LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Reproducible Statistical Analyzes Today

Torsten Hothorn
Ludwig-Maximilians-Universität München

joint work with Fritz Leisch

Case-studies in Reproducibility

THE TIMES
THE SUNDAY TIMES

Archive Article | Please enjoy this article from The Times & The Sunday Times archives.

From The Sunday Times

January 18, 2009

Wealthy men give women more orgasms

Jonathan Leake, Science and Environment Editor

Scientists have found that the pleasure women get from making love is directly linked to the size of their partner's bank balance.

They found that the wealthier a man is, the more frequently his partner has orgasms.

"Women's orgasm frequency increases with the income of their partner," said Dr Thomas Pollet, the Newcastle University psychologist behind the research.

Pollet & Nettle (2009)

Thomas Pollet and Daniel Nettle (2009, Evolution and Human Behavior) report that "partner wealth predicts self-reported orgasm frequency in a sample of Chinese women".

The study is based on the Chinese Health and Family Life Survey, data being available from

http://popcenter.uchicago.edu/data/chfls.shtml

The main conclusion is drawn from a proportional odds model linking the self-reported orgasm frequency of women with male (!) partners to sociodemographic and wealth variables of the couple.

Reproducing Pollet & Nettle (2009)

The paper is actually reproducible because

- the data are publically available,
- the data preprocessing is well-described in the manuscript, and
- the software used to fit the model and perform AIC-based model selection is cited (SPSS).

However, Esther Herberich and myself failed to reproduce the analysis in R.

It turned out that SPSS 15.0 did not exclude a model-specific constant in the multinomial log-likelihood before comparing models differing in the covariates.

Reproducing Pollet & Nettle (2009)

When calculating the AIC in a correct manner, the women's education is most strongly (positively) related to the response.

BERLINER KURIER News

Bookmarken Drucken Artikel versenden

ORGASMUS-STUDIE

Kluge Frauen kommen öfter

Klischee vom geilen Dummerchen haben Münchner Uni-Forscher widerlegt

Reproducing Pollet & Nettle (2009)

A correction was published with the authors of the original publication (Herberich et al., 2010, Evolution and Human Behavior).

What did we learn?

```
R> fortune("linear model")

If you give people a linear model function you give them something
dangerous.
   -- John Fox
      useR! 2004, Vienna (May 2004)
```

Replace 'linear' with 'proportional odds'.

Hockeysticks?

From http://www.amstat.org

Climate Science: Key Questions and Answers

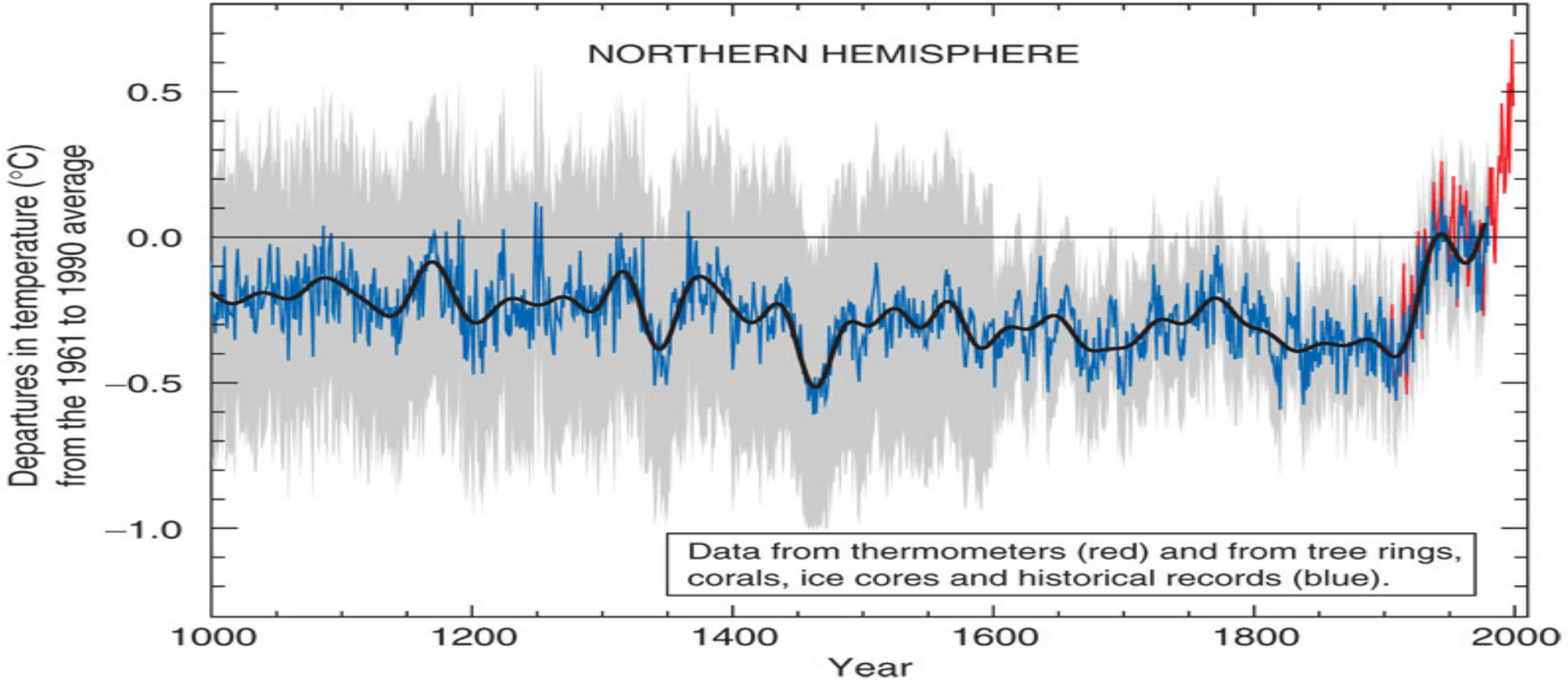
A Congressional Briefing
Tuesday, May 11, 2010

Links to slides, video and audio adjacent to names below

Recent events including the publication of private e-mail correspondence between climate scientists and the examination of the Intergovernmental Panel on Climate Change (IPCC) have led to questions about some climate change research results, the ethics of practicing scientists, and even the efficacy of scientific processes. This briefing will provide the opportunity to examine which climate change science results are well understood and where key uncertainties exist, including issues recently covered in the media such as climate impacts on glaciers and recent temperature trends. Discussion with the distinguished panelists will include examination of the peer-review process, data sources, research processes, statistical analysis, and how various bodies like the IPCC conduct their studies and assessments.

Hockeysticks?

Mann et al. (1998, Nature)



Hockeysticks?

McIntyre & McKitrick (2003, 2005, Energy & Environment) reported several problems with data preprocessing (the data policy became popular as "climate gate") and partially reproduce the results from Mann et al. (1998, Nature).

They point out problems with the statistical analysis, the most important one being the question if and how the data were centered prior to a principle component analysis (the graph essentially displays the first principle component).

The issue was discussed in various boards, including the US Congress. As a consequence of this and similar debates, Prof. Warren Washington, National Center for Atmospheric Research, in a Congressional Briefing (May 11, 2010)* demanded that "All climate data should be freely available by others" and "The scientific results must have reproducibility".

*see http://amstat.org/outreach/climatescience.cfm

Hockeysticks?

Steven McIntyre received a BSc in mathematics and an MA in philosophy, politics, and economics. He works as a mining consultant. In his spare time, he reanalyzed climate data on an old laptop, mainly using R.

Should the scientific community give someone like Mr. McIntyre access to data and the possibility to raise his voice in case of doubt in a scientific publication?

Yes! As a citizen and tax payer, he should have access to data gathered in publically founded research projects. And if his criticism is sound (peer review!) there is no reason to exclude such an opinion.

Environmental data

Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC

...

The objectives of this Directive are:
(a) to guarantee the right of access to environmental information held by or for public authorities ...

So, granting access to (environmental) data is not a matter of taste but an obligation.

OECD Pisa Study

SPIEGEL ONLINE

08. November 2006, 19:37 Uhr

Derbe Forscher-Schelte

"Pisa ist spektakulär gescheitert"

Von Carola Padtberg

In einem neuen Buch holen neun Wissenschaftler zu einem Rundumschlag gegen die Pisa-Studien aus. Kapitale Programmierfehler, unseriöse Methodik, wertlose Ergebnisse - so lauten ihre Vorwürfe. Die Pisa-Macher halten das für abwegig und wundern sich über den späten Alarm.

PISA – Programme for International Student Assessment

OECD Pisa Study

At least in Austria and Germany there was huge media interest, but also growing criticism from the scientific community:

- Grossmann & Neuwirth (2005): Talk at Austrian Statistical Society
- Wuttke (2006): Book "Pisa & Co – Kritik eines Programms"

Common tone: The results of PISA 2000 cannot be reproduced based on the freely available raw data and the technical description of the analysis.

OECD Pisa Study

Analysis was done centrally by the Australian Council of Educational Research using proprietary software.

- In order to compare students (schools, countries, . . .), the difficulty of problems and competences of students were mapped onto a common scale using a Rasch model. This is the basis for all rankings.
- Documentation is too sparse to reprogram the model: several independent research teams have failed.
- Reaction of PISA authors: analysis of other researchers "not correct".

OECD Pisa Study

The main point here is not who is right, but that the complete discussion should not have started in the first place.

- As "scientific insight" academia accepts since centuries only results that can be independently reproduced.
- Sufficient documentation for reproduction is responsibility of authors, not of peers.
- Note: Strongest confirmation of PISA would be a replica of the complete study, here we cannot even get the same results from the same primary data!

Protein Data Bank

http://www.wwpdb.org is an archive for protein structures, mostly obtained from X-ray crystallography. Storage of detected structures is mandatory prior to publication. X-ray images are 'raw' data to this analysis.

The project recently proposed an "X-ray Validation Task Force" responsible to "collect recommendations and develop consensus on additional validation that should be performed on PDB entries, and to identify software applications to perform validation tasks."

Also other communities discuss standardized ways to store (and publish) experimental data.

Is data sharing new?

Charles Spearman (1904, The American Journal of Psychology)

"GENERAL INTELLIGENCE," OBJECTIVELY DETERMINED AND MEASURED.

By C. SPEARMAN.

The method of "product moments," though sometimes involving lengthy calculations, is so simple in principle that it can be worked by any moderately intelligent schoolboy. Explanation and illustration are given in the above article; here, nothing more than the general formula can be stated, which is as follows:

$$r = \frac{Sxy}{\sqrt{Sx^2.Sy^2}}$$

Is data sharing new?

APPENDIX.

Experimental Series I.

Village School, 24 Oldest Children.

A. Original Data.

Sex.	Age.		Discriminative Threshold.			Intellectual Rank.		
	Years	Months	Pitch 1/3 v. d.	Light 1:200	Weight 1:200	Common Sense out of School. (A)	Common Sense out of School. (B)	Cleverness in School.
f	11	6	8	4	4	6	5	2
m	12	11	15	3	4	11	7	22
f	12	8	14	6	4	16	10	7
f	13	8	13	4	9	1	1	1
m	11	4	5	14	7	3	2	3

What about us?

From the case studies discussed above it is clear that we should aim at

- publishing data (as raw as possible) AND
- source code

needed to reproduce and, potentially, improve statistical analyzes.

When it comes to making data available to other scientists, it seems that our 'clients' outperform us clearly.

What we have to add is knowledge about making statistical analyzes reproducible.

The rest of the talk focuses on the state of affairs of reproducibility in statistics and bioinformatics today.

Biometrical Journal

Total numbers of papers presenting simulation studies or example analyzes and giving access to data or code in issues 1–4 and 6 of volume 50.

	Simulation	Example	Data	Code
no	17 (30.4%)	8 (14.3%)	39 (69.6%)	48 (85.7%)
yes	39 (69.6%)	48 (85.7%)	17 (30.4%)	8 (14.3%)

Biometrical Journal

Since 2008, I serve as "Reproducible Research Editor". My duty is to review code that is submitted as supplementary material.

The majority of authors submit R code, some C or FORTRAN, hardly anybody still uses SAS. My experiences are:

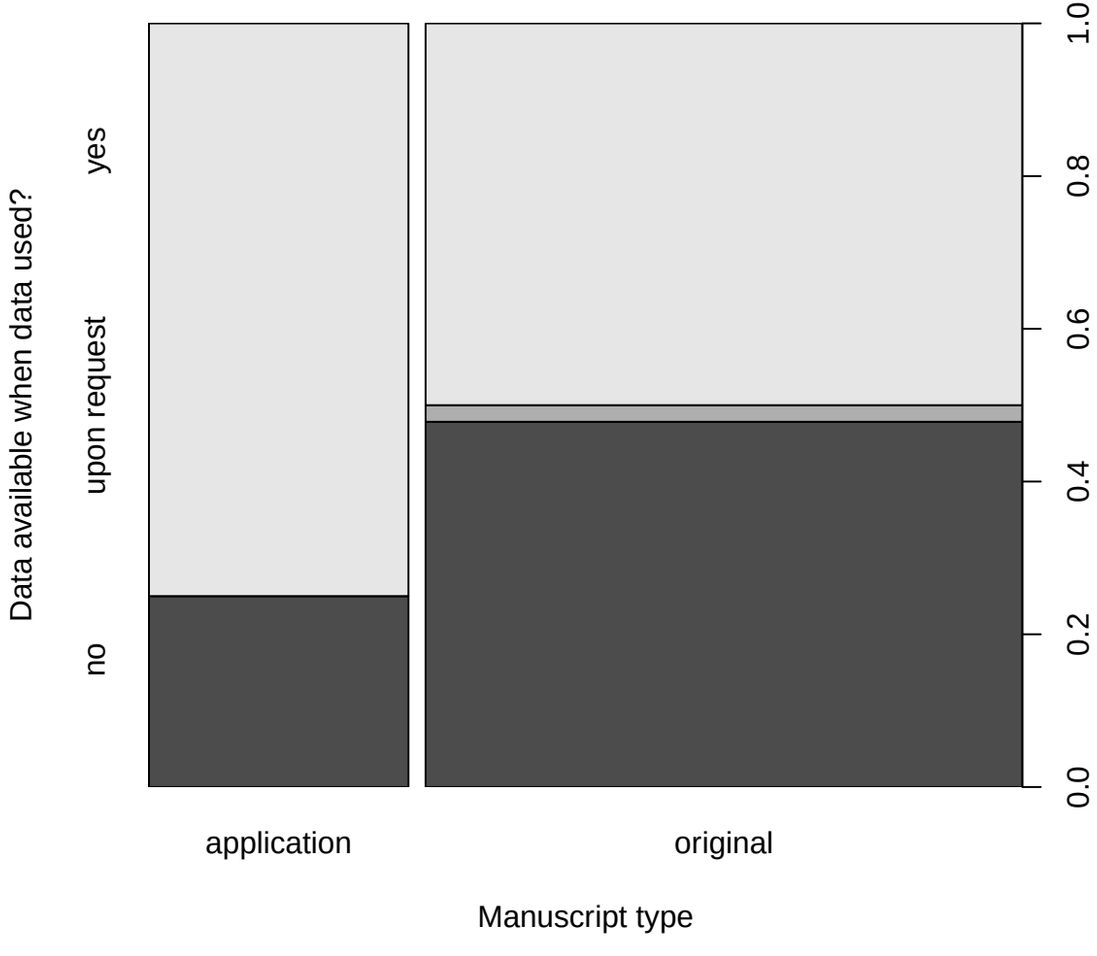
- 1/2 of the submissions can't be compiled or immediately give an error that is not easy to fix for me.
- 1/4 of the submissions has problems that I'm able to fix (but others might not).
- Only a small proportion of submissions exactly reproduces the numbers/figures given in the manuscript.
- Source code of simulations is hardly ever submitted.
- Nobody knows about `set.seed()`.

Bioinformatics

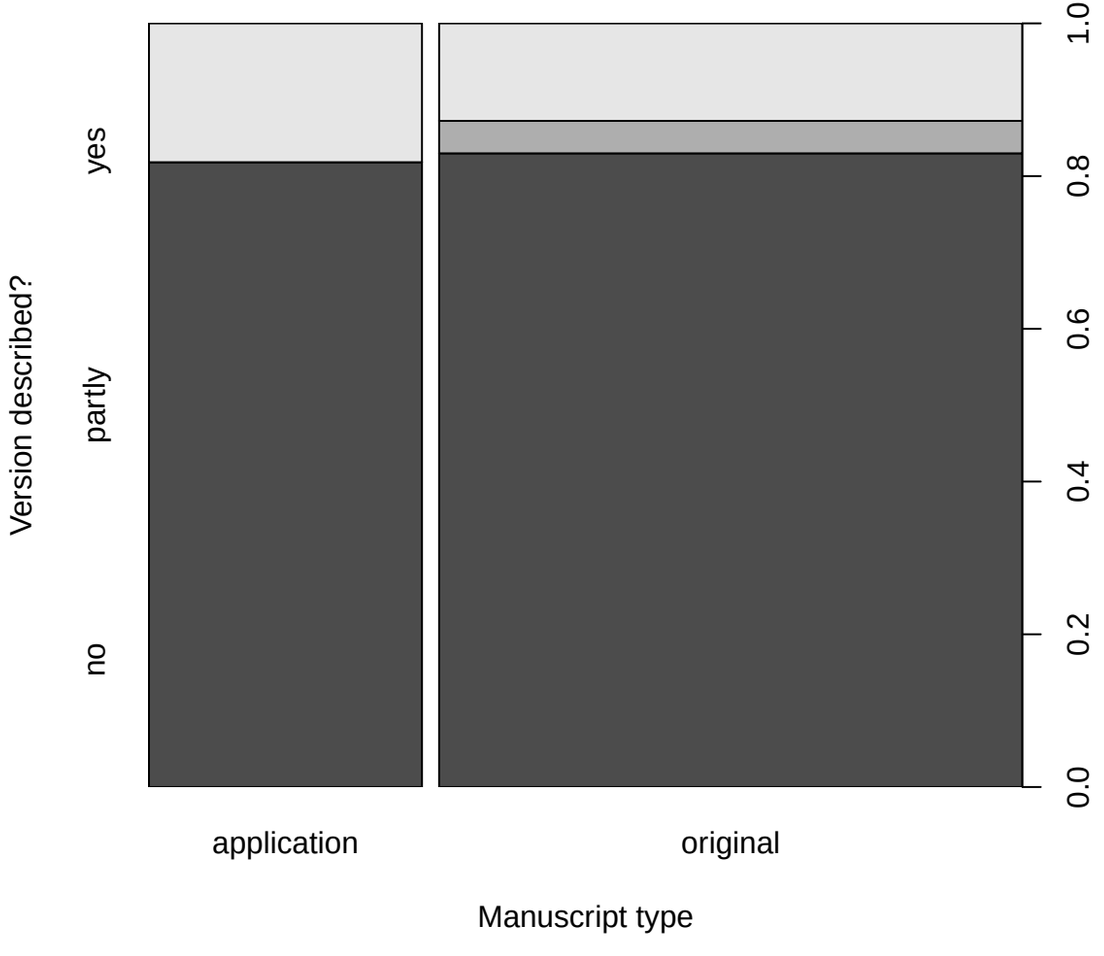
Fritz Leisch and myself sampled 100 of 209 papers published in numbers 1–7 of volume 26 of *Bioinformatics* and recorded if data, analysis code, and simulation code is available.

We distinguish between Application Notes and Original Papers.

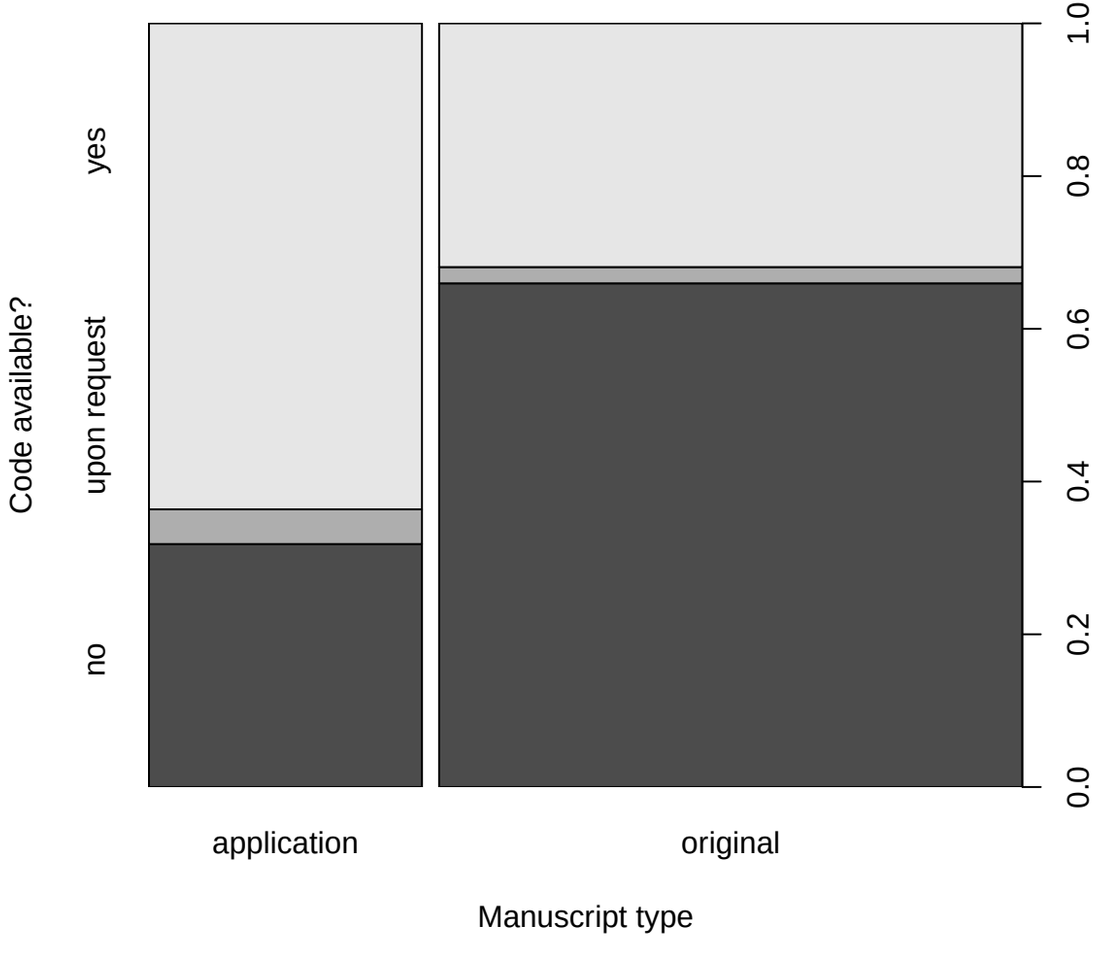
Bioinformatics



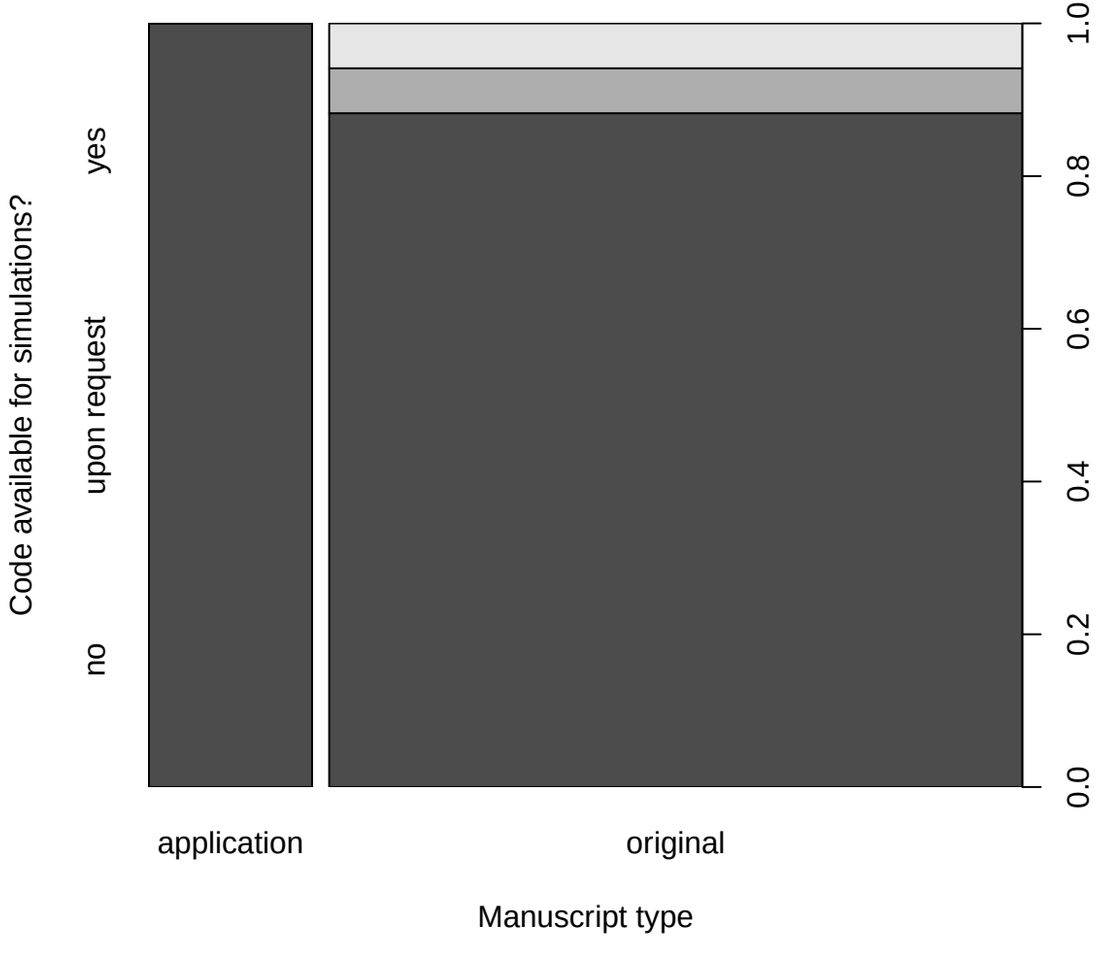
Bioinformatics



Bioinformatics



Bioinformatics



Two Extremes

Hanczar et al. (2010) investigate the small-sample performance of estimates in receiver operator characteristics via simulation.

Only very briefly are the classifiers introduced (linear discriminant analysis, support vector machines and radial basis function support vector machine). There is no hope to reproduce the findings because

- the description of the simulation model is insufficient,
- a lack of information how the classifiers were tuned,
- which software was used for fitting the classifiers.

Allowing users to access the source code of this simulation experiment would be an appropriate way to solve these issues.

Two Extremes

Kirchner et al (2010) introduce a random forest and discrete mapping approach to the analysis of mass spectrometry data. The methods are evaluated and compared based on results obtained from analyzes of two proteomics data sets. The interested reader is referred to a web page offering access to the data and the R source code along with the necessary information needed to re-perform the analysis. This electronic material makes this paper fully reproducible.

Problems

Reproducibility over time

In 2006, Brian Everitt and myself published the "Handbook of Statistical Analyses Using R". A dedicated R add-on package **HSAUR** contains all data sets used and, for each chapter, a package vignette reproduces the analyzes presented in the book.

As of December 2005, the output of the analyzes matched what was printed in the book. Today, the code still runs without errors (see http://CRAN.R-project.org/package=HSAUR). However, the results changed in approx. 170 instances due to changes/updates in R or contributed packages.

However, the book is no longer reproducible–well, at least not in a very strict sense.

Problems

- Data might be static, but reproducibility is a moving target.
- There is a need for maintenance of code.
- Publishers provide only inadequate infrastructure for storing data and code.
- Even if not published, data and code should at least be available to referees but hardly anybody is willing to review extensive source code.
- Checking code is actually less work than checking a mathematical proof:
  - If the code runs, it is a copy and paste exercise.
  - If the code does not run, reject.
- Checking that the code makes sense is of course a different question.
- What about proprietory software?
- The problem is getting more urgent all the time because computational methods and environments are getting more complicated.

Problem? Feature?

- If you give away your code, people might actually start to use your methods. They might start asking questions or even criticize you.
- People will cite your papers.

Problem? Feature?

- If you give away your code, people might actually start to use your methods. They might start asking questions or even criticize you.
- People will cite your papers.

Rank in journal-specific citation hit lists 2001–2010:

Software	Publication	Rank
multcomp	Hothorn et al. (2008, Biometrical Journal)	2/776
mboost	Bühlmann & Hothorn (2007, Statistical Science)	36/379
party	Hothorn et al. (2006, JCGS)	6/488
coin	Hothorn et al. (2006, TAS)	10/742