use separate variances and so say that the (estimated) standard error of $(\bar{D}_1 - \bar{D}_2)/2$ is

$$\sqrt{\frac{1}{4}(\frac{S^2_{\bar{D}_1}}{n_1} + \frac{S^2_{\bar{D}_2}}{n_2})}$$

  and use the approximate formula for (say) a 95% CI of $(\bar{D}_1 - \bar{D}_2)/2 \pm 2 \times s.e.((\bar{D}_1 - \bar{D}_2)/2)$. Using 2 rather than 1.96 is adequate, given the other approximations made.

- 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 even simpler non-parametric sign test, is essentially identical to the case of binary responses considered below.

- Crossover trials can be extended to > 2 treatments and periods, usually when intervals between treatments can be very short. See Example 2.1.

- 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 Design and Sampling course).

- 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.

8.7 Analysis with Linear Models

appreciation only ↓

Since the underlying theory for this section is not covered until later courses, and even then only for some students, this section will not be examinable.

8.7.1 Introduction

The analyses presented above using carefully chosen t-tests provide an illustration of how complex collections of hypotheses may be tested with a series of simple tests in certain circumstances. However, to extend the ideas to more complicated crossover trials 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 $\varepsilon_{ijk} \sim N(0, \sigma^2),\ \alpha_k \sim N(0, \phi^2),\ \Sigma\tau_i = \Sigma\pi_j = \Sigma\lambda_{ij} = 0$ and where λ_{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 Section 8.2 where the suffix i was used to indicate which group a patient belonged to while here it denotes the treatment received; additionally, we have imposed constraints on the parameters to ensure identifiability.

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 α_k is specified as a random variable this is strictly a random effects model which is a topic covered in Extended Linear Models, so we present first an approximate analysis with a fixed effects model which alters the assumption that the α_k are random variables and instead adds fixed constants and changes the identifiability constraint to $\Sigma\alpha_k = 0$.

8.7.2 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 the second semester courses.

8.7.3 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).

appreciation only ends ↑

8.8 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.

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 investigator's overall assessment. Subjects were randomized to one of two groups: group 1 received the treatments in the order formoterol $\longrightarrow$ salbutamol; group 2 in the order salbutamol $\longrightarrow$ formoterol.

The results are given below:

group	subject	formoterol	salbutamol	preference
group 1 for ⟶ sal	1	+	+	—
	2	-	-	—
	3	+	-	f
	4	+	-	f
	5	+	+	—
	6	+	-	f
	7	+	-	f
	8	+	-	f
	9	+	-	f
	10	+	-	f
	11	+	-	f
	12	+	-	f
group 2 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 12.34, which is clearly significant at a level p <0.001 and so the data provide strong evidence of superiority of the treatment by formoterol.

[It might be noted here that the entries in this table are rather small. More relevantly, the expected values of the cell values are small, with two of them less than 5. This means that the chi-squared distribution is not an adequate approximation to the null distribution of the test statistic and so we might prefer to simulate the p-value or use a Fisher exact test.]

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 1.37 and we conclude that there is no evidence of a period effect.

8.9 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 1 - period 2 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 1, in which case treatment comparisons are between subjects and the sample size may be too low for good power. A full crossover analysis gives a within subject comparison.

Use of a preliminary test for carryover is not recommended by some authors 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.

Chapter 9

Combining Trials

9.1 Small Trials

Some trials are too small to have much chance of picking up differences when they exist, perhaps because of insufficient care over power and sample size considerations. This problem is becoming less severe with generally better regulatory procedures, but it can still happen and it is also the case in the historical record of trials we have currently available. This causes problems of two main types:

Problem 1
Non-significant test results are (inevitably) interpreted by clinicians as 'two treatments are the same', even though the test may have been so low in power that it was not able to detect a real difference.

Problem 2
Small trials giving non-significant results are hardly ever published i.e./because of **publication bias**. The effect is that the medical literature contains all large trials and (only) the significant small trials.

Solutions

1. Ideally: do not conduct any small trials.
2. Do not publish any small trials
3. Combine trials

9.2 Pooling Trials and Meta Analysis

We may have results from several similar trials, though perhaps some of these are smaller than would be desirable. It may be worthwhile making some compromise over the risk of using trials which are only 'similar' (and not identical, or part of a single holistic programme) to increase patient numbers to a level for making a suitably powerful assessment of treatment effect. That is, we may decide to pool the trial results or to conduct a **meta analysis**. A full meta analysis is beyond the scope of this course; it is often part of a full **systematic review** and involves careful identification of potential trials to include, consideration of the similarities and differences in the trials and then combining the assessments of the evidence of a treatment effect made in each relevant trial in a particular way. The topic is covered more fully in Further Clinical Trials.

We will look at more advanced methods in Chapter 11, but in this chapter we introduce a simple method of combining data from multiple trials or centres in the simple case of binary response data from two-arm trials of treatment *versus* placebo. That is, we have data of the form

N trials, with data for trial j, $j = 1, 2, \ldots, N$

	Successes	Failures	Total
Treatment	Y_{1j}	$n_{1j} - Y_{1j}$	n_{1j}
Placebo	Y_{2j}	$n_{2j} - Y_{2j}$	n_{2j}
Total	t_j	$n_j - t_j$	n_j

It is worth stating clearly at the outset that it is rarely appropriate simply to collapse the data from N separate 2×2 tables into a single 2×2 table, because of the risk of **Simpson's paradox** or a treatment by trial interaction, as illustrated below.

Example 9.1 Simpson's Paradox

Suppose we have data from two centres

Centre I

	S	F	
Trt	30	70	30%S
Placebo	120	180	40%S
	150	250	

Centre II

	S	F	
Trt	210	90	70%S
Placebo	80	20	80%S
	290	110	

For both centres, it looks as though Placebo may be better (higher percentage of successes), but χ^2 tests prove non-significant (for Centre I, X^2=3.20, while for Centre II, X^2=3.76).

However, if we collapse the two tables into one

Centres I & II

	S	F	
Trt	240	160	60%
Placebo	200	200	50%
	440	360	

it now seems possible that Treatment is better, and $X^2 = 8.08$ which is 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 or in the magnitudes of the effects.

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

9.3 Mantel-Haenszel Test

One way of combining data from such trials is using the Mantel-Haenszel test. This does not necessarily overcome Simpson's Paradox, but it avoids differences **between** trials by assessing evidence **within** trials and **then combining these assessments**.

Consider a single 2 × 2 table:

	Successes	Failures	Total
Treatment	Y_1	$n_1 - Y_1$	n_1
Placebo	Y_2	$n_2 - Y_2$	n_2
Total	t	$n - t$	n

Assume $Y_i \sim Bi(n_i, \theta_i)$, i=1,2 and we are interested in testing the hypothesis $H_0 : \theta_1 = \theta_2$.

Fisher's exact test considers

$$P(y_1, y_2 | y_1 + y_2 = t)$$

i.e. it conditions on the total number of successes. If $\theta_1 = \theta_2$, then the probability may be evaluated by considering the hypergeometric distribution as

$$P(y_1, y_2 | y_1 + y_2 = t) = \frac{{}^{n_1}C_{y_1} \, {}^{n_2}C_{t-y_1}}{{}^{n}C_t}.$$

The test can be evaluated directly, though it is computationally intensive for all except very small data sets. An alternative way of proceeding is to make an approximation. Recognizing that

$$E(Y_1) = \frac{n_1 t}{n} \text{ and } V(Y_1) = \frac{n_1 n_2 t(n-t)}{n^2(n-1)}$$

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

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

We can view this just as an alternative to the usual (Pearson) χ^2 test (and indeed the two are asymptotically equivalent), but our main interest lies in the way that this version can be extended to handling several tables.

This extension is simple. Keeping the N tables separate, we calculate $E(Y_{1j})$ and $V(Y_{1j})$ from each of the tables, $j = 1, ..., N$. We use

$$W = \Sigma Y_{1j}$$

and under $H_0 : \theta_1 = \theta_2$ in each table, i.e. $\theta_{1j} = \theta_{2j}$, we have

$$E(W) = \Sigma E(Y_{1j}) \text{ and } V(W) = \Sigma V(Y_{1j})$$

and so, once again,

$$\frac{[W - E(W)]^2}{V(W)} \sim \chi_1^2 \text{ under } H_0.$$

Comments

1. The test is known as the Mantel-Haenszel or, rather misleadingly, as a randomization test.

2. It does not matter whether you use Y_1, Y_2, $n - Y_1$, or $n - Y_2$.

3. This test is most appropriate when treatment differences are consistent across tables (we can test this, but it is easier in the logistic regression framework of Chapter 11).

4. In R the test can be conducted using the command `mantelhaen.test()`. It is advised to use it without a continuity correction. Setting up the data in the correct tabular form, is however, complex and so is not illustrated here.

5. Possible limitations of the M-H test

   - Randomness dubious
   - Reporting bias
   - Not clear that θ_i is the same for all trials.

6. Relative merits of M-H & logistic regression approaches
   The Mantel-Haenszel test is simpler if one has just two qualitative prognostic factors to adjust for and wishes only to assess significance, not magnitude, of a treatment difference. The logistic approach 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).

Example 9.2
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'. 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 Rewriting the data in standard form we have:

Study 1

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

Y_{11} =number of response 'severe' who were bottle-fed.
Under H_0 response ratios equal:

$E(Y_{11}) = 20\text{x}26/40 = 13$

$V(Y_{11}) = 20\text{x}20\text{x}26\text{x}14/40\text{x}40\text{x}39 = 2.333$

So Mantel-Haenszel test statistic is

$$(16-13)^2/2.333 = 3.86 > \chi^2_{1;0.95} = 3.84$$

and so is just significant at 5% level, i.e. more severe cases were bottle-fed.

Study 2

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

$E(Y_{12}) = 50\text{x}64/100 = 32$

$V(Y_{12}) = 5.8182$

Mantel-Haenszel test statistic is 0.687, p $>$ 0.05, n.s.

Study 3

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

$E(Y_{13}) = 68.83$

$V(Y_{13}) = 12.667$

Mantel-Haenszel test statistic is 9.85, p $<$ 0.005

Combining all 3 studies

Use $W = Y_{11} + Y_{12} + Y_{13}$.

Under H_0: response ratios equal,

$W = 130,\ E(W) = 113.83,\ V(W) = 20.8183$

so Mantel-Haenszel 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).

9.4 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 (logistic regression) modelling allowing for a separate 'trial effect'.
- The Mantel-Haenszel test provides an alternative way of analyzing 2 × 2 tables which makes it easier to combine results from different trials. It does not overcome Simpson's Paradox but avoids it.

Chapter 10

Comparing Methods of Measurement

10.1 Introduction

Many situations arise where two (or more) techniques have been used to measure some quantity on the same subject. For example, a new instrument for measuring blood pressure is introduced and compared with an old instrument by taking simultaneous measurements on the same subjects. Another example is when two (or more) observers rate some feature by assigning a category (e.g. good/medium/bad). The first requires the comparison of methods on the basis of continuous measurements, the second on the basis of (ordered) categorical responses.

It is inappropriate to base the analyses on calculating a correlation coefficient or a χ^2 statistic for independence. In the first case, you obviously expect there to be a strong correlation between the measurements on the two instruments and although it may be of interest to quantify it using a correlation coefficient, it is of no interest at all to test whether the correlation is 'significantly different from zero', the only test routinely available on correlation coefficients. In the second case, again, you already know that the categorizations cannot be independent, so it is of no interest to conduct a test of independence. Of much more interest is whether there is some consistent bias by one instrument with respect to the other (does it consistently provide a higher reading?) or whether the observers show reasonable agreement or not. Two techniques used in these contexts are **Bland & Altman Plots** and calculation of the **kappa statistic**. Neither of these produce any formal statistical assessment and it is a clinical decision, not a statistical one, whether the degree of agreement is acceptable or not.

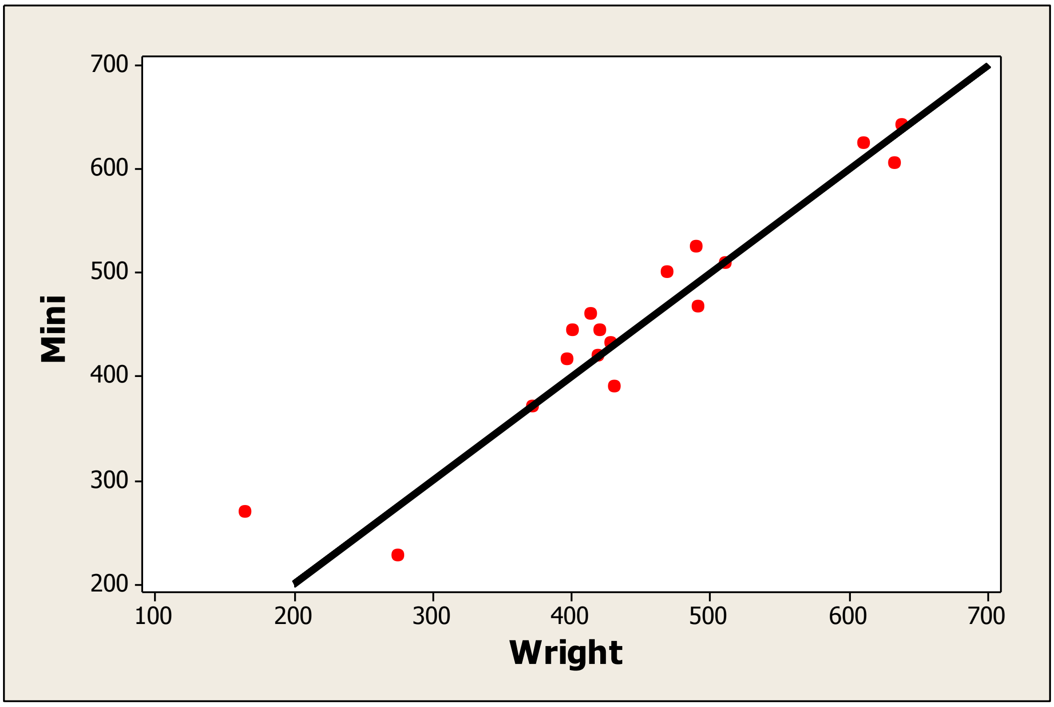
A core source for references on this topic is Martin Bland's webpage at http://www-users.york.ac.uk/ mb55/

10.2 Bland & Altman Plots

We illustrate these by means of an example using data from Bland (2000) available from the website referenced above. The table below, gives the PEFR in litres/min of 17 subjects measured by two instruments: a Wright meter and a Mini meter.

Subject	Wright	Mini	Mean	Difference
1	490	525	507.5	-35
2	397	415	406.0	-18
3	512	508	510.0	4
4	401	444	422.5	-43
5	470	500	485.0	-30
6	611	625	618.0	-14
7	415	460	437.5	-45
8	431	390	410.5	41
9	638	642	640.0	-4
10	429	432	430.5	-3
11	420	420	420.0	0
12	633	605	619.0	28
13	275	227	251.0	48
14	492	467	479.5	25
15	165	268	216.5	-103
16	372	370	371.0	2
17	421	443	432.0	-22

A sensible first step is to plot a scatterplot of the two measurements, as shown below. Note that the line is NOT the (unhelpful) regression line, but the line of equality, i.e. the ideal line if the two instruments agreed perfectly with each other.

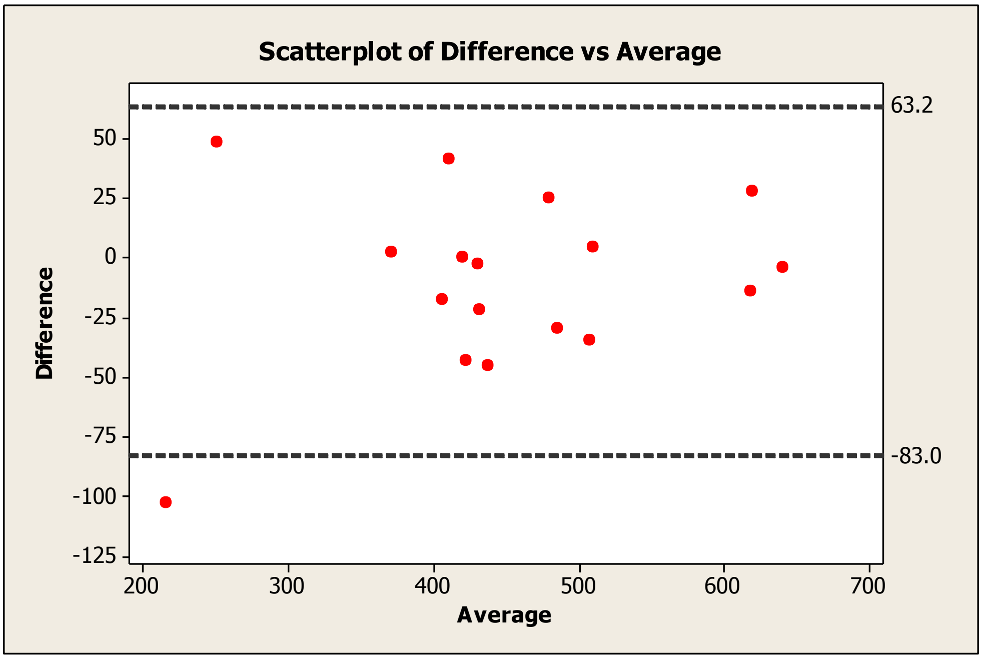


There is a suggestion that there are more points above the line than below it but this is not easy to see. More effective is a **Bland & Altman Plot** which plots the difference against the average of the two measurements. The mean of the differences is -9.9 with standard deviation 36.54, so a 95% confidence interval for the **mean difference** is (-27.6, 7.8). The difference is a measure of the bias between the two measuring methods, so (loosely speaking) 'there could be a bias of as much as 27.6 litres per minute'. Whether this is unacceptably large is a clinical question, not a statistical issue.

Also shown on the graph are what are conventionally known as the **limits of agreement** which is

mean difference ± 2× standard deviation of differences,

i.e. $-9.9 \pm 2 \times 36.54$, and can be thought of as an approximate 95% confidence interval for an **individual** difference between the measurements made by the instruments (NB not the narrower interval for the **mean difference** calculated above).



Note that Bland & Altman plots do not show which instrument is the more accurate (they may both be wrong!) but only whether they agree between themselves. It is possible that one of the methods is the 'Gold Standard' and the other is a cheaper or more convenient alternative. It is then up to the clinicians involved to decide whether the alternative is acceptably close to the gold standard.

10.3 The Kappa Statistic for Agreement of Categorical Variables

Suppose two observers rate objects into a set of categories. The **kappa statistic** is based upon comparing the observed proportion of agreement (A_{obs}) between the two observers with the proportion of agreement (A_{exp}) expected purely by chance. The kappa statistic is then defined as

$$\kappa = \frac{A_{obs} - A_{exp}}{1 - A_{exp}}$$

This statistic is not assessed in statistical terms, but there is a conventional scale of interpretation

$\kappa > 0.75$	excellent agreement
$0.4 < \kappa < 0.75$	fair to good agreement
$\kappa < 0.4$	moderate or poor agreement

The observed agreements are those down the diagonal of the two-way table of assessments made by the two observers and so the observed proportion of agreements is the total of the diagonals divided by the overall total. The expected numbers of agreements are the expected diagonal terms calculated as the product of the marginal totals divided by the overall total (as done in calculating the expected numbers for a χ^2 test on a contingency table).

Example
The table below (Kirkwood & Stone, 2003) give the classification of the 'dominant style' of 179 people on two occasions.

	Second classification				
First classification	Normalizer	Somatizer	Psychologizer	None	Total
Normalizer	76	0	7	10	93
Somatizer	2	0	3	1	6
Psychologizer	17	1	15	8	41
None	20	3	5	11	39
Total	115	4	30	30	179

We calculate A_{obs} = (76+0+15+11)/179 = 0.57

The 'expected' numbers of interest (NB we do not need to calculate these for off-diagonals) are

	Second classification				
First classification	Normalizer	Somatizer	Psychologizer	None	Total
Normalizer	59.7				93
Somatizer		0.1			6
Psychologizer			6.9		41
None				6.5	39
Total	115	4	30	30	179

This gives $A_{exp} = (59.7 + 0.1 + 6.9 + 6.5)/179 = 0.409$ and $\kappa = 0.27$ indicating poor agreement.

Note that as the number of categories increases the value of κ is likely to decrease since there are more 'opportunities' for misclassification.

10.4 Modifications

Two modifications to the kappa statistic are possible, but are not detailed here. The first is when there are several ordered categories where it may be felt that there is a 'partial' or 'limited' agreement for cases classified as only one or two categories apart rather than several. In this case the proportion of agreement could be modified by allowing such partial agreements to contribute to the total with less weight. This could be useful for comparative purposes with other κ values calculated with the same system of weighting but does not provide any absolute measure of agreement. The second modification is when there are more than two observers. In this case an average of all the pairwise κ values will provide an overall measure of consistency within the group of observers but there are other possibilities.

10.5 Summary and Conclusions

- It is not helpful to calculate a correlation coefficient between two methods of measurement to assess the degree of agreement or reproducibility.
- It is appropriate to plot the difference in measurements against their average. This is termed a Bland & Altman plot.
- Limits of agreement are given by mean difference ± 2×st.dev. of differences
- It is not appropriate to calculate a chi-squared statistic for a two-way table of results from two observers to assess the level of agreement.
- A kappa statistic measures the level of agreement.
- Extensions to ordered categories, limited disagreement and several observers are possible.

Chapter 11

Binary Response Data

11.1 Introduction

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 ith patient by $y_i = 1$ (success) or $y_i = 0$ (failure).

Often we summarize the data by cross-classifying it in an $r \times c$ table of counts of responses falling in some particular category. We already know how to use the standard **Pearson χ^2 test** for handling such data to answer questions about **goodness-of-fit**, **independence** or **homogeneity** of classification. Recall that in a goodness-of-fit scenario we have an expected probability distribution (specified by some external hypothesis) and a collection of observational data and we wish to test whether the observational data fit the supposed pattern acceptably well, giving no evidence against plausibility of the underlying hypothesis. Here $r = 2$ and we examine numbers in any specified category of possible values. In the case of homogeneity, we are examining whether two or more rows ($r \geq 2$) have the same distribution of observations over the possible c categories. In the case of independence we are examining whether the allocation of an observation to a particular cell (i.e to both row r and column c) depends on, or is independent of, the particular row or column. Somewhat surprisingly, all these tests boil down to the same test statistic

$$X^2 \;=\; \Sigma \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

under the null hypothesis.

However, not all cross-classified tables are appropriate for application of these hypothesis tests of goodness-of-fit, 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. In the same way as we used the Mainland-Gart test for binary results from a crossover trial, there are further techniques for special designs (e.g. paired observations) or if we have additional data, e.g. on covariates (such as different centres). These are what we consider here.

11.2 Observational Studies

11.2.1 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 disease 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) are followed up for a period of time. 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 χ^2 test or a 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 can 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 analyzed 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.

11.2.2 Prospective Studies and 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 a feature which occurs at some considerable 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.

We begin by presenting the data in the following tabular format.

	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 is $c/(c+d)$. The **relative risk** is the ratio of these two

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

To conduct formal tests, we compare this with the value 1 (the value the RR would take if there is no difference in risks for the two groups) by using its standard error. In fact, to improve Normality assumptions, we need to work on the natural logarithmic scale and then back-transform.

The formula for the standard error of $log_e(RR)$ is

$$s.e[(log_e(RR)] = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}.$$

Example 11.1
The data are taken from a study of 'small-for-dates' babies who were classified as having symmetric or asymmetric growth retardation for whom the outcome is whether or not they have a low Apgar score.

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

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

An approximate 90% CI for $log_e(RR)$ is then $-1.0651 \pm 1.645 \times 0.6759$ = (-2.1769, 0.0467). Taking exponentials of the endpoints 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.

11.2.3 Retrospective Studies and Odds Ratios

Retrospective studies identify a collection of cases (i.e. those 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 on the characteristic defining the group as it is in prospective studies. In this case, our table might have the subtly different layout:

	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' (i.e. $a/(a+b)$) 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. Recall that the odds of an event E are given by the ratio P[E occurs]/P[E does not occur].

Thus we define the **odds ratio** as

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

If exposure is not a risk factor for being a case, then this odds ratio will be close to 1. As before, we can conduct tests by noting that there is a simple formula for the standard error of the log_e of the odds ratio

$$s.e[(log_e(OR)] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

and that on that scale approximate Normality is an acceptable assumption.

Example 11.2
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	
	Yes	No
Swim $\geq$ 6 hours/week	32	118
Swim $<$ 6 hours/week	17	127

The calculations give $OR = 2.0259$, $s.e.[log_e(OR)] = 0.3262$ 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). This excludes the value 1 and so provides evidence at the 5% level of an effect of exposure to chlorine. Since the interval lies entirely above 1, we infer a raised risk of dental erosion in those swimming more than 6 hours a week.

11.3 Matched Pairs

11.3.1 Introduction

In the comparison of two treatments, A and B, suppose each patient receives both treatments (in random order), e.g. a crossover or matched-pair trial. We then observe data which are ordered pairs, (y_{i1}, y_{i2}), where the first element is the response of patient i to A and the second is their response to B. Thus our data will be of the form

(0, 0), (0, 1), (0, 1), (1, 1), (1, 0), (1, 1), ...

It is very common to summarize these results as in the following example.

Example 11.3
Rheumatoid arthritis study, two treatments A and B. Response coded as 1=yes, 0=no.

Could present results in tabular form as

		Response		
		yes	no	
Treatment	A	11	37	48
	B	20	28	48

It is then tempting to analyse this as an ordinary 2 × 2 table using a χ^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

Example 11.3 (cont.)

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

and a suitable test for what is really of interest, i.e.treatment differences not ‘no association’, is **McNemar’s Test**.

11.3.2 McNemar’s Test

In this test, in a similar way to the Mainland-Gart test, the like pairs are uninformative as they indicate no treatment difference. Thus we ignore any responses of the form (1,1) and (0,0), and use the unlike pairs only. This means that we do not use the data from subjects where the responses are the same, i.e. subjects for whom both treatments produced successes or both failures, even though intuitively the results from these subjects might actively suggest that the two treatments are equivalent. This is essentially because these subjects provide no evidence on treatment **differences**, our declared core interest.

If no treatment differences exist, then the proportion 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 $\theta = 1/2$.

Example 11.3 (cont.)
There are 3 (1,0)’s out of a total of 15 unlike pairs. Using our null binomial distribution, we can thus calculate our significance probability as

$$p \;=\; 2 \times \Sigma_{x=0}^{3}\ {}^{15}C_x\ (1/2)^{15} = 0.035$$

which is significant at the 5% level.

For larger n use the Normal approximation, which results in another χ^2 statistic with null distribution

$$\frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}} \sim \chi^2_1.$$

11.4 Logistic Modelling

11.4.1 Introduction

Logistic modelling has become a very popular way of handling binary data and the analyses can be handled in most standard statistical packages. A more detailed and technical discussion is given in Extended Linear Models, but an informal rationale for the models is as follows.

In our simple introduction to binary response data so far, we have only really considered the possibility of one treatment factor affecting whether the response is a success or failure. However, in many cases, there are many other factors, or **covariates**, which distinguish the patients in our study and which might affect whether they respond successfully to the treatment. Ideally, we would like to be able to identify the separate effects of all these factors on the response, or, if we regard them as only 'nuisance variables', to assess the effect of the treatment **after allowing for** these covariates.

To set up our problem in the same way as ordinary linear modelling (multiple regression), we want to identify a suitable linear combination of the explanatory variables that explains or predicts the responses well. However, it is infeasible that we could find a combination of variables that would have as its prediction a value of 1 or 0, the only possible responses. So is it possible instead to predict not the response Y_i directly, but some function of it? We could consider the issue of modelling the probability of success, $P[Y_i = 1]$, but this is constrained to lie in [0,1], so again, the computational effort in finding a well-fitting model which is guaranteed to give only plausible probabilities as its predictions is substantial. We could consider modelling the odds of success, $P[Y_i = 1]/P[Y_i = 0]$, but even then, solutions must be constrained to be positive. The usual solution is to estimate the **(natural) logarithm of the odds of success** (or **logit**)

$$\ln(\frac{P[Y_i = 1]}{P[Y_i = 0]}).$$

Thus for **covariates** or **prognostic factors** $X_1,\ X_2, \ldots, X_p$ we have for patient i,

$$\begin{aligned}\ln(\frac{P[Y_i=1]}{P[Y_i=0]}) &= \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip} \\ &= \beta_0+\underline{\beta}'\underline{x}_i\end{aligned}$$

if we write in vector notation, where the x_{ij} can be continuous or discrete or dummy.

NB For clarity, we have omitted here either the expectation operator or the error term.

NB In some notations the β_0 term will be considered part of the overall parameter vector $\underline{\beta}$ and $\underline{x_i}$ extended to start with a constant first term.

We can rearrange the model to give the odds

$$\frac{P[Y_i=1]}{P[Y_i=0]}=\exp(\beta_0+\underline{\beta}'\underline{x_i})$$

and also to give P[success] itself

$$P[Y_i=1]=\theta_i \text{ say}=\frac{\exp(\beta_0+\underline{\beta}'\underline{x_i})}{1+\exp(\beta_0+\underline{\beta}'\underline{x_i})}.$$

11.4.2 Interpretation

As for the simple linear model, to assess the importance of the various explanatory variables we consider the size and sign of the (estimated) coefficients. In general

$\beta_j>0 \Longrightarrow$ P(success) $\nearrow$ as $x_j \nearrow$ and P(success) $\searrow$ as $x_j \searrow$
$\beta_j<0 \Longrightarrow$ P(success) $\searrow$ as $x_j \nearrow$ and P(success) $\nearrow$ as $x_j \searrow$.

In particular, to consider a potential treatment effect, assume that the treatment indicator x_{i1} is coded as

$$\begin{aligned}x_{i1} &= 0 \text{ on placebo} \\ &= 1 \text{ on treatment}\end{aligned}$$

then our model becomes

$$\begin{aligned}\ln(\frac{P[Y_i = 1]}{P[Y_i = 0]}) &= \beta_0 + \quad \mathbf{0} \quad + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad \text{if } x_{i1} = 0, \\ &= \beta_0 + \quad \beta_{\mathbf{1}} \quad + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad \text{if } x_{i1=1},\end{aligned}$$

so if $\beta_1 > 0$, the odds in favour of success are greater in the treatment group and if $\beta_1 < 0$, the odds in favour of success are greater in the placebo group.

Another useful way of describing the importance of explanatory variables, particularly binary factors, 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. Again, as an example, consider its use to describe the importance of our treatment x_1. We have the odds ratio of getting the disease on treatment compared with placebo as the ratio

$$\frac{P[Y = 1|x_1 = 1]}{P[Y = 0|x_1 = 1]} \Big/ \frac{P[Y = 1|x_1 = 0]}{P[Y = 0|x_1 = 0]} = exp(\beta_1).$$

Similar calculations can be performed under the two values of other binary factors, or at two specified settings of continuous covariates.

11.4.3 Inference

The parameters β_0 and $\underline{\beta}$ are estimated by Maximum Likelihood, since we can write our likelihood as (proportional to)

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

Standard iterative methods (e.g. Newton-Raphson) give m.l.e's and estimated standard errors of these estimates can be obtained from the diagonal of the estimated variance matrix.

R will fit the model and give estimates and standard errors. We can test significance using either of the methods below.

1. **Partial z-test**
   For $H_0 : \beta_j = 0$, test compares

$$\frac{\hat{\beta}_j}{\sqrt{\hat{var}(\hat{\beta}_j)}}$$

   with N(0,1) % points, ignoring the strict need for a t-test.

2. **Likelihood ratio test**
   Here we compare the **deviance**

   $$2\ |l_{\text{full model}} - l_{\text{reduced model with } \beta_j=0}|$$

   where l is the maximized log likelihood, with χ^2_1 % points. Note, that there is some ambiguity over the definition of deviance, as some authors reserve it for the case of comparing against a null model, while others use it to compare any two nested models; it should be clear in context. Also, by using additivity of χ^2 variables, we can modify this test to compare models with more complex differences than just removal of a single term. This is sensible, for example, when we have had to use several dummy variables to code a multilevel factor or interaction and want to assess their overall efffect. An example of its use (analogous to use of RSS in comparing linear models) is given in the table below.

```
Analysis of Deviance Table

Binomial model

Response: Kyphosis

Terms added sequentially (first to last)

      Df  Deviance   Resid.Df  Resid.Dev  Pr(Chi)
NULL                    80    83.23447
Age    1     1.30198    79    81.93249  0.2538510
Number 1    10.30593    78    71.62656  0.0013260
Start  1    10.24663    77    61.37993  0.0013693
```

11.4.4 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. Thus, $p = 6$, with

$x_1 =$	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 mmHg
$x_6 =$	cholesterol in mg/dl

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

Maximum likelihood estimation gives the following summary information

x_j	$\hat{\beta}_j$	z-value
1: treatment	-0.32	-2.9
2: age	3.0	6.3
3: smoking	0.83	6.8
4 father's hist	0.64	3.6
5: systolic BP	0.011	3.7
6: cholesterol	0.0095	5.6

Constant term $\hat{\beta}_0 = -19.60$

Recall the following 2-sided 5% and 1% critical values

$$\Phi^{-1}(0.025) = -1.96 \quad \Phi^{-1}(0.005) = -2.58$$

Treatment effect

Significant, p <0.01

$\beta_1 < 0$, so probability of IHD is smaller on treatment than on placebo, i.e. treatment is beneficial.

Odds ratio of getting IHD on clofibrate compared with placebo is

$$\exp(\beta_1) = \exp(-0.32) = 0.73 < 1.$$

That is, the odds of getting IHD are 27% lower on clofibrate after allowing for the other prognostic factors.

The standard error of β_1 is 0.11 (can be deduced from output as -0.32/-2.9, but actually obtained direct from diagonal of information matrix (not given here)). So approximate 95% confidence limits for β_1 are $-0.32 \pm 2 \times 0.11$, i.e. -0.10 and -0.54. Hence $\exp(\beta_1)$ has 95% confidence limits $\exp(-0.10) = 0.90$ and $\exp(-0.54) = 0.58$ so that 95% confidence limits for the reduction due to clofibrate in the odds of getting IHD are 10% and 42%.

Prognostic factors
All five are significant (p <0.01); all have positive coefficients, thus probability of IHD increases with age, smoking, 'poorer heredity', high blood pressure, high cholesterol. Similar calculations to above show, for example, for smoking the 95% limits for the increase in odds of getting IHD for smokers are 80% and 193%.

11.4.5 Additional Explanatory Variables

Interactions
Interaction terms are handled, as usual, by creating a new variable as the product of the two or more existing variables. Most usually, we might want to consider whether the treatment behaves differently for different types of patients, i.e. interacts with patient characteristics. In the above example, the treatment (x_1) 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. For example, if we consider an interaction between treatment and covariate x_2, ln(age), we create a new variable $x_7 = x_1 \times x_2$ which reflects the interaction effect (note that x_7 is identical to x_2 for those on clofibrate and 0 for those on placebo) and then our model is

$$\begin{aligned} ln(\frac{\theta_i}{1-\theta_i}) &= \beta_0 + \beta_2 x_{i2} + \ldots \beta_6 x_{i6} \quad \text{for placebo} \\ &= \beta_0 + \beta_1 + (\beta_2 + \beta_7) x_{i2} + \ldots \beta_6 x_{i6} \quad \text{for clofibrate} \end{aligned}$$

Exactly the same method is appropriate for handling interactions between two continuous covariates or 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 $(k-1)(l-1)$ degrees of freedom for an interaction between a k-level and a l-level factor. This also means that the separate parts of the chi-squared statistic must be combined before assessing significance (see 11.4.3 point 2).

Combining Trials
Within the context of combining trials (see Chapter 9), we might decide on a particular model structure to reflect between-trial differences. For example, that we want a model which keeps the treatment effect β_1 the same in each trial, but

allows β_0 to vary to reflect possible differences in trial j conditions.

$$\ln(\frac{P[Y_{ij}=1]}{P[Y_{ij}=0]}) = \beta_j + \beta_1 x_{ij}$$

Thus, for an example with 3 clinics, we might code the clinics using two dummy variables as

$$\begin{array}{rl} (x_{i2}, x_{i3}) = & \text{(0,0) for clinic 1} \\ & \text{(1,0) for clinic 2} \\ & \text{(0,1) for clinic 3} \end{array}$$

and so have the model

$$\ln(\frac{P[Y_{ij}=1]}{P[Y_{ij}=0]}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

which gives

$$\begin{array}{rl} \beta_0 + \beta_1 x_{i1} & \text{for clinic 1} \\ (\beta_0 + \beta_2) + \beta_1 x_{i1} & \text{for clinic 2} \\ (\beta_0 + \beta_3) + \beta_1 x_{i1} & \text{for clinic 3.} \end{array}$$

11.5 Summary and Conclusions

- Observational studies may be prospective or retrospective
- Binary responses in a prospective study are usefully summarized by relative risks.
- Binary responses in a retropective study are usefully summarized by odds ratios.
- Care needs to be taken in analyzing matched pairs binary responses. McNemar's test, using only the information from unlike pairs, may be suitable.
- 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 odds ratios (or 'relative risks') to be estimated (including confidence intervals).
- Positive coefficients in a logistic model indicate that the factor increases the risk of the 'success'.