Ton J. Cleophas · Aeilko H. Zwinderman

Statistical Analysis of Clinical Data on a Pocket Calculator

Statistics on a Pocket Calculator

Springer

Statistical Analysis of Clinical Data on a Pocket Calculator

Ton J. Cleophas • Aeilko H. Zwinderman

Statistical Analysis of Clinical Data on a Pocket Calculator

Statistics on a Pocket Calculator

Springer

Prof. Ton J. Cleophas
Department of Medicine
Albert Schweitzer Hospital
Dordrecht, The Netherlands
and
European College of Pharmaceutical
Medicine, Lyon, France
ajm.cleophas@wxs.nl

Prof. Aeilko H. Zwinderman
Department of Epidemiology
and Biostatistics
Academic Medical Center
Amsterdam, The Netherlands
and
European College of Pharmaceutical
Medicine, Lyon, France
a.h.zwinderman@amc.uva.nl

ISBN 978-94-007-1210-2 e-ISBN 978-94-007-1211-9
DOI 10.1007/978-94-007-1211-9
Springer Dordrecht Heidelberg London New York

Preface

The time that statistical analyses, including analysis of variance and regression analyses, were analyzed by statistical laboratory workers, has gone for good, thanks to the availability of user-friendly statistical software. The teaching department, the educations committee, and the scientific committee of the Albert Schweitzer Hospital, Dordrecht, Netherlands, are pleased to announce that since November 2009 the entire staff and personal is able to perform statistical analyses with help of SPSS Statistical Software in their offices through the institution's intranet.

It is our experience as masters' and doctorate class teachers of the European College of Pharmaceutical Medicine (EC Socrates Project) that students are eager to master adequate command of statistical software for carrying out their own statistical analyses. However, students often lack adequate knowledge of basic principles, and this carries the risk of fallacies. Computers cannot think, and can only execute commands as given. As an example, regression analysis usually applies independent and dependent variables, often interprets as causal factors and outcome factors. E.g., gender and age may determine the type of operation or the type of surgeon. The type of surgeon does not determine the age and gender. Yet, software programs have no difficulty to use nonsense determinants, and the investigator in charge of the analysis has to decide what is caused by what, because a computer can not do a thing like that, although it is essential to the analysis.

It is our experience that a pocket calculator is very helpful for the purpose of studying the basic principles. Also, a number of statistical methods can be performed more easily on a pocket calculator, than using a software program.

Advantages of the pocket calculator method include the following.

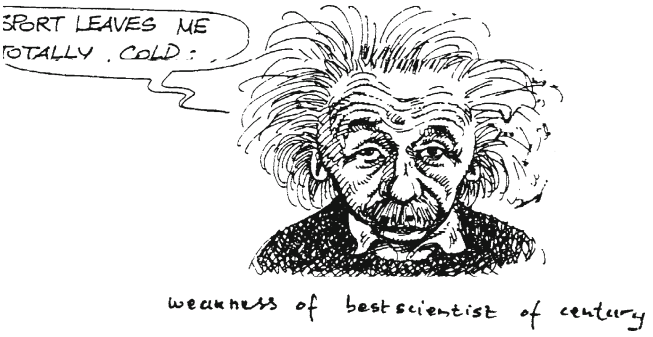
1. You better understand what you are doing. The statistical software program is kind of black box program.
2. The pocket calculator works faster, because far less steps have to be taken.
3. The pocket calculator works faster, because averages can be used.
4. With statistical software all individual data have to be included separately, a time-consuming activity in case of large data files.

Also, some analytical methods, for example, power calculations and required sample size calculations are difficult on a statistical software program, and easy on

a pocket calculator. The current book reviews the pocket calculator methods together with practical examples. This book was produced together with the similarly sized book "SPSS for Starters" from the same authors (edited by Springer, Dordrecht 2010). The two books complement one another. However, they can be studied separately as well.

Lyon
December 2010

Ton J. Cleophas
Aeilko H. Zwinderman



Contents

1 Introduction 1

2 Standard Deviations 3

3 t-Tests 5
1 Sample t-Test 5
Paired t-Test 6
Unpaired t-Test 7

4 Non-Parametric Tests 9
Wilcoxon Test 9
Mann-Whitney Test 10

5 Confidence Intervals 15

6 Equivalence Tests 17

7 Power Equations 19

8 Sample Size 23
Continuous Data, Power 50% 23
Continuous Data, Power 80% 24
Continuous Data, Power 80%, 2 Groups 24
Binary Data, Power 80% 25
Binary Data, Power 80%, 2 Groups 25

9 Noninferiority Testing 27
Step 1: Determination of the Margin of Noninferiority, the Required Sample, and the Expected p-Value and Power of the Study Result 27
Step 2: Testing the Significance of Difference Between the New and the Standard Treatment 28

Step 3: Testing the Significance of Difference Between the New Treatment and a Placebo 28
Conclusion 28

10 Z-Test for Cross-Tabs 29

11 Chi-Square Tests for Cross-Tabs 31
First Example Cross-Tab 31
Chi-Square Table (χ^2-Table) 31
Second Example Cross-Tab 33
Example for Practicing 1 33
Example for Practicing 2 34

12 Odds Ratios 35

13 Log Likelihood Ratio Tests 37

14 McNemar's Tests 39
Example McNemar's Test 39
McNemar Odds Ratios, Example 40

15 Bonferroni t-Test 41
Bonferroni t-Test 41

16 Variability Analysis 43
One Sample Variability Analysis 43
Two Sample Variability Test 44

17 Confounding 47

18 Interaction 49
Example of Interaction 50

19 Duplicate Standard Deviation for Reliability Assessment of Continuous Data 51

20 Kappas for Reliability Assessment of Binary Data 53

Final Remarks 55

Index 57

Chapter 1
Introduction

This book contains all statistical tests that are relevant to starting clinical investigators. It begins with standard deviations and t-tests, the basic tests for the analysis of continuous data. Next, non-parametric tests are reviewed. They are, particularly, important to investigators whose affection towards medical statistics is little, because they are universally applicable, i.e., irrespective of the spread of the data. Then, confidence intervals and equivalence testing as methods based on confidence intervals are explained.

In the next chapters power-equations that estimate the statistical power of data samples are reviewed. Methods for calculating the required sample size for a meaningful study, are the next subject. Non-inferiority testing including comparisons against historical data and sample size assessments are, subsequently, explained. The methods for assessing binary data include: z-tests, chi-square for cross-tabs, log likelihood ratio tests and odds ratio tests. Mc Nemar's tests for the assessment of paired binary data is the subject of Chap. 14. Then, the Bonferroni test for adjustment of multiple testing is reviewed, as well as chi-square en F-tests for variability analysis of respectively one and two groups of patients.

In the final chapters the assessment of possible confounding and possible interaction is assessed. Also reliability assessments for continuous and binary data are reviewed.

Each test method is reported together with (1) a data example from practice, (2) all steps to be taken using a scientific pocket calculator, and (3) the main results and their interpretation. All of the methods described are fast, and can be correctly carried out on a scientific pocket calculator, such as the Casio fx-825, the Texas TI-30, the Sigma AK222, the Commodoor and many other makes. Although several of the described methods can also be carried out with the help of statistical software, the latter procedure will be considerably slower.

In order to obtain a better overview of the different test methods each chapter will start on an uneven page. The pocket calculator book will be applied as a major help to the workshops "Designing and performing clinical research" organized by the teaching department of Albert Schweitzer STZ (collaborative top clinical)

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_1,

Hospital Dordrecht, and the statistics modules at the European College of Pharmaceutical Medicine, Claude Bernard University, Lyon, and Academic Medical Center, Amsterdam.

The authors of this book are aware that it consists of a minimum of text and do hope that this will enhance the process of mastering the methods. Yet we recommend that for a better understanding of the test procedures the book be used together with the same authors' textbook "Statistics Applied to Clinical Trials" 4th edition edited 2009, by Springer Dordrecht Netherlands. More complex data files like data files with multiple treatment modalities or multiple predictor variables can not be analyzed with a pocket calculator. We recommend that the in 2010 by the same editor published book "SPSS for Starters" (Springer, Dordrecht, 2010) from the same authors be used as a complementary help for the readers' benefit.

The human brain excels in making hypotheses, but hypotheses have to be tested with hard data.

Chapter 2
Standard Deviations

Standard deviations (SDs) are often being used for summarizing the spread of the data from a sample. If the spread in the data is small, then the same will be true for the standard deviation. Underneath the calculation is illustrated with the help of a data example.

	55	
	54	
	51	
	55	
	53	
	53	
	54	
	52+	
Mean	=>	…/8 = 53.375
SD=		
	55	$(55–53.375)^2$
	54	$(54–53.375)^2$
	51	$(51–53.375)^2$
	55	$(55–53.375)^2$
	53	$(53–53.375)^2$
	53	$(53–53.375)^2$
	54	$(54–53.375)^2$
	52	$(52–53.375)^2$+
SD=		……. =>…./ n−1=> √….=> 1.407885953

Each scientific pocket calculator has a modus for data-analysis. It is helpful to calculate in a few minutes the mean and standard deviation of a sample.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_2,

Calculate standard deviation: mean = 53.375 SD = 1.407885953

The next steps are required:

Casio fx-825 scientific

On … mode … shift … AC … 55 … M+ … 54 … M+ … 51 … M+ … 55 … M+ … 53 … M+ … 53 … M+ … 54 … M+ … 52 … M+ … shift … [x] … shift … σxn–1

Texas TI-30 scientific

On … 55 … Σ+ … 54 … Σ+ … 51 … Σ+ … 55 … Σ+ … 53 … Σ+ … 53 … Σ+ … 54 … Σ+ … 52 … Σ+ … 2nd … x … 2nd … σxn–1

Sigma AK 222 and Commodoor

On … 2ndf … on … 55 … M+ … 54 … M+ … 51 … M+ … 55 … M+ … 53 … M+ … 53 … M+ … 54 … M+ … 52 … M+ … x=>M … MR

Calculator: Electronic Calculator

On … mode … 2 … 55 … M+ … 54 … M+ … 51 … M+ … 55 … M+ … 53 … M+ … 53 ... M+ … 54 … M+ … 52 … M+ … Shift … S-var … 1 … = … (mean) … Shift … S-var … 3 … (sd)

Example:

What is the mean value, what is de SD?

5
4
5
4
5
4
5
4

Chapter 3
t-Tests

1 Sample t-Test

As an example, the mean decrease in blood pressure after treatment is calculated with the accompanying p-value. A p-value <0.05 indicates that there is less than 5% probability that such a decrease will be observed purely by the play of chance. There is, thus, >95% chance that the decrease is the result of a real blood pressure lowering effect of the treatment. We call such a decrease statistically significant.

Patient	mm Hg decrease
1	3
2	4
3	−2
4	3
5	1
6	−2
7	4
8	3

Is this decrease statistically significant?

$$\text{Mean decrease} = 1.75\ \text{mm Hg}$$
$$\text{SD} = 2.49\ \text{mm Hg}$$

From the standard deviation the standard error (SE) can be calculated using the equation

$$\text{SE} = \text{SD} / \sqrt{\text{n}} \quad (\text{n} = \text{sample size})$$
$$\text{SE} = 2.49 / \sqrt{8} = 0.88$$

De t-value is the test-statistic of the t-test and is calculated as follows:

$$t = 1.75 / 0.88 = 1.9886$$

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_3,

Because the sample size is 8, the test has here 8−1 = 7 degrees of freedom.

The t-table on the pages 7–8 shows that with 7 degrees of freedom the p-value should equal: $0.05 < p < 0.10$. This result is close to statistically significant, and is called a trend to significance.

Paired t-Test

Two rows of observations in ten persons are given underneath:

Observation 1:

6.0, 7.1, 8.1, 7.5, 6.4, 7.9, 6.8, 6.6, 7.3, 5.6

Observation 2:

5.1, 8.0, 3.8, 4.4, 5.2, 5.4, 4.3, 6.0, 3.7, 6.2

Individual differences

0.9, −0.9, 4.3, 3.1, 1.2, 2.5, 2.5, 0.6, 3.8, −0.6

A. not significant
B. $0.05 < p < 0.10$
C. $P < 0.05$
D. $P < 0.01$

Is there a significant difference between the observation 1 and 2, and which level of significance is correct?

$$\begin{aligned} &\text{Mean difference} && = 1.59 \\ &\text{SD of mean difference} && = 1.789 \\ &\text{SE} = \text{SD} / \sqrt{10} && = 0.566 \\ &t = 1.59 / 0.566 && = 2.809 \end{aligned}$$

10−1 = 9 degrees of freedom, because we have 10 patients and 1 group of patients.

According to the t-table of page XXX the p-value equals <0.05, and we can conclude that a significant difference between the two observations is in the data: the values of row 1 are significantly higher than those of row 2. The answer C is correct.

Unpaired t-Test

Two matched groups of patients are compared with one another.

Group 1:
6.0, 7.1, 8.1, 7.5, 6.4, 7.9, 6.8, 6.6, 7.3, 5.6
Group 2:
5.1, 8.0, 3.8, 4.4, 5.2, 5.4, 4.3, 6.0, 3.7, 6.2

Mean Group 1 = 6.93 SD = 0.806 SE = SD/√10 = 0.255
Mean Group 2 = 5.21 SD = 1.299 SE = SD/√10 = 0.411

A. not significant
B. $0.05 < p < 0.10$
C. $p < 0.05$
D. $P < 0.01$

Is there a significant difference between the two groups, which level of significance is correct?

Mean	Standard deviation (SD)
6.93	0.806
5.21–	1.299
1.72	pooled SE $= \sqrt{\left(\frac{0.806^2}{10} + \frac{1.299^2}{10}\right)} = 0.483$

The t-value = (6.93−5.21)/0.483 = 3.56.

20−2 = 18 degrees of freedom, because we have 20 patients and 2 groups.

According to the t-table of page the p-value is <0.01, and we can conclude that that a very significant difference exists between the two groups. The values of group 1 are higher than those of group 2. The answer D is correct.

t-Table

df	0.1	0.05	0.01	0.002
1	6.314	12.706	63.657	318.31
2	2.920	4.303	9.925	22.326
3	2.353	3.182	5.841	10.213
4	2.132	2.776	4.604	7.173
5	2.015	2.571	4.032	5.893
6	1.943	2.447	3.707	5.208
7	1.895	2.365	3.499	4.785
8	1.860	2.306	3.355	4.501
9	1.833	2.262	3.250	4.297

(continued)

t-Table (continued)

df	0.1	0.05	0.01	0.002
10	1.812	2.228	3.169	4.144
11	1.796	2.201	3.106	4.025
12	1.782	2.179	3.055	3.930
13	1.771	2.160	3.012	3.852
14	1.761	2.145	2.977	3.787
15	1.753	2.131	2.947	3.733
16	1.746	2.120	2.921	3.686
17	1.740	2.110	2.898	3.646
18	1.734	2.101	2.878	3.610
19	1.729	2.093	2.861	3.579
20	1.725	2.086	2.845	3.552
21	1.721	2.080	2.831	3.527
22	1.717	2.074	2.819	3.505
23	1.714	2.069	2.807	3.485
24	1.711	2.064	2.797	3.467
25	1.708	2.060	2.787	3.450
26	1.706	2.056	2.779	3.435
27	1.701	2.052	2.771	3.421
28	1.701	2.048	2.763	3.408
29	1.699	2.045	2.756	3.396
30	1.697	2.042	2.750	3.385
40	1.684	2.021	2.704	3.307
60	1.671	2.000	2.660	3.232
120	1.658	1.950	2.617	3.160
∞	1.645	1.960	2.576	3.090

The rows give t-values adjusted for degrees of freedom. The numbers of degrees of freedom largely correlate with the sample size of a study. With large samples the frequency distribution of the data will be a little bit narrower, and that is corrected in the table. The t-values are to be looked upon as mean results of studies, but not expressed in mmol/l, kilograms, but in so-called SE-units (Standard error units), that are obtained by dividing your mean result by its own standard error. A t-value of 3.56 with 18 degrees of freedom indicates that we will need the row no. 18 of the table. The upper row gives the area under the curve of the Gaussian-like t-distribution. The t-value 3.56 is left from 3.610. Now look right up to the upper row: we are right from 0.01. The p-value equals <0.01

Chapter 4
Non-Parametric Tests

Wilcoxon Test

The t-tests reviewed in the previous chapter are suitable for studies with normally distributed results. However, if there are outliers, then the t-tests are not sensitive and non-parametric tests have to be applied. We should add that non-parametric are also adequate for testing normally distributed data. And, so, these tests are, actually, universal, and are, therefore, absolutely to be recommended.

Calculate the p-value with the paired Wilcoxon test.

Observation 1:									
6.0,	7.1,	8.1,	7.5,	6.4,	7.9,	6.8,	6.6,	7.3,	5.6
Observation 2:									
5.1,	8.0,	3.8,	4.4,	5.2,	5.4,	4.3,	6.0,	7.3,	6.2
Individual differences:									
0.9,	−0.9,	4.3,	3.1,	1.2,	2.5,	2.5,	0.6,	3.6,	−0.6
Rank number:									
3.5,	3.5,	10,	7,	5,	8,	6,	2,	9,	1

A. not significant
B. $0.05 < p < 0.10$
C. $p < 0.05$
D. $P < 0.01$

Is there a significant difference between observation 1 and 2? Which significance level is correct?

The individual differences are given a rank number dependent on their magnitude of difference. If two differences are identical, and if they have for example the rank numbers 3 and 4, then an average rank number is given to both of them, which

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_4,

means 3.5 and 3.5. Next, all positive and all negative rank numbers have to be added up separately. We will find 4.5 and 50.5. According to the Wilcoxon table underneath the smaller one of the two add-up numbers must be smaller than 8 in order to be able to speak of a p-value <0.05. This is true in our example.

Be careful with type of data
Unless suffer serious damage!!!!!

Wilcoxon test table

Number of pairs	P<0.05	P<0.01
7	2	0
8	2	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	19

Mann-Whitney Test

Like the Wilcoxon test, being the non-parametric alternative for the paired t-test, the Mann-Whitney test is the non-parametric alternative for the unpaired t-test. Also this test is applicable for all kinds of data, and, therefore, particularly, to be recommended for investigators with little affection for medical statistics.

Calculate the p-value of the difference between two groups of ten patients with the help of this test.

Group 1:
6.0 7.1, 8.1, 7.5, 6.4, 7.9, 6.8, 6.6, 7.3, 5.6
Group 2:
5.1, 8.0, 3.8, 4.4, 5.2, 5.4, 4.3, 6.0, 3.7, 6.2

A. not significant
B. $0.05 < p < 0.10$
C. $p < 0.05$
D. $p < 0.01$

Is there a significant difference between the two groups? What significance level is correct?

All values are ranked together in ascending order of magnitude. The values from group 1 are printed thin, those from group 2 are printed fat. Add a rank number to each value. If there are identical values, for example, the rank numbers 9 and 10, then replace those rank numbers with average rank numbers, 9.5 and 9.5.

Subsequently, all fat printed rank numbers are added up, and so are the thin printed rank numbers. We will find the values 142.5 for fat print, and 67.5 for thin print.

According to the Mann-Whitney table of page 13, the difference should be larger than 71 in order for the significance level of difference to be <0.05. We find a difference of 75, which means that there is a p-value <0.05 and that the difference between the two groups is, thus, significant.

3.7	1
3.8	2
4.3	3
4.4	4
5.1	5
5.2	6
5.4	7
5.6	**8**
6.0	**9.5**
6.0	9.5
6.2	11
6.4	**12**
6.6	**13**
6.8	**14**
7.1	**15**
7.3	**16**
7.5	**17**
7.9	**18**
8.0	19
8.1	**20**

Mann-Whitney test

$P < 0.01$ levels

$n_2\downarrow$ / $n_1\rightarrow$	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4			10											
5		6	11	17										
6		7	12	18	26									
7		7	13	20	27	36								
8	3	8	14	21	29	38	49							
9	3	8	15	22	31	40	51	63						
10	3	9	15	23	32	42	53	65	78					
11	4	9	16	24	34	44	55	68	81	96				
12	4	10	17	26	35	46	58	71	85	99	115			
13	4	10	18	27	37	48	60	73	88	103	119	137		
14	4	11	19	28	38	50	63	76	91	106	123	141	160	
15	4	11	20	29	40	52	65	79	94	110	127	145	164	185
16	4	12	21	31	42	54	67	82	97	114	131	150	169	
17	5	12	21	32	43	56	70	84	100	117	135	154		
18	5	13	22	33	45	58	72	87	103	121	139			
19	5	13	23	34	46	60	74	90	107	124				
20	5	14	24	35	48	62	77	93	110					
21	6	14	25	37	50	64	79	95						
22	6	15	26	38	51	66	82							
23	6	15	27	39	53	68								
24	6	16	28	40	55									
25	6	16	28	42										
26	7	17	29											
27	7	17												
28	7													

The values are the minimal differences that are statistically significant with a p-value <0.01. The upper row gives the size of Group 1, the left column the size of Group 2

Mann-Whitney test

P < 0.05 levels

$n_2\downarrow$ / $n_1\rightarrow$	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5				15										
6			10	16	23									
7			10	17	24	32								
8			11	17	25	34	43							
9		6	11	18	26	35	45	56						
10		6	12	19	27	37	47	58	71					
11		6	12	20	28	38	49	61	74	87				
12		7	13	21	30	40	51	63	76	90	106			
13		7	14	22	31	41	53	65	79	93	109	125		
14		7	14	22	32	43	54	67	81	96	112	129	147	
15		8	15	23	33	44	56	70	84	99	115	133	151	171
16		8	15	24	34	46	58	72	86	102	119	137	155	
17		8	16	25	36	47	60	74	89	105	122	140		
18		8	16	26	37	49	62	76	92	108	125			
19	3	9	17	27	38	50	64	78	94	111				
20	3	9	18	28	39	52	66	81	97					
21	3	9	18	29	40	53	68	83						
22	3	10	19	29	42	55	70							
23	3	10	19	30	43	57								
24	3	10	20	31	44									
25	3	11	20	32										
26	3	11	21											
27	4	11												
28	4													

The values are the minimal differences that are statistically significant with a p-value <0.01. The upper row gives the size of Group 1, the left column the size of Group 2

Chapter 5
Confidence Intervals

The 95% confidence interval of a study represents an interval covering 95% of many studies similar to our study. It tells you something about what you can expect from future data: if you repeat the study, you will be 95% sure that the outcome will be within the 95% confidence interval. The 95% confidence of a study is found by the equation

$$95\%\ \text{confidence interval} = \text{mean} \pm 2 \times \text{standard error (SE)}$$

The SE is equal to the standard deviation (SD)/$\sqrt{n}$, where n = the sample size of your study. The SD can be calculated from the procedure reviewed in the Chap. 2.

With an SD of 1.407885953 and a sample size of n = 8,

$$\text{your SE} = 1.407885953 / \sqrt{8}$$
$$= 0.4977$$

With a mean value of your study of 53.375

$$\text{your 95\% confidence interval} = 53.375 \pm 2 \times 0.4977$$
$$= \text{between 52.3796 and 54.3704.}$$

The mean study results are often reported together with 95% confidence intervals. They are also the basis for equivalence studies, which will be reviewed in the next chapter. Also for study results expressed in the form of numbers of events, proportion of deaths, odds ratios of events, etc., 95% confidence intervals can be readily calculated. Plenty software on the Internet is available to help you calculate the correct confidence intervals.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_5,

Chapter 6
Equivalence Tests

Equivalence testing is important, if you expect a new treatment to be equally efficaceous as the standard treatment. This new treatment may still be better suitable for practice, if it has fewer adverse effects or other ancillary advantages.

For the purpose of equivalence testing we need to set boundaries of equivalence prior to the study. After the study we check whether the 95% confidence interval of the study is entirely within the boundaries.

As an example, in a blood pressure study a difference between the new and standard treatment between −10 and +10 mm Hg is assumed to smaller than clinically relevant. The boundary of equivalence is, thus, between −10 and +10 mm Hg. This boundary is a priori defined in the protocol.

Then, the study is carried out, and both the new and the standard treatment produce a mean reduction in blood pressure of 10 mm Hg (parallel-group study of 20 patients) with standard errors 10 mm Hg.

$$\begin{aligned}\text{The mean difference} &= 10 - 10 \text{ mm Hg}\\ &= 0 \text{ mm Hg}\end{aligned}$$

$$\text{The standard errors of the mean differences} = 10 \text{ mm Hg}$$

$$\begin{aligned}\text{The pooled standard error } (n = 10) &= \sqrt{(100/10 + 100/10)} \text{ mm Hg}\\ &= \sqrt{20} \text{ mm Hg}\\ &= 4.47 \text{ mm Hg}\end{aligned}$$

$$\begin{aligned}\text{The 95\% confidence interval of this study} &= 0 \pm 2 \times 4.47 \text{ mm Hg}\\ &= \text{between } -8.94 \text{ and } +8.97 \text{ mm Hg}\end{aligned}$$

This result is entirely within the a priori defined boundary of equivalence, which means that equivalence is demonstrated in this study.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_6,

Chapter 7
Power Equations

Power can be defined as statistical conclusive force. A study result is often expressed in the form of the mean result and its standard deviation (SD) or standard error (SE). With the mean result getting larger and the standard error getting smaller, the study obtains increasing power.

What is the power of the underneath study?

A blood pressure study shows a mean decrease in blood pressure of 10.8 mm Hg with a standard error of 3.0 mm Hg. Results from study samples are often given in grams, liters, Euros, mm Hg etc. For the calculation of power we have to standardize our study result, which means that the mean result has to be divided by its own standard error:

$$\begin{aligned} &\text{Mean} \pm \text{SE} \\ &= \text{mean} / \text{SE} \pm \text{SE} / \text{SE} \\ &= \text{t-value} \pm 1. \end{aligned}$$

The t-values are found in the t-table, can be looked upon as standardized results of all kinds of studies.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_7,

In our blood pressure study the t-value = 10.8/3.0 = 3.6. The unit of the t-value is not mm Hg, but rather SE-units. The question is: what power does the study have, if we assume a type I error (alpha) = 5% and a sample size of n = 20.

The question is: what is the power of this study if we assume a type I error (alpha) of 5%, and will have a sample size of n = 20.

A. 90% < power < 95%,
B. power > 80%,
C. power < 75%,
D. power > 75%.

n = 20 indicates 20–2 = 18 degrees of freedom in the case of two groups of ten patients each.

We will use the following power equation (prob = probability, z = value on the z-line (the x-axis of the t-distribution)

$$\text{Power} = 1 - \text{prob}\ (z < t - t^1)$$

t = the t-value of your results,

t^1 = the t-value, that matches a p-value of 0.05 = 2.1;

$t = 3.6;\ t^1 = 2.1;\ t - t^1 = 1.5;$

$\text{prob}\ (z < t - t^1) = \text{beta} = \text{type II error} = 0.05 - 0.1$

1-beta = power = 0.9 – 0.95 = between 90% and 95%.

So, there is a very good power here. See below for explanation of the calculation.

Explanation of the above calculation.

The t-table on the next page is a more detailed version of the t-table of page 21, and is adequate for power calculations. The degrees of freedom are in the left column and correlate with the sample size of a study. With large samples the frequency distribution of the data will be a little bit narrower, and that is corrected in the table. The t-values are to be looked upon as mean results of studies, but not expressed in mmol/l, kilograms, but in so-called SE-units (Standard error units), that are obtained by dividing your mean result by its own standard error. With a t-value of 3.6 and 18 degrees of freedom $t–t^1$ equals 1.5. This value is between 1.330 and 1.734. Look right up at the upper row for finding beta (type II error = the chance of finding no difference where there is one). We are between 0.1 and 0.05 (10% and 5%). This is an adequate estimate of the type II error. The power equals 100% – beta = between 90% and 95% in our example.

t-Table

	Q = **0.4**	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
v	2Q = **0.8**	**0.5**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.002**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326

(continued)

t-Table (continued)

3	0.277	0.765	1.638	2.353	3.182	4.547	5.841	10.213
4	0.171	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.269	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.269	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.685	1.319	1.714	2.069	2.600	2.807	3.485
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.654	1,315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.684	1,314	1.701	2.052	2.473	2.771	3.421
28	0.256	0.683	1,313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.254	0.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090

The upper row shows p-values = Areas under the curve (AUCs) of t-distributions. The second row gives two-sided p-values, it means that left and right end of the AUCs of the Gaussian-like curves are added up. The left column gives the adjustment for the sample size. If it gets larger, then the corresponding Gaussian-like curves will get a bit narrower. In this manner the estimates become more precise and more in agreement with reality. The t-table is empirical, and has been constructed in the 1930s of the past century with the help of simulation models and practical examples. It is till now the basis of modern statistics, and all modern software makes extensively use of it

Chapter 8
Sample Size

Continuous Data, Power 50%

An essential part of clinical studies is the question, how many subjects need to be studied in order to answer the studies' objectives. As an example, we will use an intended study that has an expected mean effect of 5, and a standard deviation (SD) of 15.

What required sample size do we need to obtain a significant result, or, in other words, a p-value of at least 0.05.

A. 16,
B. 36,
C. 64,
D. 100.

A suitable equation to assess this question can be constructed as follows.

With a study's t-value of 2.0 SEM-units, a significant p-value of 0.05 will be obtained. This should not be difficult for you to understand when you think of the 95% confidence interval of study being between – and + 2 SEM-units (Chap. 5).

We assume

$$\begin{aligned} \text{t-value} &= 2 \text{ SEMs} \\ &= (\text{mean study result}) / (\text{standard error}) \\ &= (\text{mean study result}) / (\text{standard deviation} / \sqrt{n}) \\ &\quad (n = \text{study's sample size}) \end{aligned}$$

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_8,

From the above equation it can be derived that

$$\sqrt{n} = 2 \times \text{standard deviation (SD)} / (\text{mean study result})$$
$$n = \text{required sample size}$$
$$= 4 \times (\text{SD} / (\text{mean study result}))^2$$
$$= 4 \times (15 / 5)^2 = 36$$

Answer B is correct.

You are testing here whether a result of 5 is significantly different from a result of 0. Often two groups of data are compared and the standard deviations of the two groups have to be pooled (see page 25). As stated above, with a t-value of 2.0 SEMs a significant result of $p = 0.05$ is obtained. However, the power of this study is only 50%, indicating that you will have 50% chance of an insignificant result the next time you perform a similar study.

Continuous Data, Power 80%

What is the required sample size of a study with an expected mean result of 5, and SD of 15, and that should have a p-value of at least 0.05 and a power of at least 80% (power index $= (z_\alpha + z_\beta)^2 = 7.8$).

A. 140,
B. 70,
C. 280,
D. 420.

An adequate equation is the following.

$$\text{Required sample size} = \text{power index} \times (\text{SD} / \text{mean})^2$$
$$= 7.8 \times (15 / 5)^2 = 70$$

If you wish to have a power in your study of 80% instead of 50%, you will need a larger sample size. With a power of only 50% your required sample size was only 36.

Continuous Data, Power 80%, 2 Groups

What is the required sample size of a study with two groups and a mean difference of 5 and SDs of 15 per Group, and that will have a p-value of at least 0.05 and a power of at least 80%. (Power index $= (z_\alpha + z_\beta)^2 = 7.8$).

A. 140,
B. 70,
C. 280,
D. 420.

The suitable equation is given underneath.

$$\text{Required sample size} = \text{power index} \times (\text{pooled SD})^2 / (\text{mean difference})^2$$
$$(\text{pooled SD})^2 = \text{SD}_1{}^2 + \text{SD}_2{}^2$$
$$\text{Required sample size} = 7.8 \times (15^2 + 15^2) / 5^2 = 140.$$

The required sample size is 140 patients per group. And so, with two groups you will need considerably larger samples than you do with 1 group.

Binary Data, Power 80%

What is the required sample size of a study in which you expect an event in 10% of the patients and wish to have a power of 80%.

10% events means a proportion of events of 0.1.

The standard deviation (SD) of this proportion is defined by the equation

$$\sqrt{[\text{proportion} \times (1 - \text{proportion})]} = \sqrt{(0.1 \times 0.9)}.$$

The suitable formula is given.

$$\begin{aligned}\text{Required sample size} &= \text{power index} \times \text{SD}^2 / \text{proportie}^2 \\ &= 7.8 \times (0.1 \times 0.9) / 0.1^2 \\ &= 7.8 \times 9 = 71.\end{aligned}$$

We conclude that with 10% events you will need about 71 patients in order to obtain a significant number of events for a power of 80% in your study.

Binary Data, Power 80%, 2 Groups

What is the required sample size of a study of two groups in which you expect.

A difference in events between the two groups of 10%, and in which you wish to have a power of 80%.

10% difference in events means a difference in proportions of events of 0.10.

Let us assume that in Group one 10% will have an event and in Group two 20%. The standard deviations per group can be calculated.

$$\text{For group 1: SD} = \sqrt{[\text{proportion} \times (1 - \text{proportion})]} = \sqrt{(0.1 \times 0.9)} = 0.3.$$

$$\text{For group 2: SD} = \sqrt{[\text{proportion} \times (1 - \text{proportion})]} = \sqrt{(0.2 \times 0.8)} = 0.4$$

$$\text{The pooled standard deviation of both groups} = \sqrt{(SD_1{}^2 + SD_2{}^2)}$$

$$= \sqrt{(0.3^2 + 0.4^2)}$$

$$= \sqrt{0.25} = 0.5$$

The adequate equation is underneath.

$$\text{Required sample size} = \text{power index} \times (\text{pooled SD})^2 / (\text{difference in proportions})^2$$

$$= 7.8 \times 0.5^2 / 0.1^2$$

$$= 7.8 \times 25 = 195.$$

Obviously, with a difference of 10% events between two groups we will need about 195 patients per group in order to demonstrate a significant difference with a power of 80%.

Chapter 9
Noninferiority Testing

Just like equivalence studies noninferiority studies are very popular in modern clinical research with many treatments at hand and new compounds being mostly only slightly different from the old ones. Unlike equivalence studies (Chap. 6), noninferiority studies have a single boundary, instead of two boundaries, with an interval of equivalence in between. Noninferiority studies have been criticized for their wide margin of inferiority making it virtually impossible to reject noninferiority.

As an example, two parallel-groups of patients with rheumatoid arthritis are treated with either a standard or a new nonsteroidal anti-inflammatory drug (NSAID). The reduction of gamma globuline levels (g/l) after treatment is used as the primary estimate of treatment success. The underneath three steps constitute an adequate procedure for noninferiority analysis.

Step 1: Determination of the Margin of Noninferiority, the Required Sample, and the Expected p-Value and Power of the Study Result

1. The left boundaries of the 95% confidence intervals of previously published studies of the standard NSAID versus various alternative NSAIDS were never lower than -8 g/l. And, so, the margin was set at -8 g/l.
2. Based on a pilot-study with the novel compound the expected mean difference was 0 g/l with an expected standard deviation of 32 g/l. This would mean a required sample size of

$$\text{n} = \text{power index} \times (\text{SD} / (\text{margin} - \text{mean}))^2$$

$$\text{n} = 7.8 \times (32 / (-8 - 0))^2 = 125 \text{ patients per group.}$$

A power index of 7.8 takes care that noninferiority is demonstrated with a power of about 80% in this study (see also Chap. 8).

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_9,

3. The mean difference between the new and standard NSAID was calculated to be 3.0 g/l with a standard error (SE) of 4.6 g/l. This means that the t-value of the study equaled t = (margin − mean)/SE = (−8 − 3)/4.6 = −2.39 SE-units or SEM-units. This t-value corresponds with a p-value of <0.05 (page 21 bottom row, why the bottom row can be applied is explained in the next Chapter). Non-inferiority is, thus, demonstrated at $p<0.05$.

Step 2: Testing the Significance of Difference Between the New and the Standard Treatment

The mean difference between the new and standard treatment equaled 3.0 g/l with an SE of 4.6 g/l. The 95% confidence of this result is 3.0 ± 2*4.6, and is between −6.2 and 12.2 g/l (* = sign of multiplication). This interval does cross the zero value on the z-axis, which means no significant difference from zero ($p>0.05$).

Step 3: Testing the Significance of Difference Between the New Treatment and a Placebo

A similarly sized published trial of the standard treatment versus placebo produced a t-value of 2.83, and thus a p-value of 0.0047. The t-value of the current trial equals 3.0/4.6 = 0.65 SE-units. The add-up sum 2.83 + 0.65 = 3.48 is an adequate estimate of the comparison of the new treatment versus placebo. A t-value of 3.48 corresponds with a p-value of <0.002 (see page 21, bottom row, the use of bottom row will be explained in the next Chapter). This would mean that the new treatment is significantly better than placebo at $p<0.002$.

Conclusion

We can now conclude that

(1) noninferiority is demonstrated at $p<0.05$, that
(2) a significant difference between the new and standard treatment is rejected at $p>0.05$, and that
(3) the new treatment is significantly better than placebo at $p<0.002$. Non-inferiority has, thus, been unequivocally demonstrated in this study.

Chapter 10
Z-Test for Cross-Tabs

Two groups of patients are assessed for being sleepy through the day. We wish to estimate whether group 1 is more sleepy than group 2. The underneath cross-tab gives the data.

	Sleepiness	No sleepiness
Treatment 1 (group 1)	5 (a)	10 (b)
Treatment 2 (group 2)	9 (c)	6 (d)

$$z = \frac{\text{difference between proportions of sleepers per group (d)}}{\text{pooled standard error difference}}$$

$$z = \frac{d}{\text{pooled SE}} = \frac{(9/15 - 5/15)}{\sqrt{(SE_1^2 + SE_2^2)}}$$

$$SE_1 \text{ (or } SEM_1) = \sqrt{\frac{p_1(1-p_1)}{n_1}} \text{ where } p_1 = 5/15 \text{ etc.....,}$$

$z = 1.45$, not statistically significant from zero, because for a $p < 0.05$ a z-value of at least 1.96 is required. This means that no significant difference between the two groups is observed. The p-value of the z-test can be obtained by using the bottom row of the t-table from page 21.

Note:

For the z-test a normal distribution approach can be used. The t-distributions are usually a bit wider than the normal distributions, and therefore, adjustment for study size using degrees of freedom (left column of the t-table) is required. With a large study size the t-distribution is equal to the normal distribution, and the t-values are equal to the z-values. They are given in the bottom row of the t-table.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_10,

Note:

A single group z-test is also possible. For example in ten patients we have four responders. We question whether four responders is significantly more than zero responders.

$$\begin{aligned} z &= \text{proportion} / (\text{its SE}) \\ SE &= \sqrt{[(4/10 \times (1 - 4/10)) / n]} \\ &= \sqrt{(0.24/10)} \\ z &= 0.4 / \sqrt{(0.24/10)} \\ z &= 0.4 / 0.1549 \\ z &= 2.582 \end{aligned}$$

According to the bottom row of the t-table from page 21 the p-value is < 0.01. The proportion of 0.4 is, thus, significantly larger than a proportion of 0.0.

Chapter 11
Chi-Square Tests for Cross-Tabs

First Example Cross-Tab

The underneath table shows two separate groups with patients assessed for suffering from sleepiness through the day. We wish to know whether there is a significant difference between the proportions of subjects being sleepy.

	Sleepiness	No sleepiness	
Group 1	5 (a)	10 (b)	15 (a+b)
Group 2	9 (c)	6 (d)	15 (c+d)
	14 (a+c)	16 (b+d)	30 (a+b+c+d)

The chi-square pocket calculator method is used for testing these data.

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} = \frac{(30-90)^2(30)}{15\times15\times16\times14} = \frac{3{,}600\times30}{15\times15\times16\times14}$$

$$= \frac{108.000}{50.400} = 2.143$$

The chi-square value equals 2.143. The chi-square table can tell us whether or not the difference between the groups is significant. See next page for the procedure to be followed.

Chi-Square Table (χ^2-Table)

The underneath chi-square table gives columns and rows: the upper row gives the p-values. The first column gives the degrees of freedom which is here largely in agreement with the numbers of cells in a cross-tab. The simplest cross-tab has 4 cells, which means $2\times2=4$ cells. The table has been constructed such that we have here $(2-1)\times(2-1)=1$ degree of freedom. Look at the row with 1 degree of freedom: a chi-square value of 2.143 is left from 2.706. Now look from here right up at the

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_11,

Chi-squared distribution

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.321
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.527
14	21.064	23.685	29.141	36.124
15	22.307	24.996	30.578	37.698
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.791
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.819
20	28.412	31.410	37.566	45.314
21	29.615	32.671	38.932	46.796
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.619
26	35.563	38.885	45.642	54.051
27	36.741	40.113	46.963	55.475
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.702
40	51.805	55.758	63.691	73.403
50	63.167	67.505	76.154	86.660
60	74.397	79.082	88.379	99.608
70	85.527	90.531	100.43	112.32
80	96.578	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45

upper row. The corresponding p-value is larger than 0.1 (10%). There is, thus, no significant difference in sleepiness between the two groups. The small difference observed is due to the play of chance.

Second Example Cross-Tab

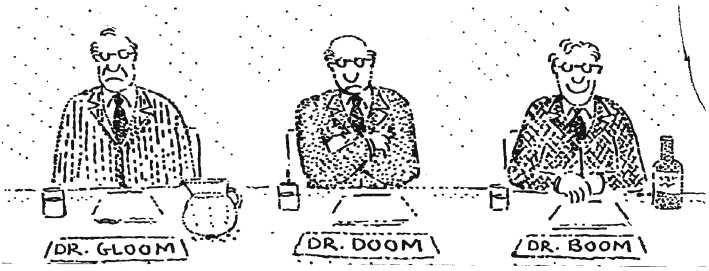
Two partnerships of internists have the intention to associate. However, in one of the two a considerable number of internists has suffered from a burn-out.

	Burn out	No burn out	
Partnership 1	3 (a)	7 (b)	10 (a+b)
Partnership 2	0 (c)	10 (d)	10 (c+d)
	3 (a+c)	17 (b+d)	20 (a+b+c+d)

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} = \frac{(30-0)^2(20)}{10\times10\times17\times3} = \frac{900\times20}{.....} = 3.529$$

According to the chi-square table of the previous page a p-value is found of <0.10.

This means that no significant difference is found, but a p-value between 0.05 and 0.10 is looked upon as a trend to significance. The difference may be due to some avoidable or unavoidable cause. We should add here that values in a cell lower than 5 is considered slightly inappropriate according to some, and another test like the log likelihood ratio test (Chap. 13) is more safe.



Example for Practicing 1

Example	2 × 2 table	Events	No events	
	Group 1	15 (a)	20 (b)	35 (a+b)
	Group 2	15 (c)	5 (d)	20 (c+d)
		30 (a+c)	25 (b+d)	55 (a+b+c+d)

Pocket calculator

$$\frac{(ad-bc)^2\,(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} = \qquad p = ...$$

Example for Practicing 2

Another example	2×2 table	Events	No events	
	Group 1	16 (a)	26 (b)	42 (a+b)
	Group 2	5 (c)	30 (d)	35 (c+d)
		21 (a+c)	56 (b+d)	77 (a+b+c+d)

Pocket calculator

$$\frac{(ad-bc)^2\,(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} = \qquad p = ...$$

Chapter 12
Odds Ratios

The odds ratio test is just like the chi-square test applicable for testing cross-tabs. The advantage of the odds ratio test is that a odds ratio value can be calculated. The odds ratio value is just like the relative risk an estimate of the chance of having an event in group 1 compared to that of group 2. An odds ratio value of 1 indicates no difference between the two groups.

Example 1

	Events	No events	
	Numbers of patients		
Group 1	15 (a)	20 (b)	35 (a+b)
Group 2	15 (c)	5 (d)	20 (c+d)
	30 (a+c)	25 (b+d)	55 (a+b+c+d)

The odds of an event = the number of patients in a group with an event divided by the number without. In group 1 the odds of an event equals = a/b.

The odds ratio (OR) of group 1 compared to group 2

$$= (a / b) / (c / d)$$
$$= (15 / 20) / (15 / 5)$$
$$= 0.25$$
$$\ln OR = \ln 0.25 = -1.386 \text{ (ln = natural logarithm)}$$

The standard error (SE) of the above term

$$= \sqrt{(1 / a + 1 / b + 1 / c + 1 / d)}$$
$$= \sqrt{(1 / 15 + 1 / 20 + 1 / 15 + 1 / 5)}$$
$$= \sqrt{0.38333}$$
$$= 0.619$$

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_12,

The odds ratio can be tested using the z-test (Chap. 10).

$$\text{The test-statistic} = \text{z-value}$$

$$= (\text{odds ratio}) / \text{SE}$$

$$= -1.386 / 0.619$$

$$= -2.239$$

If this value is smaller than −2 or larger than +2, then the odds ratio is significantly different from 1 with $p < 0.05$. An odds ratio of 1 means that there is no difference in events between group 1 and group 2. The bottom row of the t-table (page 21) gives the z-values matching Gaussian distributions. Look at a z-value of 1.96 right up at the upper row. We will find a p-value here of 0.05. And, so, a z-value larger than 1.96 indicates a p-value of <0.05. There is a significant difference in event between the two groups.

Example 2

	Events	No events	
	Number of patients		
Group 1	16 (a)	26 (b)	42 (a+b)
Group 2	5 (c)	30 (d)	35 (c+d)
	21 (a+c)	56 (b+d)	77 (a+b+c+d)

Test with OR whether there is a significant difference between group 1 and 2.
See for procedure also example 1.

$$\text{OR} = (16 / 26) / (5 / 30)$$

$$= 3.69$$

$$\text{lnOR} = 1.3056 \text{ (ln = natural logarithm see the above example)}$$

$$\text{SE} = \sqrt{(1/16 + 1/26 + 1/5 + 1/30)}$$

$$= \sqrt{0.334333}$$

$$= 0.578$$

$$\text{z-value} = 1.3056 / 0.578$$

$$= 2.259$$

Because this value is larger than 2, a p-value of <0.05 is observed, 0.024 to be precise (numerous "p-calculator for z-values" sites in Google will help you calculate an exact p-value if required.

Chapter 13
Log Likelihood Ratio Tests

The sensitivity of the chi-square test (Chap. 11) and the odds ratio test (Chap. 12) for testing cross-tabs is limited, and not entirely accurate if the values in one or more cells is smaller than 5. The log likelihood ratio test is an adequate alternative with generally better sensitivity, and, so, it must be absolutely recommended.

Example 1
A group of citizens is taking a pharmaceutical company to court for misrepresenting the danger of fatal rhabdomyolysis due to statin treatment.

	Patients with rhabdomyolysis	Patients without
Company	1 (a)	309,999 (b)
Citizens	4 (c)	300,289 (d)

p_{co} = proportion given by the pharmaceutical company = a/(a + b) = 1/310,000
p_{ci} = proportion given by the citizens = c/(c + d) = 4/300,293

We make use of the z-test (Chap. 10) for testing log likelihood ratios.

As it can be shown that −2 log likelihood ratio equals z^2, we can test the significance of difference between the two proportions.

$$\text{Log likelihood ratio} = 4 \log \frac{1/310{,}000}{4/300{,}293} + 300289 \log \frac{1-1/310{,}000}{1-4/300{,}293}$$

$$= -2.641199$$

$$-2 \text{ log likelihood ratio} = -2 \times -2.641199$$

$$= 5.2824 \ (p < 0.05, \text{ because } z > 2).$$

$$= z^2$$

A z-value larger than 2 means a significant difference in your data (Chap. 10). Here the z-value equals $\sqrt{5.2824} = 2.29834$. The "p-calculator for z-values" in Google tells you that the exact p-value = 0.0215, much smaller than 0.05.

We should note here that both the odds ratio test and chi-square test produced a non-significant result here (p > 0.05). Indeed, the log likelihood ratio test is much

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_13,

more sensitive than the other tests for the same kind of data, which might once in a while be a blessing for desperate investigators.

Example 2
Two group of 15 patients at risk for arrhythmias were assessed for the development of torsade de points after calcium channel blockers treatment.

	Patients with torsade de points	Patients without
Calcium channel blocker 1	5	10
Calcium channel blocker 2	9	6

The proportion of patients with event from calcium channel blocker 1 is 5/15, from blocker 2 it is 9/15.

$$\text{Log likelihood ratio} = 9\ \log\frac{5/15}{9/15} + 6\ \log\frac{1-5/15}{1-9/15}$$

$$= -2.25$$

$$-2 \text{ log likelihood ratio} = 4.50$$

$$= z^2$$

$$\text{z-value} = \sqrt{4.50} = 2.1213$$

$$\text{p-value} < 0.05, \text{ because } z > 2.$$

Both odds ratio test and chi-square test were again non-significant ($p > 0.05$).

Example 3
Two groups of patients with stage IV New York Heart Association heart failure were assessed for clinical admission while on two beta-blockers.

	Patients with clinical admission	Patients without
Beta blocker 1	77	62
Beta blocker 2	103	46

The proportion of patients with event while on beta blocker 1 is 77/139, while on beta blocker 2 it is 103/149.

$$\text{Log likelihood ratio} = 103\ \log\frac{77/139}{103/149} + 46\ \log\frac{1-77/139}{1-103/149}$$

$$= -5.882$$

$$-2 \text{ log likelihood ratio} = 11.766$$

$$= z^2$$

$$\text{z-value} = \sqrt{11.766} = 3.43016$$

$$\text{p-value} < 0.002, \text{ because } z > 3.090$$

(see the t-table on page 21).

Both the odds ratio test and chi-square test were also significant. However, at lower levels of significance, both p-values $0.01 < p < 0.05$.

Chapter 14
McNemar's Tests

The past four Chapters have reviewed four methods for analyzing cross-tabs of two groups of patients. Sometimes a single group is assessed twice, and, then, we obtain a slightly different cross-tab. McNemar's test must be applied by analyzing these kind of data.

Example McNemar's Test

315 subjects are tested for hypertension using both an automated device (test-1) and a sphygmomanometer (test-2).

		Test 1		
		+	−	Total
Test 2	+	184	54	238
	−	14	63	77
Total		198	117	315

$$\text{Chi - square McNemar} = \frac{(54-14)^2}{54+14} = 23.5$$

184 subjects scored positive with both tests and 63 scored negative with both tests. These 247 subjects, therefore, give us no information about which of the tests is more likely to score positive.

The information we require is entirely contained in the 68 subjects for whom the tests did not agree (the discordant pairs). The above table also shows how the chi-square value is calculated. The chi-square table (page 32) is used for finding the appropriate p-value. Here we have again 1 degree of freedom. The 1 degree of freedom row of the chi-square table shows that our result of 23.5 is a lot larger than 10.827. When looking up at the upper row we will find a p-value < 0.001. The two devices produce significantly different results at $p < 0.001$.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_14,

McNemar Odds Ratios, Example

Just like with the usual cross-tabs (Chap. 12) odds ratios can be calculated with the single group cross-tabs. So far we assessed two groups, one treatment. two antihypertensive treatments are assessed in a single group of patients

		Normotension with drug 1	
		Yes	No
Normotension with drug 2	Yes	(a) 65	(b) 28
	No	(c) 12	(d) 34

Here the OR = b/c, and the SE is not $\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}$, but rather $\sqrt{\left(\frac{1}{b}+\frac{1}{c}\right)}$.

$$\begin{aligned} \text{OR} &= 28/12 \\ &= 2.33 \\ \text{lnOR} &= \ln 2.33 \ (\ln = \text{natural logarithm}) \\ &= 0.847 \\ \text{SE} &= \sqrt{\left(\frac{1}{b}+\frac{1}{c}\right)} = 0.345 \\ \text{lnOR} \pm 2\ \text{SE} &= 0.847 \pm 0.690 \\ &= \text{between } 0.157 \text{ and } 1.537, \end{aligned}$$

Turn the ln numbers into real numbers by the anti-ln button (the invert button, on many calculators called the 2ndF button) of your pocket calculator.

$$\begin{aligned} &= \text{between } 1.16 \text{ and } 4.65 \\ &= \text{significantly different from } 1.0. \end{aligned}$$

A p-value can be calculated using the z-test (Chap. 10).

$$\begin{aligned} z &= \text{lnOR} / \text{SEM} \\ &= 0.847 : 0.345 \\ &= 2.455. \end{aligned}$$

The bottom row of the t-table (page 21) shows that this z-value is smaller than 2.326, and this means the corresponding p-value of < 0.02. The two drugs, thus, produce significantly different results at $p < 0.02$.

Chapter 15
Bonferroni t-Test

The t-test can be used to test the hypothesis that two group means are not different (Chap. 3). When the experimental design involves multiple groups, and, thus, multiple tests, we increase our chance of finding a difference. This is, simply, due to the play of chance rather than a real effect. Multiple testing without any adjustment for this increased chance is called data dredging, and is the source of multiple type I errors (chances of finding a difference where there is none). The Bonferroni t-test (and many other methods) are appropriate for the purpose of adjusting the increased risk of type I errors.

Bonferroni t-Test

The underneath example studies three groups of patients treated with different hemoglobin improving compounds. The mean increases of hemoglobin are given.

	Sample size	Mean hemoglobin (mmol / l)	Standard deviation (mmol / l)
Group 1	16	8.725	0.8445
Group 2	10	10.6300	1.2841
Group 3	15	12.3000	0.9419

An overall analysis of variance test produced a p-value of < 0.01. The conclusion is that we have a significant difference in the data, but we will need additional testing to find out where exactly the difference is, between group 1 and 2, between group 1 and 3, or between group 2 and 3. The easiest approach is to calculate the t–test for each comparison. It produces a highly significant difference at $p < 0.01$ between group 1 versus 3 with no significant differences between the other comparisons. This highly significant result is, however, unadjusted for multiple comparisons. If one analyzes a set of data with three t-tests, each using a 5% critical value for concluding that there is a significant difference, then there is about $3 \times 5 = 15\%$ chance of finding it. This mechanism is called the Bonferroni inequality.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_15,

Bonferroni recommended a solution for the inequality, and proposed to follow in case of three t-tests to use a smaller critical level for concluding that there is a significant difference:

$$\text{With 1 t-test: critical level} = 5\%$$

$$\text{With 3 t-tests: critical level} = 5 / 3 = 1.6\%.$$

The above equations lead rapidly to very small critical values, otherwise called p-values, and is, therefore, considered to be over-conservative. A somewhat less conservative version of the above equation was also developed by Bonferroni., and it is called the Bonferroni t-test.

In case of three comparisons the rejection p-value will be $0.05 \times \frac{2}{3(3-1)} = 0.0166.$

In the given example a p-value of 0.0166 is still larger than 0.01, and, so, the difference observed remained statistically significant, but using a cut-off p-value of 0.0166, instead of 0.05, means that the difference is not *highly* significant anymore.

Chapter 16
Variability Analysis

In some clinical studies, the spread of the data may be more relevant than the average of the data. E.g., when we assess how a drug reaches various organs, variability of drug concentrations is important, as in some cases too little and in other cases dangerously high levels get through. Also, variabilities in drug response may be important. For example, the spread of glucose levels of a slow-release-insulin is important.

One Sample Variability Analysis

For testing whether the standard deviation (or variance) of a sample is significantly different from the standard deviation (or variance) to be expected the chi-square test with multiple degrees of freedom is adequate. The test statistic, the chi-square-value (= χ^2–value) is calculated according to

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \text{ for } n-1 \text{ degrees of freedom}$$

(n = sample size, s = standard deviation, s^2 = variance sample, σ = expected standard deviation, σ^2 = expected variance).

For example, the aminoglycoside compound gentamicin has a small therapeutic index. The standard deviation of 50 measurements is used as a criterion for variability. Adequate variability is accepted if the standard deviation is less than 7 μg/l. In our sample a standard deviation of 9 μg/l is observed.

The test procedure is given.

$$\chi^2 = (50-1)9^2 / 7^2 = 81$$

The chi-square table (page 32) shows that, for $50-1=49$ degrees of freedom, we will find a p-value < 0.01. This sample's standard deviation is significantly larger than that required. This means that the variability in plasma gentamicin concentrations is larger than acceptable.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_16,

Two Sample Variability Test

F-tests can be applied to test if the variabilities of two samples are significantly different from one another. The division sum of the samples' variances (larger variance/smaller variance) is used for the analysis. For example, two formulas of gentamicin produce the following standard deviations of plasma concentrations.

	Patients (n)	Standard deviation (SD) (µg/l)
Formula-A	10	3.0
Formula-B	15	2.0

$$\text{F-value} = SD_A^2 / SD_B^2$$

$$= 3.0^2 / 2.0^2$$

$$= 9 / 4 = 2.25$$

with degrees of freedom (dfs) for

$$\text{formula-A of } 10 - 1 = 9$$

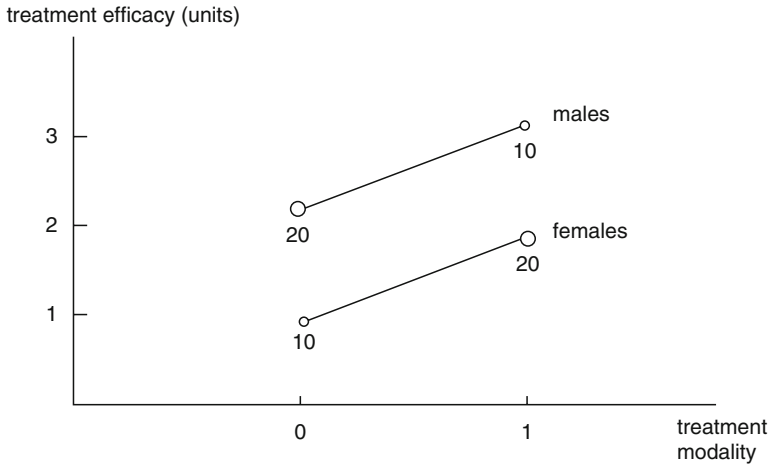
$$\text{formula-B of } 15 - 1 = 14.$$

The F-table on the next page shows that an F-value of at least 3.01 is required not to reject the null - hypothesis. Our F-value is 2.25 and, so, the p-value is > 0.05. No significant difference between the two formulas can be demonstrated. This F-test is given on the next page.

F-Table

df of denominator	2-tailed *P*-value	1-tailed P-value	Degrees of freedom (*df*) of the numerator												
			1	2	3	4	5	6	7	8	9	10	15	25	500
1	0.05	0.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.6	963.3	968.6	984.9	998.1	1017.0
1	0.10	0.205	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	245.9	249.3	254.1
2	0.05	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.46	39.50
2	0.10	0.05	18.51	19.00	19.16	19.25	19.13	19.33	19.35	19.37	19.38	19.40	19.43	19.46	19.49
3	0.05	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.12	13.99
3	0.10	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.63	8.53
4	0.05	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.50	8.27
4	0.10	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.77	5.64
5	0.05	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.27	6.03
5	0.10	0.05	6.61	5.79	5.41	5.19	5.05	4.95	6.88	4.82	4.77	4.74	4.62	4.52	4.37
6	0.05	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.11	4.86
6	0.10	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.83	3.68
7	0.05	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.40	4.16
7	0.10	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.40	3.24
8	0.05	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	3.94	3.68
8	0.10	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.11	2.94
9	0.05	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.60	3.35
9	0.10	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.89	2.72
10	0.05	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.35	3.09
10	0.10	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.73	2.55
15	0.05	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.69	2.41
15	0.10	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.28	2.08
20	0.05	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.40	2.10
20	0.10	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.07	1.86
30	0.05	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.12	1.81
30	0.10	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.88	1.64
50	0.05	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11	1.92	1.57
50	0.10	0.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.73	1.46
100	0.05	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97	1.77	1.38
100	0.10	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.62	1.31
1000	0.05	0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06	1.85	1.64	1.16
1000	0.10	0.05	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.68	1.52	1.13

Chapter 17
Confounding



In the above study the treatment effects are better in the males than they are in the females. This difference in efficacy does not influence the overall assessment as long as the numbers of males and females in the treatment comparison are equally distributed. If, however, many females received the new treatment, and many males received the control treatment, a peculiar effect on the overall data analysis is observed as demonstrated by the difference in magnitudes of the circles in the above figure: the overall regression line will become close to horizontal, giving rise to the erroneous conclusion that no difference in efficacy exists between treatment and control. This phenomenon is called confounding, and may have a profound effect on the outcome of the study.

Confounding can be assessed by the method of subclassification. In the above example an overall mean difference between the two treatment modalities is calculated.

For treatment zero

$$\text{Mean effect} \pm \text{standard error (SE)} = 1.5 \text{ units} \pm 0.5 \text{ units}$$

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_17,

For treatment one

$$\text{Mean effect} \pm \text{SE} = 2.5 \text{ units} \pm 0.6 \text{ units}$$

The mean difference of the two treatments

$$= 1.0 \text{ units} \pm \text{pooled standard error}$$

$$= 1.0 \pm \sqrt{(0.5^2 + 0.6^2)}$$

$$= 1.0 \pm 0.61$$

$$\text{The t-value as calculated} = 1.0 / 0.61 = 1.639$$

With $100 - 2$ (100 patients, 2 groups) = 98 degrees of freedom the p-value of this difference is calculated to be

$$= p > 0.10 \text{ (according to t-table page 21).}$$

In order to assess the possibility of confounding, a weighted mean has to be calculated. The underneath equation is adequate for the purpose.

$$\text{Weighted mean} = \frac{\text{Difference}_{\text{males}} / \text{ its SE}^2 \pm \text{Difference}_{\text{females}} / \text{ its SE}^2}{1 / \text{ SE}^2_{\text{males}} + 1 / \text{ SE}^2_{\text{females}}}$$

For the males we find means of 2.0 and 3.0 units, for the females 1.0 and 2.0 units. The mean difference for the males and females separately are 1.0 and 1.0 as expected from the above figure. However, the pooled standard errors are different, for the males 0.4, and for the females 0.3 units.

According to the above equation a weighted t-value is calculated

$$\text{Weighted mean} = \frac{(1.0 / 0.4^2 + 1.0 / 0.3^2)}{(1 / 0.4^2 + 1 / 0.3^2)}$$

$$= 1.0$$

$$\text{Weighted SE}^2\ 1 / (1 / 0.4^2 + 1 / 0.3^2)$$

$$= 0.576$$

$$\text{Weighted SE} = 0.24$$

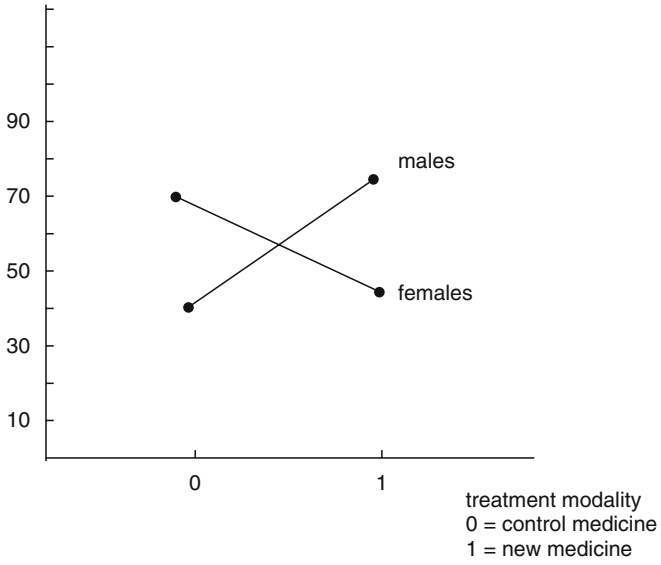
$$\text{t-value} = 1.0 / 0.24 = 4.16$$

$$\text{p-value} < 0.001$$

The weighted mean is equal to the unweighted mean. However, its SE is much smaller. It means that after adjustment for confounding a very significant difference is observed.

Other methods for assessing confounding include multiple regression analysis and propensity score assessments. Particularly, with more than a single confounder these two methods are unavoidable, and they can not be carried out on a pocket calculator.

Chapter 18
Interaction



The medical concept of interaction is synonymous to the terms heterogeneity and synergism. Interaction must be distinguished from confounding. In a trial with interaction effects the parallel groups have similar characteristics. However, there are subsets of patients that have an unusually high or low response. The above figure gives an example of a study in which males seem to respond better to the treatment 1 than females. With confounding things are different. For whatever reason the randomization has failed, the parallel groups have asymmetric characteristics. E.g., in a placebo-controlled trial of two parallel-groups asymmetry of age may be a confounder. The control group is significantly older than the treatment group, and this can easily explain the treatment difference as demonstrated in the previous chapter.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_18,

Example of Interaction

A parallel-group study of verapamil versus metoprolol for the treatment of paroxysmal atrial tachycardias. The numbers of episodes of paroxysmal atrial tachycardias per patient are the outcome variable.

	Verapamil	Metoprolol	
Males	52	28	
	48	35	
	43	34	
	50	32	
	43	34	
	44	27	
	46	31	
	46	27	
	43	29	
	49	25	
	464	302	766
Females	38	43	
	42	34	
	42	33	
	35	42	
	33	41	
	38	37	
	39	37	
	34	40	
	33	36	
	34	35	
	368	378	746
	832	680	

Overall metoprolol seems to perform better. However, this is only true only for one subgroup (males).

	Males	Females
$\text{Mean}_{\text{verapamil}}$ (SD)	46.4 (3.23866)	36.8 (3.489667)
$\text{Mean}_{\text{metoprolol}}$ (SD)	30.2 (3.48966) −	37.8 (3.489667) −
Difference means (SE)	16.2 (1.50554)	−1.0 (1.5606)
Difference between males and females	17.2 (2.166)	

$$\text{t - value} = 17.2 / 2.166 = 8...$$

$$p < 0.0001$$

There is a significant difference between the males and females, and, thus, a significant interaction between gender and treat-efficacy. Interaction can also be assessed with analysis of variance and regression modeling. These two methods are the methods of choice in case you expect more than a single interaction in your data. They should be carried out on a computer.

Chapter 19
Duplicate Standard Deviation for Reliability Assessment of Continuous Data

The reliability, otherwise called reproducibility of diagnostic tests is an important quality criterion. A diagnostic test is very unreliable, if it is not well reproducible.

Example 1

Test 1	Test 2	Difference	(Difference)^2
Result			
1	11	−10	100
10	0	10	100
2	11	−9	81
12	2	10	100
11	1	10	100
1	12	−11	121
Mean			
6.17	6.17	0	100.3

$$\text{Duplicate standard deviation} = \text{duplicate standard deviation (SD)}$$
$$= \sqrt{(1/2 \times \text{mean (difference)}^2)}$$
$$= \sqrt{(1/2 \times 100.3)}$$
$$= 7.08$$

The proportional duplicate standard deviation%

$$= \frac{\text{duplicate standard deviation}}{\text{overall mean}} \times 100\%$$
$$= \frac{7.08}{6.17} \times 100\%$$
$$= 115\%$$

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_19,

An adequate reliability is obtained with a proportional duplicate standard deviation of 10–20%. In the current example, although the mean difference between the two tests equals zero, there is, thus, a very poor reproducibility.

Example 2

Question is this test well reproducible?

Test 1	Test 2
Result	
6.2	5.1
7.0	7.8
8.1	3.9
7.5	5.5
6.5	6.6

Analysis:

Test 1	Test 2	Difference	Difference^2
Result			
6.2	5.1	1.1	1.21
7.0	7.8	–0.8	0.64
8.1	3.9	4.2	17.64
7.5	5.5	2.0	4.0
6.5	6.6	–0.1	0.01
Mean			
7.06	5.78		4.7

Grand mean 6.42

$$\text{Duplicate standard deviation} = \sqrt{½} \times 4.7$$
$$= 1.553$$

Proportional duplicate standard deviation %

$$= \frac{\text{duplicate standard deviation}}{\text{overall mean}} \times 100\%$$

$$= \frac{1.533}{6.42} \times 100\%$$

$$= 24\%$$

A good reproducibility is between 10% and 20%. In the above example reproducibility is, thus, almost good.

Chapter 20
Kappas for Reliability Assessment of Binary Data

The reproducibility of continuous data can be estimated with duplicate standard deviations (Chap. 19). With binary data Cohen's kappas are used for the purpose. Reliability assessment of diagnostic procedures is an important part of the validity assessment of scientific research.

Example

Positive (pos) or negative (neg) laboratory tests of 30 patients are assessed. All patiënts are tested a second time in order to estimate the level of reproducibility of the test.

		1st time		
		pos	neg	
2nd time	pos	10	5	15
	neg	4	11	15
		14	16	30

If the test is not reproducible at all, then we will find twice the same result in 50% of the patients, and a different result the second time in the other 50% of the patients.

Overall	30 tests have been carried out twice.
We observe	10 times 2 × positive and 11 times 2 × negative.

And thus, twice the same is found in

21 patients which is considerable more than in half of the cases, which should have been 15 times.

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9_20,

Minimal indicates the number of duplicate observations if reproducibility were zero, maximal indicates the number of duplicate observations if the reproducibility were 100%.

$$\text{Kappa} = \frac{\text{observed} - \text{minimal}}{\text{maximal} - \text{minimal}}$$

$$= \frac{21 - 15}{30 - 15}$$

$$= 0.4$$

A kappa-value of 0.0 means that reproducibility is very poor.

A kappa of 1.0 would have meant excellent reproducibility.

In our example we observed a kappa of 0.4, which means reproducibility is very moderate.

Final Remarks

Statistics is no bloodless algebra. It is a discipline at the interface of biology and mathematics. Mathematics is used to answer biological questions. Biological processes are full of variations, and statistics gives no certainties, only chances. What kind of chances: chances that your prior hypotheses are true or untrue. The human brain hypothesizes all the time. And we currently believe that hypotheses must be assessed with hard data.

When it comes to statistical data analyses, clinicians and clinical investigators soon get very nervous, and tend to leave their data to a statistician who runs the data through SAS of SPSS or any other software program to see if there are significant p-values. This practice is called data dredging and is the source of multiple type I errors of finding a difference where there is none.

The best defense against this practice is the use of simple tests. These tests, generally, provide the best power for confirmative research, because this research is based on sound arguments. Multiple variable tests are not always in place here, as they tend to enhance the risk of power loss, data dredging, and type I errors producing a host of irrelevant p-values. Also multiple variable tests, although interesting, are considered exploratory rather than confirmatory, in other words they, generally, prove nothing, and have to be confirmed.

The current book was written for various reasons:

1. To review the basic principles of statistical testing which tends to be increasingly forgotten in the current computer era.
2. To serve as a primer for nervous investigators who would like to perform their own data analyses but feel inexpert to do so.
3. To make investigators better understand what they are doing, when analyzing clinical data.
4. To facilitate data analysis by use of a number of *rapid* pocket calculator methods.
5. As a primer for those who wish to master more advanced statistical methods. More advanced methods are reviewed by the same authors in the books "SPSS

T.J. Cleophas and A.H. Zwinderman, *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*, DOI 10.1007/978-94-007-1211-9,

for Starters" 2010, "Statistics Applied to Clinical Trials" fourth edition, 2009, "Statistics Applied to Clinical Trials: Self-Assessment Book, 2002, all of them edited by Springer, Dordrecht. These books closely fit and complement the format and contents of the current book.

The current book is very condensed, but this should be threshold lowering to readers. As a consequence, however, the theoretical background of the methods described are not sufficiently explained in the text. Extensive theoretical information is also given in the above mentioned books from the same authors.

Index

A
Alpha, 20
Analysis of variance, 41, 50
Areas under the curve, 21

B
Beta, 20
Bloodless algebra, 55
Bonferroni inequality, 41
Bonferroni t-test, 41–42
Boundaries of equivalence, 17

C
Chi-square table, 31–33
Chi-square test, 31–35, 37, 38, 43
Chi-square test for cross-tabs, 31–34
Cohen's kappa, 53
Confidence intervals, 1, 15
Confounding, 1, 47–49
Cross-tabs, 29–35, 37, 39, 40

D
Data dredging, 41
Degrees of freedom, 6–8, 20, 29, 31, 43–45, 48
Dependent variables, v
Diagnostic tests, 51
Duplicate standard deviation, 51–52

E
Equivalence tests, 17

F
Frequency distribution, 8
F-table (Fisher), 44, 45
F-test (Fisher), 44

G
Gaussian distribution, 36

I
Independent variables, v
Interaction, 49–50
Irrelevant p-values, 42

K
Kappa, 53–54
Kappa-values, 54

L
LnOR, 35, 36, 40
Ln values, 40
Log likelihood ratio, 37–38
Log likelihood ratio tests, 37–38

M
Mann Whitney tables, 11
Mann Whitney test, 11–13
Margin of inferiority, 27
Matched groups, 7
McNemar odds ratios, 40
McNemar's test, 39–40

Means, 10, 41, 48
Multiple regression analysis, 48
Multiple testing, 1, 41
Multiple variable tests, 55

N
Noninferiority testing, 27–28
Non-parametric tests, 1, 9–13
Normal distribution, 29

O
Odds ratios, 15, 35–36
Odds ratio test for cross-tabs, 35–36
One-sample t-test, 5–6

P
Paired t-test, 6
Parallel groups, 27, 49
Parallel-group study, 17, 50
P-calculator for z-values, 36, 37
Pocket calculator method, 31
Pocket calculators, 1–3, 31, 34, 40, 48
Pooled SE. *See* Pooled standard error
Pooled standard deviation, 26
Pooled standard error, 17, 29, 48
Power, 19–21, 23–27
Power equations, 19–21
Power index, 23–27
Prior hypothesis, 55
Propensity scores, 48
Proportions, 31–33
Proportional duplicate standard deviation, 51, 52
P-values, 5–13, 21, 23, 24, 27–33, 36–44, 48

R
Rank numbers, 9–11
Regression modeling, 50
Reliability assessment, 51–54
Reproducibility, 51–54

S
Sample size, 1, 6, 8, 15, 20, 23–27, 43
Sample size and binary data, 25–26
Sample size and continuous data, 23–25
SAS statistical software, 55
SD. *See* Standard deviation (SD)
SE. *See* Standard error (SE)
SEM. *See* Standard error of the mean (SEM)
SEM-unit, 23, 28
Sensitivity of tests for cross-tabs, 37
SE-unit, 8, 20, 28
SPSS for Starters, 2
SPSS statistical software, v
Standard deviation (SD), 3–7, 15, 19, 23–27, 44, 51
Standard error (SE), 5–7, 15, 19, 20, 28–30, 35, 36, 40, 47, 48
Standard error of the mean (SEM), 17, 23, 24, 28, 29, 40
Subclassification, 47

T
T-distribution, 20, 21, 29
T-table, 6–8, 19–21, 29, 30, 36, 40, 48
T-tests, 5–8, 41–42
Two-sided p-values, 21
Type I error, 20, 41
Type II error, 20

U
Unpaired t-test, 7–8

V
Variability analysis, 1, 43–45
Variability test one sample, 43
Variability test two samples, 44–45

W
Weighted mean, 48
Weighted standard error, 48
Wilcoxon table, 10
Wilcoxon test, 9–10

Z
Z-distribution,
Z-test for cross-tabs, 29–30
Z-values, 29, 36, 37