

Statistics for Non-Statisticians

Birger Stjernholm Madsen

Statistics for Non-Statisticians

Second Edition

 Springer

Birger Stjernholm Madsen
Novozymes A/S
Bagsvaerd, Denmark

The Danish edition was published in 2012 as “Statistik for ikke-statistikere” by Samfundslitteratur, Frederiksberg, Denmark.

ISBN 978-3-662-49348-9 ISBN 978-3-662-49349-6 (eBook)
DOI 10.1007/978-3-662-49349-6

Library of Congress Control Number: 2016940499

© Springer-Verlag Berlin Heidelberg 2011, 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Berlin Heidelberg

Preface

Never have so many organizational decisions been taken based on statistics as today! Everything is supported by numbers. This applies to marketing, economics, social sciences, natural sciences, industry, and administrative work within organizations, businesses, and institutions. It is, therefore, important to have insight into basic statistical concepts, when assessing statistical data material, as well as when preparing an investigation, so that it produces useful statistical results.

There are several books on elementary statistics. So why write another? The simple answer is: because it is needed! This book fills a gap in the existing literature about statistics. Most existing short introductions to statistics take one of the following approaches:

- An approach based primarily on descriptive statistics (charts, tables, etc.)
- A purely verbal approach without any mathematical formula, but also without practical guidelines
- An approach based primarily on probability theory

In contrast, this book is intended to be a “first course for the practitioner,” giving a lot of useful details for, e.g., planning of surveys. Comparing this book to standard 500–600 pages textbooks on statistics, you will actually find a lot of practical information in this book that is not available in the standard textbooks!

I have for some decades taught statistics at all levels. It is my experience that the most important concepts of statistics can be explained, so that “ordinary” people can understand it. I have experienced this through hundreds of courses for many different audiences. Now, I have put my words on paper!

Who Is this Book Written for?

The book is written for those who need to know how to collect, analyze, and present data. You may be working with administrative data, financial data, or data from the social sciences or natural sciences. Maybe you plan to collect data through sample surveys, such as customer surveys or similar.

You do not know much about statistics. Maybe you have learned a little about the topic earlier, but forgotten most of it again. Maybe you never learned anything about the topic, but you are curious!

Although the book does not require knowledge of statistics, I assume that you are not totally unfamiliar with numbers! You are able to perform simple calculations with a calculator. And you don't panic, when you see a simple formula containing a square root! Don't worry: This book is not loaded with mathematical formulas. But it is unfortunately impossible to introduce statistical concepts without a minimum of mathematical calculations.

It is an advantage, if you have a basic understanding of spreadsheets. This book is not a course in the use of spreadsheets—the easiest way to learn spreadsheets is by reading a computer booklet or taking a course!

Neither is this a “How to do statistics with Excel” book—you can use the references in the literature list, if you need this. There are many books of this kind, often occupying hundreds of pages. . .

Yet it may be useful to know how the most important statistical calculations can be performed using the features of a spreadsheet. Spreadsheets have nowadays made numbers and graphs accessible to most people. This also applies to statistical calculations!

If you do not have access to a spreadsheet, I can recommend the spreadsheet Calc from Open Office (a free Office suite). See links to software in Appendix. Virtually everything discussed in this book can be done with this spreadsheet!

I recommend that, while reading the book, you work with some simple data, which you enter in a spreadsheet. It is easier to learn statistics if you work a little with the substance!

The beginner may be satisfied with a spreadsheet as a tool for statistical analysis. In professional work with statistics, however, you will very quickly discover the limitations of a spreadsheet. Then it is time to consider a better tool for the purpose! Therefore, the Appendix presents some of the major programs for statistics as well as links, where you can find more about them.

It is my hope that the book can be used for private study and as supplementary reading at business colleges, technical schools, high schools, and the initial training of statistics at business schools and social sciences at a university level. The book is not written for any specific education.

After reading this book, you should have the ability to dig further into some of the many other books on statistics available on the market. It is my hope that this book can ease the transition to the reading of the (many) more advanced books on

the subject. The number of books on statistics grows dramatically as the professional level rises!

The mathematically oriented reader has to accept that the book does not achieve 100% mathematical precision everywhere. Focus is on an understandable rather than mathematical precise presentation.

Some topics in the book are a bit more “technical” than the rest of the book. These issues can be skipped, without thereby losing coherence. Some of these items are placed in a text frame and entitled “Technical note: . . .” Some topics provide a clear indication that they may be skipped. They are often put at the end of a chapter.

There are also many examples of using spreadsheets. If you do not use spreadsheets, you can just read the examples, without bothering about how the results were obtained in the spreadsheet.

Structure of the Book

The book is structured in such a way that what you learn in one chapter is used in the following chapters. This means that you should read it from the beginning, at least up to and including Chap. 5.

Chapters 1 and 2 are about the collection and presentation of data. These are crucial issues for most people working with statistics.

Chapters 3–5 are the core of the book. They introduce the basic statistical concepts, including descriptive statistics, the normal distribution, and statistical tests.

When you have read Chap. 5, Chaps. 6–8 can be read independently of each other.

Chapter 6 supplements Chap. 1; it is about the planning of sample surveys and experiments.

Chapters 7 and 8 supplement Chap. 4 on the normal distribution. Chapter 8 is probably the “heaviest” material of the book and appropriately placed at the end!

The Appendices of the book contain a lot of hopefully useful information: review of probability theory, bibliography, glossary of statistical terms, list of statistical functions in spreadsheets, list of statistical software, useful links, and various useful tables. All words in this book, which are marked with an *asterisk* (*), are explained in the glossary.

At the publisher’s website, you will find additional material for the book: useful worksheets, further explanation, examples, etc. Of course, there is also a spreadsheet with the example dataset “Fitness Club,” which is used as recurrent example.

I wish you a pleasant reading!

Preface to the 2nd edition

During the years since the first edition of this book, I have received many comments. Some pointed out that a few useful sections could be added in order to make the book available to a wider audience, including people working in industry. The main enhancements are as follows:

1. New sections in Chap. 4 about the lognormal distribution, control charts and process capability have been added.
2. A new section in Chap. 8 about ANOVA (with 1 factor) has been added.

The appendices about references, links and software have also been updated.

It is my hope that these changes and additions have improved the book substantially. I urge the readers to further comment on the book!

Copenhagen, Denmark

Birger Stjernholm Madsen

Acknowledgements

I send my warm thanks to my editors at Springer, Barbara Fess and Johannes Gläser, who came with a lot of good suggestions.

I also want to thank my editor on the Danish editions, Peter Byriel from Samfundslitteratur. He gave a lot of constructive criticism along the way.

In addition, I want to thank the following Danish statisticians for valuable comments: Leif Albert Jørgensen, Niels Landvad and Anders Milhøj.

Finally, I thank my wife, Yrsa, for being extremely patient with me in the busy periods!

Contents

1	Data Collection	1
1.1	Sample Surveys	2
1.2	Fitness Club: Example of a Sample Survey	4
1.3	Experiments	4
1.4	Experiments: An Example	5
1.5	Data Collection	6
1.6	Registers	6
1.7	Questionnaire Surveys	7
	1.7.1 Background Questions	7
	1.7.2 Study Questions	7
1.8	Sources of Errors in Surveys	10
1.9	Comparing Methods of Data Collection	11
1.10	Example Continued	13
2	Presentation of Data	15
2.1	Bar Charts	15
2.2	Histograms	17
2.3	Pie Charts	19
2.4	Scatter Plots	20
2.5	Line Charts	21
2.6	Bubble Plots	22
2.7	Tables	22
	2.7.1 The Ingredients of a Table	23
	2.7.2 Percentages	24
3	Description of Data	27
3.1	Systematic and Random Variation	27
3.2	Measures of Location	29
	3.2.1 Average	29
	3.2.2 Median	31

3.2.3	Mode	32
3.2.4	Choosing a Measure of Location	33
3.3	Measures of Dispersion	35
3.3.1	Range	35
3.3.2	Variance and Standard Deviation	36
3.3.3	Interquartile Range	38
3.3.4	Choosing a Measure of Dispersion	40
3.3.5	Relative Spread (Dispersion)	40
3.4	Example: Statistical Functions in Spreadsheets	41
3.5	Data Type and Descriptive Statistics	43
3.5.1	Data Types	44
3.5.2	Descriptive Statistics and Type of Data	44
4	The Normal Distribution	47
4.1	Characteristics of the Normal Distribution	47
4.2	Density Function and Distribution Function	49
4.3	Fractiles	50
4.4	Calculations in the Normal Distribution	51
4.5	The Normal Distribution and Spreadsheets	52
4.5.1	NORMDIST (X; Mean; Stdev; Cumulative)	53
4.5.2	NORMINV (Probability; Mean; Stdev)	53
4.5.3	Example	53
4.6	Testing for the Normal Distribution	54
4.6.1	Simple Methods	55
4.6.2	Skewness and Kurtosis	56
4.6.3	Normal Plot	59
4.7	Random Numbers	60
4.8	Confidence Intervals	62
4.8.1	Confidence Interval for the Mean	63
4.8.2	Confidence Interval for the Mean in Case of a Small Sample	66
4.8.3	Confidence Interval for the Standard Deviation	70
4.9	More About the Normal Distribution	72
4.10	Lognormal Distribution	74
4.10.1	Introduction	74
4.10.2	Example	75
4.11	Control Charts: Everything in Control?	76
4.11.1	Introduction	76
4.11.2	Statistical Control	77
4.11.3	Construction of a Simple Control Chart	78
4.11.4	Unusual Patterns	78
4.11.5	Practical Use of Control Charts	79
4.11.6	Example	79
4.12	Process Capability	80
4.12.1	Introduction	81
4.12.2	Example	81

4.12.3	Process Capability Indices	81
4.12.4	Process Capability Indices: Important!	83
4.12.5	Example, continued	84
5	Analysis of Qualitative Data	85
5.1	The Binomial Distribution	85
5.1.1	Example	86
5.2	The Binomial Distribution and the Normal Distribution	87
5.3	The Binomial Distribution in Spreadsheets	89
5.3.1	Example	89
5.4	Statistical Uncertainty in Sample Surveys	90
5.4.1	Example	91
5.5	Is the Sample Representative?	95
5.6	Statistical Tests	96
5.6.1	Example	97
5.6.2	Approximation with the Normal Distribution	98
5.6.3	Significance Level	99
5.6.4	Statistical Test or Confidence Interval	99
5.7	Frequency Tables	100
5.7.1	Introduction to Chi-Squared Test	100
5.7.2	Confidence Interval for Difference Between Two Proportions	103
5.7.3	Several Rows and/or Columns	104
5.7.4	Calculations in Spreadsheets	107
5.7.5	Calculations by Calculator	108
6	Error Sources and Planning	111
6.1	Two Kinds of Errors	111
6.2	Random Error and Sample Size	111
6.2.1	A Qualitative Variable	113
6.2.2	A Quantitative Variable	115
6.3	Bias (Systematic Errors)	116
6.3.1	Errors in the Sampling (Sample Selection)	117
6.3.2	Errors in the Definition of the Sample	117
6.3.3	What Is a Representative Sample?	118
6.4	Sampling (Sample Selection)	119
6.4.1	Simple Random Sampling	119
6.4.2	Stratified Sampling	120
6.4.3	Cluster Sampling	121
6.4.4	Systematic Sampling	123
6.4.5	Quota Sampling	123
6.4.6	Purposive Sampling	124
6.4.7	Convenience Sampling	125

- 7 Assessment of Relationship 127**
 - 7.1 Example 128
 - 7.2 Linear Regression with Spreadsheets 131
 - 7.3 Is There a Relationship? 134
 - 7.3.1 Note 135
 - 7.4 Multiple Linear Regression 136
 - 7.5 Final Remarks 137
- 8 Comparing Two Groups 139**
 - 8.1 Matched Pairs: The Paired t-Test 139
 - 8.1.1 Example 139
 - 8.1.2 Description 140
 - 8.1.3 Calculation 141
 - 8.1.4 Spreadsheets 143
 - 8.2 Comparing Two Groups Means 144
 - 8.2.1 Example 144
 - 8.2.2 Description 144
 - 8.2.3 Calculation 145
 - 8.2.4 Spreadsheets 147
 - 8.2.5 Size of an Experiment 148
 - 8.3 Other Statistical Tests for Two Groups 148
 - 8.3.1 Test for the Same Variance in the Two Groups 148
 - 8.3.2 Comparing Two Group Means: Two Samples with Equal Variances 149
 - 8.4 Analysis of Variance, ANOVA 150
 - 8.4.1 Introduction 150
 - 8.4.2 Example 150
 - 8.5 Final Remarks 151
- 9 Appendices 153**
 - 9.1 Probability Theory 153
 - 9.1.1 Sample Space, Events, and Probability 154
 - 9.1.2 Random Variables; the Binomial Distribution 159
 - 9.1.3 Random Variables: Mean and Variance 161
 - 9.1.4 Technical Note: The Binomial Coefficient 163
 - 9.2 Summary of Statistical Methods 164
 - 9.2.1 Quantitative Data 164
 - 9.2.2 Qualitative Data 166
 - 9.3 Statistical Functions in Spreadsheets 168
 - 9.4 Statistical Tables 169
 - 9.4.1 Fractiles in the Normal Distribution 169
 - 9.4.2 Probabilities in the Normal Distribution 169
 - 9.4.3 Table of the t-Distribution 169

- 9.4.4 Table of the Chi-Squared Distribution 170
- 9.4.5 Statistical Uncertainty in Sample Surveys 172
- 9.5 Fitness Club: Data from the Sample Survey 174
- 9.6 Where to Go from Here 175
 - 9.6.1 Literature 175
 - 9.6.2 Useful Links 176
 - 9.6.3 Overview of Statistical Software 177
- 9.7 Glossary 178

- Index 183**

Abbreviations

\bar{x}	Sample average
<i>ANOVA</i>	Analysis of Variance
$B(n,p)$	Binomial distribution (n observations, probability p)
<i>C_p</i>	Process Capability Index
<i>C_{pk}</i>	Minimum Process Capability Index
<i>CV</i>	Coefficient of Variation
<i>DF</i>	Degrees of Freedom
<i>DOE</i>	Design of Experiments
$E(X)$	Mean of X . E = "Expectation"
H_0	Null hypothesis
H_1	Alternative hypothesis
<i>LCL</i>	Lower Control Limit
<i>LSL</i>	Lower Specification Limit
$N(0,1)$	Standardized normal distribution (mean 0, variance 1)
$N(\mu,\sigma^2)$	Normal distribution with mean μ and variance σ^2
<i>R</i>	Sample range
<i>s</i>	Sample standard deviation
s^2	Sample variance
<i>SPC</i>	Statistical Process Control
<i>UCL</i>	Upper Control Limit
<i>USL</i>	Upper Specification Limit
$V(X)$	Variance of X
μ	Population mean
σ	Population standard deviation
σ^2	Population variance
Σ	Sum
χ^2	Chi-squared (distribution or test)

About the Author

Birger Stjernholm Madsen has a Master of Science in Statistics and Mathematics.

He has several years of experience as a statistician at major Danish companies within several industries as well as within national statistics.

He has also taught statistics for several years at University of Copenhagen and has held hundreds of statistics courses for various audiences throughout decades. This book is a direct result of his teaching.