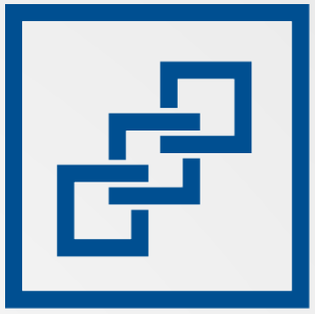






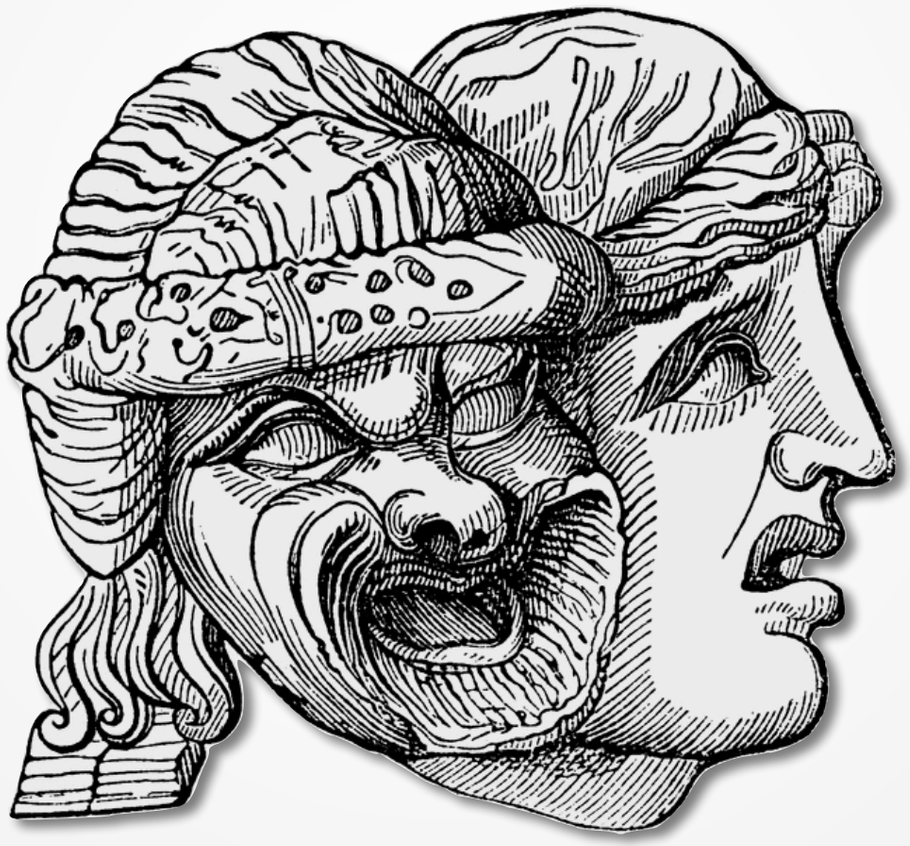

p-hacking
or the attractiveness of dei ex machinis

Wolfgang Karl Härdle
Ilyas Agakishiev
Raphael C. G. Reule

Ladislaus von Bortkiewicz Professor of Statistics
Humboldt-Universität zu Berlin
BRC Blockchain Research Center
lvb.wiwi.hu-berlin.de
Charles University, WISE XMU, NCTU 玉山学者

Dei Ex Machinis

Unexpected powers or events,
saving a seemingly *hopeless situation*,
especially as a *contrived plot device* in a play or novel.



Personae
social role(s) or character(s) played by an actor;
originally referred to theatrical masks.

Scientific progress ! (?)

"No editor wants to be the one
that rejects the next Black Scholes paper."

Campbell R. Harvey

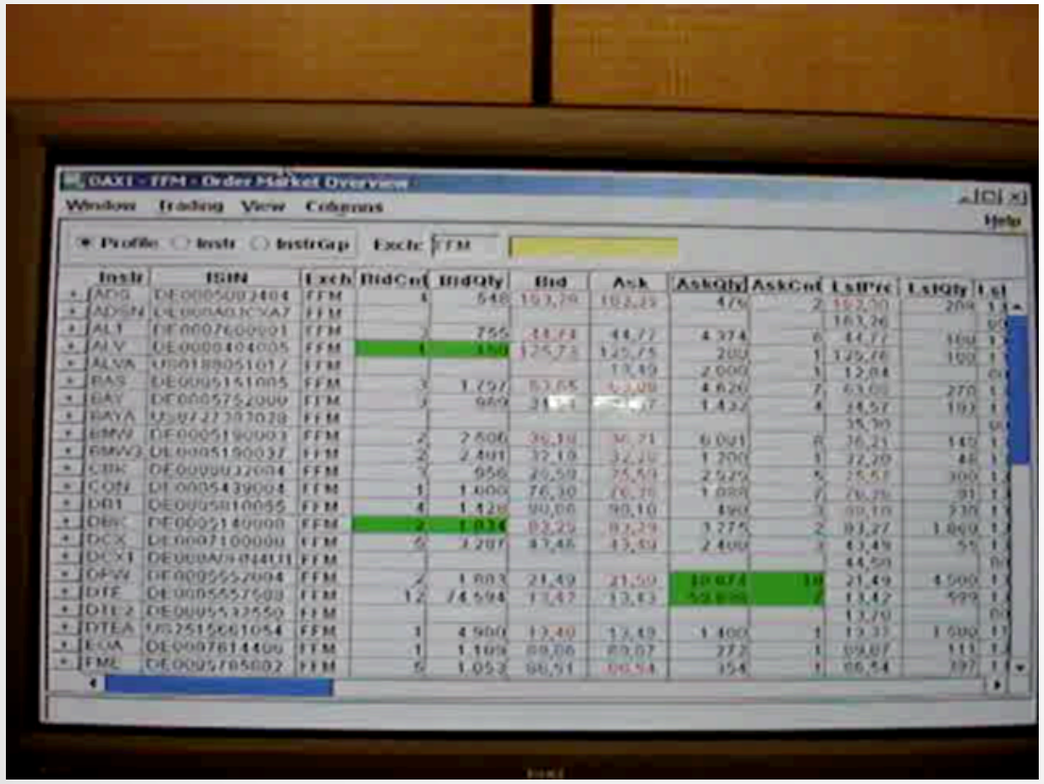
- Black & Scholes formula for pricing a stock option 1965/1969-1970
- J Political Econ: reject without review
- Review of Economics and Statistics: analogous rejection
- J Finance published article 1972
- J Political Econ hesitantly published revised article in 1973

Black, Scholes

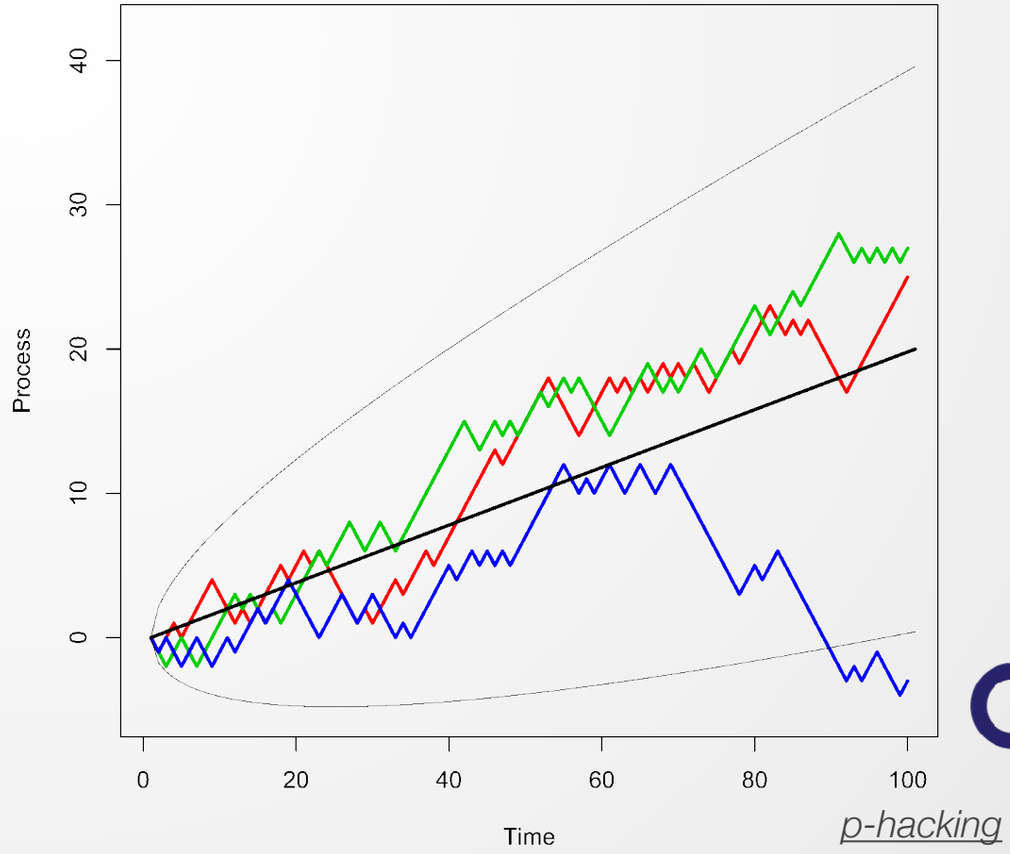
http://garfield.library.upenn.edu/classics1987/A1987J461500001.pdf

p-hacking vs. $$$

- construct portfolii via 1st, 2nd, 3rd letter of ticker
- time frame 1926 - today; 1963 - today
- do value weighting and equal weighting
- id best long-short portfolio based on smallest *p*-value
- # portfolii = 25 280 (max)

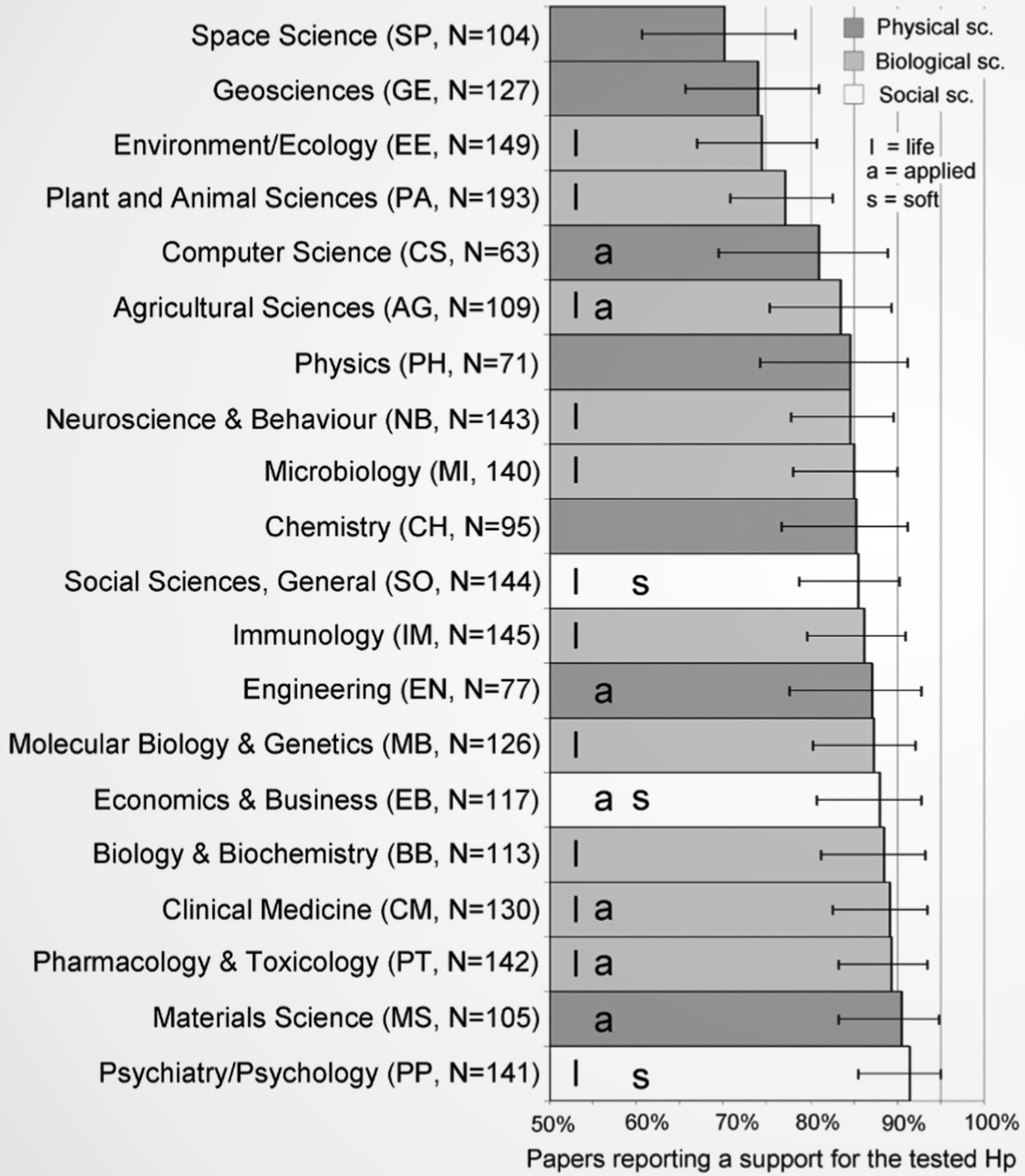


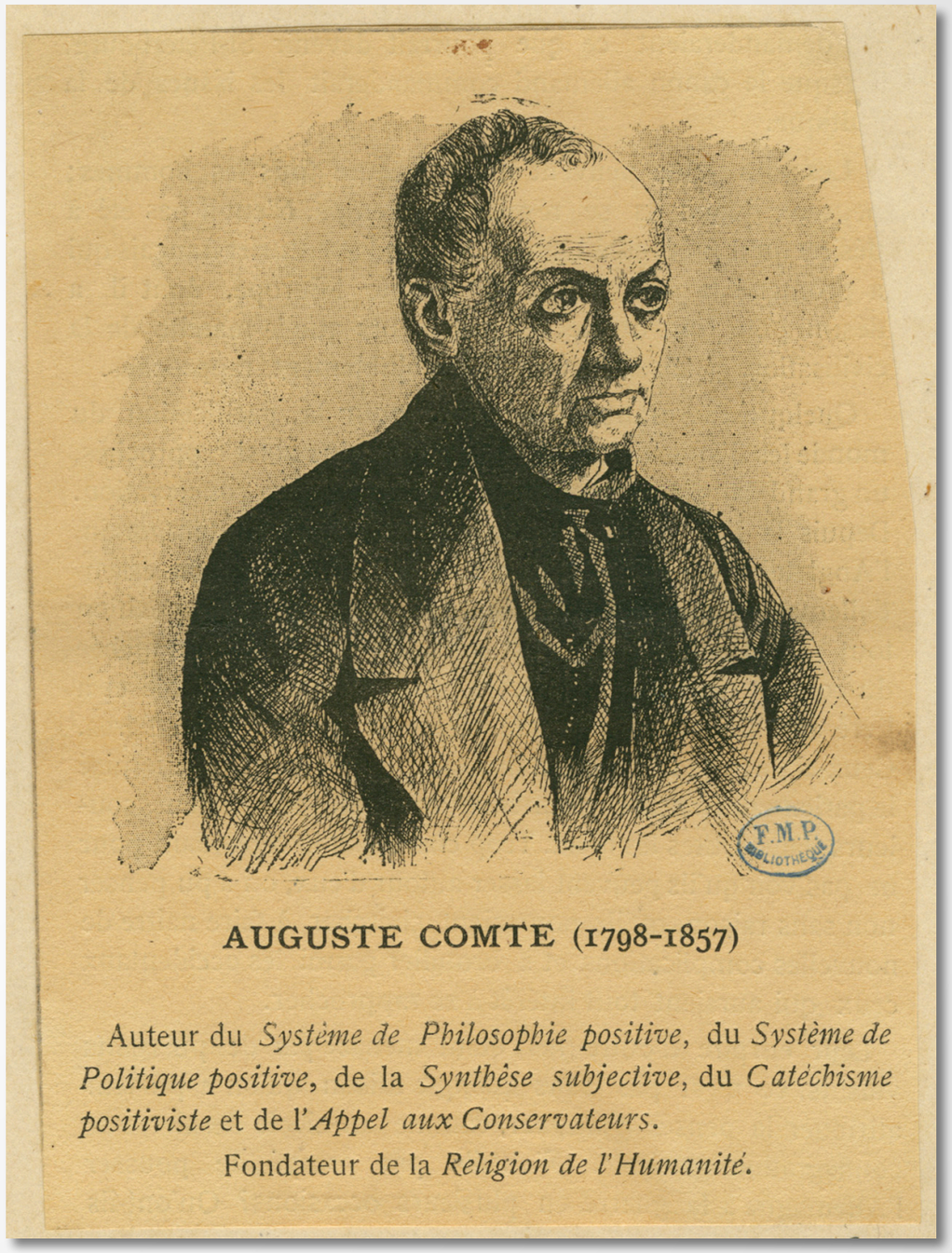
http://eoddata.com/stocklist/NYSE/P.htm



p-hacking

p-hacking = outstanding results / citations





Percent of research papers reporting support for the tested hypothesis by field (Fanelli, 2010)

Chocolate

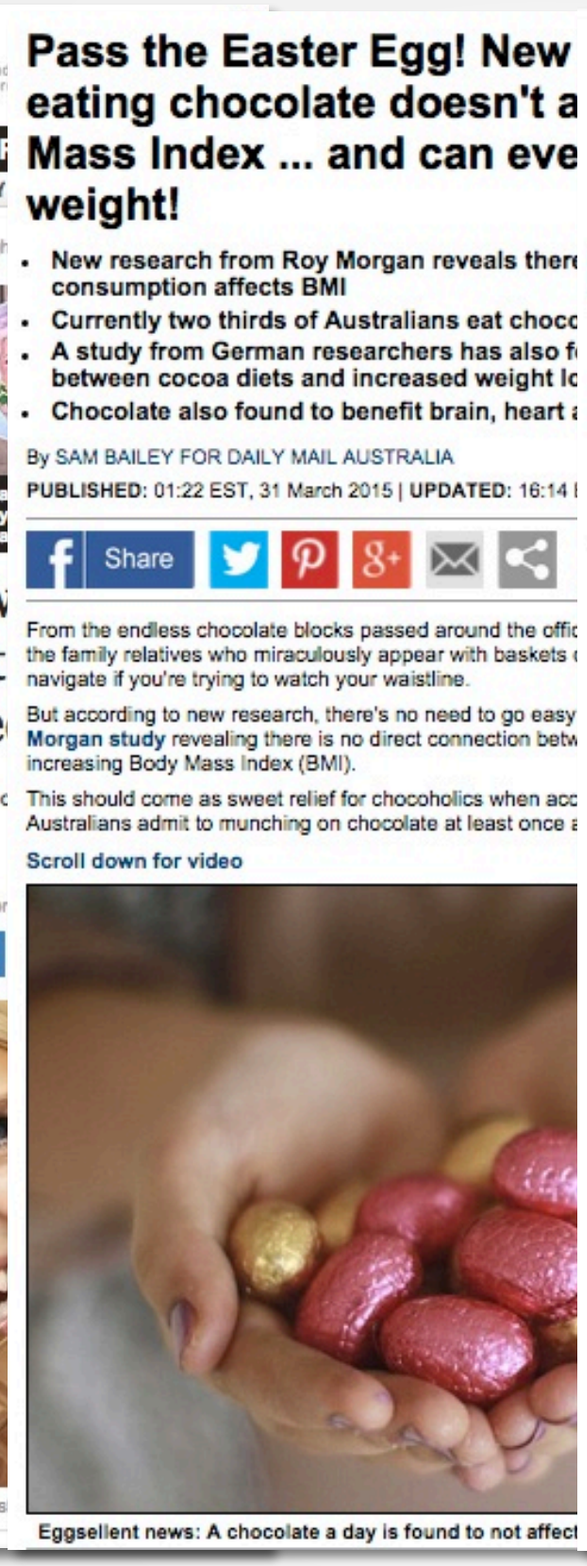
p-hacking chocolate

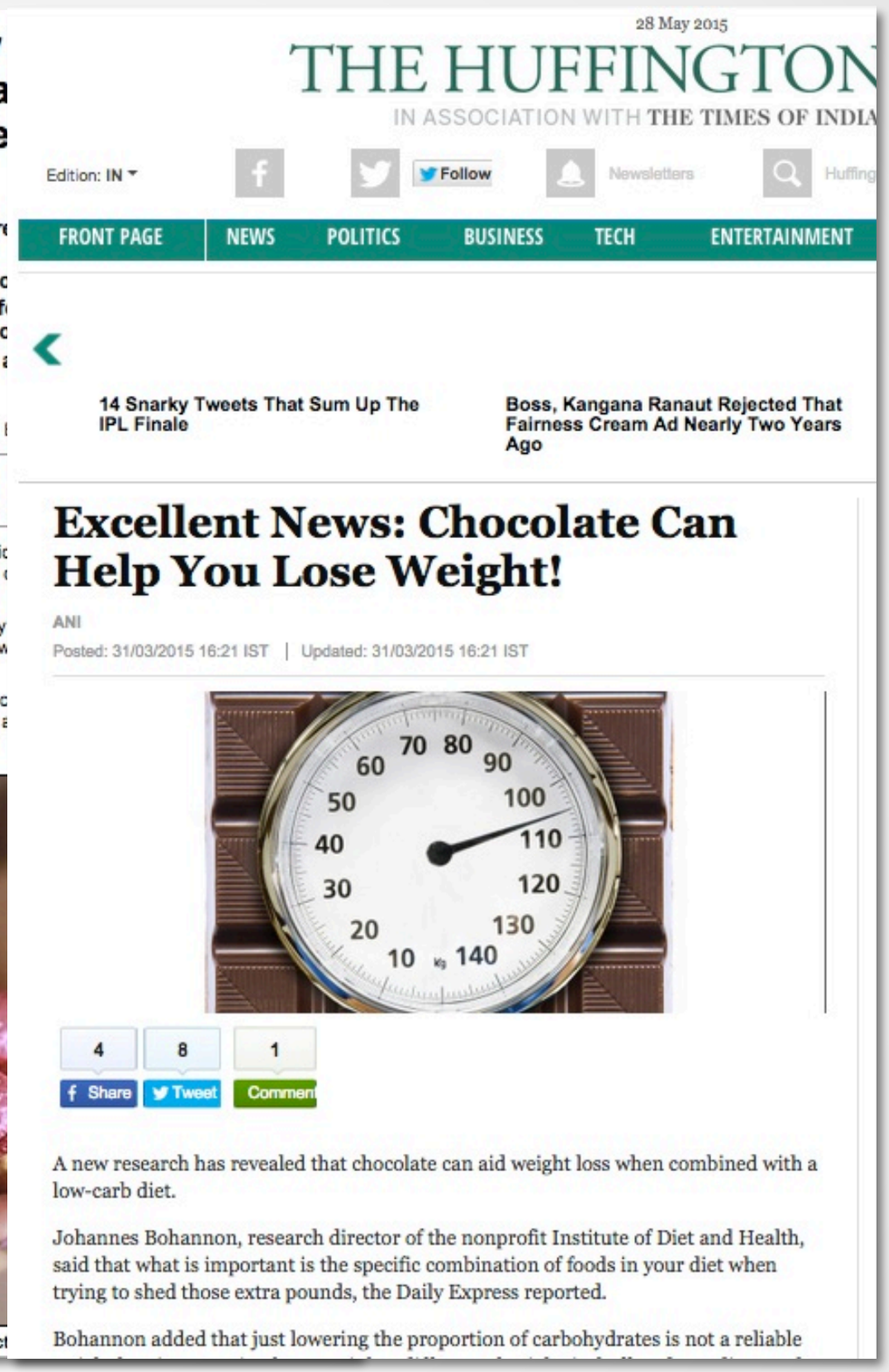
- case study with low carb diet +/- chocolate
- small *p*-values (spv)
- conclusion: Chocolate in, weight loss out !

> Consumption of chocolate with a high cocoa content can *significantly* increase the success of weight-loss diets. The weight-loss effect of this diet occurs with a certain delay. Long-term weight loss, however, seems to occur easier and more successfully by adding chocolate. *The effect of the chocolate, the so-called "weight loss turbo",* seems to go hand in hand with personal well-being, which was *significantly* higher than in the control groups.

Details, Source ; Spring 2015


EXPRESS
Home of the Daily and Sunday Expr
HOME NEWS ELECTION SPORT COMMENT
HEALTH LIFE DIETS GARDEN FOOD STYLE PROPERTY
Home Life & Style Health Chocolate accelerates weigh
Pollen threat: How foreign ragweed will increase risks for...
Blood pressure breakthrough: Jab every six months could
Chocolate accelerates w
loss: Research claims it
cholesterol and aids sle
CAN you indulge your sweet tooth and lose weight? If it's choc
the answer seems to be yes.
By SARAH BARNS
PUBLISHED: 10:31, Mon, Mar 30, 2015 | UPDATED: 20:28, Sat, Apr
SHARE TWEET
Chocolate can aid weight loss when combined with a low-carb diet, s


Pass the Easter Egg! New
eating chocolate doesn't a
Mass Index ... and can eve
weight!
- New research from Roy Morgan reveals there consumption affects BMI
- Currently two thirds of Australians eat choco
- A study from German researchers has also f between cocoa diets and increased weight lo
- Chocolate also found to benefit brain, heart a
By SAM BAILEY FOR DAILY MAIL AUSTRALIA
PUBLISHED: 01:22 EST, 31 March 2015 | UPDATED: 16:14
Share
From the endless chocolate blocks passed around the offic the family relatives who miraculously appear with baskets navigate if you're trying to watch your waistline.
But according to new research, there's no need to go easy Morgan study revealing there is no direct connection betw increasing Body Mass Index (BMI).
This should come as sweet relief for chocoholics when acc Australians admit to munching on chocolate at least once a
Scroll down for video
Eggsellent news: A chocolate a day is found to not affect


28 May 2015
THE HUFFINGTON
IN ASSOCIATION WITH THE TIMES OF INDIA
Edition: IN
Follow
Newsletters
FRONT PAGE NEWS POLITICS BUSINESS TECH ENTERTAINMENT
14 Snarky Tweets That Sum Up The IPL Finale
Boss, Kangana Ranaut Rejected That Fairness Cream Ad Nearly Two Years Ago
Excellent News: Chocolate Can Help You Lose Weight!
ANI
Posted: 31/03/2015 16:21 IST | Updated: 31/03/2015 16:21 IST
4 8 1
Share Tweet Comment
A new research has revealed that chocolate can aid weight loss when combined with a low-carb diet.
Johannes Bohannon, research director of the nonprofit Institute of Diet and Health, said that what is important is the specific combination of foods in your diet when trying to shed those extra pounds, the Daily Express reported.
Bohannon added that just lowering the proportion of carbohydrates is not a reliable

Challenges

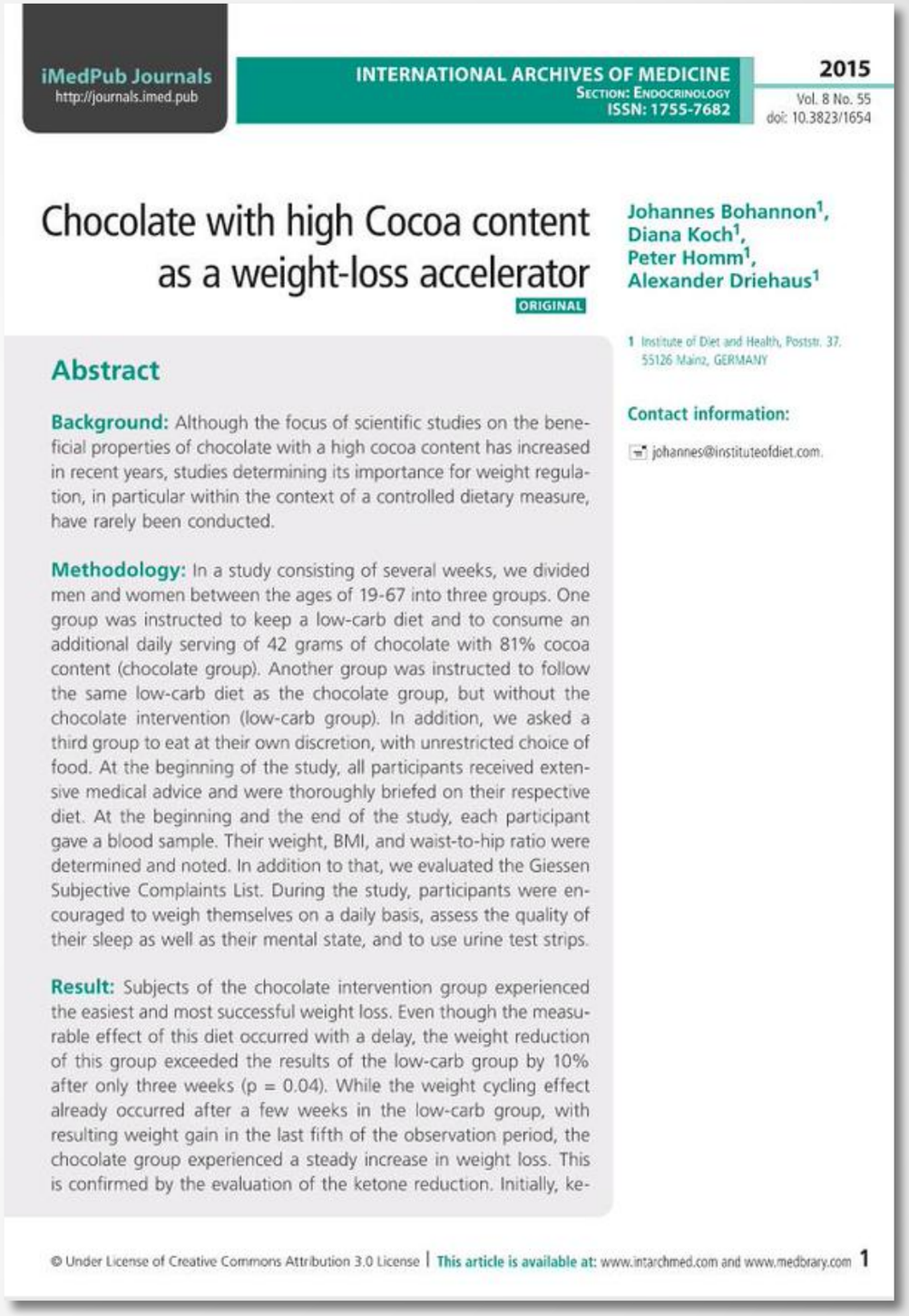
- are spv's a *win* ?
- *conclusions* from spv ?
- *unconscious p*-hacking ?

Outline

1. Motivation ✔
2. Case Study
3. Massaging, dredging, *p*-hacking
4. Asymptopia hackers
5. Critical examples
6. Redo it and show it!
7. Conclusions

Chocolate

- standard clinical trial
  - sample size $n = 15$
  - 18 variables
- 5 men, 11 women (19-67 Y) via Facebook (150 EUR p.P.), 1 dropout
- eat bitter chocolate as supplement to low-carb diet for 3W
- create +/- chocolate groups
- measure weight, cholesterol, sodium, blood protein levels, sleep quality, well-being, etc.



iMedPub Journals
http://journals.imed.pub

INTERNATIONAL ARCHIVES OF MEDICINE
SECTION: ENDOCRINOLOGY
ISSN: 1755-7682

2015
Vol. 8 No. 55
doi: 10.3823/1654

Chocolate with high Cocoa content as a weight-loss accelerator

ORIGINAL

Johannes Bohannon[1],
Diana Koch[1],
Peter Homm[1],
Alexander Driehaus[1]

1 Institute of Diet and Health, Poststr. 37, 55126 Mainz, GERMANY

Contact information:

johannes@instituteofdiet.com.

Abstract

Background: Although the focus of scientific studies on the beneficial properties of chocolate with a high cocoa content has increased in recent years, studies determining its importance for weight regulation, in particular within the context of a controlled dietary measure, have rarely been conducted.

Methodology: In a study consisting of several weeks, we divided men and women between the ages of 19-67 into three groups. One group was instructed to keep a low-carb diet and to consume an additional daily serving of 42 grams of chocolate with 81% cocoa content (chocolate group). Another group was instructed to follow the same low-carb diet as the chocolate group, but without the chocolate intervention (low-carb group). In addition, we asked a third group to eat at their own discretion, with unrestricted choice of food. At the beginning of the study, all participants received extensive medical advice and were thoroughly briefed on their respective diet. At the beginning and the end of the study, each participant gave a blood sample. Their weight, BMI, and waist-to-hip ratio were determined and noted. In addition to that, we evaluated the Giessen Subjective Complaints List. During the study, participants were encouraged to weigh themselves on a daily basis, assess the quality of their sleep as well as their mental state, and to use urine test strips.

Result: Subjects of the chocolate intervention group experienced the easiest and most successful weight loss. Even though the measurable effect of this diet occurred with a delay, the weight reduction of this group exceeded the results of the low-carb group by 10% after only three weeks ($p = 0.04$). While the weight cycling effect already occurred after a few weeks in the low-carb group, with resulting weight gain in the last fifth of the observation period, the chocolate group experienced a steady increase in weight loss. This is confirmed by the evaluation of the ketone reduction. Initially, ke-

Chocolate

" […] the subjects of the chocolate group found a significant improvement in their well-being (physically and mentally). The controlled improvement compared to the results of the low-carb group was **highly significant (p<0.001)**.

[…] A higher amount of ketones could be detected in the participants of the chocolate group than in the low-carb group. The measured results were found to be **highly significant (p<0.01)**.

[…] Exhaustion symptoms in particular, such as fatigue or the sensation of heavy legs, significantly decreased in the chocolate group. The significance of this survey was **p<0.001**.

- erroneous statistically significant result (false positive)
- massaged (e.g. winsorized) data and *p*-hacking
- Bohannon: "You have to know how to read a scientific paper"

Sidekick: Simpson's paradox

- phenomenon in probability and statistics
- trend appears in several different groups of data
- but disappears or reverses when these groups are combined

Treatment X	**Dr. med. A**	**Dr. med. B**
Success Rate	83 %	46 %

Success Rate for	**Dr. med. A**	**Dr. med. B**
M	90 %	100 %
F	?	?

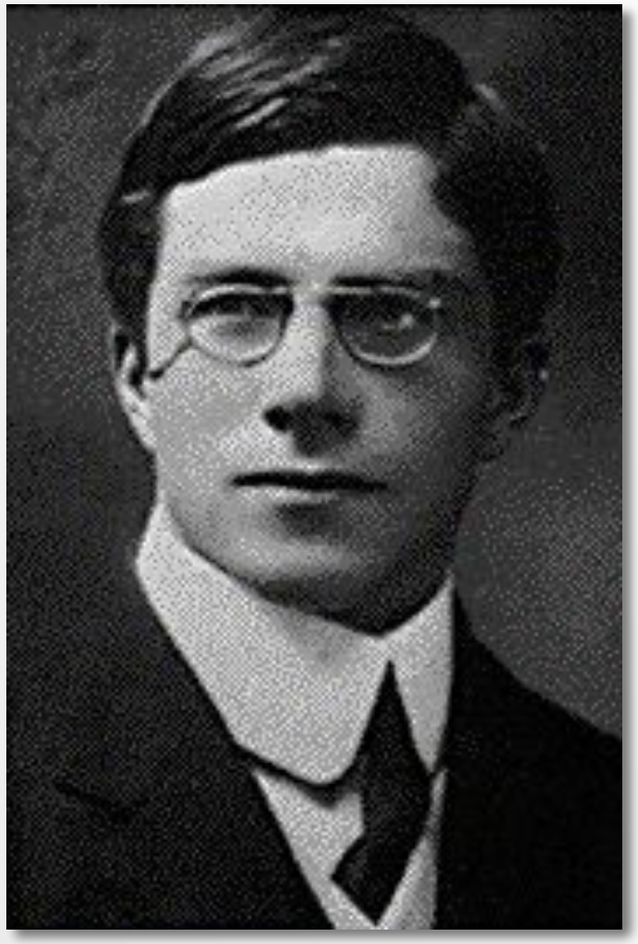
Dr. med. A		
#	**M**	**F**
Success	810	20
Failure	90	80
Total	900	100

Dr. med. B		
#	**M**	**F**
Success	100	360
Failure	-	540
Total	100	900

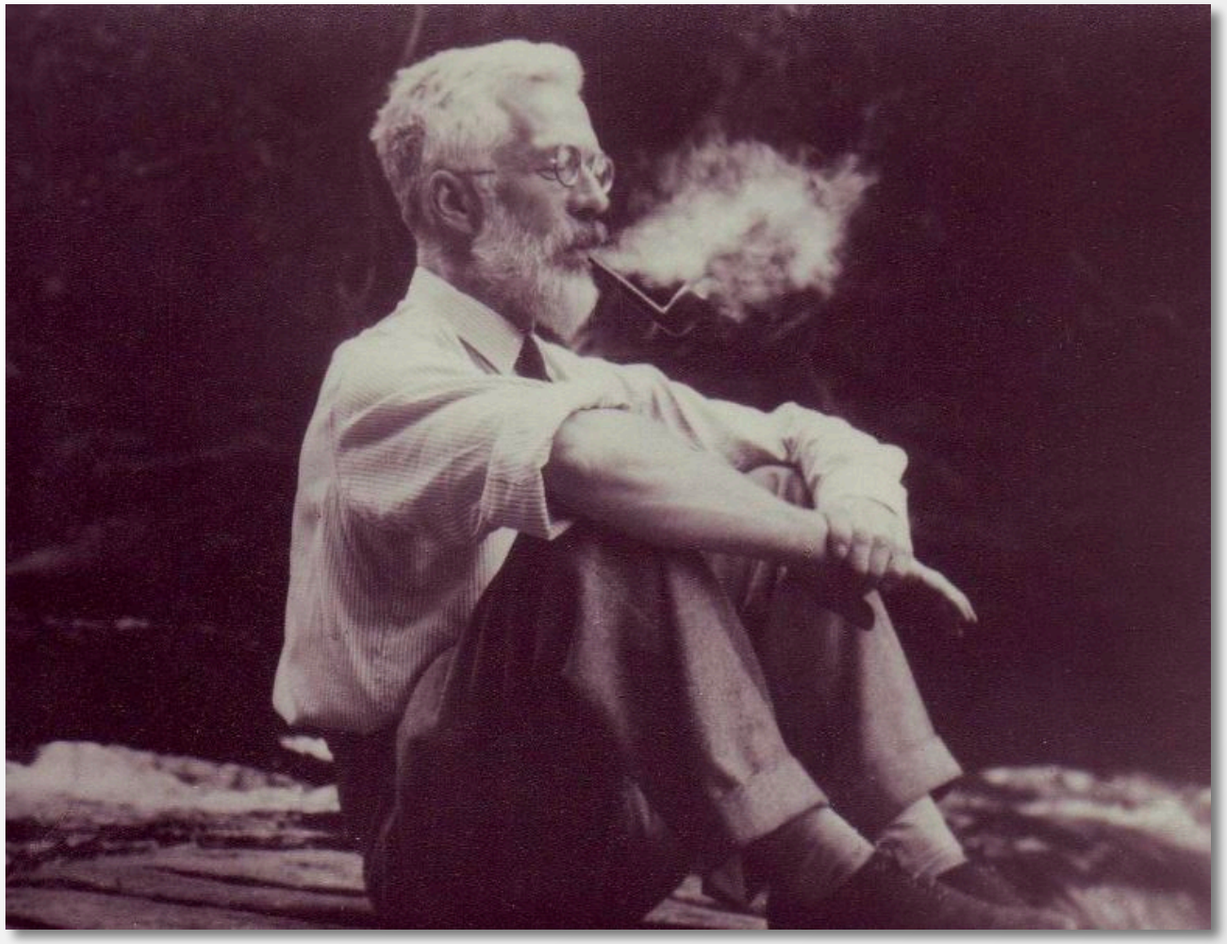
Theorem 1: Always question your results!

p-value (R. A. Fisher, 1925)

- "measure of the strength of the evidence against H_0" (*originally* no H_1)
- *objectively* separate findings of interest from noise
- get the observed value of the test statistic, or a value with even greater evidence against H_0, *if* H_0 *is actually true.*



Starting STAT studies



A few weeks later

"Most efficient" H-tests (Neyman & E. Pearson, 1933)

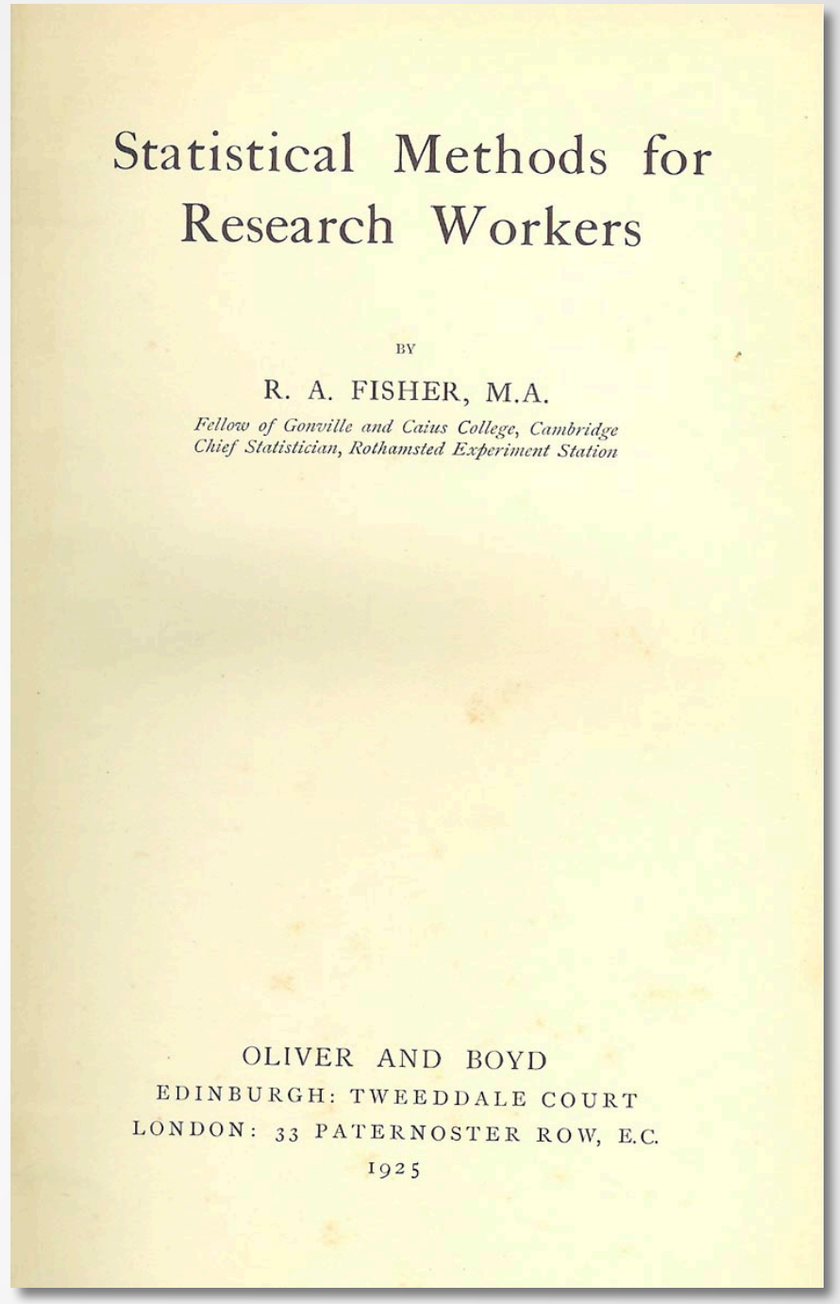
- focuses on *behavior/decision-making regime* between H_0 / H_1
- "[…] divide set by a system of ordered boundaries […] such that as we pass across one boundary and proceed to the next, we come to a class of results *which makes us more and more inclined*, on the information available, to reject the hypothesis tested in favour of alternatives which differ from it by increasing amounts […]"
- "[…] associate with each contour level the chance that, if [H_0] is true, a result will occur in random sampling lying beyond that level […] tells us nothing as to whether a *particular case* [H_0] is true […]"

Neyman & E. Pearson (son of K. Pearson)

Neyman; Pearson

Neyman–Pearson vs. Fisher

- Fisher (1925) contrasted use of the *p*-value for statistical inference with NP method ("Acceptance Procedures")
- Fisher (1925 & 1971): fixed *significance levels* especially $\alpha = 0.05$, $\alpha = 0.02$, $\alpha = 0.01$ *convenient*
- Fisher (1935): "Lady tasting tea"-experiment



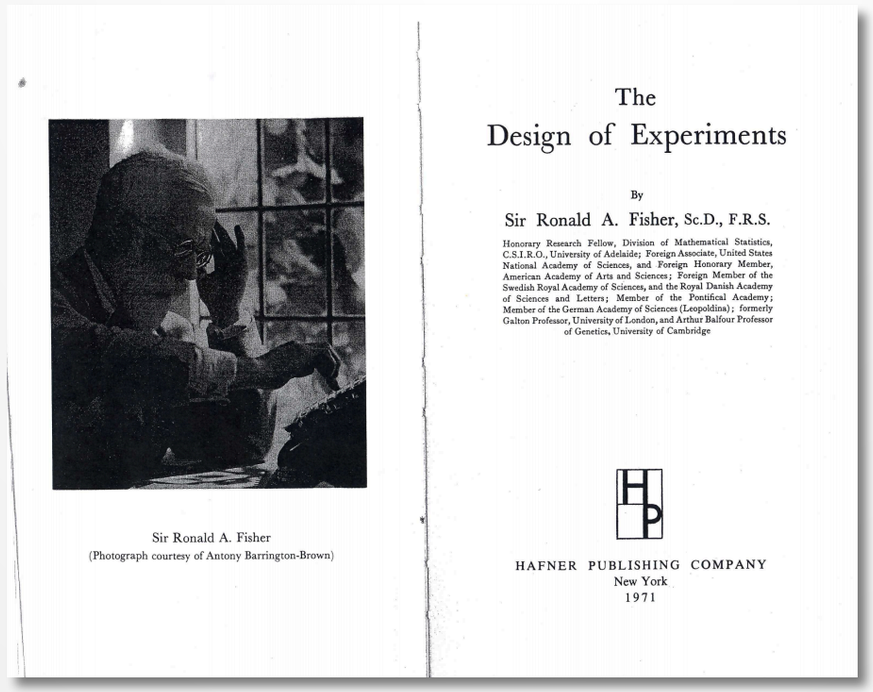
Statistical Methods for Research Workers

BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925



Sir Ronald A. Fisher
(Photograph courtesy of Antony Barrington-Brown)

The Design of Experiments

By

Sir Ronald A. Fisher, Sc.D., F.R.S.

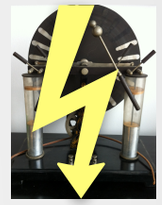
Honorary Research Fellow, Division of Mathematical Statistics, C.S.I.R.O., University of Adelaide; Foreign Associate, United States National Academy of Sciences, and Foreign Honorary Member, American Academy of Arts and Sciences; Foreign Member of the Swedish Royal Academy of Sciences, and the Royal Danish Academy of Sciences and Letters; Member of the Pontifical Academy; Member of the German Academy of Sciences (Leopoldina); formerly Galton Professor, University of London, and Arthur Balfour Professor of Genetics, University of Cambridge

HAFNER PUBLISHING COMPANY
New York
1971

Fisher

Purpose of the *p*-value

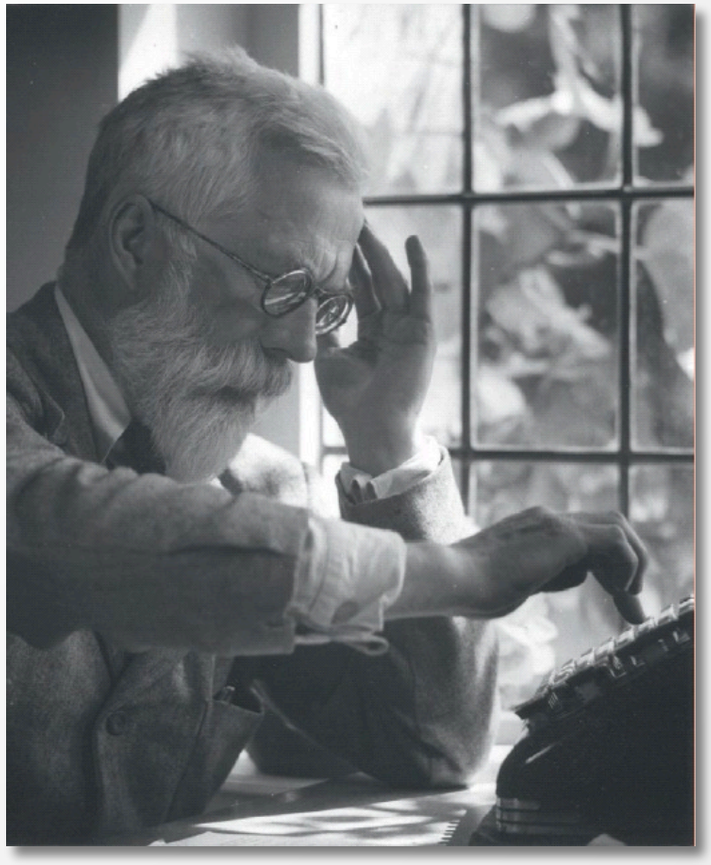
- "H_0 : *There is no difference in eating chocolate*, yet how *surprising* could be the experimental evidence in real life?"
- *p*-value is a quantitative tool to challenge our initial belief (H_0).
- an *spv* experimental scenario is called *statistically significant*



- standard experiment: fair coin flips with H_0: $\theta = 0.5$
- data = 1, 0, 1, 0, 1, 1, 1, 0, 0, … *B(n, θ)* distributed

Lady tasting tea

- Tea party in Cambridge in the *twenties*
- Muriel Bristol: can taste whether milk or tea first poured in cup
- Fisher: Sure, let's check for *spv*!



Lady tasting tea

- 8 randomly ordered cups of tea (70 variations in total)
- preparation methods
  - 4 prepared by *first* pouring the *tea*, *then* adding *milk*
  - 4 prepared by *first* pouring the *milk*, *then* adding the *tea*
- Select *only 4 cups* of *one preparation method*, e.g. "*all Tea #1*"

$$P(\text{"all correct"}) = \frac{1}{\text{"number of ways to guess"}} = \frac{1}{\binom{8}{4}} = \frac{1}{70} \approx 0.014$$

$k \geq 3 \succ$  (*k* = # *correct guesses*)

Truth \ Answer	Tea #1	Milk #1	Margin
Tea #1	3	1	4
Milk #1	1	3	4
Margin	4	4	8

Lady tasting tea

- H_0: lady *not able* to discriminate tea/milk first poured; $\theta = 0.5$
- chance of result at least $k \geq 3$ as favourable towards her claim = 24%
- all x_{ij} determined by x_{11}
- row & column sums n_1, n_0 & s_1, s_0 fixed; Truth/Answer independent
- x_{11} follows *hypergeometric distribution*

$$x_{11} \sim \text{Hypergeometric}(n, n_1, s_1)$$

$k \geq 3$ ≻ 

Truth \ Answer	Tea #1	Milk #1	Margin
Tea #1	x_{11}	x_{10}	n_1
Milk #1	x_{01}	x_{00}	n_0
Margin	s_1	s_0	n

Lady tasting tea

- Tea-Tasting distribution assuming H_0, e.g. "*choose all 4* tea #1"
  (o = correct cup chosen, x = correct not chosen)

Success	Selection Combinations	# Combinations
0	oooo	1 × 1 = **1**
1	ooox, ooxo, oxoo, xooo	4 × 4 = **16**
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = **36**
3	oxxx, xoxx, xxox, xxxo	4 × 4 = **16**
4	xxxx	1 × 1 = **1**
Total		**70**

- 5 possible outcomes

$$p(0) = \frac{1}{70}, \qquad p(1) = \frac{16}{70}, \qquad p(2) = \frac{36}{70}, \qquad p(3) = \frac{16}{70}, \qquad p(4) = \frac{1}{70}$$

- # successes follows *hypergeometric distribution*

Sidekick: Yates' continuity correction

- small samples ($n = 8$!): approximation not quite correct (CLT)
- Yates correction

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{n} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

- correction may tend to overcorrect (Type II error)

$$\chi^2_{\text{Yates}} = \frac{n(|ad - bc| - n/2)^2}{(a+c)(b+d)(a+b)(c+d)} = \frac{8(|9-1|-4)^2}{(4)(4)(4)(4)} = \frac{128}{256} = 0.5$$

Yates: $p = 0.23975$ Fisher: $p = 0.24286$

Significance testing

1. define data stochastics
2. formulate initial belief H_0
3. create a test statistic *t*
4. fix significance level α
5. define critical values = $F^{-1}(\alpha)$ under the "Null" H_0
6. Compare *t* with critical value, spv corresponds to " $F(t) \leq \alpha$ "
7. Does it make sense ?

Theorem 1: Always question your results!

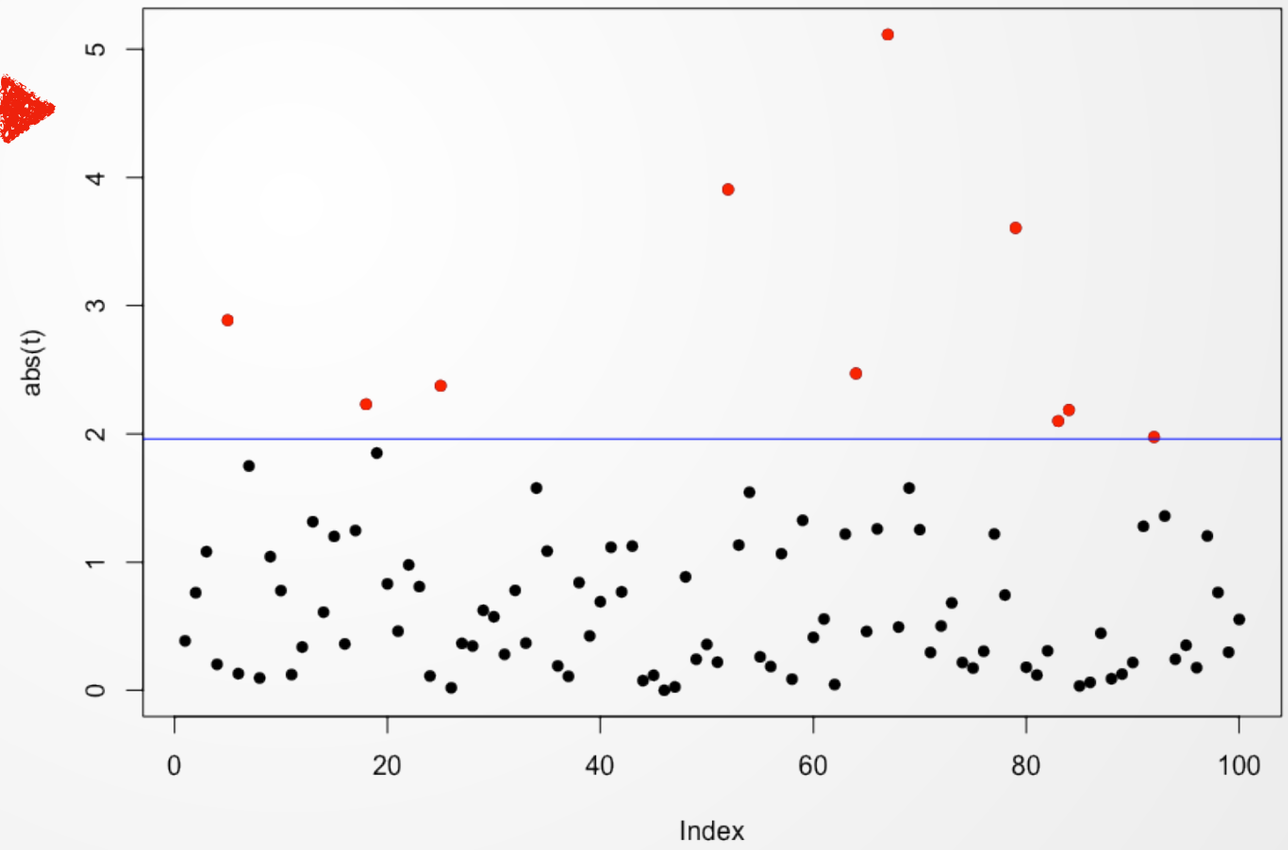
False Positives Experiment

```
n       = 10
J       = 100
p       = 0.025
q_p     = -qnorm(0.025, 0, 1)
t       = c( rep(0, J) )

for (j in 1:J){
  x   = runif(n)
  t[j]= sqrt(n) * ((mean(x) - 0.5)/sd(x))

  if (abs(t[j]) > q_p)
  {
    p_Value = pnorm(abs(t[j]))
    print(t[j])
    print(1 - p_Value)
  }
}

plot( abs(t), pch = 16 )
abline(h = q_p, col = "blue")
```



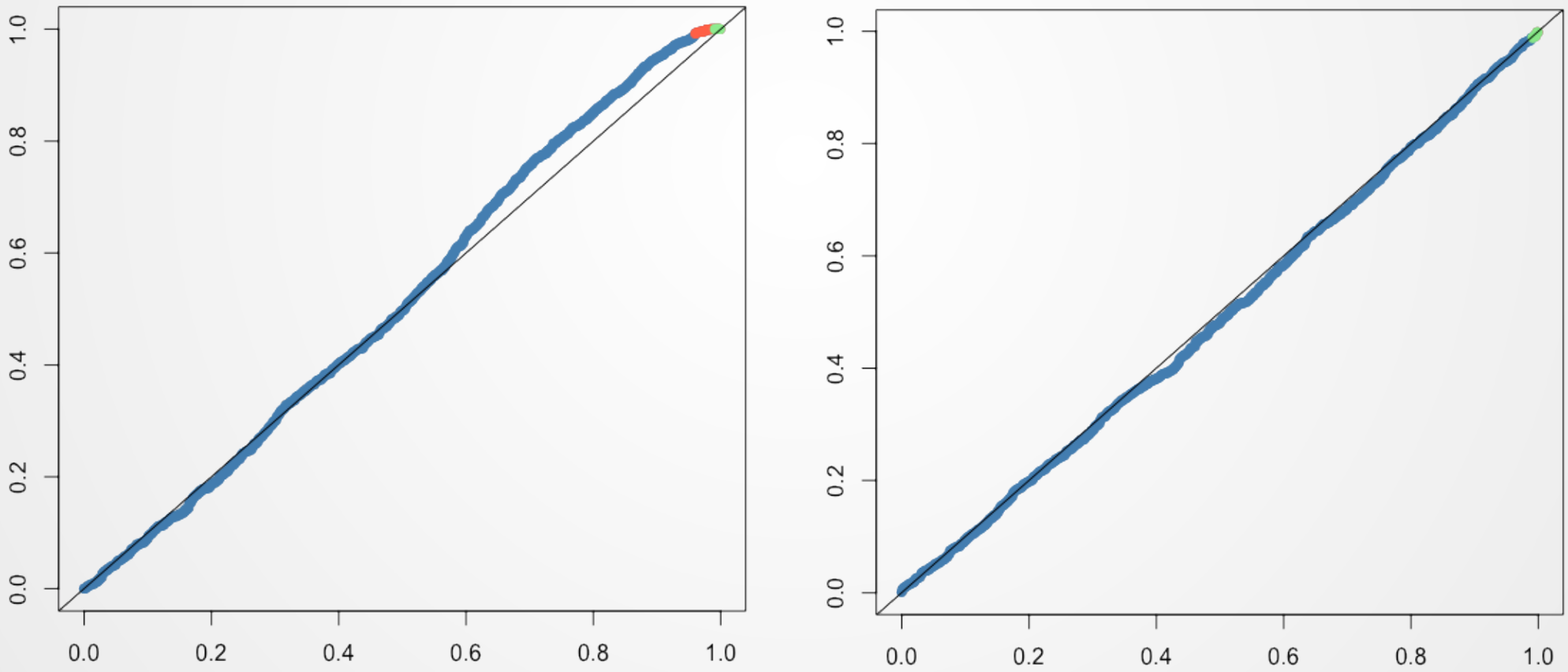
With higher n, P of crossing the q_p approaches 5%

Sidekick: Edgeworth Expansion

- Distribution of studentized t is approximated via normal cdf:

$$F_n(x) = \Phi(x) + n^{-1/2}\gamma(x)\varphi(x) + \mathcal{O}(n^{-1})$$

As $n \to \infty$, $F_n(x)$ "resembles" the Gaussian CDF more and more



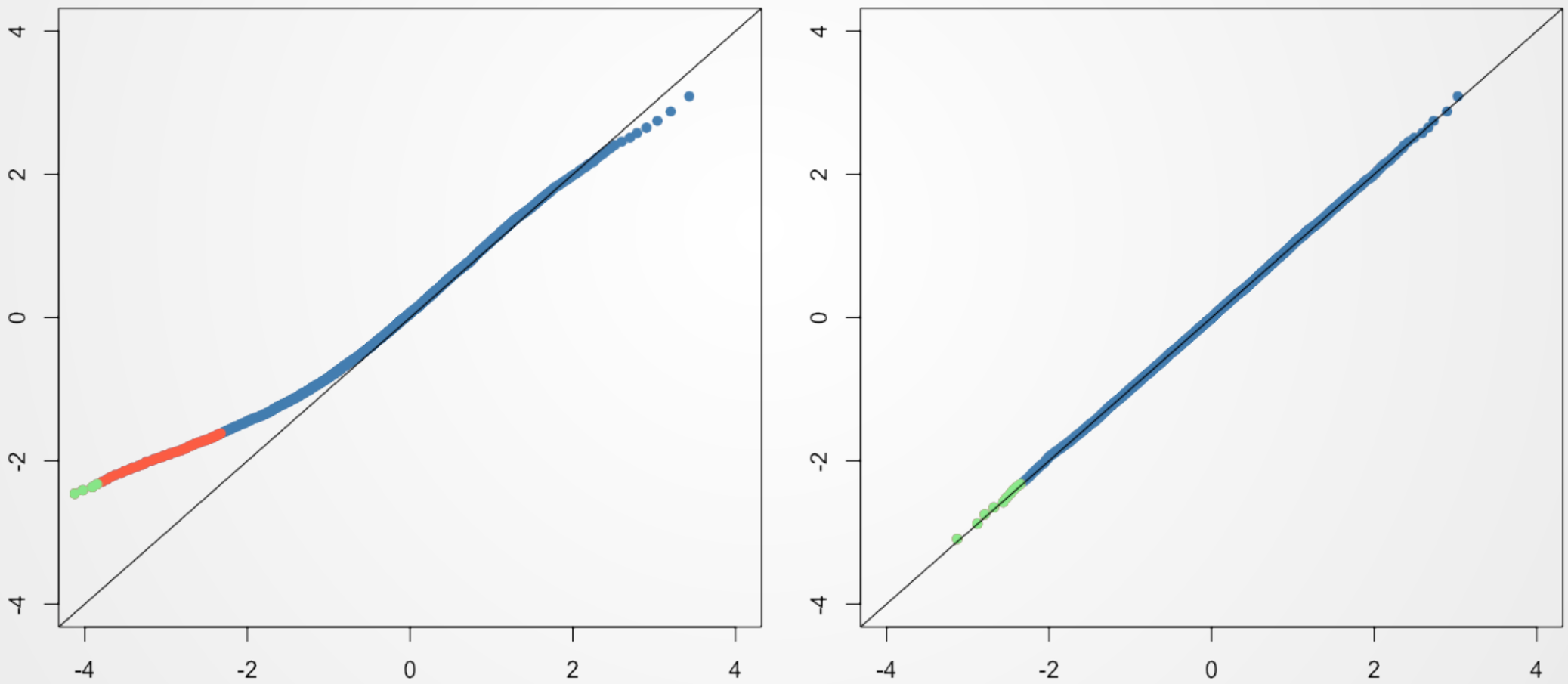
Figures: P-P plots for n=10 (left) and n=1000 (right)

For n=10, 4.9% of the points have a value above 0.99

Sidekick: Cornish-Fisher Expansion

- Critical values from the quantile function:

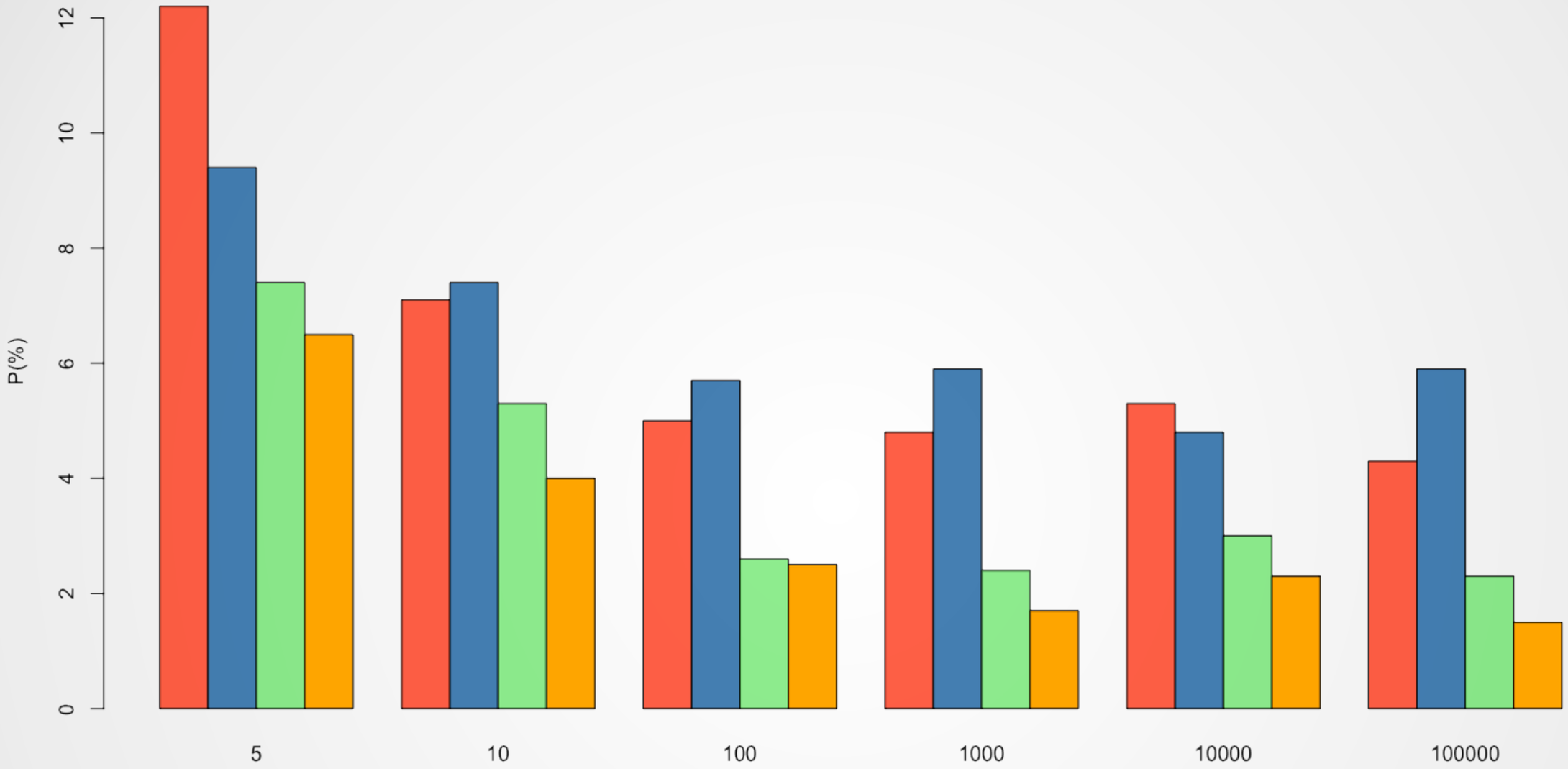
$$F_n^{-1}(\alpha) = \Phi^{-1}(\alpha) + n^{-1/2}\delta(x)\Phi^{-1}(\alpha) + \mathcal{O}(n^{-1})$$



Figures: Q-Q plots for n=10 (left) and n=1000 (right)

For n=10, 6.3% of the points are smaller than 99% critical value

R-based Experiment



Probabilities to randomly reject H_0 ($\mathsf{E}[X] = 0.5$ if uniform or $\mathsf{E}[X] = 0$ otherwise) in %

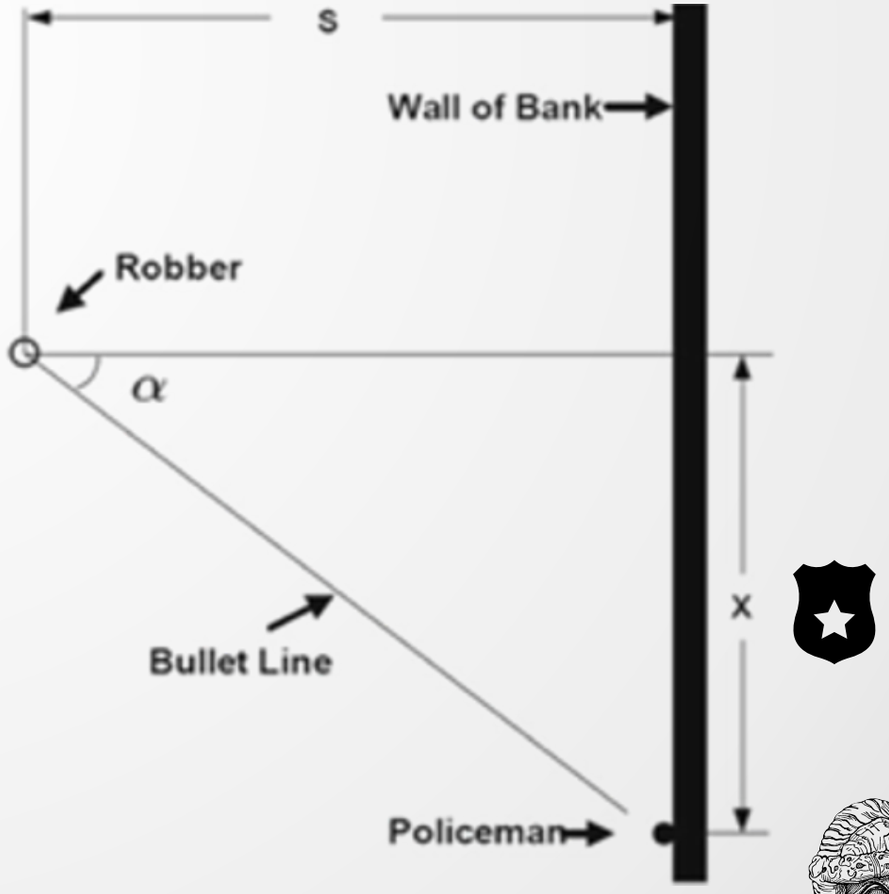
Uniform, Laplace, Student *t* (df=1) and Cauchy distributions given *n, J=1000*

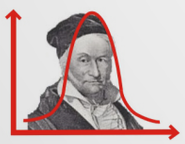
- Student *t* and Cauchy probabilities do not converge to 5%, as their variance is undefined

Sidekick: Cauchy Robber

- A bank robber shoots in uniform U(0,pi) angles on a wall
- A non STAT trained cop calc the mean of the bullet locations
- Cop is not happy, bullet hits are Cauchy distributed
- Cauchy pdf has no mean, moment,…
- The distance and the angle then follow the relationship:

$$\alpha = \arctan(\frac{x}{s})$$





Cauchy

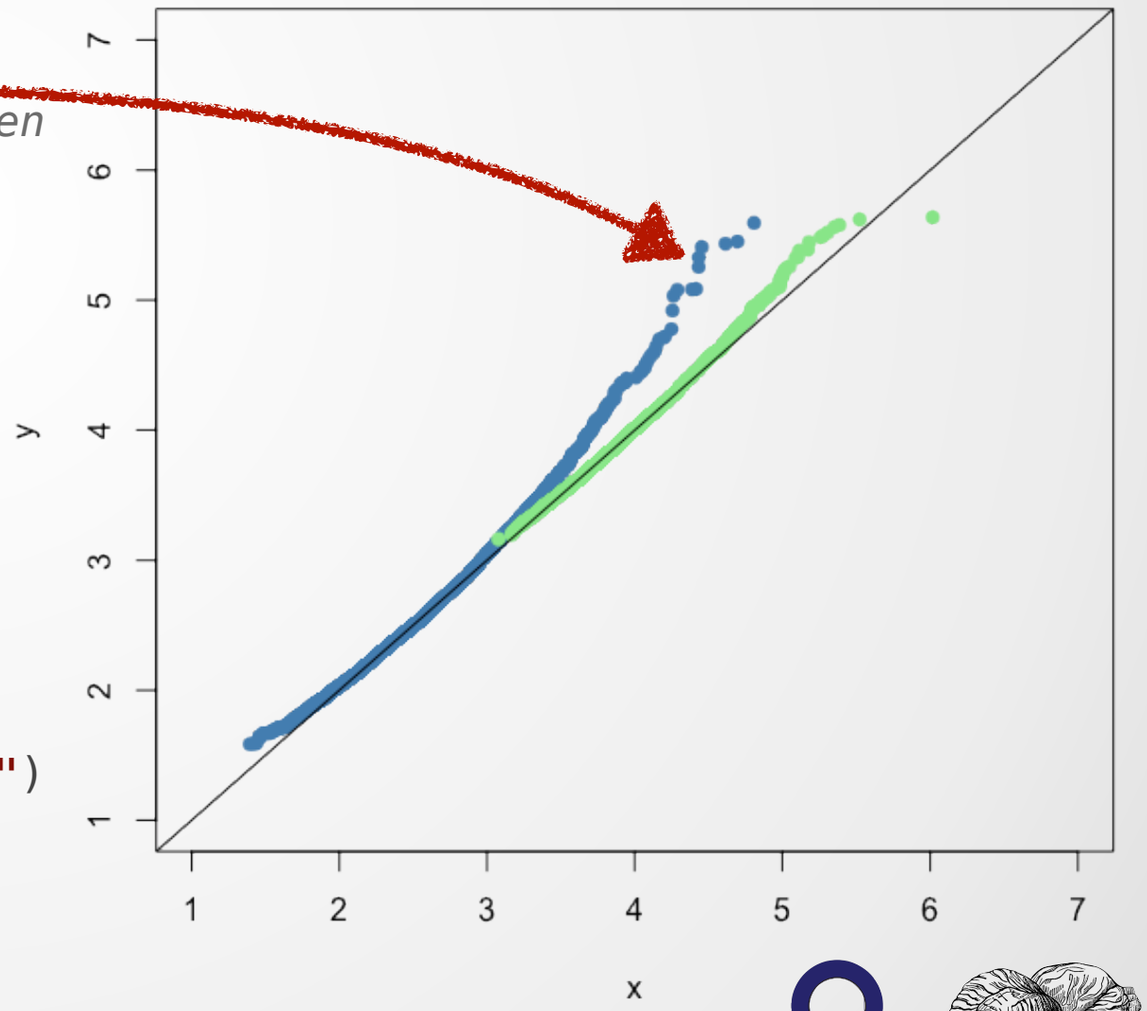
Sidekick: Extreme Value Theory

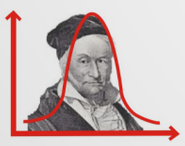
- The Gumbel distribution can be used to estimate the distribution of maxima/minima of n samples of certain distributions, however, convergence is slow

$$F\{x, \mu(n), \beta(n)\} = \exp\left\{ -\exp(-\frac{x - \mu(n)}{\beta(n)}) \right\}$$

```
library("evd")

n    = 100
J    = 10000
x_c = c( rep(0, J) )
for (i in 1:J){
    x       = max(rnorm(n))
    x_c[i] = x
}
mu_n    = qnorm(1 - (1/n))
beta_n = qnorm(1 - (1/n) * exp(-1)) - mu_n
y       = rgumbel(J, mu_n, beta_n)
plot(sort(x_c), sort(y), pch=16, col="steelblue")
lines(c(-10,10), c(-10,10))
```





Gumbel

Uniform confidence bands (CB)

- *CB* reflects uncertainty about the curve $m(\,.\,)$
- *confidence intervals (CI)* idem but for a point $m(x)$ on the curve
- *CI* = covers with $1-\alpha$ probability the true $m(x)$, x fixed (*narrower*)
- *CB* = covers the whole curve m(.) with $1-\alpha$ probability (*broader*)
- *CB* $\hat{m}(x) \pm w(x)$ with coverage probability $1-\alpha$

$$P\left\{ \sup_x |\hat{m}(x) - m(x)| / w(x) < c_\alpha \right\} = 1 - \alpha$$

Uniform confidence bands (cont.)

- *M-smoothers* $m_n(x)$: nonparametric curve estimators
- defined as zero (w.r.t. θ)

$$G_N(\theta) = (nh_n)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h_n\}\psi(Y_i - \theta)$$

- with $\psi : \mathbb{R} \to \mathbb{R}$ denoting bounded monotone, antisymmetric function with δ, $r(t)$, d_n being suitable scaling parameters

$$P\{(2\delta \log n)^{1/2} [\sup_t r(t)|\hat{m}(t) - m(t)|/\lambda(K)^{1/2} - d_n] < x\}$$

$$\underset{n\to\infty}{\to} \exp\{-2\exp(-x)\}$$

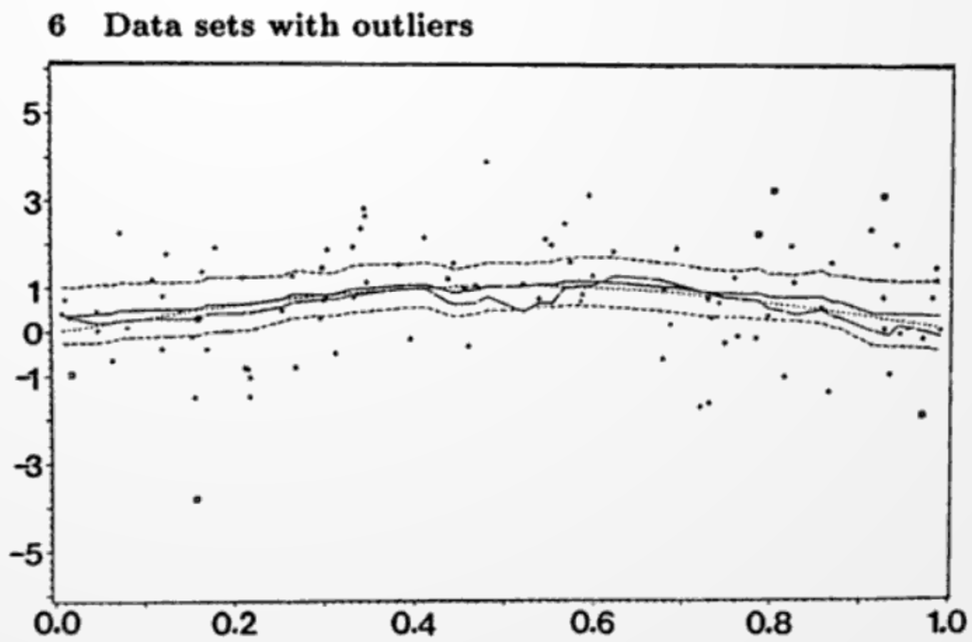
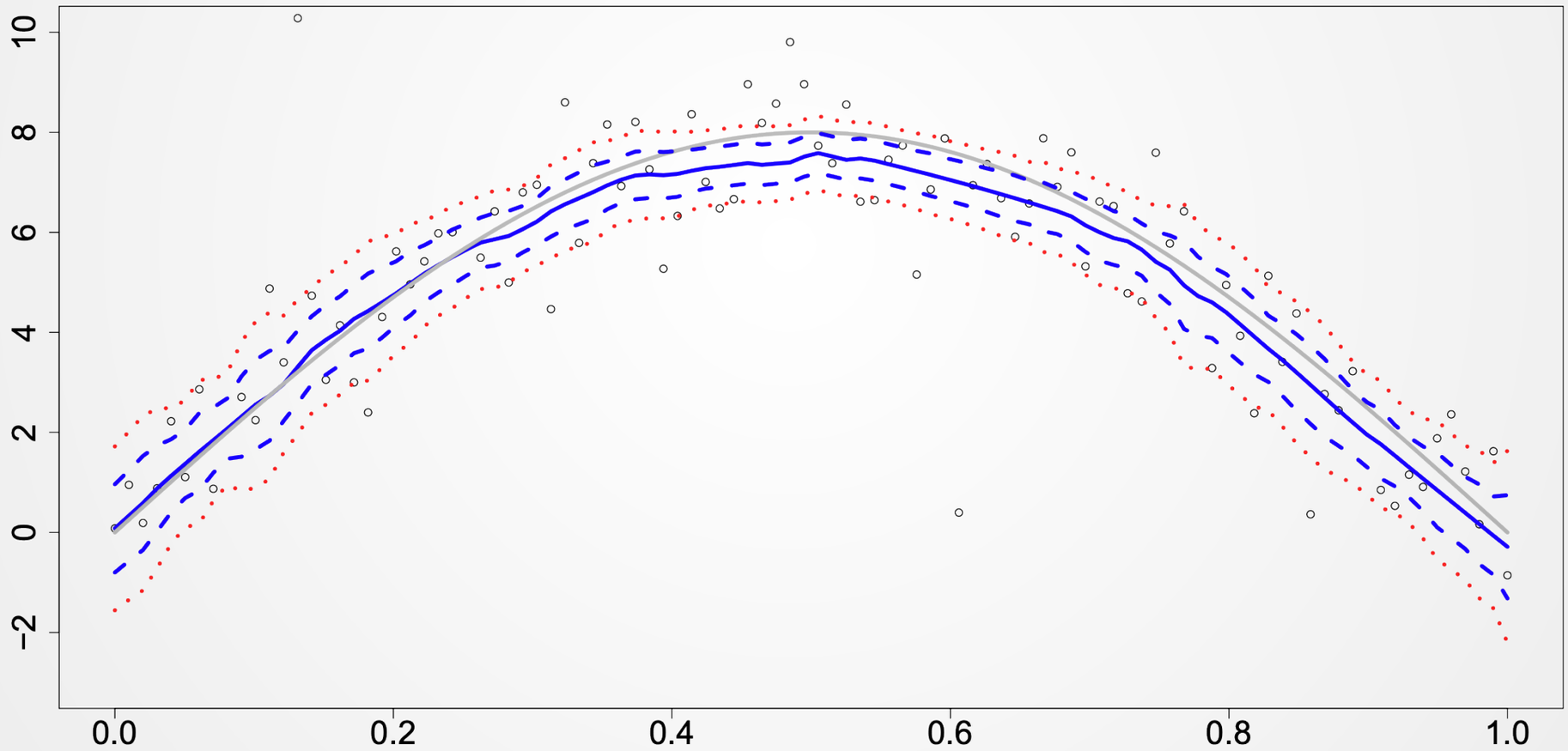


Figure 6.6. The kernel M-smoother with uniform confidence bands and the kernel smoother $\hat{m}_h(x)$. The original data are those of Figure 6.1. From Härdle (1989).

CB & Bootstrap

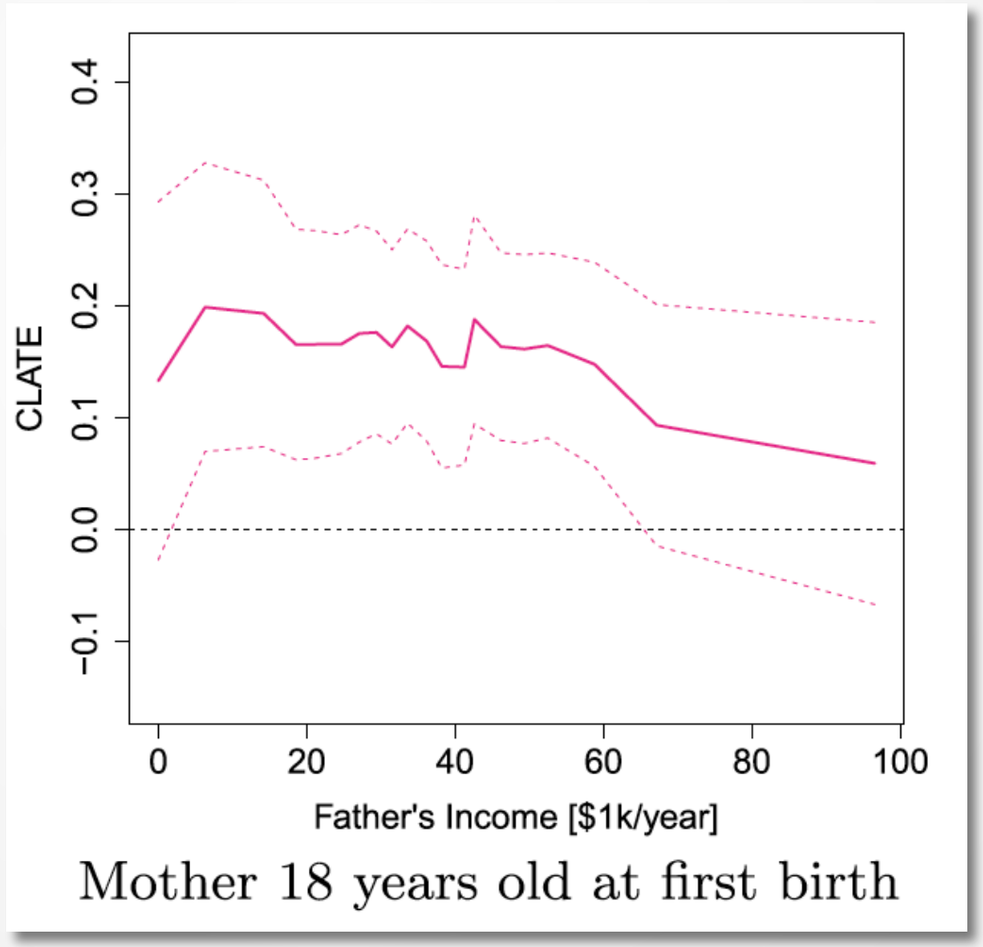
- CB precision improved by bootstrap (*even broader*)
- asymptotic theory methods = natural drawbacks w/ finite samples
- smoothed asymptotic CB = much lower coverage probability



Theoretical signal curve: **true curve**, robust estimate with Tukey biweight loss and 5% confidence bands (dashed), **local polynomial estimate**, bootstrap band, *n = 150*.

Modern application = Generalized RF's

- Susan Athey CI's represented / interpreted as CB's
- CB's could include 0 (approximated function insignificant)
- no existing framework to compute CB's

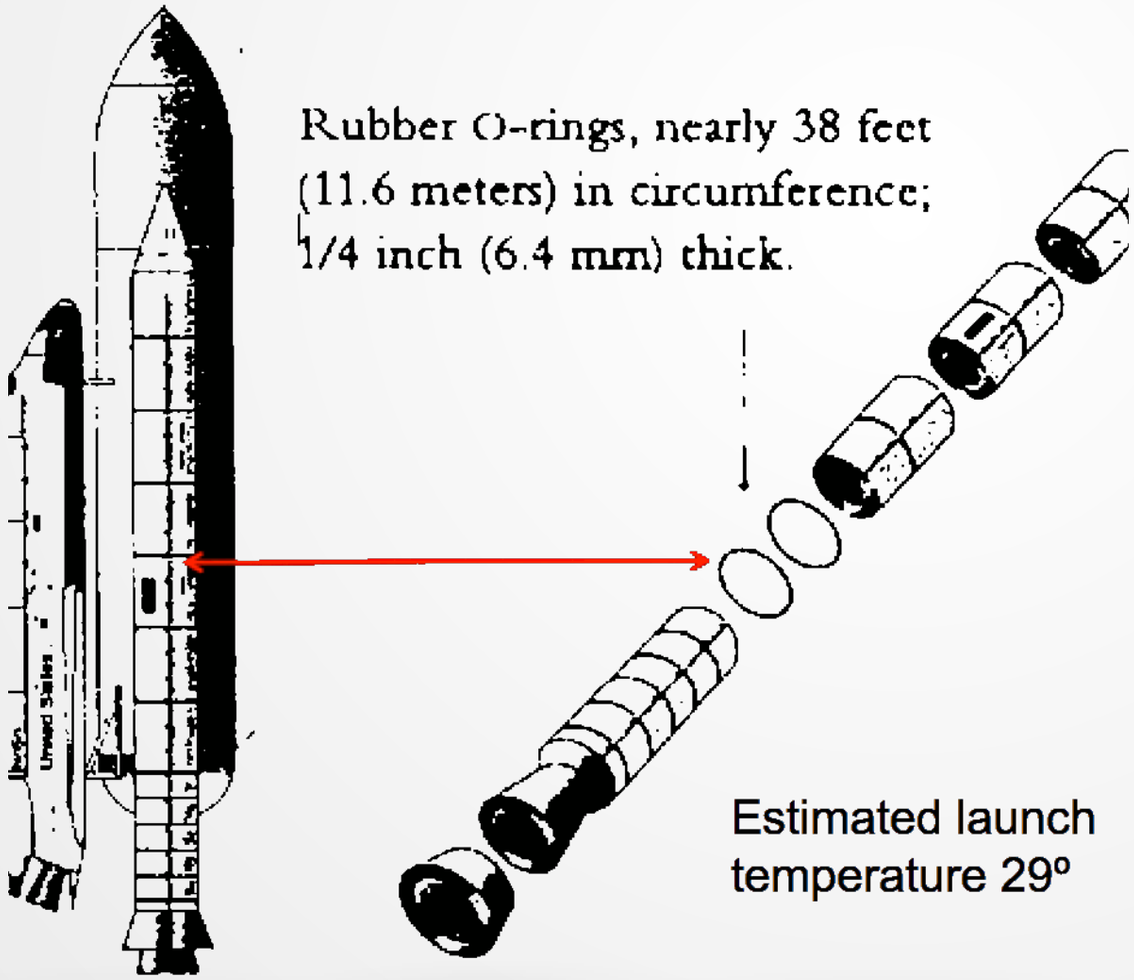


GRF-based treatment effect estimates with **pointwise** 95% CI that mother works for pay the year after third child

p-value vs. wetware survival

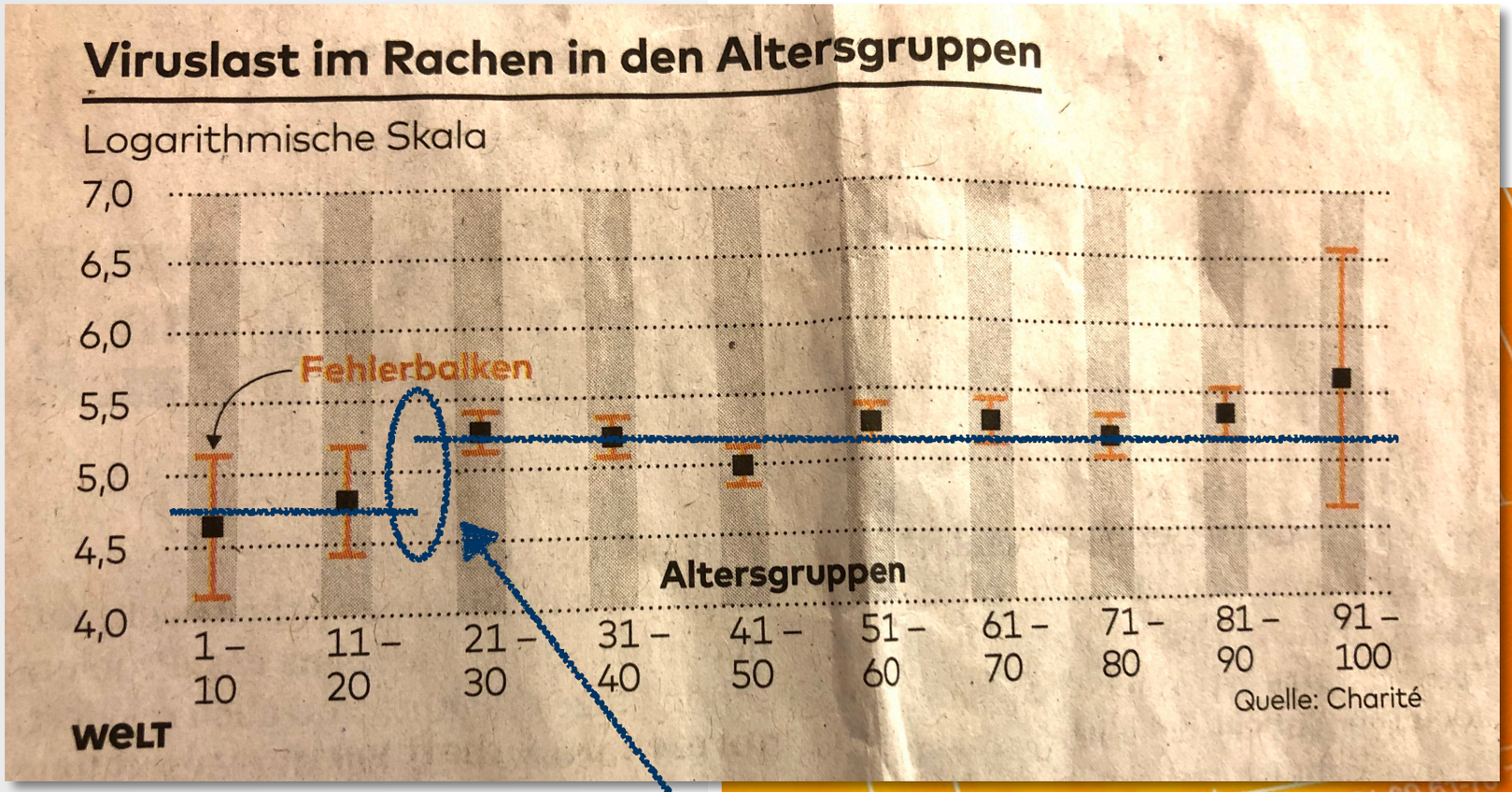
"It is *usual and convenient* for experimenters *to take 5 per cent as a standard level of significance*, […] *they are prepared to ignore all results which fail to reach this standard* […] *eliminate* […] *further discussion* […]"

R. A. Fisher



The Space Shuttle Challenger Explosion and the O-ring

p-value vs. nCOV-19 data controversy ?



WELT+ DEBATTE UM CORONA-STUDIE

Kinder, Viren, Schule. Die Drosten-Kontroverse

- „However, the authors' very own statistical analysis contradicts their central conclusion." (Prof. Dominik Liebl, Uni Bonn)
- „There are many good arguments against a quick reopening of schools, but the Charit´e study does not add to them." (Prof. Stoye, Cornell University)
- „The original analysis by Jones et al. (2020) suffers from small sample sizes among children and adolescents." (Prof. Leonhard Held, Uni Zürich)
- „Kinder haben in dieser Coronavirus Studie im Schnitt 67-85% weniger Viruslast als Erwachsene. Dass derart große Unterschiede von den Autoren als "nicht signifikant" eingestuft werden, liegt daran dass die verwendeten statistischen Methoden sehr schwach sind." (Prof. Christoph Rothe, Uni Mannheim)

Details, Source ; Deutschlandfunk Kultur, 26.05.2020

Details, Source ; Die Welt, 27.05.2020

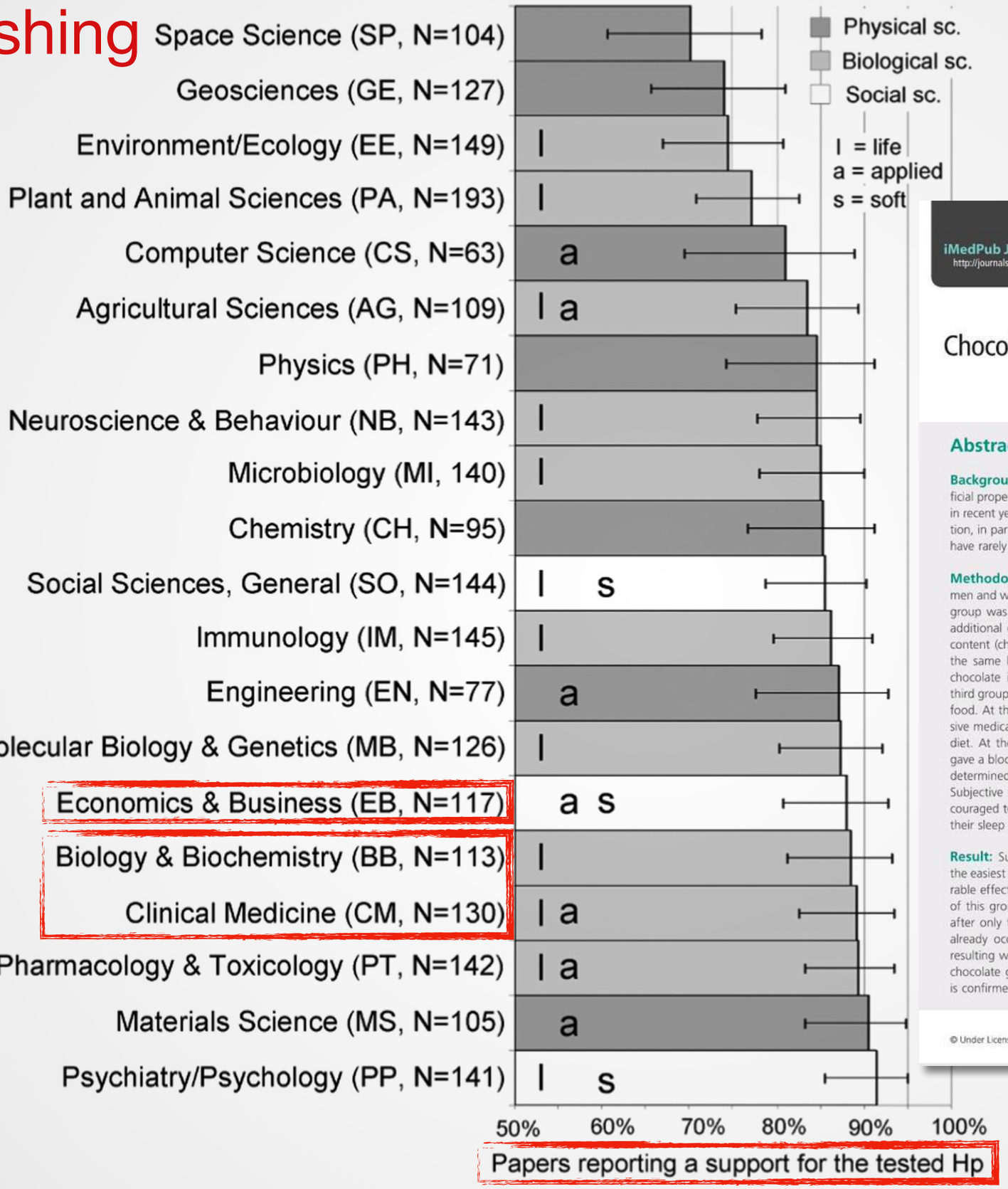
p-value vs. model selection

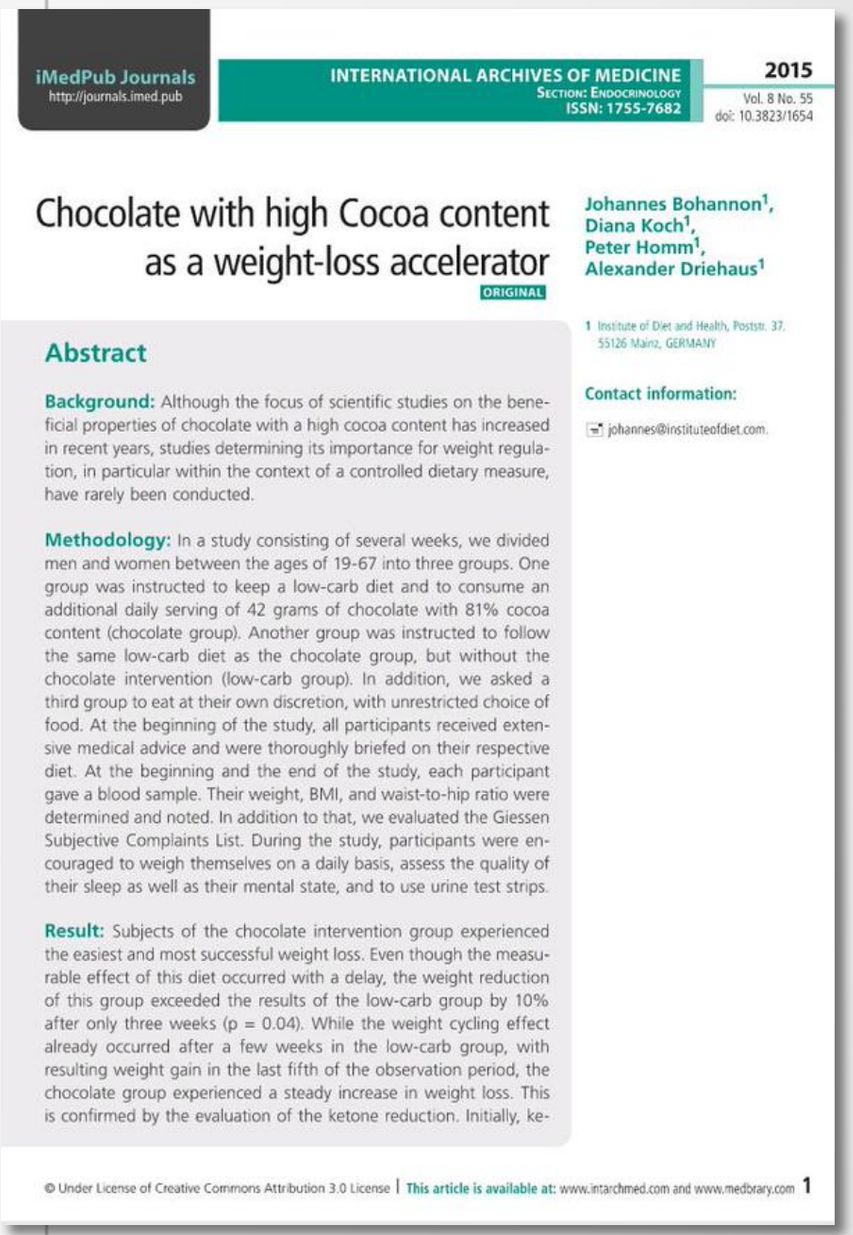
- variable selection using *p*-values *is* a bad idea
- *significant* variables get included, *insignificant* ones do not
- variables that are collinear (high correlation) = big *p*-values
  (does *not* mean the variables are not important)
- variable coefficient ≠ 0 will eventually ($n \to \infty$) have arbitrarily large test statistic and arbitrarily small *p*-value

Chocolate *p*-hacking

- 18 measurements = 60% chance of getting some spv result
- measurements not independent: probability can be even higher
- fiddling with the test's design (e.g. males/females, fiat/cryptocurrency etc.) and data to push p under 0.05

Publishing





iMedPub Journals
http://journals.imed.pub

INTERNATIONAL ARCHIVES OF MEDICINE
SECTION: ENDOCRINOLOGY
ISSN: 1755-7682

2015
Vol. 8 No. 55
doi: 10.3823/1654

Chocolate with high Cocoa content as a weight-loss accelerator

ORIGINAL

Johannes Bohannon[1], Diana Koch[1], Peter Homm[1], Alexander Driehaus[1]

1 Institute of Diet and Health, Poststr. 37, 55126 Mainz, GERMANY

Contact information:

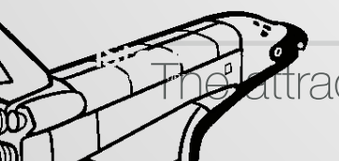
johannes@instituteofdiet.com.

Abstract

Background: Although the focus of scientific studies on the beneficial properties of chocolate with a high cocoa content has increased in recent years, studies determining its importance for weight regulation, in particular within the context of a controlled dietary measure, have rarely been conducted.

Methodology: In a study consisting of several weeks, we divided men and women between the ages of 19-67 into three groups. One group was instructed to keep a low-carb diet and to consume an additional daily serving of 42 grams of chocolate with 81% cocoa content (chocolate group). Another group was instructed to follow the same low-carb diet as the chocolate group, but without the chocolate intervention (low-carb group). In addition, we asked a third group to eat at their own discretion, with unrestricted choice of food. At the beginning of the study, all participants received extensive medical advice and were thoroughly briefed on their respective diet. At the beginning and the end of the study, each participant gave a blood sample. Their weight, BMI, and waist-to-hip ratio were determined and noted. In addition to that, we evaluated the Giessen Subjective Complaints List. During the study, participants were encouraged to weigh themselves on a daily basis, assess the quality of their sleep as well as their mental state, and to use urine test strips.

Result: Subjects of the chocolate intervention group experienced the easiest and most successful weight loss. Even though the measurable effect of this diet occurred with a delay, the weight reduction of this group exceeded the results of the low-carb group by 10% after only three weeks (p = 0.04). While the weight cycling effect already occurred after a few weeks in the low-carb group, with resulting weight gain in the last fifth of the observation period, the chocolate group experienced a steady increase in weight loss. This is confirmed by the evaluation of the ketone reduction. Initially, ke-

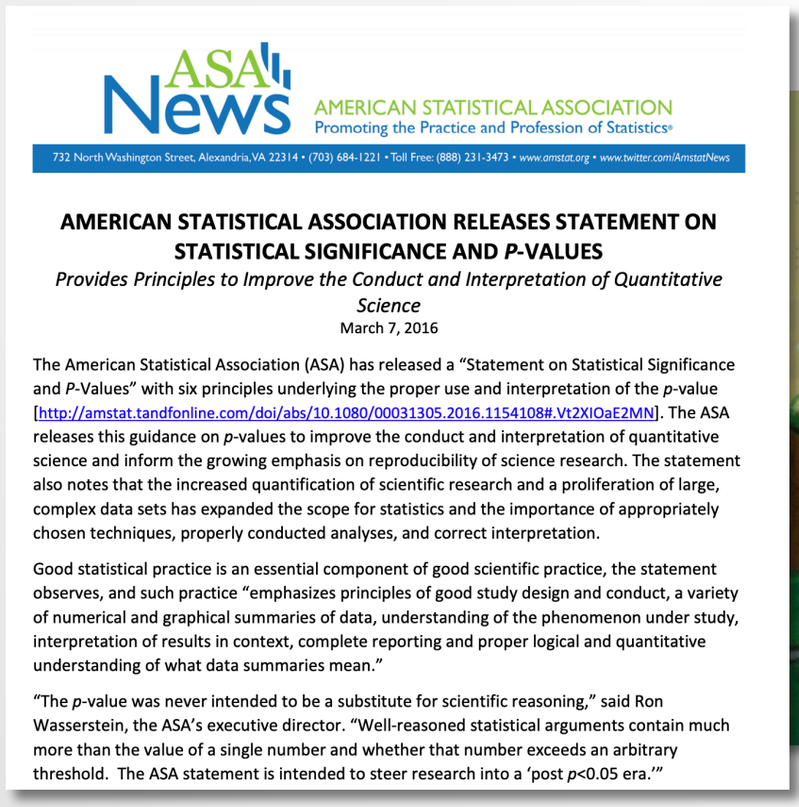


Why most FinE articles "support" stated H_0's ?

- better theories & hypotheses = better empirical findings ?
  - unlikely & hard to compare, c.f. particle physics
  - *HARKing* (Hypothesizing After the Results are Known)
- data manipulation & interaction effects researcher/hypothesis
  - un-/willingly, confirmation bias, data interpretation/handling
- lack of a replication/transparency culture
  - data disclosure rarely requested in top journals

ASA statement on *p*-values

- *p*-values can indicate problems data vs. specified statistical model
- *p*-values ≠ good measure of evidence for model/hypothesis.
- *p*-values ≠ measure P of hypothesis = true, data = produced by random chance alone.
- conclusions & decisions ≠ based *only* on *p*-values
- full reporting & transparency needed



ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

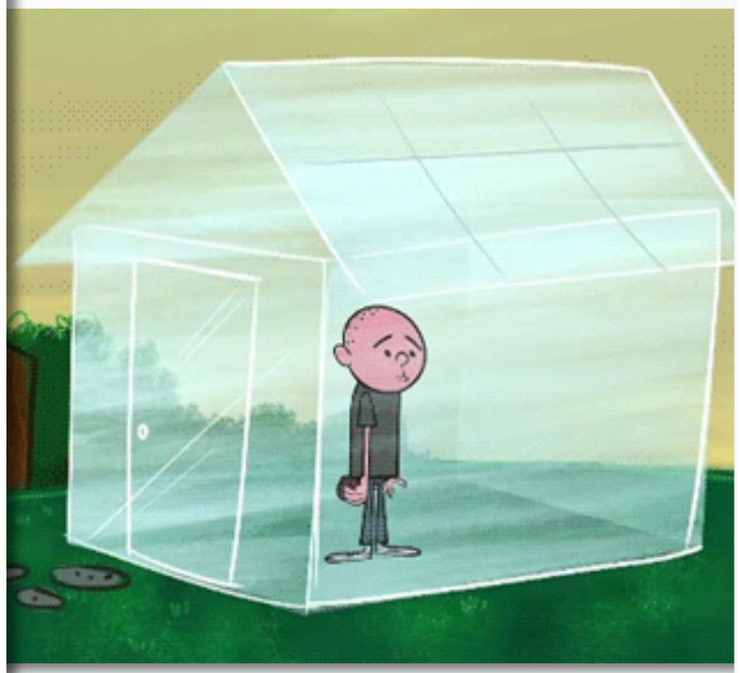
Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice "emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean."

"The *p*-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post p<0.05 era.'"



IN FOCUS NEWS

REPRODUCIBILITY

Statisticians issue warning on *P* values

Statement aims to halt missteps in the quest for certainty.

BY MONYA BAKER

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warned on 8 March. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the *P* value was being misapplied, in ways that cast doubt on statistics generally, he adds.

In its statement, the ASA advises researchers to avoid drawing scientific conclusions or making policy decisions purely on the basis of *P* values (R. L. Wasserstein and N. A. Lazar *Am. Stat.* http://doi.org/bc4d; 2016). Researchers should describe not only the data analyses that produce statistically significant results, the society says, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust.

Véronique Kiermer, executive editor of the Public Library of Science journals, says that the ASA's statement lends weight and visibility to longstanding concerns over undue reliance on the *P* value. "It is also very important in that it shows statisticians, as a profession, engaging cannot indicate the importance of a finding; for instance, a drug can have a statistically significant effect on patients' blood glucose levels without having a therapeutic effect.

Giovanni Parmigiani, a biostatistician at the Dana Farber Cancer Institute in Boston, Massachusetts, says that misunderstandings about what information a *P* value provides often crop up in textbooks and practice manuals. A course correction is long overdue, he adds. "Surely if this happened twenty years ago, biomedical research could be in a better place now."

FRUSTRATION ABOUNDS

Criticism of the *P* value is nothing new. In 2011, researchers trying to raise awareness about false positives gamed an analysis to reach a statistically significant finding: that listening to music by the Beatles makes undergraduates younger (J. P. Simmons *et al. Psychol. Sci.* **22,** 1359–1366; 2011). More controversially, in 2015, a set of documentary filmmakers published conclusions from a purposely shoddy clinical trial — supported by a robust *P* value — to show that eating chocolate helps people to lose weight. (The article has since been retracted.)

But Simine Vazire, a psychologist at the University of California, Davis, and editor of the journal *Social Psychological and Personality Science*, thinks that the ASA statement could help to convince authors to disclose all of the statistical analyses that they run. "To the extent that people might be sceptical, it helps to have statisticians saying, 'No, you can't interpret *P* values without this information," she says.

Correct interpretation of *p*-values

- *p*-values do not answer the question “Given the data, what is the probability that H_0 is true”, but rather “Given H_0, what is the probability of this or more extreme data”:

$$P(D \mid H_0) \neq P(H_0 \mid D)$$

- Example: Schizophrenia detection test
  - 1000 cases, H_0 - normal case, H_1 - schizophrenia
  - 2% of cases are schizophrenic
  - Test accuracy: 95% true positives, 97% true negatives
  - $P(D \mid H_0) = 0.05$
  - However:
  - $P(H_0 \mid D) = \dfrac{0.98 \times 0.03}{0.98 \times 0.03 + 0.02 \times 0.95} = 0.607$
  - The test performance is much worse than the *p*-value would suggest

"The Questionnaire" - testing for significance

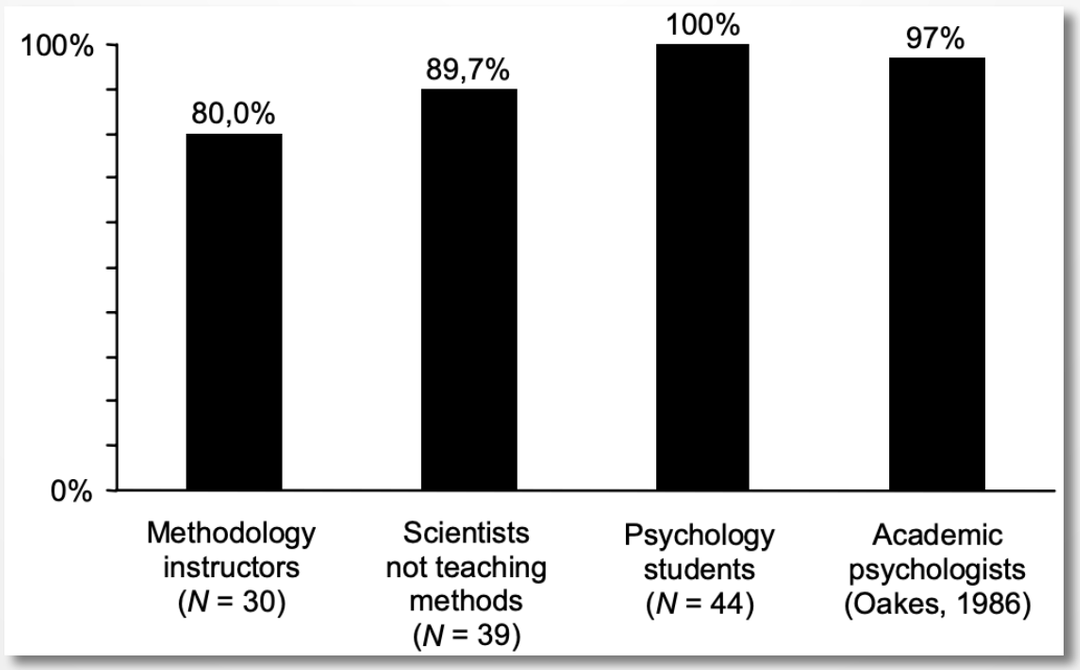
- treatment that may alter performance on a certain task.
  - *control & experimental groups* ($n_1 = n_2 = 20$)
  - independent means *t*-test ($t = 2.7$, $df = 18$, $p = 0.01$)
- which statements are "true", "false", or neither
  1. absolutely disproved the null hypothesis
  2. found the probability of the null hypothesis being true
  3. absolutely proved experimental/alternative hypothesis
  4. can deduce the prob of the experimental hypothesis being true
  5. know, if one decides to reject the null, the probability that one is making the wrong decision
  6. have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions
- several/none statements may be correct

“The Questionnaire” (cont.)

- 1, 3: significance tests can never *prove* (or *disprove*) H’s; only provide probabilistic information, at best *corroborate* theories.
- 2, 4, 5: generally impossible to assign a probability to any hypothesis by applying significance tests
  - (1, 3) can not assign a probability of 1
  - (2, 4) can not assign any other probability
  - (1, 2, 3, 4) statements about probabilities of H’s only possible in alternative approach of Bayesian statistics.
  - (5) similar to definition of *Type I error*, but having rejected H_0, decision would be wrong, if and only if the H_0 is true
  - (2, 5) P in (5) “making the wrong decision” = $p(H_0)$; P as (2) cannot be derived with H_0 significance testing (NHST)
- 6: “replication fallacy” …

“The Questionnaire” (cont.)

- 6: “replication fallacy” …
  - NP: interpret $p = 0.01$ = relative frequency of rejections of H_0 if H_0 is true, *here*: no evidence of H_0 being true
  - “In the minds of many, $1 - p$ erroneously turned into the relative frequency of rejections of H_0, that is, into the probability that significant results could be replicated” (Gigerenzer, 1993)
  - “The level of significance measures the confidence that the results of the experiment would be repeatable under the conditions described” (A.W. Melton, 1962)



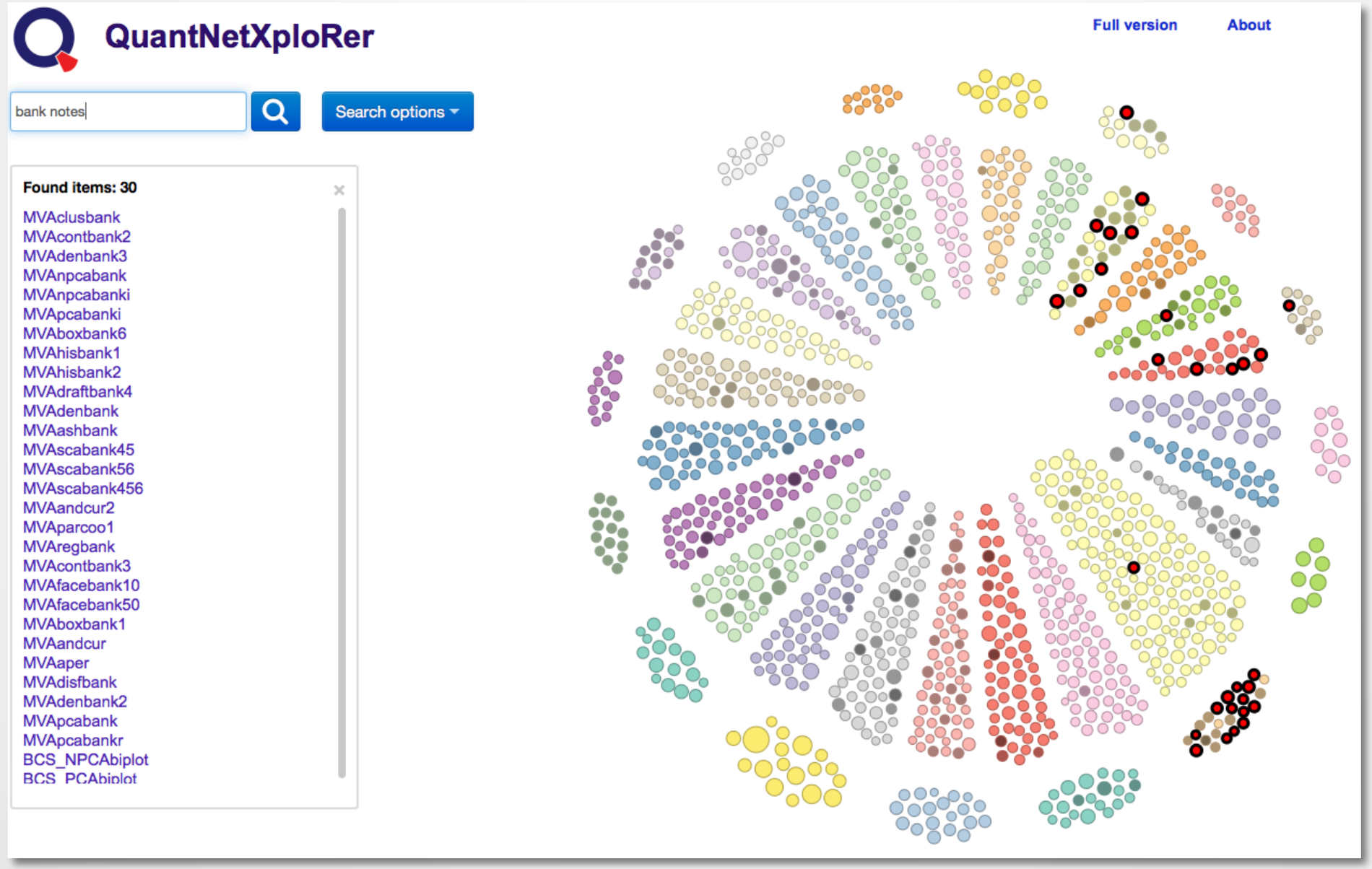
Participants % who made at least one mistake; comparison to Oakes’ original study (1986)

Further literature

- Loftus, G.R. (1991):
  On the tyranny of hypothesis testing in the social sciences.
- Gigerenzer, G. (1993):
  Über den mechanischen Umgang mit statistischen Methoden.
- Cohen, J. (1994):
  The earth is round (p < .05).
- Loftus, G.R. (1994):
  Why psychology will never be a real science until we change the way we analyze data.
- Dar, R., Serlin, D., & Omer, H. (1994):
  Misuse of statistical tests in three decades of psychotherapy research.
- Falk, R., & Greenbaum, W. (1995):
  Significance tests die hard.

Quantlets move science: Be Q-able!

- Quantnet: share *data* and *programs - easy!*
- *Collaboration* via seamless GitHub integration
- Boosting *transparent* and *reproducible* science



QuantLet / pHacking

Watch 3 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Actions | Projects 0 | Wiki | Security 0 | Insights

Branch: master ▾ pHacking / phacking_visualisation / Create new file | Upload file

QuantLetTeam created README.md (automatically) Latest commit 4b

..	
Metainfo.txt	init upload
README.md	created README.md (automatically)
phacking.R	init upload
phacking.png	init upload

README.md

phacking

```
Name of Quantlet : phacking

Published in :  'Dei ex machinis or the attractiveness of p-hacking'

Description : 'Data visualisation for phacking talk'

Keywords : phacking, data, noise, noisy data, hacking, hack, high frequency data, cryptocurrency

See also : quantlet
```

QuantNetXploRer

phack

Search options

Found items: 1

phacking

phacking

Cluster Name: data, regression, linear, mining, text
Elements in Cluster: 95

Size of Metainfo in Byte: 425

Software: R
Book/Project: Dei ex machinis or the attractiveness of p-hacking - pHacking

We choose to go to the Moon …

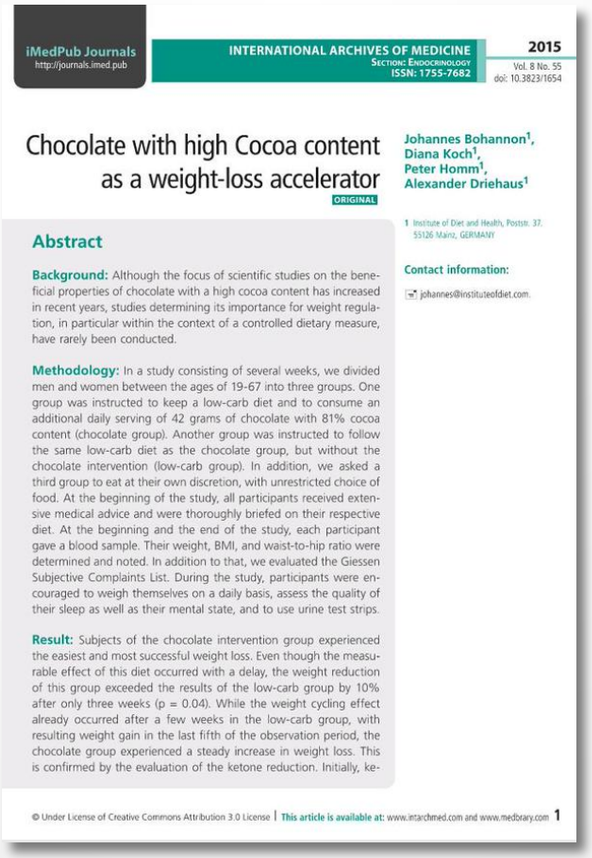
" […] *not because [it is] easy, but because [it is] hard*; because that goal will serve to organize and measure the best of our energies and skills, because that challenge is one that we are willing to accept, one we are unwilling to postpone, and one we intend to win […] ."

The *p*'ers

IA

WKH

R2



iMedPub Journals
http://journals.imed.pub

INTERNATIONAL ARCHIVES OF MEDICINE
SECTION: ENDOCRINOLOGY
ISSN: 1755-7682

2015
Vol. 8 No. 55
doi: 10.3823/1654

Chocolate with high Cocoa content as a weight-loss accelerator

ORIGINAL

Johannes Bohannon[1], Diana Koch[1], Peter Homm[1], Alexander Driehaus[1]

1 Institute of Diet and Health, Postst. 37, 55126 Mainz, GERMANY

Contact information:

johannes@institutediet.com

Abstract

Background: Although the focus of scientific studies on the beneficial properties of chocolate with a high cocoa content has increased in recent years, studies determining its importance for weight regulation, in particular within the context of a controlled dietary measure, have rarely been conducted.

Methodology: In a study consisting of several weeks, we divided men and women between the ages of 19-67 into three groups. One group was instructed to keep a low-carb diet and to consume an additional daily serving of 42 grams of chocolate with 81% cocoa content (chocolate group). Another group was instructed to follow the same low-carb diet as the chocolate group, but without the chocolate intervention (low-carb group). In addition, we asked a third group to eat at their own discretion, with unrestricted choice of food. At the beginning of the study, all participants received extensive medical advice and were thoroughly briefed on their respective diet. At the beginning and the end of the study, each participant gave a blood sample. Their weight, BMI, and waist-to-hip ratio were determined and noted. In addition to that, we evaluated the Giessen Subjective Complaints List. During the study, participants were encouraged to weigh themselves on a daily basis, assess the quality of their sleep as well as their mental state, and to use urine test strips.

Result: Subjects of the chocolate intervention group experienced the easiest and most successful weight loss. Even though the measurable effect of this diet occurred with a delay, the weight reduction of this group exceeded the results of the low-carb group by 10% after only three weeks ($p = 0.04$). While the weight cycling effect already occurred after a few weeks in the low-carb group, with resulting weight gain in the last fifth of the observation period, the chocolate group experienced a steady increase in weight loss. This is confirmed by the evaluation of the ketone reduction. Initially, ke-

The attractiveness of *p*-hacking

Thank you!

谢谢

Ladislaus von Bortkiewicz Professor of Statistics
Humboldt-Universität zu Berlin
BRC Blockchain Research Center
lvb.wiwi.hu-berlin.de
Charles University, WISE XMU, NCTU 玉山学者

References

- Wang et al. (2020) Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA. 2020 Feb 7. doi: 10.1001/jama.2020.1585. [Epub ahead of print] https://www.ncbi.nlm.nih.gov/pubmed/32031570
- iMedPub Journal “Chocolate” https://web.archive.org/web/20181004164627/https://www.scribd.com/doc/266969860/Chocolate-causes-weight-loss
- ARTE Chocolate Documentary https://info.arte.tv/de/schlank-durch-schokolade
- ASA on P-Values https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf
- Nature on P-Values https://www.nature.com/news/polopoly_fs/1.19503!/menu/main/topColumns/topLeftColumn/pdf/nature.2016.19503.pdf
- Cocoa Bean & Machete https://chocolatephayanak.com/the-chocolate-making-process/
- Fanelli, Daniele, 2010, “Positive” results increase down the Hierarchy of the Sciences, PLoS ONE 5, e10068
- Auguste Compte https://commons.wikimedia.org/wiki/Category:Auguste_Comte
- Tea Tasting https://www.eastamericangroup.com/products/beverages/tea-tisanes/
- Neyman & Pearson https://errorstatistics.com/2018/12/01/neyman-pearson-tests-an-episode-in-anglo-polish-collaboration-excerpt-from-excursion-1-3-2/

- Silke Janitza, Harald Binder, Anne-Laure Boulesteix (2014) Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications. https://epub.ub.uni-muenchen.de/21889/7/TR.pdf
- Wolfgang Härdle. 9. April 1992. Applied Nonparametric Regression. Econometric Society Monographs, Band 19.
- Härdle, W, Ritov, Y, Wang, W (2013) Tie the Straps: Uniform Bootstrap Confidence Bands for Bounded Influence Curve Estimators. J. Multivariate Analysis, 134, 129-145, doi: https://doi.org/10.1016/j.jmva.2014.11.003