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

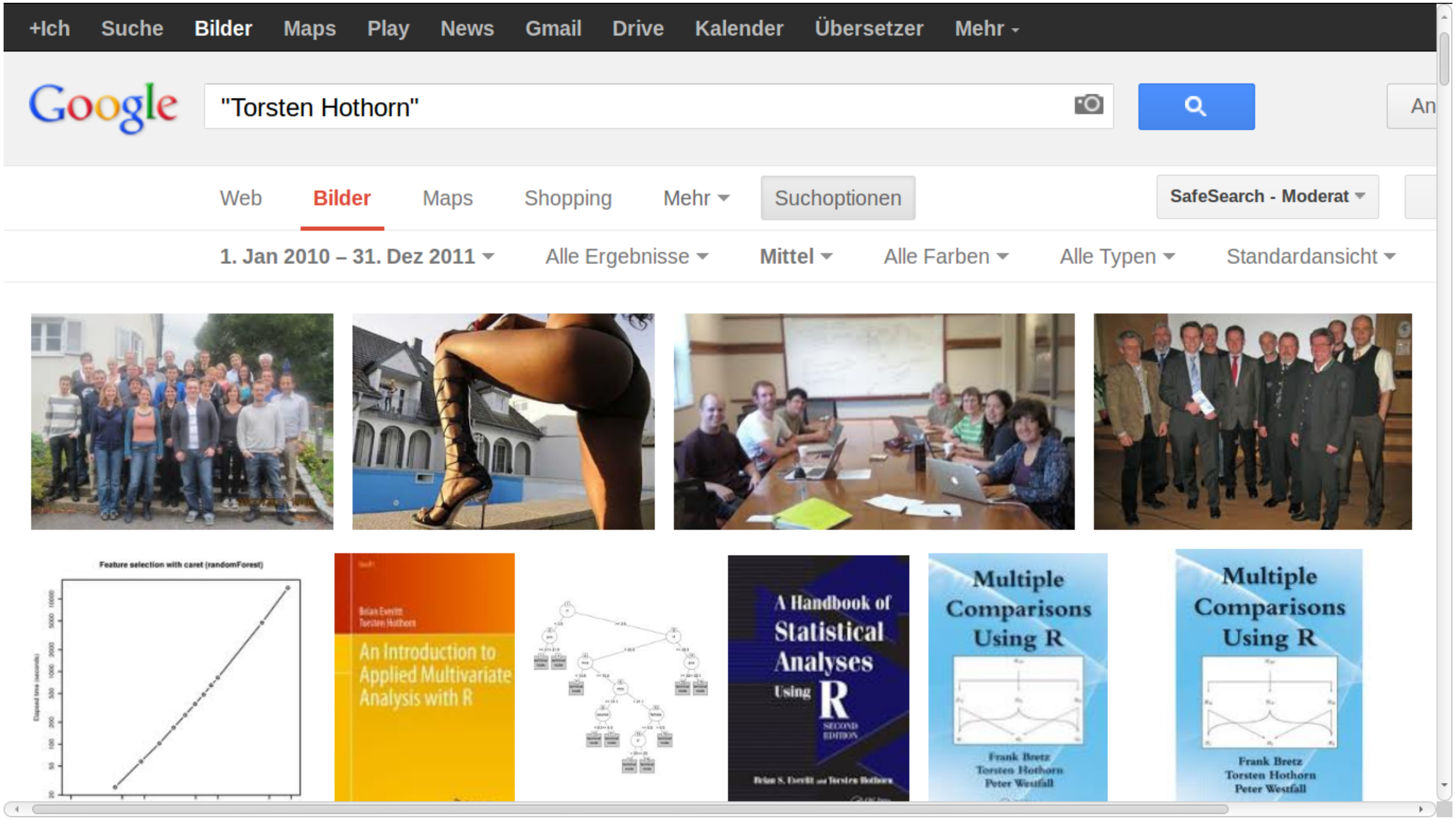
What did we learn?

```
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'.

Unintended Consequences...



"Torsten Hothorn"

SafeSearch - Moderat

1. Jan 2010 – 31. Dez 2011

Alle Ergebnisse

Mittel

Alle Farben

Alle Typen

Standardansicht

Does Tamiflu Work?

British Medical Journal, Dec 2009:

ANALYSIS

What can we learn from observational studies of oseltamivir to treat influenza in healthy adults?

Nick Freemantle, Mel Calvert

BMJ 2009;339:b5248 (Published 08 December 2009)

FEATURE

Complications: tracking down the data on oseltamivir

Deborah Cohen

BMJ 2012;339:b5387 (Published 08 December 2009)

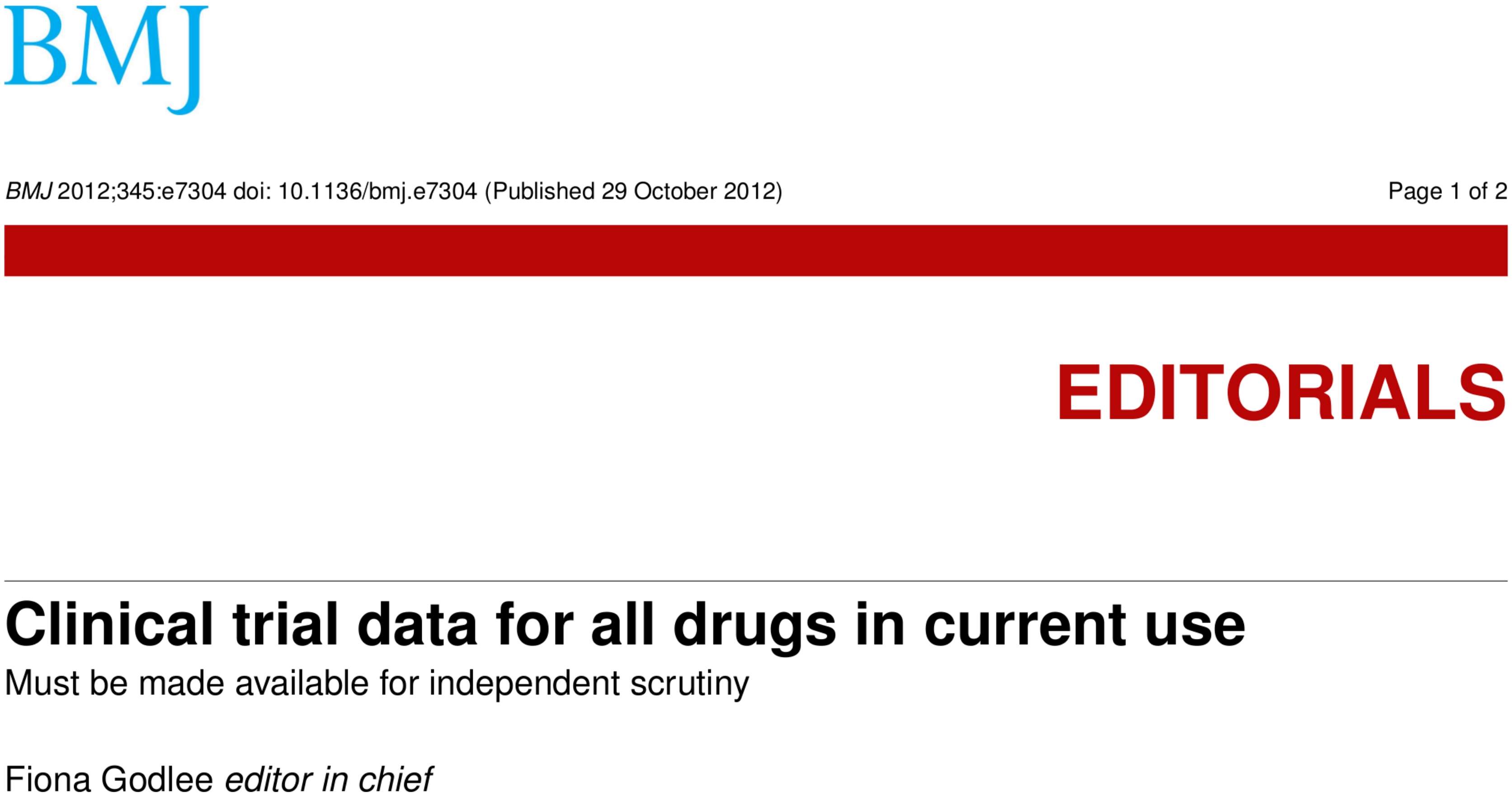
EDITORIAL

Why don't we have all the evidence on oseltamivir?

Fiona Godlee, Mike Clarke

BMJ 2009;339:b5351 (Published 08 December 2009)

BMJ's Open Data Policy



BMJ

BMJ 2012;345:e7304 doi: 10.1136/bmj.e7304 (Published 29 October 2012)

Page 1 of 2

EDITORIALS

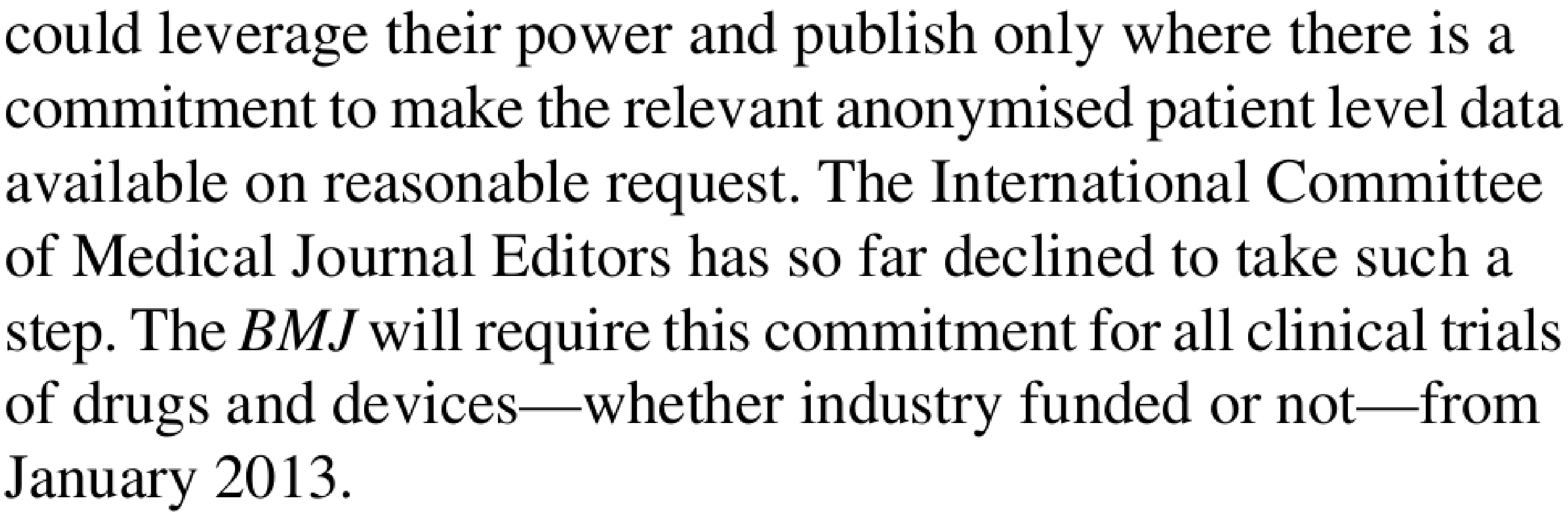
Clinical trial data for all drugs in current use

Must be made available for independent scrutiny

Fiona Godlee *editor in chief*

BMJ's Open Data Policy

(Medical journals ..., TH)



could leverage their power and publish only where there is a commitment to make the relevant anonymised patient level data available on reasonable request. The International Committee of Medical Journal Editors has so far declined to take such a step. The *BMJ* will require this commitment for all clinical trials of drugs and devices—whether industry funded or not—from January 2013.

Protein Data Bank

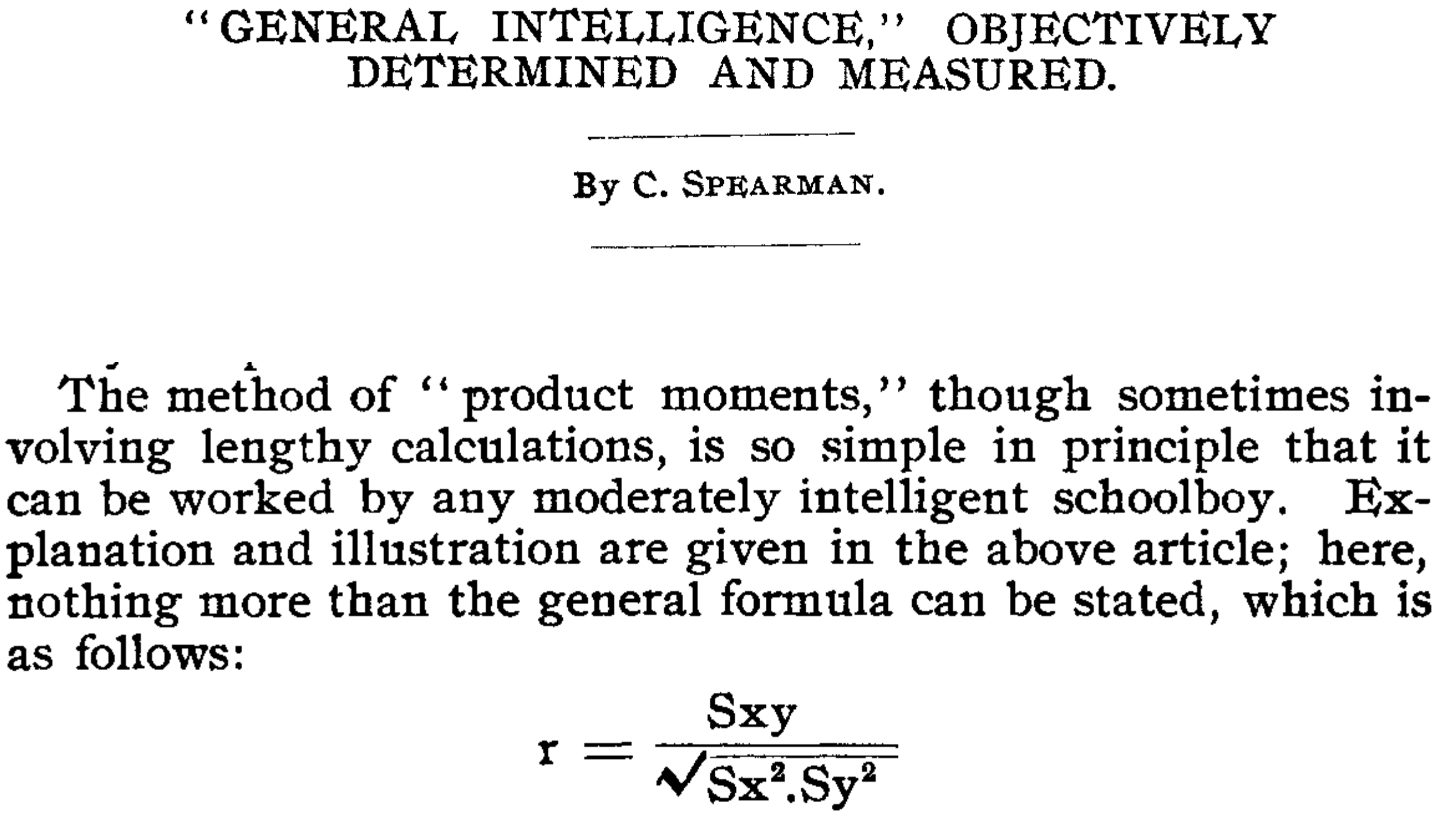
`http://www.wwpdb.org` is an archive for protein structures, mostly obtained from X-ray crystallography. Storage of detected structures is mandatory prior to publication. X-ray images are ‘raw’ data to this analysis.

The project proposed an “X-ray Validation Task Force” in 2011, 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)



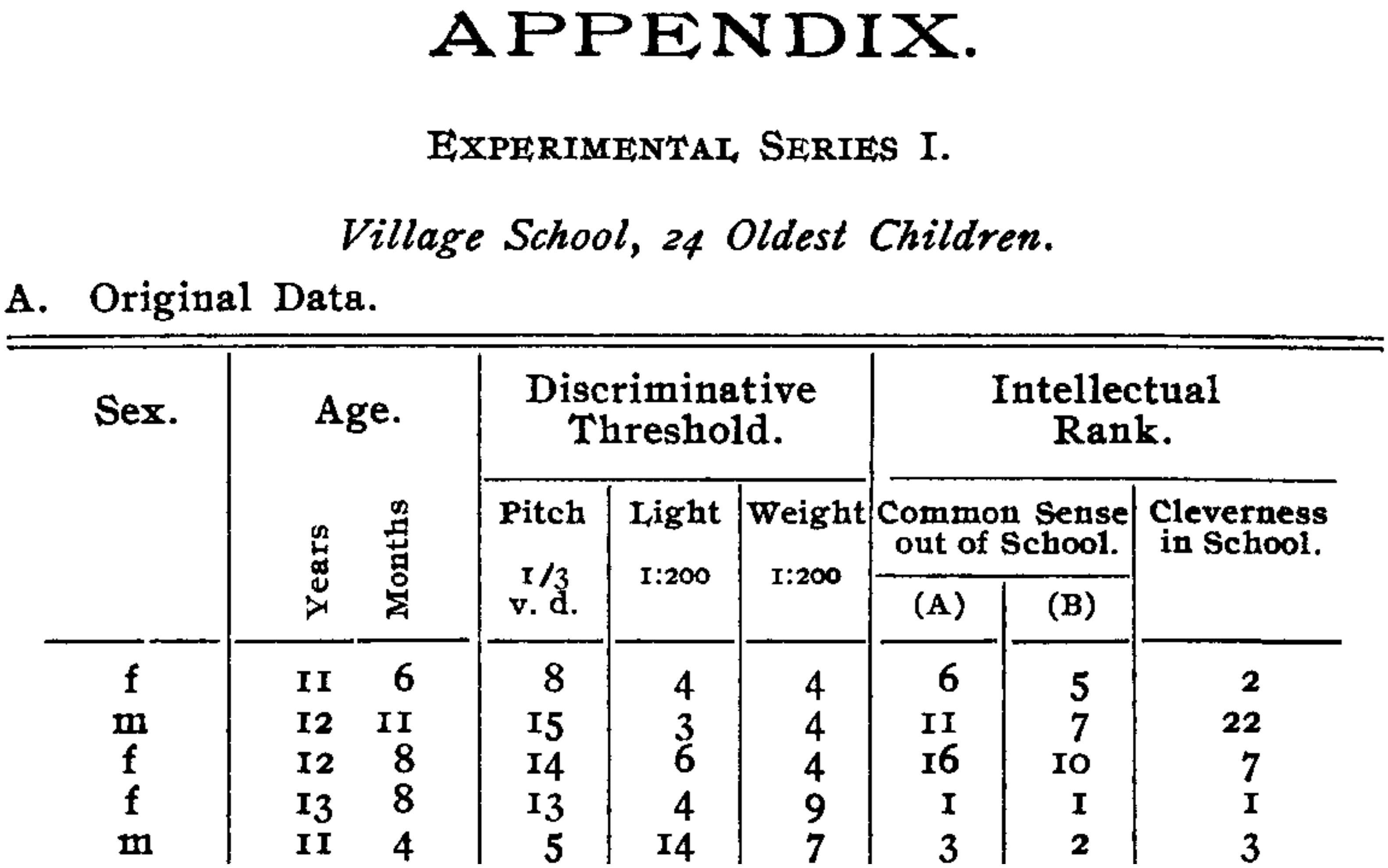
"GENERAL INTELLIGENCE," OBJECTIVELY DETERMINED AND MEASURED.

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.		Cleverness in School.
						(A)	(B)	
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

From 2008 to 2011, I served as “Reproducible Research Editor”. My duty was to review code that was submitted as supplementary material.

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

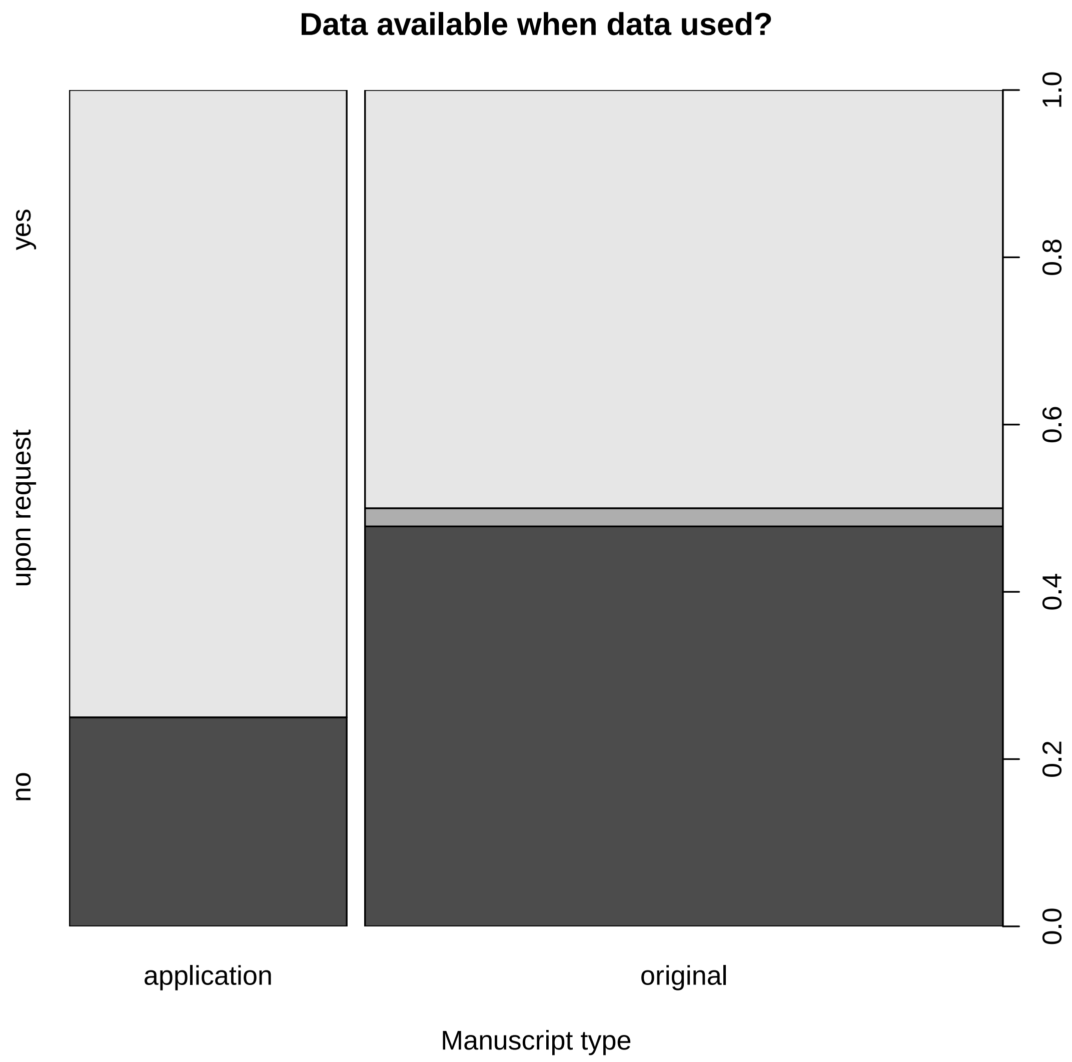
- 1/2 of the submissions could not be compiled or immediately gave an error that was not easy to fix for me.
- 1/4 of the submissions had problems that I wasn’t able to fix.
- Only a small proportion of submissions exactly reproduced the numbers/figures given in the manuscript.
- Source code of simulations was hardly ever submitted.
- Nobody knows about `set.seed()`.

Bioinformatics

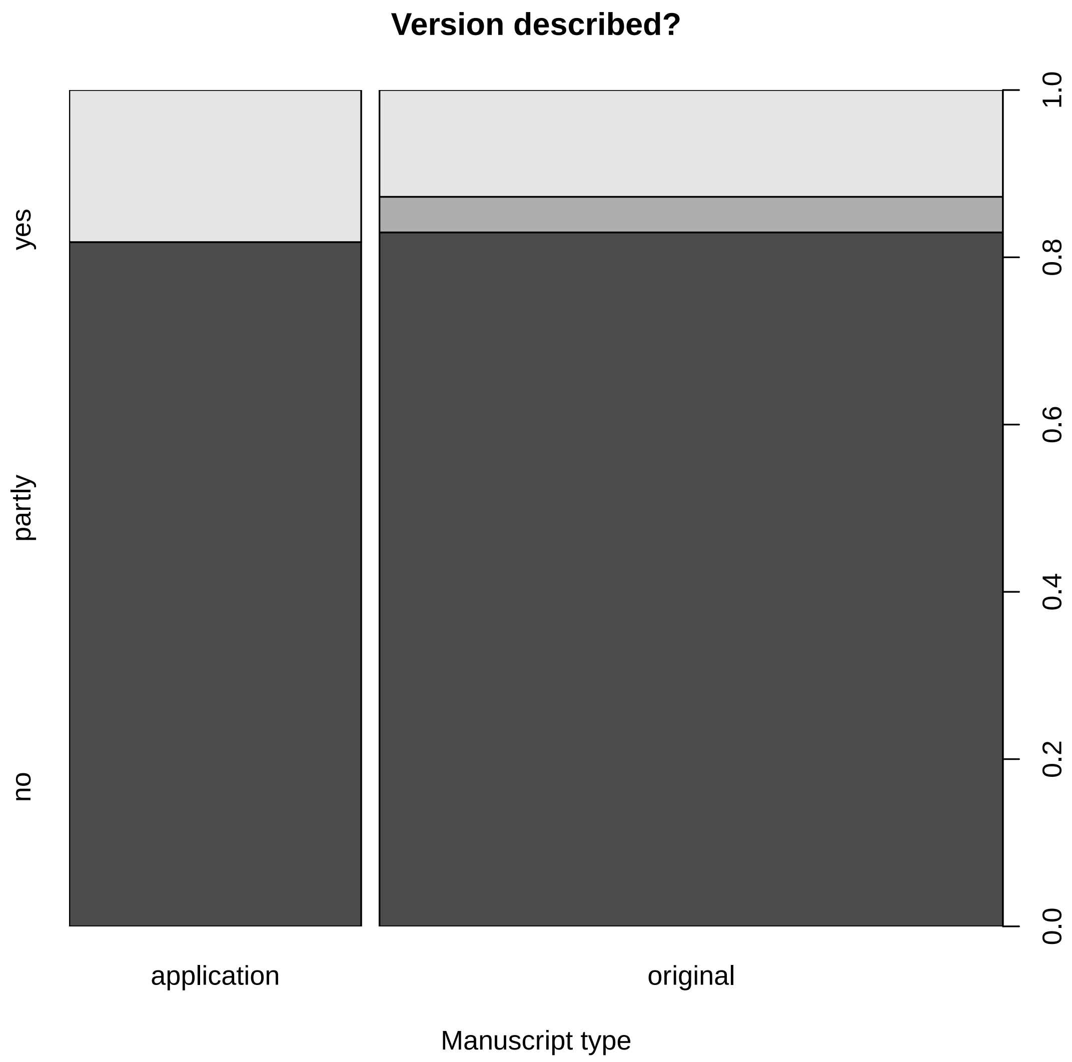
Fritz Leisch and myself sampled 100 of 209 papers published in numbers 1–7 of volume 26 of *Bioinformatics* and recorded if data, analysis code, and simulation code is available.

We distinguish between Application Notes and Original Papers.

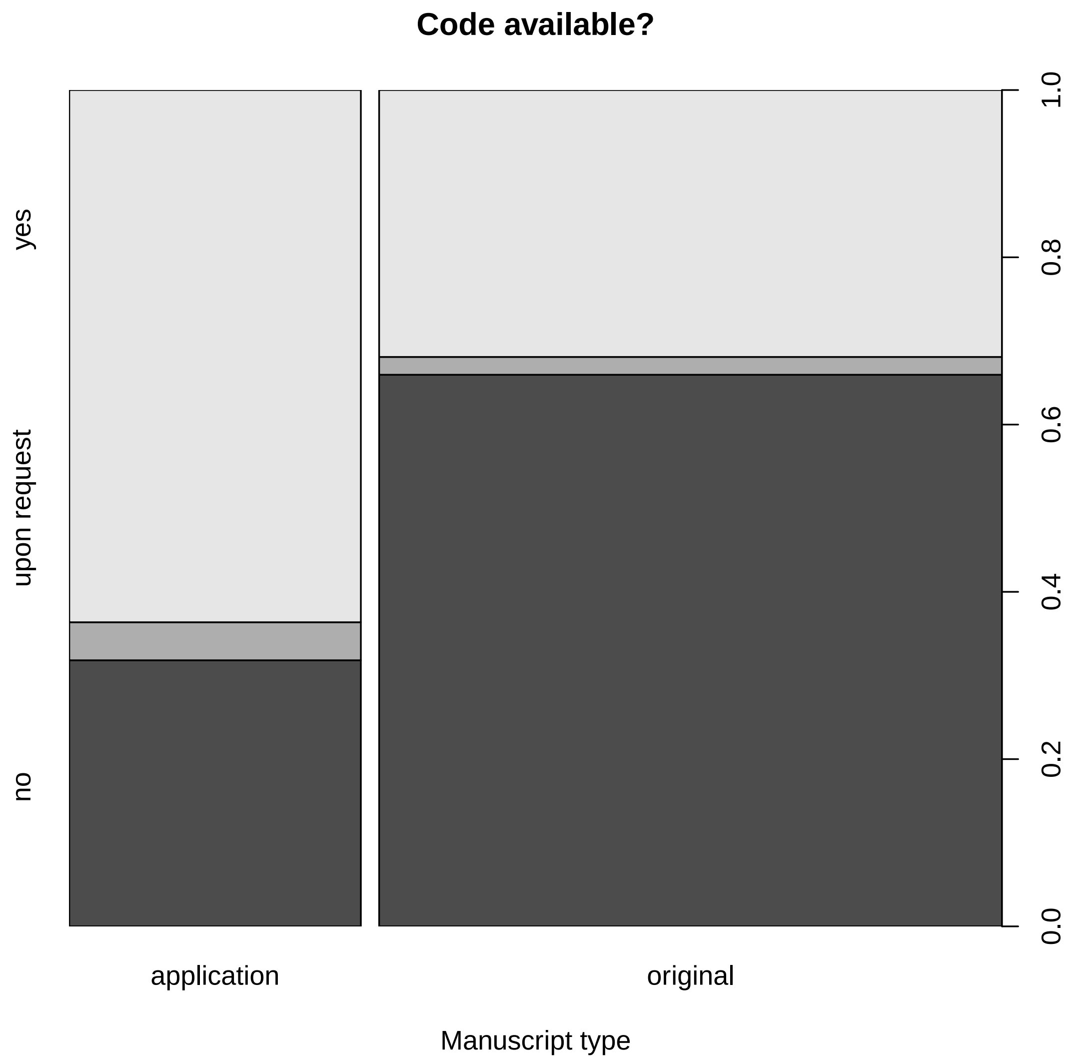
Bioinformatics



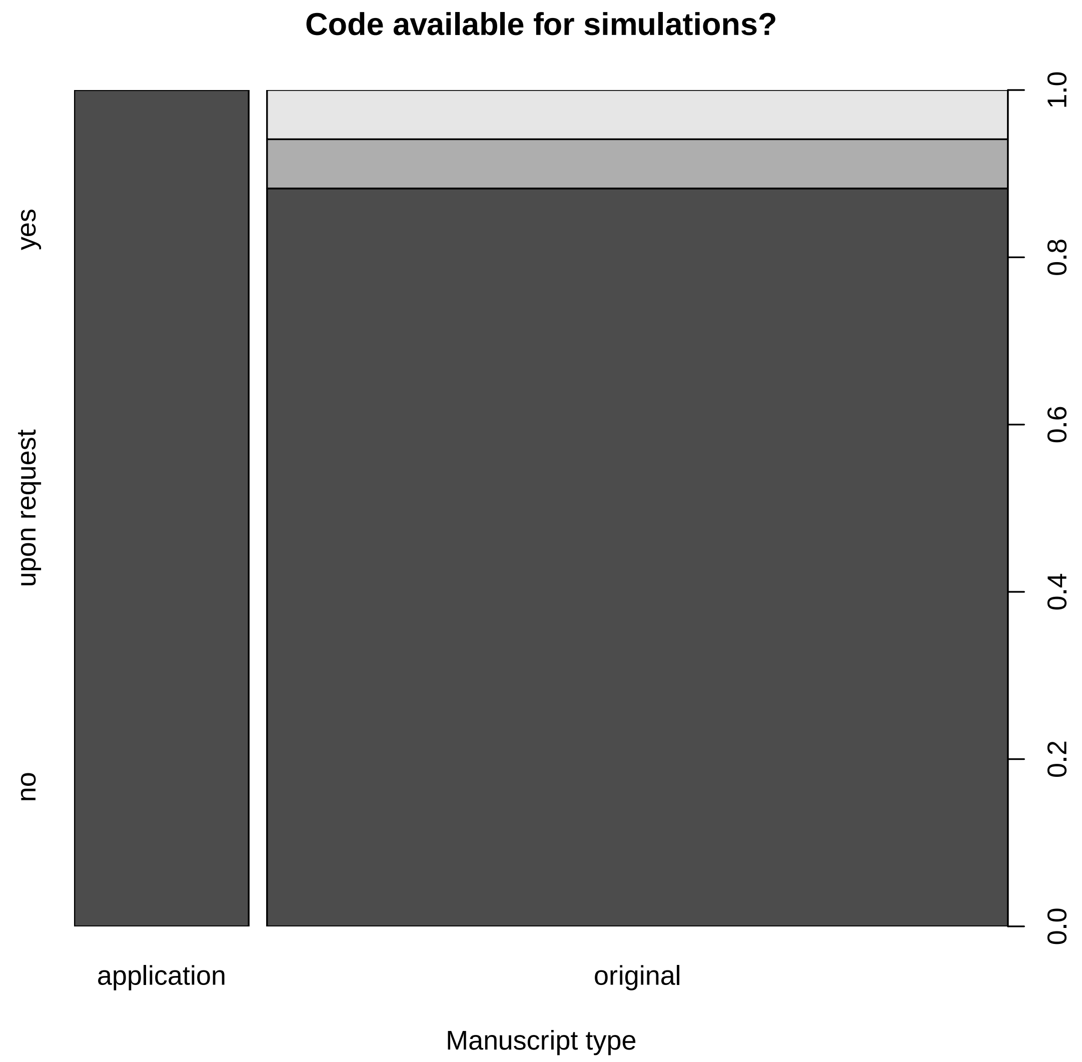
Bioinformatics



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 proprietary software?
- The problem is getting more urgent all the time because computational methods and environments are getting more complicated.

We Need...

- to overcome current restrictions in the review and publishing process which aims at publishing condensed “advertisements” (aka papers) instead of the whole scholarship (protocols, data, code, additional results).
- to make sure that current computational environments will remain functional in the future (“clouds” of ancient OSes?). Open source software is key here: R-2.0.0 (2004-10-04) compiles and runs on Ubuntu Precise but `make check` fails.
- make sure published data is well understood, so there is a need to publish protocols (see also Keiding, 2010, Biostatistics).