UNIVERSITÄT BONN

Einführung in Bayes Statistik

Replizierbarkeit und Alternativen zu *p*-Werten

Wintersemester 2017/18

Christopher Harms

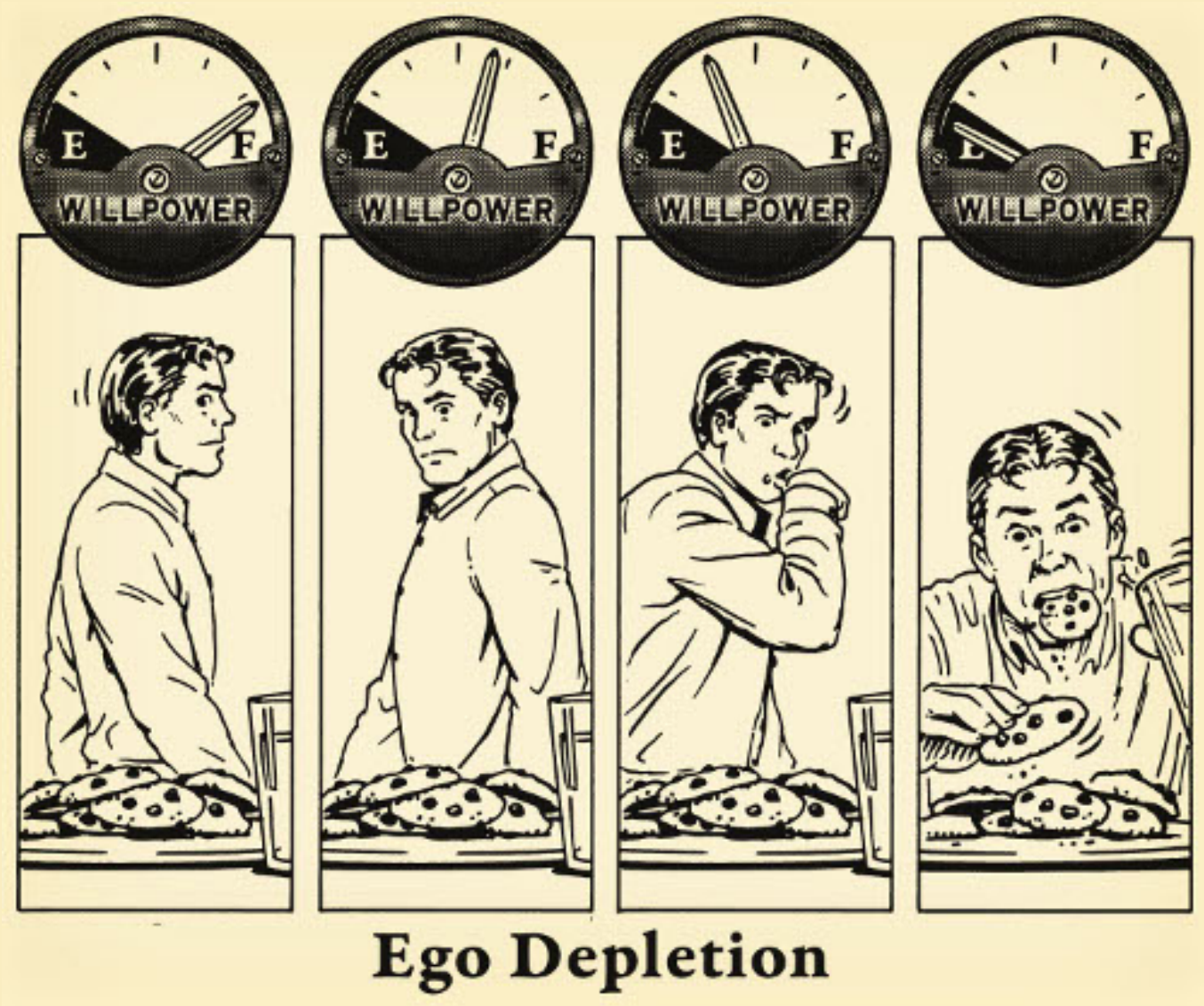
christopher.harms@uni-bonn.de

Institut für Psychologie, Abteilung Methodenlehre, Diagnostik, Evaluation

Elderly Priming

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology, 71, 230-244.
Replication Attempts: http://www.psychfiledrawer.org/replication.php?attempt=MTU%3D

Ego Depletion



Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*, 823. http://doi.org/10.3389/fpsyg.2014.00823

Power Posing

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance. *Psychological Science*, *21*(10), 1363–1368. http://doi.org/10.1177/0956797610383437

Replication Attempts & Re-Analyse: http://datacolada.org/37

Wir haben ein Problem.

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

100 Studien aus drei Top-Journals im Jahrgang 2008 wurden untersucht.

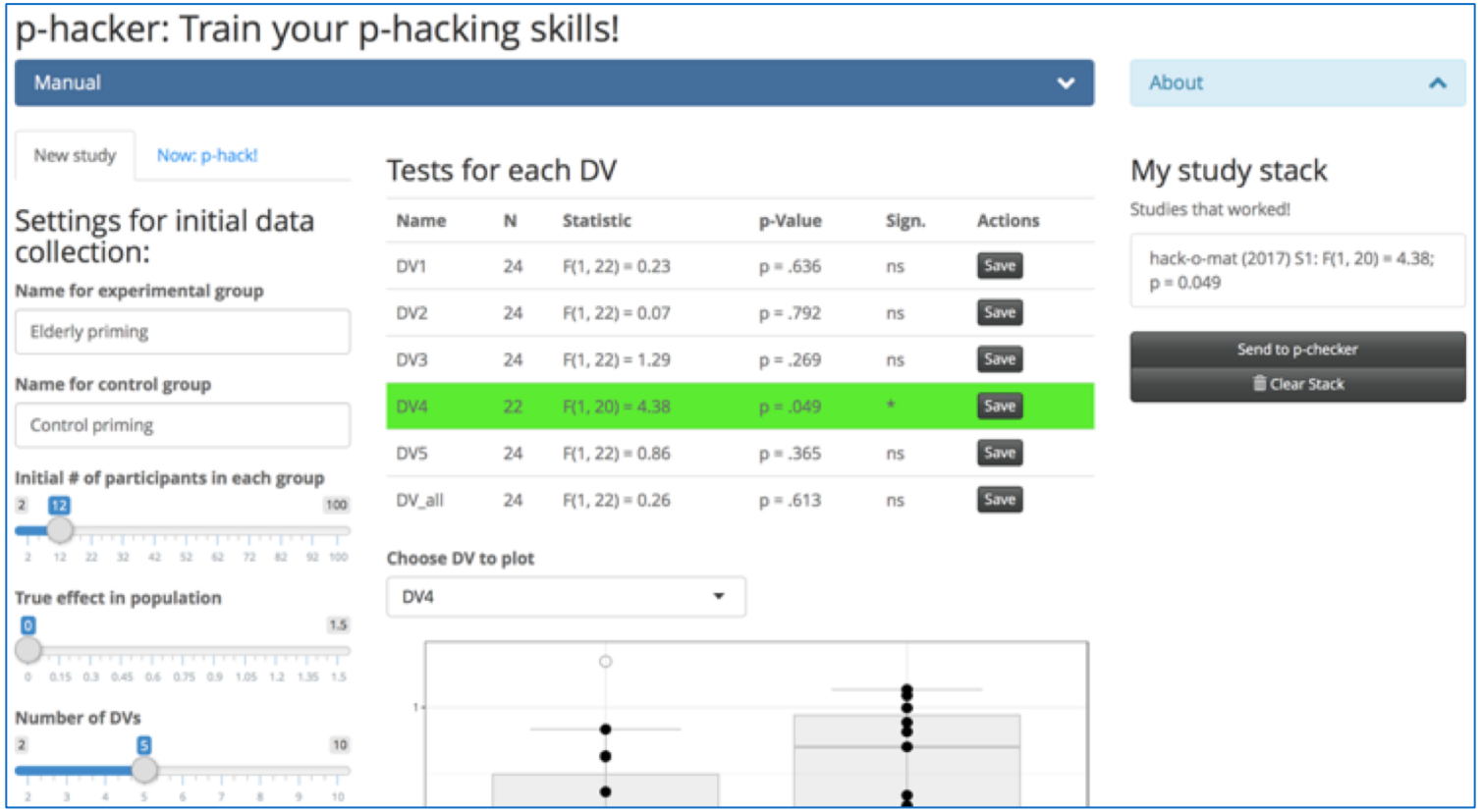
Nur 39 ließen sich erfolgreich replizieren.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-aac4716. http://doi.org/10.1126/science.aac4716

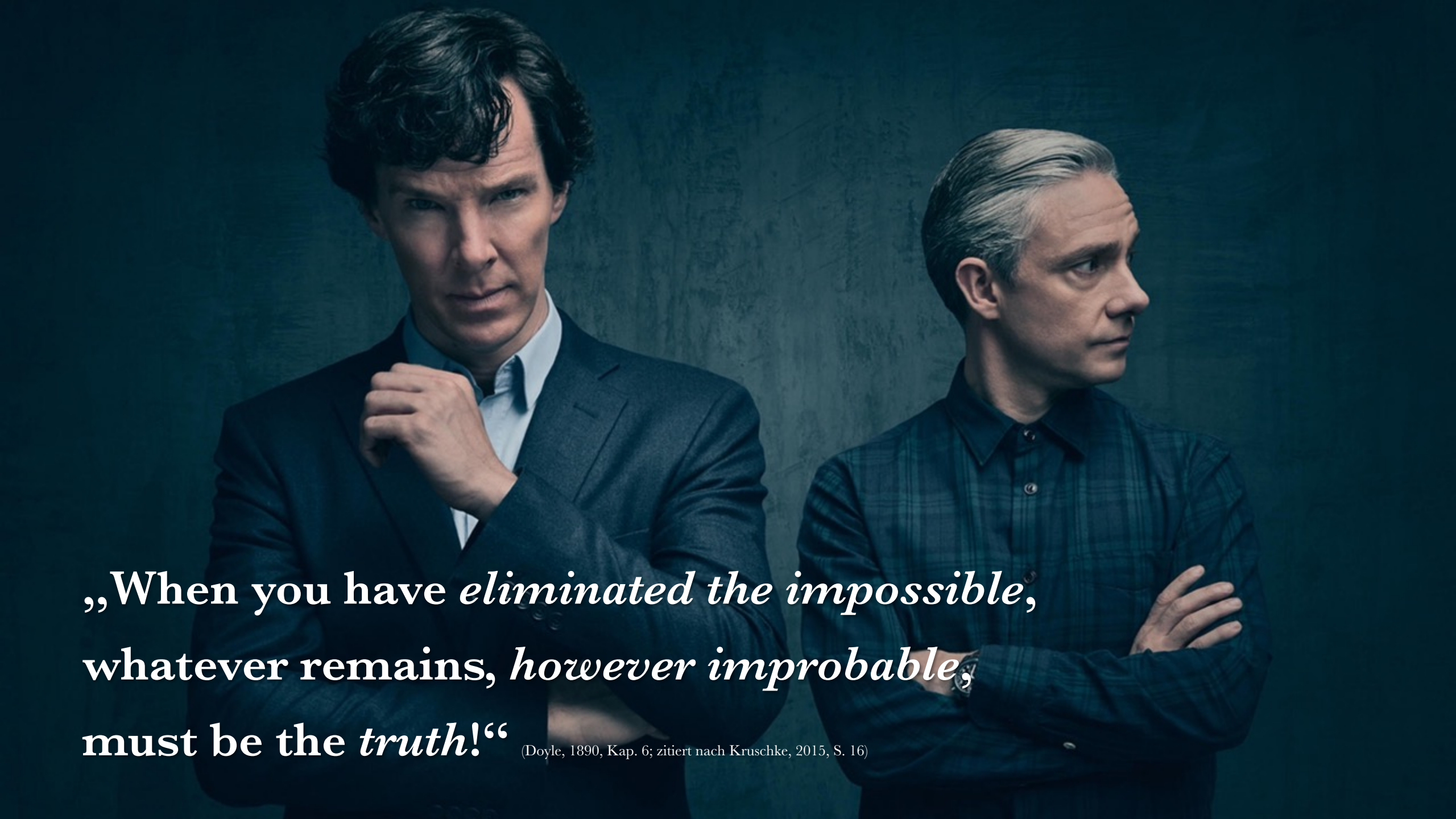
Probleme mit Signifikanztests und *p*-Werten (Auswahl)

- **Signifikanzniveau** von .05 ist zu liberal (aber .005 wird es nicht verbessern!)
- **Power-Analysen** und Fallzahlberechnungen werden in der Praxis nur sehr selten durchgeführt
- Geschickte Transformationen erlauben beliebige, signifikante Ergebnisse (**p-Hacking**, **Questionable Research Practices**)
- *und vieles mehr …*

Hack Your Way to Scientific Glory!

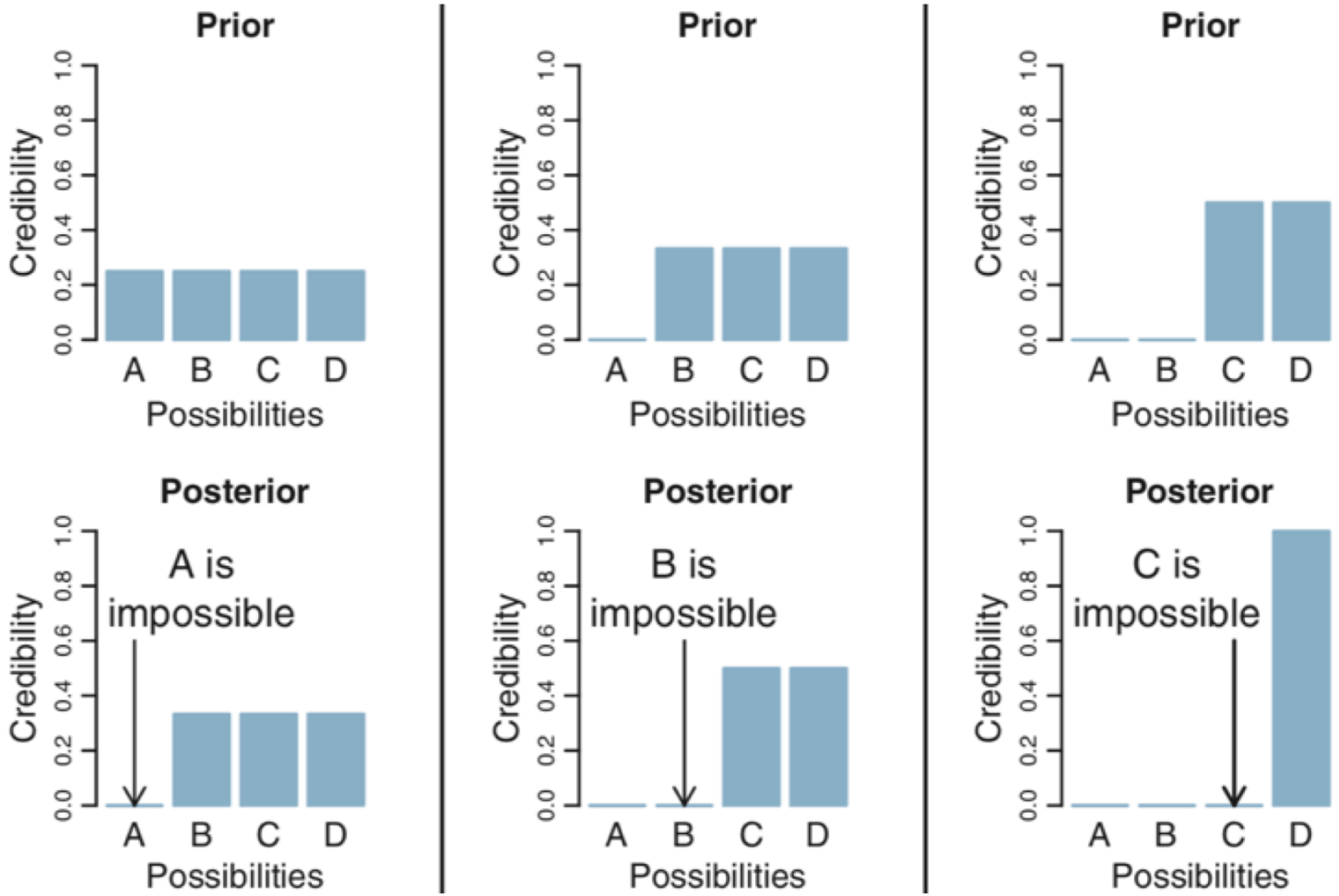


http://shinyapps.org/apps/p-hacker/



„When you have *eliminated the impossible*, whatever remains, *however improbable*, must be the *truth*!“ (Doyle, 1890, Kap. 6; zitiert nach Kruschke, 2015, S. 16)

Auf den Spuren von Sherlock Holmes



Quelle: Kruschke, J. K. (2015). Introduction: Credibility, Models, and Parameters. In*Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*(2nd Ed., pp. 15–32). Waltham, MA: Academic Press / Elsevier.

Wahrscheinlichkeitsbegriffe

Frequentistische Perspektive	Subjektive Perspektive
W. ist relative Häufigkeit (**Frequenz**) eines Ereignis bei sehr häufiger (unendlich oft) Wiederholung	W. ist der Ausdruck eines **Grad der Überzeugung** („*degree of personal belief*“)

Primär philosophische Unterscheidung.

Das Sherlock-Holmes-Beispiel zeigt aber, dass im (auch: im wissenschaftlichen) Alltag Wahrscheinlichkeit häufig „subjektiv“ gebraucht wird.

Es geht immer darum **Unsicherheit** (*uncertainty*) auszudrücken.

Mathematische Formalisierung

- Kolmogorov-Axiome beschreiben mathematische Wahrscheinlichkeitsrechnung:

1. Die Wahrscheinlichkeit ist eine nicht-negative reelle Zahl: $p(A) \geq 0$
2. Ein sicheres Ereignis hat die Wahrscheinlichkeit 1: $p(S) = 1$
3. Schließen sich die Ereignisse A und B einander aus, dann gilt: $p(A + B) = p(A) + p(B)$
4. Für ein Ereignis A und das Gegenteil zu A gilt: $p(A) + p(\neg A) = 1$
5. Für beliebige A und B gilt: $p(A + B) = p(A) + p(B) - p(AB)$
6. Schließen sich A und B aus, dann gilt: $p(AB) = 0$
7. Sind A und B voneinander unabhängig, dann gilt: $p(AB) = p(A)p(B)$

- „Funktionieren“ auch für subjektive Perspektive

Bedingte Wahrscheinlichkeiten

- Die **bedingte Wahrscheinlichkeit**, dass A eintritt, wenn B bereits eingetreten ist, lässt sich aus dem Verhältnis der Wahrscheinlichkeit des gemeinsamen Auftretens von A und B und der Wahrscheinlichkeit von B bestimmen:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- **Beispiele:**
  - *P*(Die Straße ist nass | Es hat geregnet) ≈ 1
  - *P*(Klausurnote = 1,3 | gelernt)
  - *P*(positives Testergebnis | krank)

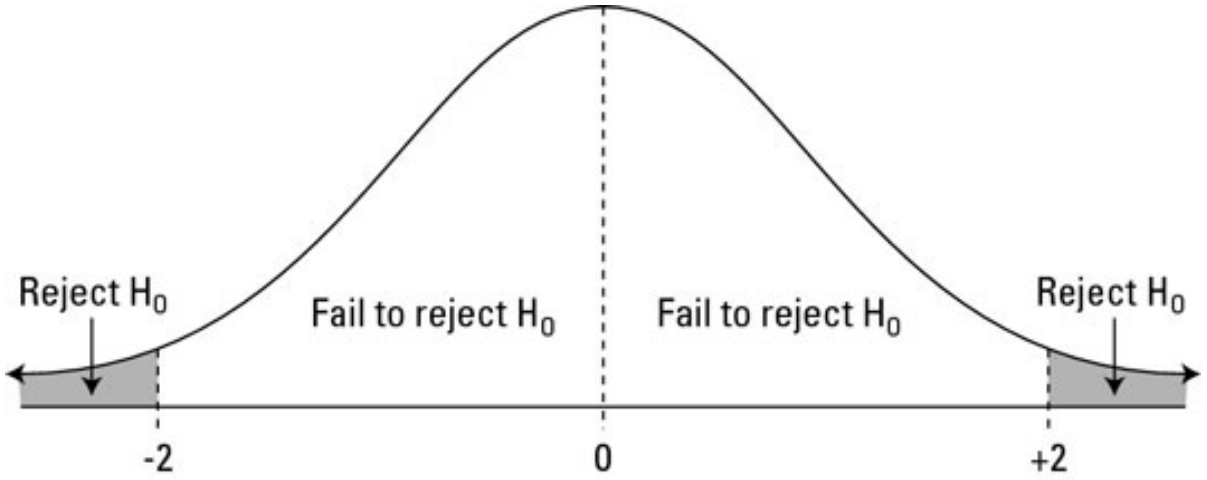
  P(Es hat geregnet | Die Straße ist nass) < 1

  P(gelernt | Klausurnote = 1,3)

  P(krank | positives Testergebnis)

Nochmal zu diesen *p*-Werten...

› *p*-Werte liefern uns eine bedingte Wahrscheinlichkeit: $P(|t| \geq |t_{obs}| \; | H_0)$



› Was aber ist $P(H_1 \mid Daten)$?

Satz von Bayes

- Aus der Definition der bedingten Wahrscheinlichkeit lässt sich ableiten:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Einige neue Begrifflichkeiten

Likelihood der Daten (E)
unter der Hypothese H

Prior der Hypothese H

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Posterior unserer Hypothese (H)
gegeben der Daten (Evidenz, E)

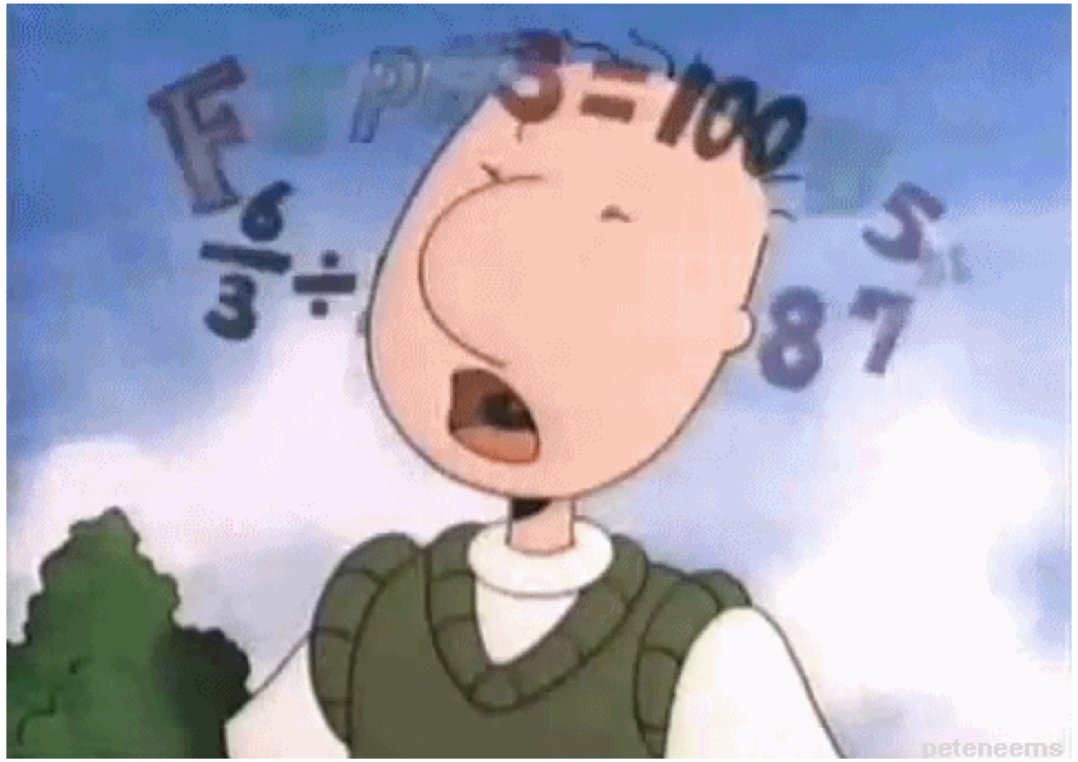
Marginal Likelihood oder auch
„prior predictive distribution“

Andere Darstellungsweisen

$$P(H|E) \propto P(E|H) \times P(H)$$

$$Posterior \propto Likelihood \times Prior$$

$$P(\theta|y) \propto P(y|\theta) \times P(\theta)$$



- Bedeuten das gleiche, aber unterschiedliche Artikel / Lehrbücher verwenden unterschiedliche Darstellungen

Es gibt unterschiedliche Ansätze Statistik zu betreiben…

Klassische Herangehensweise

- Testen von Hypothesen mittels Signifikanztests
- Je nach Daten und Art der Fragestellung werden Tests ausgewählt und durchgeführt, z.B.
  - *t*-Test für intervallskalierte Mittelwertsunterschiede
  - *ANOVA* für Mehrgruppenvergleiche bei metrischen abhängigen Variablen
- Annahmen für Tests müssen gegeben sein

Statistische Modellierung

- Daten werden durch ein mathematisches Modell beschrieben
- Einfachstes Beispiel: **lineare Regression**

$$y = \theta_0 + \theta_1 x_1 + \epsilon$$

- Solche Modelle können unterschiedliche Verteilungen und Ebenen abbilden (vgl. z.B. Strukturgleichungsmodelle)
- Aus Modell ergibt sich eine *Likelihood*-Funktion

Beispiel: Statische Modellierung

- Wir wollen **Einkommen** (y) durch Intelligenz (x_1), Alter (x_2), Familienstand (x_3; 1 = Single, 0 = Nicht-Single) und Interaktion von Familienstand und Alter vorhersagen
  - $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 (x_3 x_2) + \epsilon$
- Einkommen sei in unserem Modell normalverteilt
  - $y \sim \mathcal{N}(\mu, \sigma)$ $\quad p(y|\theta_i, \sigma)$
  - $\mu = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 (x_3 x_2)$
- Wir benötigen Prior-Verteilungen für unsere Parameter θ (Koeffizienten) und σ.
  - $\theta_i \sim \mathcal{N}(0, 1000)$ $\quad p(\theta_i)$
  - $\sigma \sim Cauchy(0, 25)$ $\quad p(\sigma)$
- Wir finden wir nun Werte für unsere Parameter?
  - Bayes!

Wie findet man geeignete Prior?

- **Noninformative priors** vs. **Informative priors**
- Je mehr Wahrscheinlichkeitsmasse auf einen geringen Bereich konzentriert wird, desto informativer ist ein Prior, d.h. desto größer ist der Einfluss auf die Posterior-Verteilung
- Priors sind Teil des Modells: Sie können daher auf Ihre Validität und prädiktiven Möglichkeiten überprüft werden
- Wesentliche Funktion von Priors: **Regularization**, um einen „overfit“ des Modells an die Daten zu verhindern
  - Betragsmäßig sehr große Parameter sind *im Allgemeinen* sehr unwahrscheinlich

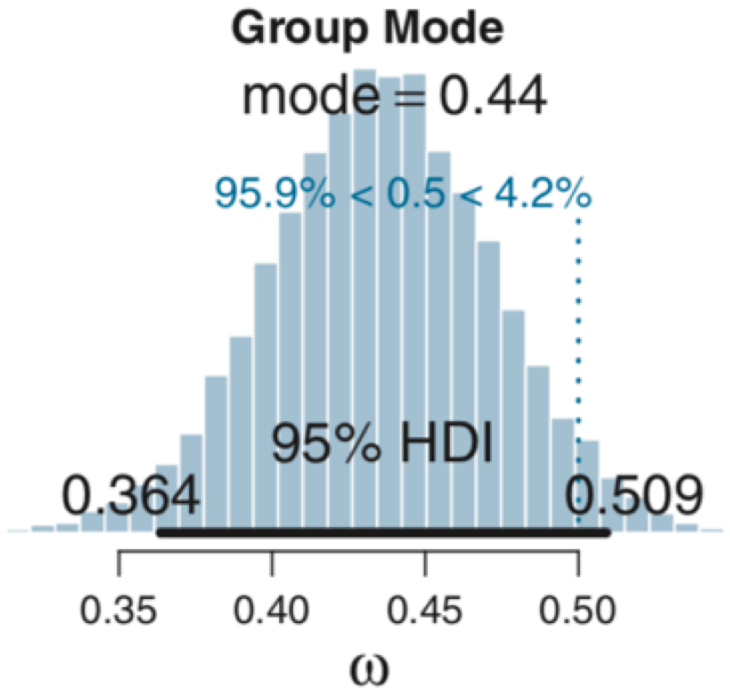
Anwendung von Bayes...

- Mittels unseres Modells, unserer Likelihood, unserer Prior und unseren Daten errechen wir die *Posterior*-Verteilung der Parameter:

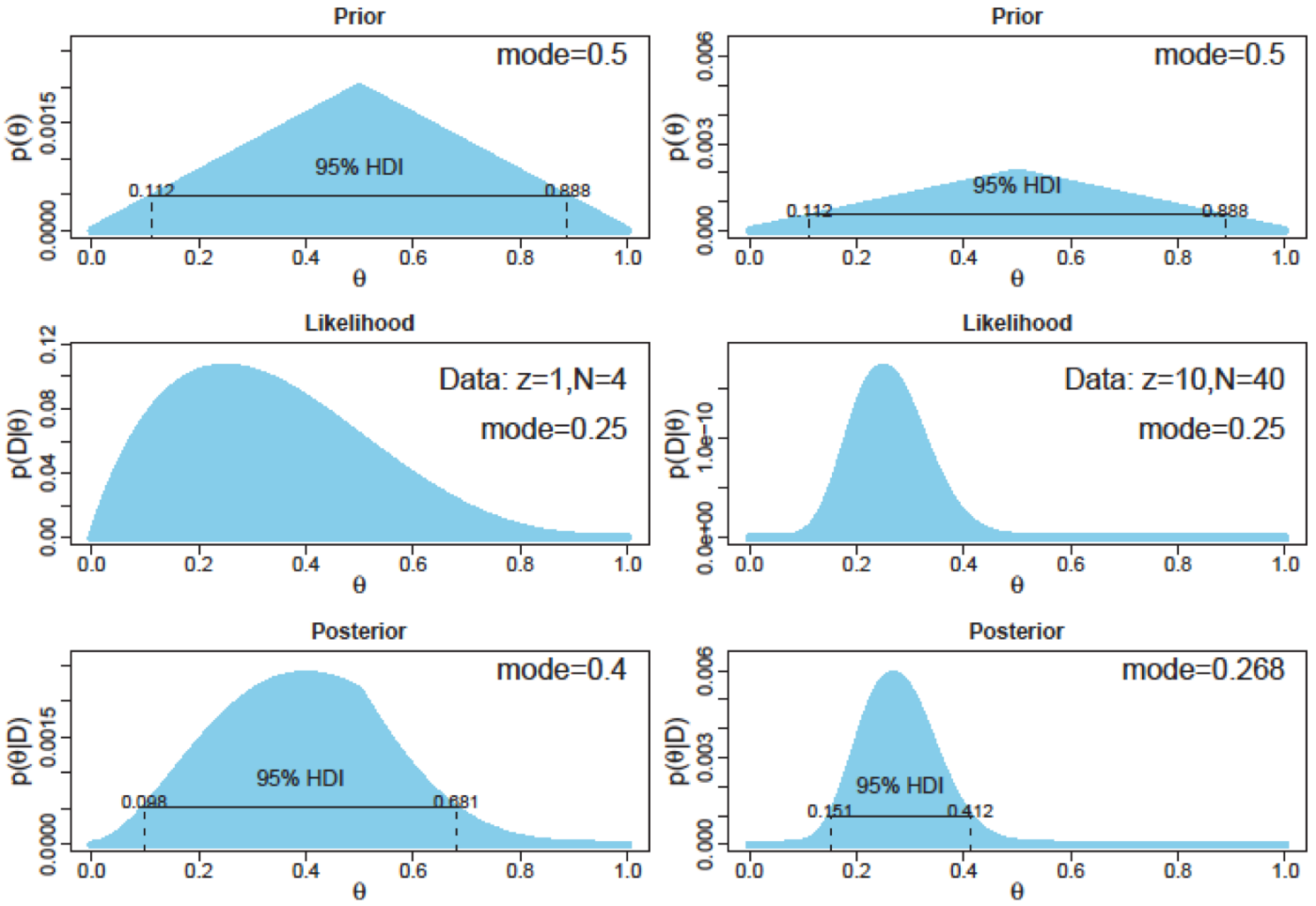
$$p(\Theta|y) = \frac{p(y|\Theta)p(\Theta)}{p(y)} = \frac{p(y|\Theta)p(\Theta)}{\int p(y|\Theta)p(\Theta)\, d\Theta}$$

Die Posterior-Verteilung

- Das Ergebnis unserer Analyse ist die **Posterior-Verteilung**:
  - Für jeden möglichen Parameterwert erhalten wir eine **Wahrscheinlichkeit**
  - Enthält nicht nur einen einzelnen Punktschätzer (z.B. Effektstärke), sondern auch Informationen über die Genauigkeit / (Un-)Sicherheit unserer Schätzung
  - Posterior ist Verteilung über alle Parameter gleichzeitig



Verhältnis von Prior, Likelihood, Posterior



In beiden Spalten ist die Prior identisch!

Links: 4 Beobachtungen
Rechts: 40 Beobachtungen

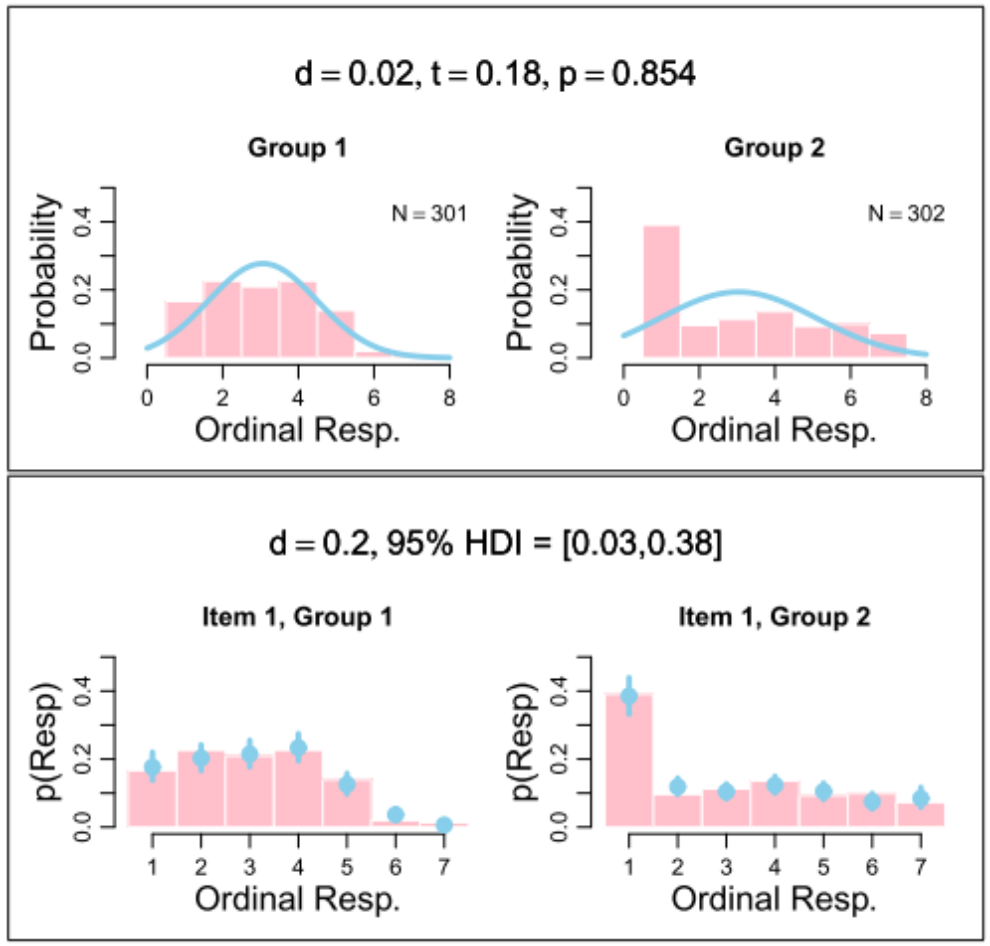
Abbildung: Copyright © Kruschke, J. K. (2014). Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition. Academic Press / Elsevier.

Wofür der ganze Aufwand?

- Statistische Modellierung ermöglicht Datenanalyse in einer **Vielzahl allgemeiner Fälle**
- **Mehrebenenmodelle** (Bayesian Multilevel Models / Hierarchical Models) sind für geschachtelte Daten besonders geeignet (z.B. wiederholte Messungen, Schüler in Klassen in Schulen in Bundesländern, ...)
- Bayes ist sinnvoller und guter Weg **Parameter eines Modells zu schätzen** und Schlussfolgerungen auf Basis der **Posterior-Verteilung** zu ziehen
- Annahmen und Prior-Verteilungen sind Teil des Modells und können getestet werden

Anwendungsbeispiel: Likert-Skalen als abhängige Variable

- Häufiger Fall in der Psychologie: Fragebogen-Antworten (Likert-Items) als abhängige Variable
- t-Test: Annahme von Intervallskalen-Niveau, d.h. equidistante Skalenpunkte
- In der Praxis wird das jedoch nie überprüft!
- **Ordered Probit Model**: Es gibt eine zugrundeliegende Normalverteilung – die Intervallgrenzen werden aber durch das Modell geschätzt



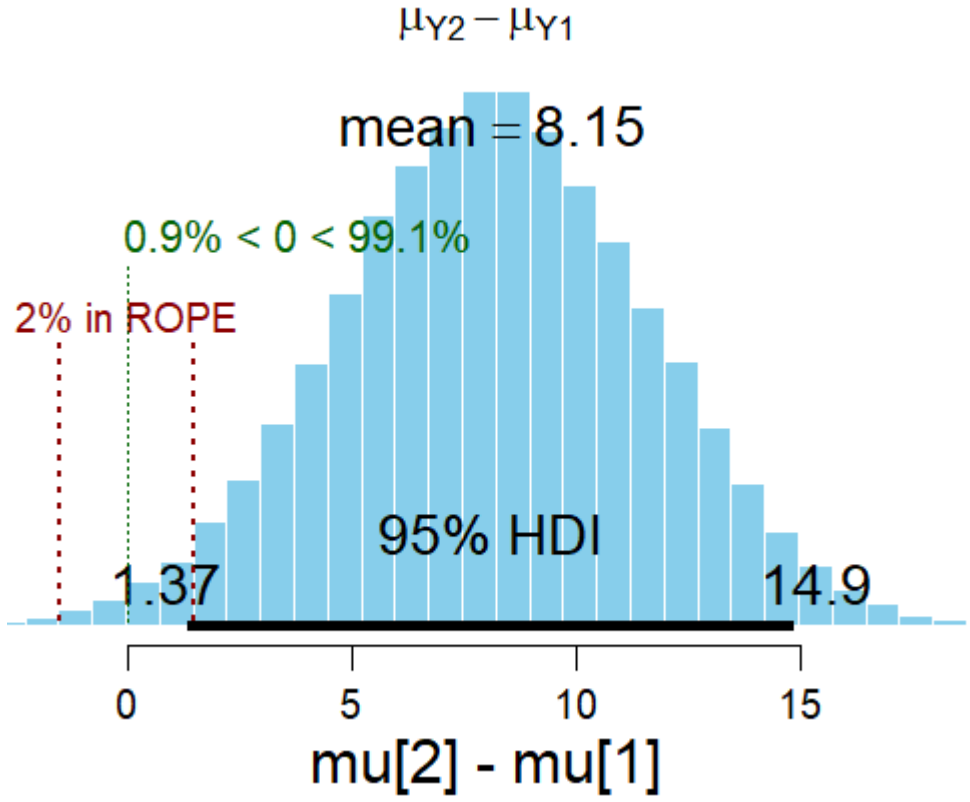
Quelle: Liddell, T. M., & Kruschke, J. K. (2017). *Analyzing ordinal data with metric models: What could possibly go wrong?* Retrieved from https://osf.io/9h3et/

Aber... Was ist mit Hypothesentests?

- Inferenzen auf Basis der **Posterior-Verteilung** sind für sich noch keine Hypothesentests oder statistischen Entscheidungen
- Es gibt verschiedene Möglichkeiten Hypothesen über Parameter in der Bayes Statistik zu testen:
  - Region of Practical Equivalence (ROPE)
  - Bayes factors (Model Selection)
  - Posterior Predictive Checks (Model Validation)

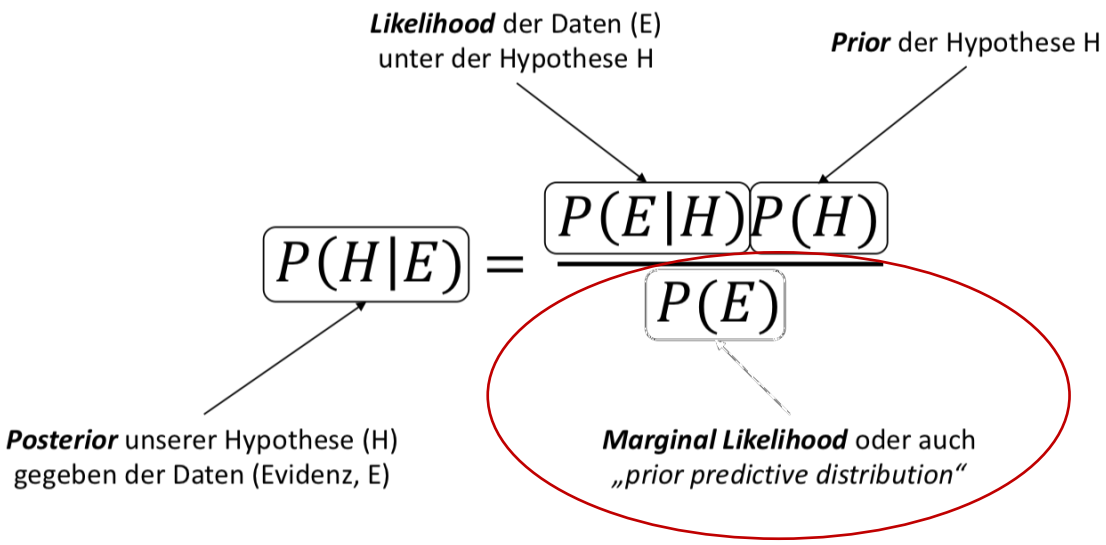
Region of Practical Equivalence (ROPE)

- Intervall für interessierende Effektgröße δ:
  - Welche Werte für δ sind für uns „quasi 0“?
- Vergleich des **95% Highest Density Interval** der Posterior-Verteilung mit dem **ROPE-Intervall**
  (1) 95% HDI **vollständig innerhalb** der ROPE
  (2) 95% HDI **vollständig außerhalb** der ROPE
  (3) 95% HDI und ROPE-Intervalls **überlappen** sich



Bayes Faktoren

- Haben wir zwei Modelle, die Hypothesen abbilden (z.B. Treatment hat einen Effekt vs. hat keinen Effekt) können wir diese über ihre **Marginal Likelihood** miteinander vergleichen
- Auch hier werden Prior-Verteilungen für die Parameter benötigt, sie wirken sich aber deutlich stärker auf das Ergebnis aus als bei der Posterior-Verteilung



Bayes Faktoren: Interpretation

- Der Bayes Faktor BF_{10} ist also ein Verhältnis und damit eine Zahl zwischen 0 und $+\infty$
  - $BF_{10} < 1$ bedeutet, dass M_0 die Daten besser erklärt als M_1
  - $BF_{10} = 1$ bedeutet, dass beide Modelle die Daten etwa gleich gut erklären
  - $BF_{10} > 1$ bedeutet, dass M_1 die Daten besser erklärt als M_0
- Hilfreiche Interpretations-Richtlinien (z.B. Jeffreys, 1961):
  - $1 < BF < 3$: Nur anekdotische Evidenz (meist: zu wenig Daten)
  - $3 < BF < 10$: Moderate Evidenz
  - $10 < BF < 30$: Starke Evidenz
  - $BF > 30$: Sehr starke Evidenz

Wie lösen wir damit nun die Replizierbarkeitskrise?

- Forscher haben Fragestellungen, die sie mittels Daten beantworten wollen
- Blind irgendwelchen Methoden, Tests oder Algorithmen zu folgen führt uns nicht zu Erkenntnisgewinn
- Statistischer Werkzeugkasten muss viele, verschiedene Methoden enthalten – auch *p*-Werte (aber dann richtig eingesetzt)!
- Zulassung eines Medikaments vs. Stroop-Effekt
  - Statistische Entscheidungen brauchen kontrollierte Fehlerraten (Neyman-Pearson und Signifikanztests) und Entscheidungstheorie (d.h. Kostenfunktionen)
  - Wollen wir probabilistische Aussagen über Effekte machen, führt an Bayes kein Weg vorbei

Wie lösen wir damit nun die Replizierbarkeitskrise?

- **ABER:** Statistik ist nicht die einzige oder wichtigste Ursache der Replizierbarkeitskrise
  - Unspezifische, nicht-falsifizierbare Theorien
  - Publication Bias (+ Measurement Error + Signifikanztests = oh boy…)
  - Intransparente Peer Reviews
  - Kein Zugriff auf Rohdaten und Materialien
  - Fehlende Direkte & Konzeptuelle Replikationen
  - Anreize im Wissenschaftssystem (Fokus auf Quantität, nicht auf Qualität)
- Pre-Registrations & Registered Reports, Replikationsprojekte, Meta-Analysen, …

Literatur: Everything in Psychology is Fucked…

PSY 607: Everything is Fucked

Prof. Sanjay Srivastava

Course Syllabus: https://hardsci.wordpress.com/2016/08/11/everything-is-fucked-the-syllabus/

Literatur: Einführung in Bayes Statistik

- McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.
  - Sehr gute, zugängliche Einführung
- Kruschke, J. K. (2015). Introduction: Credibility, Models, and Parameters. In *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2nd Ed., pp. 15–32). Waltham, MA: Academic Press / Elsevier. http://doi.org/10.1016/B978-0-12-405888-0.00002-7
  - Sehr Praxis-orientiertes Buch
- Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. *Cambridge*. Cambridge: Cambridge University Press.
  - Die Bibel für Regressionsmodelle
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd Edition). Boca Raton, FL: CRC press.
  - Sehr anspruchsvolle Lektüre, aber sehr umfassend. Die Bibel für Bayes

Literatur: Hypothesentests mit Bayes Faktoren

- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review, 16(2), 225–237. http://doi.org/10.3758/PBR.16.2.225
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. http://doi.org/10.1016/j.jmp.2012.08.001
- Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2015). Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses. Journal of Mathematical Psychology. http://doi.org/10.1016/j.jmp.2015.12.007
- Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. http://doi.org/10.1037/a0036731
- Dienes, Z. (2016). How Bayes factors change scientific pratice. *Journal of Mathematical Psychology*, 1–29. http://doi.org/10.1007/s13398-014-0173-7.2

Literatur: Replizierbarkeitskrise in der Psychologie

- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. http://doi.org/10.1177/0956797611417632
- Meehl, P. E. (1990). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, *66*(1), 195. http://doi.org/10.2466/PR0.66.1.195-244
- Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. http://doi.org/10.1371/journal.pmed.0020124
- Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, *19*(5), 640–648. http://doi.org/10.1097/EDE.0b013e31818131e7