

Contents

Preliminaries

0: Introduction

1: Background & Basic Concepts

2: Basic Trial analysis

3: Randomization

5: Multiplicity & Interim Analysis

6: Crossover Trials

7: Binary Response Data

©NRJF, University of Sheffield, 2012 Medical Statistics: Clinical Trials 145

Multiplicity &c.

Multiplicity & Interim Analysis

- **Books**
 - ◆ Andersen, B. (1990) *Methodological Errors in Medical Research*. Blackwell
- **Papers**
 - ◆ ICH E9 Expert Working Group. (1999) *Statistics in Medicine*, 18, 1905-42.
 - ◆ Philips, Alan & Haudiquet, Vincent (2003) *Statistics in Medicine*, 22, 1-11

©NRJF, University of Sheffield, 2012 Medical Statistics: Clinical Trials 146

- Multiplicity arises in
 - ◆ Multiple endpoints
 - ◆ Subgroup analyses
 - ◆ Interim testing
 - ◆ Repeated Measures
 - ◆ &c.
 - Problem of repeated significance tests
 - ◆ May inflate risk of false positive
 - i.e. overall significance level
- ©NRJF, University of Sheffield, 2012 Medical Statistics: Clinical Trials 147

- **Example:** Effect of new dietary control regime.
 - **Data:** 250 subjects:
 - ◆ Weight loss at end of week.
 - Data in kg.
 - Paired t-test gives p-value of 0.067
 - **Not quite significant at the 5% level !**
- ©NRJF, University of Sheffield, 2012 Medical Statistics: Clinical Trials 148

- Can anything be done to ‘squeeze’ a significant result out of this expensive study ?????
 - we’ve been told we cannot change our mind and use a onesided test instead!
 - subgroup data by Sign of the Zodiac:–
- ©NRJF, University of Sheffield, 2012 Medical Statistics: Clinical Trials 149

- p-value for Aries is 0.019
 - With mean weight **loss** of 0.5kg
 - p-value for Taurus is 0.099
 - With mean weight **gain** of 0.3kg
 - **Conclusions:**
 - ◆ Diet successful for those under Aries
 - ◆ Taurus subjects are perverse
- ©NRJF, University of Sheffield, 2012 Medical Statistics: Clinical Trials 150



- Clearly a **False Positive Result**
- Fallacy arises because of **selecting most significant result.**

- ◆ (data are artificial, but not very)
 - useful device to try if pressed to perform post-hoc subgroup analysis (c.f. Richard Peto)



- Statistical tests make mistakes
 - ◆ Declare a real difference exists
 - but in fact the observed difference is due to natural chance variation
 - ◆ Risk controlled
 - **for each individual single test**
 - *significance level* of the test or the *p-value*
 - ◆ if **many** separate significance tests then difficult to control overall risk of declaring **at least one false positive somewhere**



- 5% test then 95% chance of no mistake
 - ◆ Two 5% tests then 95%×95% (= 90.25%) of no mistake on either
 - ◆ So 10% risk of one or other (or both) giving a false positive
 - ◆ i.e. overall significance level is ~10%



- **c.f. Normal Ranges** in clinicochemical tests
 - A 'normal person' is one who has not been sufficiently investigated.
- A **normal range** comprise 95% of values
 - ◆ 100 normal persons evaluated then only 95 of them will 'normal'
 - ◆ If then subjected to another **independent** test only 90 will remain as 'normal'



- **Multiplicity:**
 - ◆ 10 independent tests at [nominal] 5%
 - ◆ H_0 true in all (i.e. no difference)
 - ◆ Chance of rejecting *at least 1* is 40%
 - ◆ **SO** reduce nominal level in each to control overall significance level



- **Bonferroni correction**
 - ◆ k tests, want overall level to be α
- Take nominal level on each test as α/k
- **Example:**
 - ◆ 5 separate tests
 - ◆ Overall 5% level of significance wanted
 - ◆ Declare a result if any test nominally significant at the 5%/5=1% significance level



▪ **Example:**

- ◆ 25 tests are to be performed
- ◆ overall level of 1% 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
- ◆ smallest p-value is 0.019
- ◆ What is the overall level of significance?
- ◆ Bonferroni method says overall level is $12 \times 0.019 = \mathbf{0.228}$
 - This is the Signs of the Zodiac example
 - i.e. no worthwhile evidence of any birth sign being particularly suited to dieting
 - (see again later)

▪ Bonferroni method typically **very conservative**

- ◆ i.e. less likely to be able to declare a real difference exists even if there is one
- ◆ But is 'safe'
 - i.e. you preserve your scientific reputation by avoiding making mistakes but at expense of failing to discover something scientifically interesting

▪ **Multiple Endpoints**

- ◆ e.g. pulse rate, systolic & diastolic blood pressure sitting, standing & supine before & after exercise
- ◆ Separate tests high risk of false positives

▪ **Remedies:**

- ◆ Bonferroni correction
- ◆ choose primary outcome measure
- ◆ multivariate analysis
- **NB:** Bonferroni **very conservative**
 - ◆ multiple outcome measures likely to be **highly correlated**
 - ◆ standing systolic BP will give similar evidence to sitting BP

▪ Very frustrating if you had considered 20 highly correlated measures

- ◆ each gives nominal pvalue of 0.01
- ◆ Bonferroni says can only claim an overall pvalue of 0.2
- ◆ Would have been better not to have measured the other 19



- Better is to define **primary outcome**
 - ◆ perhaps 2 or 3 secondary measures
 - ◆ **Must be stated in the protocol**
 - medical expertise
 - initial results from a pilot study
 - ◆ Other measures (e.g. lab results) should be scrutinised
 - report causes for concern



- **Multivariate Analysis**
 - ◆ Makes proper allowance in the analysis for correlated observations
 - ◆ There are multivariate equivalents of standard univariate statistical analyses
 - Student's t-test \leftrightarrow Hotelling's T^2 -test
 - ANOVA \leftrightarrow MANOVA
 - Multivariate Analysis of Variance
 - Wilks' test or Lawley-Hotelling test



- **Advantage** of multivariate analysis
 - ◆ handle all measures simultaneously
 - ◆ return a single p-value
- **Disadvantage**
 - ◆ difficulty of interpreting the nature of the difference detected
- Many MV procedures in stats packages
 - ◆ Advice must be to use them with caution unless experienced help is to hand



- **Cautionary Examples**
 - ref: Basic Clin Med 1981, 15: 445
 - ◆ doubleblind controlled clinical trial to treat rheumatoid arthritis
 - ◆ several end-points repeated at various timepoints and various subdivisions
 - ◆ 850 pairwise comparisons were made
 - t-tests and Fisher's exact test
 - ◆ 48 of these gave p-values < 0.05
 - ◆ But expect 5% of 850 = $850/20 = 42.5$ so finding 48 is not very impressive



- Andersen quotes
 - *The Lancet* (1984, ii: 1457)
- ◆ 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



- **Subgroup analyses**
 - Similar problems with subgroups
 - ◆ Need to specify which subgroups of particular interest in protocol
 - ◆ If none in particular then
 - Bonferroni adjustment
 - Analysis of Variance
 - Followup tests for multiple comparisons

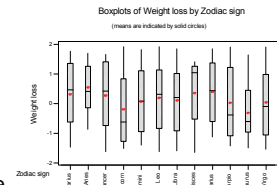


- Oneway ANOVA generalisation to several samples of a two-sample t-test
 - ◆ tests differences between subgroups
 - tests null hypothesis that all subgroups have the same mean vs one or more is different
 - ◆ If effect exhibited in only one of several subgroups then one (or more) of the subgroups is different from the rest so test this with ANOVA
 - Followup tests to identify which is of interest
 - Tukey's / Dunnett's / NeumanKeuls /



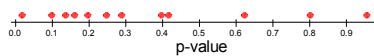
- **Example: Signs of Zodiac (see notes)**
 - ◆ pvalue for differences in weight loss between Zodiac signs is 0.405

- ◆ No evidence of difference so followup tests not really appropriate
 - (but see example in notes)



- Can also look at 12 separate p-values

Dotplot of p-values



- ◆ If any evidence that some groups were shewing an effect then some of them would be clustered towards near 0.0 and not evenly spread out



- Example (Lee et al, *Circulation*, 1980)

- ◆ 1073 subjects randomized in two groups
 - No overall significance
 - 6 [posthoc] subgroups defined
 - One of these produced significance at nominal 2.5% level ($p=0.023$)
 - Medical reason for expecting this subgroup to be different



- But:—
 - ◆ In fact, no difference between 'treatments'
 - ◆ All patients were treated in the **SAME** way
 - ◆ Groups were just random allocations
- i.e. a **false positive effect**



- **Interim analyses**

- ◆ Desirable in long trial
 - Check protocol compliance
 - Side effects?
 - Feedback
 - (maintains interest)
 - Detect big effects quickly



- However, multiplicity problems:
 - ◆ Specify in protocol
 - ◆ Adjust nominal significance levels
- Bonferroni too conservative
 - ◆ (accumulating data)

Medical Statistics: Clinical Trials 175

**Repeated significance tests
on accumulating data**

Number of repeated tests at the 5% level	overall significance level
1	0.05
5	0.14
10	0.19
100	0.37

Medical Statistics: Clinical Trials 176

Nominal significance levels required to achieve overall level		
N	$\alpha=0.05$	$\alpha=0.01$
2	0.029	0.0056
5	0.016	0.0028
10	0.0106	0.0018

Medical Statistics: Clinical Trials 177

- Comparison of drug combinations CP and CVP in nonHodgkins lymphoma.
 - ◆ Measure: tumour shrinkage
 - ◆ Trial: over 2 years, about 120 patients.
 - ◆ Five interim analyses planned, roughly after every 25th result.
 - Table gives numbers of 'successes' and **nominal** pvalues using a χ^2 test at each stage.

Medical Statistics: Clinical Trials 178

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 ($0.025<p<0.05$)

Medical Statistics: Clinical Trials 179

- **Conclusion:**
 - ◆ Not significant at end of trial (overall $p>0.05$) since $p>0.016$
 - the required nominal value for 5 repeat tests
 - ◆ If **NO** interim analyses had been done then conclusion would have been **different**
 - CVP declared significantly better at 5% level

Medical Statistics: Clinical Trials 180



▪ **Cautionary Example:**

- ref: Br J Surg, (1974), **61**: 177
- ♦ No significant difference with 49 patients
- ♦ The trial was therefore continued
- ♦ After 100 patients gave result $\chi^2 = 4.675$, d.f. = 1, $p < 0.05$
 - (and the trial was published)
- ♦ Actual nominal p-value is $0.031 > 0.029$ so cannot claim overall 5% significance



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



▪ **Repeated Measures**

- ♦ same feature on a patient measured at several time points
 - blood concentration at baseline and at 1, 3, 6, 12 and 24 hours after drug
- ♦ Must not do ttests at each time point
 - diagrams with mean values of the two treatment groups plotted against time with **error bars for each mean** invite the eye to do exactly that



▪ **Remedies**

- ♦ Bonferroni adjustments
 - Very conservative (high correlation)
- ♦ Multivariate analysis
 - Special techniques for this
- ♦ Construction of summary measures
 - Area under curve
 - Change from base line
 - Mean change
 -



▪ **Multiple Regression**

- ♦ large regression analyses many explanatory variables
 - ordinary regression
 - logistic regression for success/failure data
 - Cox regression for survival data
- ♦ Need to ensure that effects are not selected just because they are the most significant coefficients




▪ **Example:**

- ♦ men who did not shave regularly were '70% more likely to suffer a stroke and 30% more likely to suffer heart disease'
 - according to study at the University of Bristol
- ♦ Perhaps from a logistic regression analysis
- ♦ Is diligence in shaving a medically plausible feature to be investigated???
- ♦ How many other variables were included in the study???




◆ **Summary and Conclusions**

- ◆ Multiplicity can arise in
 - testing several different responses
 - subgroup analyses
 - interim analyses
 - repeated measures
 - &c.
- ◆ The effect of multiplicity is to increase the **overall risk of a false positive (i.e. the overall significance level)**


©NRJF, University of Sheffield, 2012
Medical Statistics: Clinical Trials
187

- ◆ Problems of multiplicity can be overcome by
 - Bonferroni corrections
 - Bonferroni typically **very** conservative
 - other adjustments in special cases
 - e.g. for accumulating data in interim analyses where adjusting for multiplicity can have counterintuitive effects
 - more sophisticated analyses e.g. ANOVA or multivariate methods
- ◆ **“If you torture the data often enough it will eventually confess”**


©NRJF, University of Sheffield, 2012
Medical Statistics: Clinical Trials
188


©NRJF, University of Sheffield, 2012
Medical Statistics: Clinical Trials
189


©NRJF, University of Sheffield, 2012
Medical Statistics: Clinical Trials
190


©NRJF, University of Sheffield, 2012
Medical Statistics: Clinical Trials
191


©NRJF, University of Sheffield, 2012
Medical Statistics: Clinical Trials
192

