Society, Environment and Statistics

Claus Weihs
Walter Krämer
Sarah Buschfeld *Editors*

Statistics Today

Everyday Applications, Research Questions, Insights, and Challenges

Springer

Society, Environment and Statistics

This interdisciplinary book series demonstrates the significance of statistical science in quantifying and understanding societal and environmental changes, while effectively managing resources. The series invites manuscripts that cover a wide range of themes, including original research in traditional fields, emerging topics, practical applications, and connections between novel findings and established works. Particularly encouraged are submissions that showcase methodologies, case studies, good practices and innovative solutions addressing the complex challenges of the United Nations' Sustainable Development Goals (SDGs). Rigorously peer-reviewed, all contributions adhere to the highest standards of scientific literature.

Claus Weihs • Walter Krämer • Sarah Buschfeld
Editors

Statistics Today

Everyday Applications, Research Questions, Insights, and Challenges

Springer

Editors
Claus Weihs
Department of Statistics
TU Dortmund University
Dortmund, Germany

Walter Krämer
Department of Statistics
TU Dortmund University
Dortmund, Germany

Sarah Buschfeld
Department of Cultural Studies
TU Dortmund University
Dortmund, Germany

ISSN 2948-2763 ISSN 2948-2771 (electronic)
Society, Environment and Statistics
ISBN 978-3-662-68906-6 ISBN 978-3-662-68907-3 (eBook)
https://doi.org/10.1007/978-3-662-68907-3

Translation from the German language edition: "Faszination Statistik - Einblicke in aktuelle Forschungsfragen und Erkenntnisse" by Claus Weihs et al., © Springer-Verlag GmbH Deutschland, ein Teil von Springer Nature 2019. Published by Springer Spektrum. All Rights Reserved.

This Springer imprint is published by the registered company Springer-Verlag GmbH, DE, part of Springer Nature.
The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

To Heidrun, Max, Doris, and Leo

Preface

> "I keep saying the sexy job in the next ten years will be statisticians . . . The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it is going to be a hugely important skill in the next decades."
>
> (Hal Varian, chief economist at Google, in 2009;[1] more than a decade later, we can see that Varian was right.)

Academic disciplines have their ups and downs. Some, such as medicine for obvious reasons, have experienced a stable interest. Ever since Plato and Aristotle, philosophy has enjoyed the attention of contemporaries of every century, in particular since it has never been well understood. Other disciplines have moved into the focus of general attention only sporadically. The early twentieth century was the high time of chemistry, technology, and physics: artificial fertilizers, airplanes, dyestuffs, electrification, rockets, and nuclear fission were hotly debated by experts and non-experts alike; Konrad Röntgen, the inventor of X-ray technology, was the first scientific pop star ever. Then, in the wake of Siegmund Freud, psychology was the number one party topic in certain social circles for a while; a person who could not tell about his/her last visit to a psychoanalyst was quasi socially declassed. And at the end of the 1960s, sociology was, for a short time, regarded as the key to decode the future of humanity.

Currently, statistics is one of the most popular sciences. Spurred by an enormous revolution in both data processing and data acquisition, new opportunities and challenges arise almost every day. Even the most superficial newspaper reader cannot avoid realizing that data science in general and statistics in particular are central aspects of current science and our everyday lives. Today, everyone talks about 'Big Data' and corresponding degree courses are sprouting like mushrooms. However, it has been often overlooked that the following topics are anything but new: correct sampling and handling of data, decision making under uncertainty, differentiating between chance and deterministic patterns, and extrapolating from samples to the

[1] See https://flowingdata.com/2009/02/25/googles-chief-economist-hal-varian-on-statistics-and-data/, visited 11.9.2019.

general. For decades, thousands of researchers in statistics departments, in business, and industry all over the world have been active in these fields. The current major increase in excitement and activity originates in the vast amount of data that require statistical analysis.

The data sources, however, have always existed. Ambient temperature on the northern side of the Matterhorn on Christmas Eve has been ready to be observed ever since the Matterhorn has existed. And that Mrs. X bought two bottles of Chateau Baron on January 10, 1998 at a certain Aldi-shop in Berlin was always a fact. This information was, however, limited to Mrs. X and the person working the cash register. Today, if Mrs. X payed by credit card, the whole world could know. The current 'Big Data' hype is thus the result of the recent availability of the data, for which the sources always existed, but which were simply never exploited. For most of human history, people knew comparatively little. This has changed due to the interaction of increasingly efficient computer and storage technologies and advanced methods and technologies for data exploitation. Therefore, large amounts of data are nowadays available to more and more people.

This makes it all the more important to handle and process these data carefully. For this purpose, statistics is just as indispensable today as it has always been. Of course, the current explosion in the quantity of data also brings along previously unknown problems regarding data management. In this respect, statisticians rely on the support of IT-experts. When it comes to data analysis, however, the basic principles of estimating and testing, model selection, or sampling apply to big data sets (e.g. supplied by Google) just as they do to smaller data sets (e.g. from the Federal Statistical Office).

This volume presents selected studies from the Statistics department at the TU Dortmund University, the only independent statistics department in Germany, as well as from former DFG (German Research Foundation) Collaborative Research Centers associated with the department. In addition, a number of researchers from outside the TU have contributed valuable research shedding light on how statistics can help us unveil and understand phenomena of our daily lives.

This book shows, for example, how statistics can help to assess the effect of drugs; how statistics can be utilized to analyze musical audio data to automate transcriptions; how flood catastrophes or risks on the stock market can be statistically modeled and thus better managed; and how meaningful quantitative data can be extracted from qualitative information such as texts and spoken words. Other studies in this volume deal with quality control in industry, the forecasting abilities of rating agencies, or the prevention of false alarms in intensive care. Additionally, the volume elucidates why the lottery is not purely a game of chance and why favorites are systematically overrated in horse betting.

We are reporting cutting edge research accessible to non-statisticians and non-experts in the related fields. The editors asked their participating colleagues to avoid using specialist language as much as possible. The volume is geared towards an audience that wants to gain insights into general ideas about statistical applications but does not strive to understand all the subtleties of formal statistical analysis. We hope that this book will therefore also appeal to those whose enthusiasm for

statistics has somewhat suffered in their degree courses in business administration, economics, psychology, sociology, or other subjects where statistics certificates are traditionally required. Unfortunately, in academic education the intrinsic beauty of statistics is often obscured by too much formalism. This anthology will hopefully demystify the specter of statistics and invite you to an unobstructed view of an utterly fascinating discipline.

Preparation of the English Version

In most parts, this book is a translation of the Springer book "Faszination Statistik", which was originally published in German. Only Chap. 25 ("Statistical modeling of current linguistic realities around the world: The case of Singapore") and Chap. 26 ("Linguistic manifestations of cultural differences across national varieties of English – a methodological survey") were explicitly written for the English version. A first translation of the chapters in "Faszination Statistik" was produced by the corresponding authors, partly supported by the algorithms DeepL (https://www.deepl.com/translator) or Google Translate (https://translate.google.com/). Those translations were reviewed concerning linguistic expression and clarity of content by Sarah Buschfeld, one of the book editors. She also suggested modifications and additions to the corresponding authors whenever necessary. Claus Weihs, first editor of this volume, discussed the suggested changes with the corresponding authors, checked the comprehensibility of statistical contents, and partly reformatted the chapters to unify their appearance. Finally, the texts were approved by the corresponding authors.

Acknowledgment

The editors would like to thank Brian Hess for his support in proofreading some of the chapters.

Dortmund, Germany
September 2023

Claus Weihs
Walter Krämer
Sarah Buschfeld

Contents

Contributors

Nadja Bauer [Chapter 18] FH Dortmund, Faculty of Computer Science, Dortmund, Germany nadja.bauer@fh-dortmund.de

Eva-Christina Becker-Emden [Chapter 20] FH Dortmund, Faculty of Computer Science, Dortmund, Germany eva-christina.becker-emden@fh-dortmund.de

Alexander Brinkhoff [Chapter 20] TU Dortmund, Dortmund, Germany alexander.brinkhoff@tu-dortmund.de

Gesa Brunn [Chapter 21] Deutscher Wetterdienst, Research and Development, Offenbach, Germany gesa-marie.brunn@dwd.de

Sarah Buschfeld [editor, Chapter 25] TU Dortmund, Department of Cultural Studies, Dortmund, Germany sarah.buschfeld@tu-dortmund.de

Tileman Conring [Chapter 13] TU Dortmund, Department of Statistics, Dortmund, Germany tileman.conring@tu-dortmund.de

Holger Dette [Chapter 3] RU Bochum, Stochastics, Bochum, Germany holger.dette@ruhr-uni-bochum.de

Philipp Doebler [Chapter 21] TU Dortmund, Department of Statistics, Dortmund, Germany doebler@statistik.tu-dortmund.de

Jan Feifel [Chapter 7] Ulm, Germany jan.feifel@alumni.uni-ulm.de

Svenja Fischer [Chapter 17] RU Bochum, Engineering Hydrology and Water Resources Management, Bochum, Germany svenja.fischer@ruhr-uni-bochum.de

Fritjof Freise [Chapter 21] University of Veterinary Medicine Hannover, Department for Biometry, Epidemiology and Information Processing, Hannover, Germany fritjof.freise@tiho-hannover.de

Roland Fried [Chapters 4, 17] TU Dortmund, Department of Statistics, Dortmund, Germany msnat@statistik.tu-dortmund.de

Ursula Gather [Chapter 4] TU Dortmund, Department of Statistics, Dortmund, Germany ursula.gather@tu-dortmund.de

Leo N. Geppert [Chapter 9] TU Dortmund, Department of Statistics, Dortmund, Germany geppert@statistik.tu-dortmund.de

Peter Gnändinger [Chapter 9] TU Dortmund, Department of Statistics, Dortmund, Germany peter.gnaendinger@tu-dortmund.de

Klaus Golka [Chapter 6] IfADo, Clinical Occupational Medicine, Dortmund, Germany golka@ifado.de

Andreas Groll [Chapter 8] TU Dortmund, Department of Statistics, Dortmund, Germany groll@statistik.tu-dortmund.de

Katja Ickstadt [Chapters 6, 9, 27] TU Dortmund, Department of Statistics, Dortmund, Germany ickstadt@statistik.tu-dortmund.de

Michael Imhoff [Chapter 4] RU Bochum, Faculty of Medicine, Bochum, Germany mike@imhoff.de

Carsten Jentsch [Chapter 24] TU Dortmund, Department of Statistics, Dortmund, Germany jentsch@statistik.tu-dortmund.de

Fabian Knorre [Chapter 16] TU Dortmund, Department of Statistics, Dortmund, Germany knorre@statistik.tu-dortmund.de

Christina Kopetzky [Chapter 16] University of Klagenfurt, Department of Economics, Klagenfurt, Austria christina.kopetzki@aau.at

Walter Krämer [editor, Chapters 1, 12, 13, 15] TU Dortmund, Department of Statistics, Dortmund, Germany walterk@statistik.tu-dortmund.de

Sonja Kuhnt [Chapter 20] FH Dortmund, Faculty of Computer Science, Dortmund, Germany sonja.kuhnt@fh-dortmund.de

Martin Kukuk [Chapter 11] University of Würzburg, Department of Economics, Würzburg, Germany martin.kukuk@uni-wuerzburg.de

Robert Löser [Chapter 14] TU Dortmund, Dortmund, Germany robert.loeser@tu-dortmund.de

Christine H. Müller [Chapter 19] TU Dortmund, Department of Statistics, Dortmund, Germany cmueller@statistik.tu-dortmund.de

Simon Neumärker [Chapter 15] TU Dortmund, Department of Statistics, Dortmund, Germany simon.neumaerker@tu-dortmund.de

Markus Pauly [Chapter 7] TU Dortmund, Department of Statistics, Dortmund, Germany markus.pauly@tu-dortmund.de

Andreas Quatember [Chapters 22, 23] JKU Linz, Institute of Applied Statistics, Linz, Austria andreas.quatember@jku.at

Jörg Rahnenführer [Chapters 5, 24] TU Dortmund, Department of Statistics, Dortmund, Germany rahnenfuehrer@statistik.tu-dortmund.de

Gunther Schauberger [Chapter 8] TU Munich, Epidemiology, Munich, Germany gunther.schauberger@tum.de

Edgar W. Schneider [Chapter 26] University of Regensburg, English Linguistics, Regensburg, Germany edgar.schneider@sprachlit.uni-regensburg.de

Kirsten Schorning [Chapter 3] TU Dortmund, Department of Statistics, Dortmund, Germany schorning@statistik.tu-dortmund.de

Katharina Schüller [Chapter 1] Stat-Up Statistical Consulting & Data Science GmbH, Munich, Germany katharina.schueller@stat-up.com

Andreas Schumann [Chapter 17] RU Bochum, Engineering Hydrology and Water Resources Management, Bochum, Germany andreas.schumann@hydrology.ruhr-uni-bochum.de

Silvia Selinski [Chapter 6] IfADo, Toxicology/Systemtoxicology, Dortmund, Germany selinski@ifado.de

Wolfgang Tillmann [Chapter 20] TU Dortmund, Department of Mechanical Engineering, Dortmund, Germany wolfgang.tillmann@udo.edu

Martin Wagner [Chapter 16] University of Klagenfurt, Department of Economics, Klagenfurt, Austria

Bank of Slovenia, Ljubljana, Slovenia martin.wagner@aau.at

Claus Weihs [editor, Chapters 2, 10, 18, 25, 27] TU Dortmund, Department of Statistics, Dortmund, Germany claus.weihs@tu-dortmund.de

Dominik Wied [Chapter 14] University of Cologne, Institute for Statistics and Econometrics, Köln, Germany dwied@uni-koeln.de

Part I
Human Life, Medicine, and Genetics

Chapter 1
Season of Birth and Human Longevity: A New Theory Why Children Born in November Live Longer

Walter Krämer and Katharina Schüller

Abstract In Europe, human individuals born in November live on average half a year longer than individuals born in May. The ultimate cause must involve the climate, because in the southern hemisphere, this pattern is reversed. However, it is still unclear how exactly environmental factors related to the season of birth affect longevity. We offer an explanation.

1.1 The Date of Birth Matters

Since ancient times, astrologers have tried to deduce the further fate of a person from the date of birth. The fact that such influences exist helps them a lot. However, these influences have nothing to do with planets or with stars. It has for instance been known for decades that individuals born in February or March suffer an increased risk of schizophrenia. Or consider professional sports. Psychologist Peter Jensen once recorded the month of birth of Canadian national hockey players. They were all born in January, February, March, or April. None of them was born in May, June, July, August, September, October, November, or December. In Australia, it has likewise been found that in almost all professional sports there are more successful athletes born in January than born in December. However, this does not result from the stars but from the fact that, in cohorts grouped by year of birth as is usual in many sports, children born in January are the oldest and for this reason often the best in their respective training groups. Therefore, they also receive a disproportionate amount of attention and support. If this happens many years in a row until the option

W. Krämer (✉)
TU Dortmund, Department of Statistics, Dortmund, Germany
e-mail: walterk@statistik.tu-dortmund.de

K. Schüller
Stat-Up Statistical Consulting & Data Science GmbH, Munich, Germany
e-mail: katharina.schueller@stat-up.com

C. Weihs et al. (eds.), *Statistics Today*, Society, Environment and Statistics,
https://doi.org/10.1007/978-3-662-68907-3_1

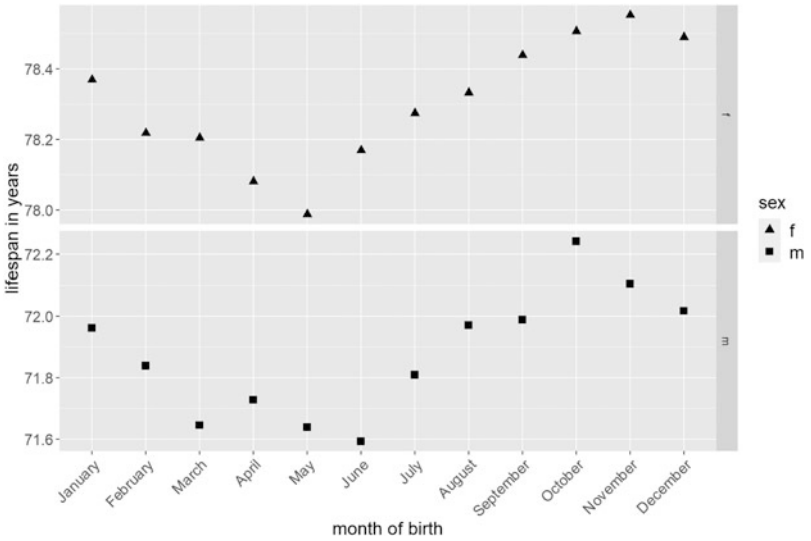

Fig. 1.1 Average length of life of all Swiss individuals who died between 1969 and 2010. Lower panel males, upper panel females

of a professional career in sports is entertained, it is no wonder that children born in January are privileged when entering the professional career.

Here, the date of birth affects success in the job. But the date of birth also affects longevity. This has been known for decades and has been documented for many countries. Figure 1.1 illustrates this for Switzerland. It shows the average age at death of all Swiss individuals who died between 1969 and 2010, grouped by month of birth. As can be seen, for men and women alike, individuals born in October/November live on average six months longer than those born in May.

Figure 1.1 does not provide a realistic picture of life expectancy in Switzerland. It is an underestimation, because the final years of our data set include a disproportionate amount of individuals who died rather young. The seasonal effect in longevity however is obvious. And it is not produced by chance. Similar results have been shown for Denmark, Sweden, or Germany, albeit based on smaller data sets. On the southern hemisphere, the opposite occurs. Here, individuals born in the fall suffer a loss in life expectancy. Therefore, it is obvious that temperature or sun exposure during conception or in-utero or early postnatal periods must somehow be involved. For example, longevity of Australians who have immigrated from Europe follows the pattern of the northern hemisphere. However, it is still not clear how exactly season of birth connects to length of life. Here we offer an explanation.

1.2 A New Look from Switzerland

Our analysis is based on a unique data set of all recorded cases of death in Switzerland from 1969 to 2010, including date, place and cause of death, kindly provided by the Swiss Federal Statistical Office. The data also include date and place of birth plus information about spouse, parents, religion, language region, profession, and economic status. Figure 1.2 shows the average length of life of individuals who have died of cancer from 1969 to 2010, and Fig. 1.3 shows the average length of life of individuals who, in the same time span, have died from cardiovascular diseases. For both causes of death, and for men and women alike, the seasonal pattern of Fig. 1.1 repeats itself.

Figure 1.4, on the other hand, shows the average length of life of individuals who have died in an accident. For obvious reasons, their lives were shorter, and, most importantly, the seasonal pattern disappears. This is what one would expect, as seasonality in longevity must somehow be encoded in the human body and will therefore be annihilated by any independent random cause of death. It might also be of interest that the average age at death of male victims of accidents is about 15 years below that of females.

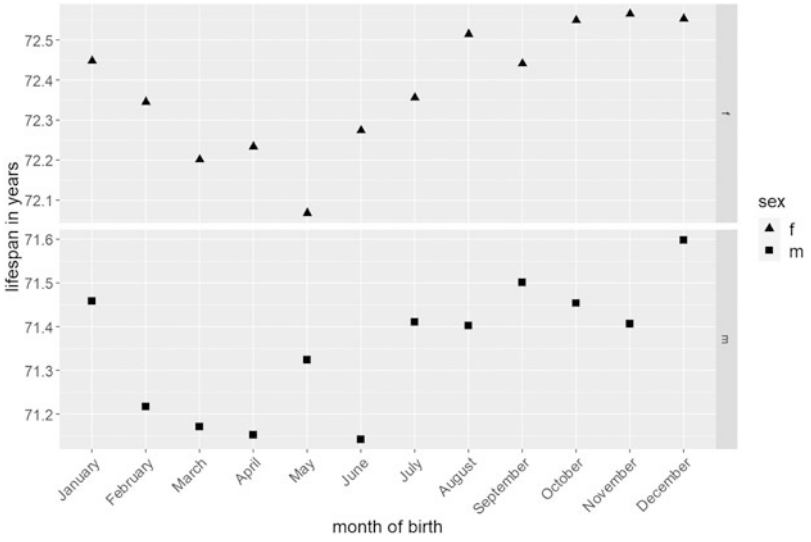

Fig. 1.2 Average length of life of all Swiss individuals who died from cancer between 1969 and 2010. Upper panel females, lower panel males

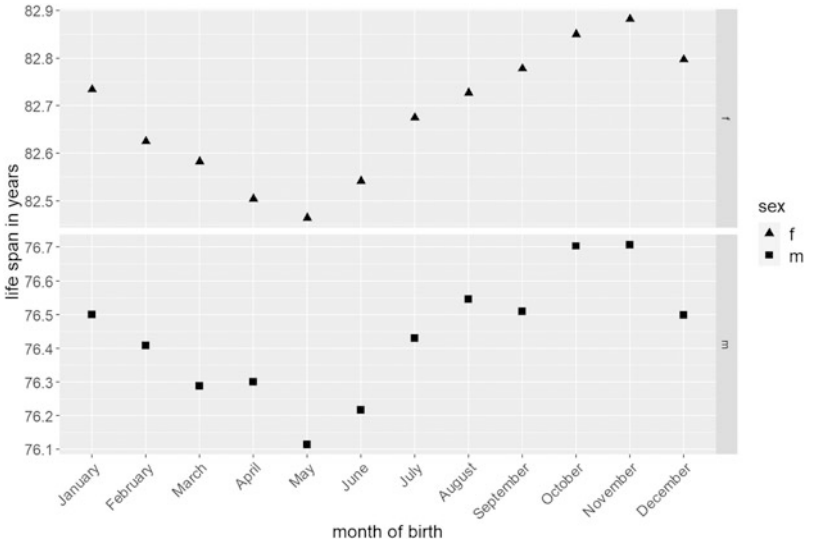

Fig. 1.3 Average length of life of all Swiss individuals who died from cardiovascular diseases between 1969 and 2010. Upper panel females, lower panel males

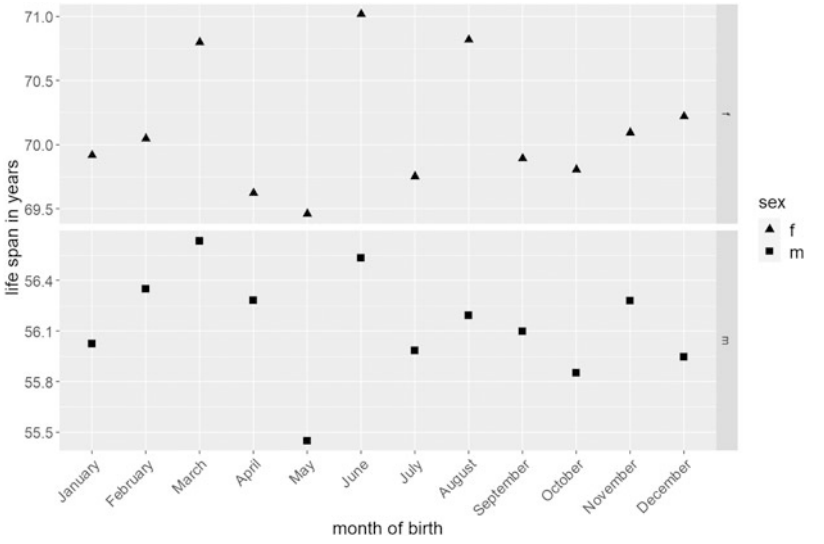

Fig. 1.4 Average length of life of all Swiss individuals who died from accidents between 1969 and 2010. Upper panel females, lower panel males