Statistical Inference in Linguistics

Roland Schäfer

Dedicated to a randomly chosen individual.

Contents

Preface

What readers can expect from this book

How did I get here?

New types of sources and references

Books and published papers are undoubtedly the primary source of knowledge about statistics and philosphy, even after the year 2015. However, I have acquired a great deal of my knowledge of statistics not just by reading books and research papers, but from respected online sources. I benefitted a lot from blogs written by renowned statisticians, from the Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/), and from *CrossValidated* (https://stats.stackexchange.com/). Especially the blogs and contributions of the following people have been very inspiring and helpful, regardless of how often I reference them in this book (in alphabetical order):

- BenBolker (https://stats.stackexchange.com/users/2126/ben-bolker)
- Andrew Gelman (http://andrewgelman.com/)
- Debora G. Mayo (https://errorstatistics.com/)
- Richard Morey (http://bayesfactor.blogspot.de/)
- Stephen Senn (https://errorstatistics.com/tag/stephen-senn/)
- as well as many contributors on R-bloggers (https://www.r-bloggers.com/)

A word on typesetting and technology

This book was created using XeLaTeX, which is in many ways superior to the old LaTeX. Mixing XeLaTeXand R code was made easy and elegant by knitr (https://yihui.name/knitr/), and I recommend anyone to use these three pieces of software to typeset their documents whenever they use statistics. Since I run an RStudio Server installation at https://webcorpora.org, I was able to work on this book in a perfectly platform- and machine-independent way. GitHub (https:

//github.com/rsling/smil) provided version control, and the available sources on GitHub can be used by anyone to see how it is done.

Acknowledgments

I grateful to the participants of many university courses I taught at Freie Universität Berlin and elsewhere (including Göttingen University and Siegen University). Among the students in my statistics classes, Elizabeth Pankratz, Johanna Rehak, and Samuel Reichert stood out as providing extremely valuable feedback through well-informed, inquisitive, and critical questions.

Also, I would like to thank Felix Bildhauer for an ongoing discussion about statistical methods and their application in linguistics. This includes the "Dahlem statistics reading circle" and a shared repository of books and papers about statistics.

Most prominently, I would like to thank Deborah G. Mayo for her work on statistical inference, so far culminating in Mayo (2018). Thanks to Mayo's works, I might at least be on the right track now with regard to statistical thinking. The suspicion lingers, however, that primarily due to my limited understanding and secondarily due to my exposure to the rich literature that propagates mindless statistics and/or asinine criticism of statistics, I got a lot of key points and details wrong.

Thus, while a lot of the good stuff that might be found in this book comes from other people, I am the sole source of all errors and misconceptions it contains. I will make a sincere effort to fix any problems that readers and critics point out in future editions.

Abbreviations and symbols

Abbreviations

ANOVA	analysis of variance
CDF	cumulative distribution function
CLT	central limit theorem
cp.	ceteris paribus (all other things being equal)
iid.	independent and identically distributed
LM	linear model
LMM	linear mixed model
GLM	linear mixed model
GLMM	generalised linear mixed model
PDF	probability density function
VCOV	variance-covariance matrix

Symbols

Symbols are overloaded ad-hoc to denote either a (possibly indexed) value such as $s_x = 1$ (for "the population mean of variable x is 1") or a function such as $s(x) = 1$ where applicable.

	Mathematical symbols
$x \sim D$	x *follows* D (x a variable, D a distribution)
$\bar{x}$	sample arithmetic mean of x
$\tilde{x}$	sample median of x
$\hat{x}$	predicted value of x

	Letter-like symbols
α	alpha level
α_i	intercept i
β	beta level
β_i	first-level coefficient i

df	degrees of freedom
e	Euler constant
ϵ	observation-level error
f	frequency
F	F statistic (see ANOVA)
γ_i	second-level coefficient i
H	Kruskal-Wallis statistic
H_0	null hypothesis
H_A	alternative hypothesis
M_M	main hypothesis
IQR	inter-quartile range
L	Likelihood
μ	population mean
μ_i	mean of modeled effect i
n	sample size
N	population size
O	Odds
p	proportion
P_i	the i-th percentile
Pr	probability
φ	dispersion parameter
Q_i	i-th quartile
r	sample covariance coefficient
r^2	coefficient of determination
R^2	multifactorial coefficient of determination
ρ	population covariance coefficient
s	sample standard deviation of x
s^2	sample variance of x
SE	standard error
SS	sum of squares
σ	population standard deviation
σ^2	population variance
U	Mann-Whitney statistic
χ^2	chi square statistic

Random distributions are denoted by bold-printed abbreviated names instead of the fancy symbols sometimes used.

Bern	Bernoulli distribution
Exp	exponential distribution
F	F distribution
Norm	normal (Gaussian) distribution
t	t distribution
Unif	uniform distribution
Chisq	χ^2 distribution

1 Science, data, and statistics

1.1 Some typestting tests

Mayo (2018)

Colorbox

Blabla

Iconcolorbox

Blabla

Sandwich

Blabla

Iconsandwich

Blabla

Framebox

Blabla

Iconframebox

Blabla

In this book, code listing are displayed as inline blocks such as the following simple code which simulates t-tests under the null pothesis in order to demonstrate that all p-values have equal probability under the null.

```
# Set simulation parameters.
nsim  <- 1000
n     <- 100
meen  <- 0
stdev <- 1

# Data structure for results.
sims <- rep(NA, nsim)
```

```
# Simulations.
for (i in 1:nsim) {
  a <- rnorm(n, mean = meen, sd = stdev)
  b <- rnorm(n, mean = meen, sd = stdev)
  p <- t.test(a,b)$p.value
  sims[i] <- p
}
```

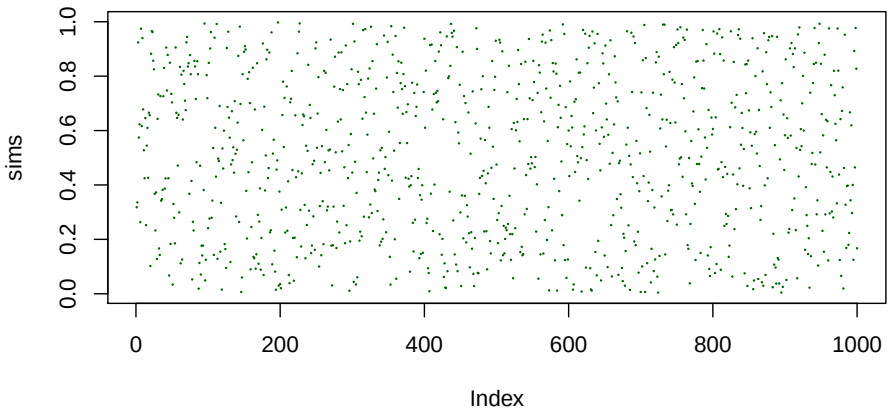


Figure 1.1: Scatterplot of p-values.

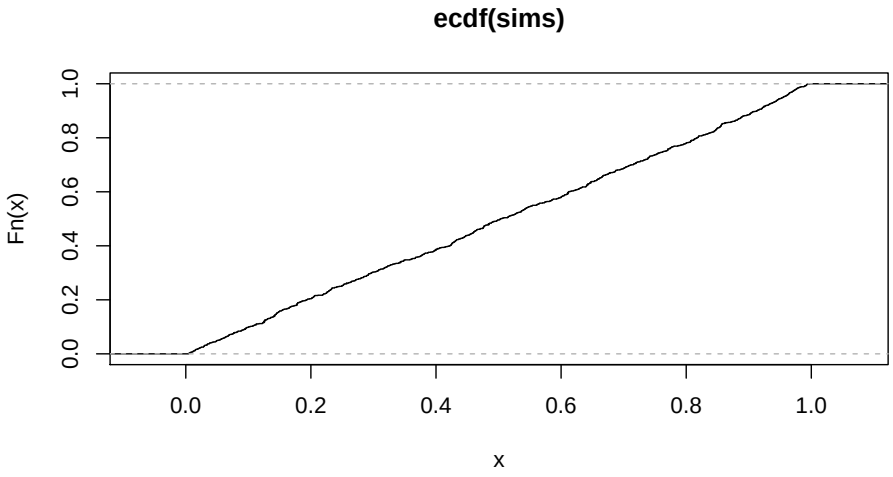
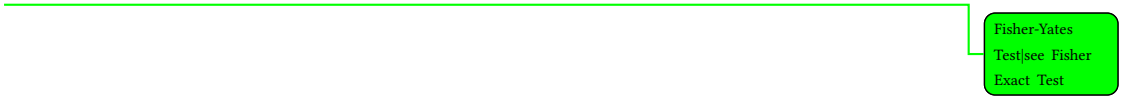


Figure 1.2: Empirical cumulative density distribution of p-values.

Figure 1.3: Histogram of p-values.

See Figure 1.2 for the cumulative density of p-values under the null in a series of 1000 t-tests. This was plotted using the following command:

```
plot(ecdf(sims))
```



References

Mayo, Deborah G. 2018. *Statistical inference as severe testing: how to get beyond the statistics wars*. Cambridge: Cambridge University Press.

Name index

Subject index