Bayesianische Statistik:

Welche Probleme löst sie und welche nicht?

Michael Höfler

Frequentistische Statistik

Interessierender Parameter **θ** (Wahrscheinlichkeit, Erwartungswert, Assoziation, Effekt)

Bevor man Daten erhebt:
Verteilungsmodell für Daten, gegeben fester, aber unbekannter Wert von **θ**:

P(Daten | **θ)**

z.B. Wahrscheinlichkeit (Prävalenz) von Depression; Anzahl der Depressionsfälle in der Stichprobe (Daten), **X**, ist binomialverteilt, Wahrscheinlichkeit für **x** Fälle:

$$P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Nach Datenerhebung:
Man hat **X** = **x** Fälle beobachtet, nun ist **x** fest. Es handelt sich jetzt um eine Funktion von **θ**, genannt Likelihood-Funktion:

$$L(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Maximum-Likelihood-Prinzip:

beste Schätzung für **θ** ist der Wert, der dann L(**θ**) maximiert;
hier: **θ** = x/n (Anteil Depressionsfälle in Stichprobe). So ein „Schätzer“ ist auch Grundlage statistischer Tests über **θ**.

Man macht keine Vorannahmen über **θ**, alle Information über **θ** kommt aus den Daten.

Wichtig:
Es gibt kein Verteilungsmodell für

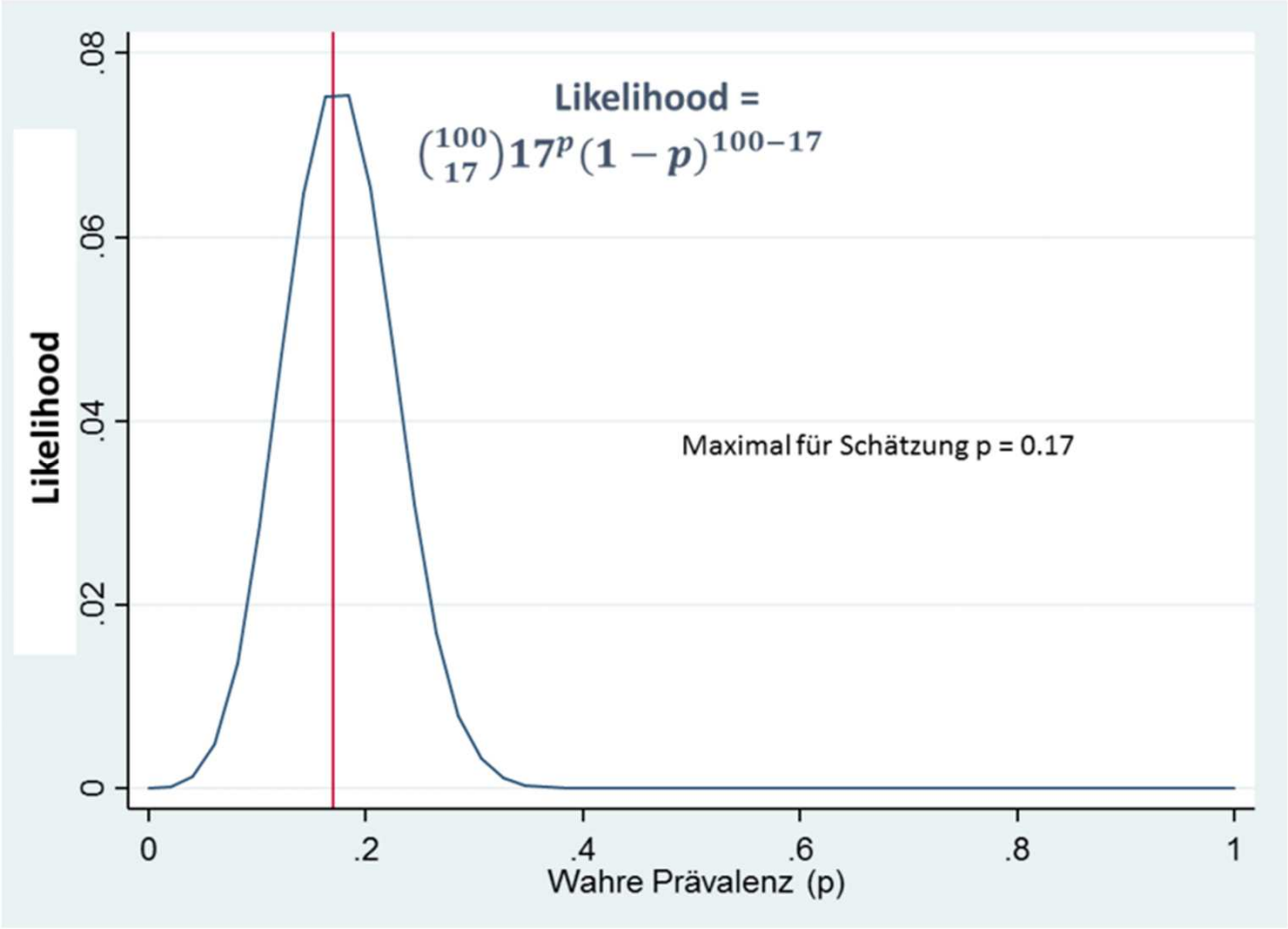
$P(\boldsymbol{\theta} \mid \text{Daten})$

θ ist ein unbekannter, aber fester Wert.

Keine Wahrscheinlichkeitsaussagen über **θ** möglich!

(Beispiel numerisch)

In einer Stichprobe von 100 Personen erfüllen 17 Personen die Kriterien einer Major Depression. Die Maximum-Likelihood-Schätzung für die Depressions-Wahrscheinlichkeit, **θ** = **p**, beträgt 0.17 (17%).



Wahrscheinlichkeit für Zutreffen von H_0/H_1 ist 0 oder 1!

Wahrscheinlichkeit, dass ein Konfidenzintervall das wahre θ enthält, ist 0 oder 1!

Für erwünschte Wahrscheinlichkeitsaussagen ($0 < P(H_1) < 1$) **Interpretationskrücke nötig!**

Bildquelle: Pixaby.com

Frequentistisches Gedankenexperiment

Häufigkeitsinterpretation: Vorstellung, man führ ganz oft (z.B. 10.000 mal) mit identischen Methoden eine Studie durch und analysiert (schätzen, testen) jedes mal auf gleiche Art. Zufall durch zufällige Stichprobenziehung/Gruppeneinteilung führt jedes mal zu anderem Ergebnis. In diesem Sinne kann man dann von Wahrscheinlichkeiten sprechen:

Falls H_0 zutrifft, verwirft ein **Test** in 100*α % der Fälle H_0 zu unrecht (5% falsch Positive, falls α = .05). Weiter keine absolute W.keitsaussage über H_1 (nur bedingt auf H_0)!

Ein (1- α)*100%-iges **Konfidenzintervall***, z.B. 95%, für einen Parameter (z.B. Regressionskoeffizient β), enthält in 95% der Fälle den wahren Parameterwert.

Problem 1: Gedankenexperiment weit hergeholt

1. Funktioniert nur, wenn Studie **Zufallskomponente** hat

- Zufällige Stichprobenziehung
- Oder zufällige Gruppeneinteilung
- (sonst noch mehr Gedankenexperimente nötig, z.B. Vorstellung, dass es hinter einer untersuchten (sehr speziellen) klinischen Stichprobe eine größere Population gebe, die sich nicht systematisch von der Stichprobe unterscheide).

(… und Analyse sollte vollständig determiniert sein)

- z.B. immer t-Test durchführen, egal wie Verteilungen in Daten
- Wünschenswert aber: auch Alternativverfahren benutzen (z.B. U-Test), das viel weniger voraussetzt (sonst “analytical bias“)
- → unterschiedliche Analyse in unterschiedlichen Datensätzen, “Garden of forking paths“, frequentistische Interpretation stimmt nicht mehr, v.a. wenn dann **unterschiedliche Schätzer** untersucht werden (U-Test basiert nicht auf Mittelwertsunterschied, sondern Unterschied im Durchschnittsrang).

Problem 2: falsche Interpretation des p-Werts

p-Wert = P(**X** > **x** | H_0): Wahrscheinlichkeitsaussage,
aber nur über **X** und bedingt auf H_0

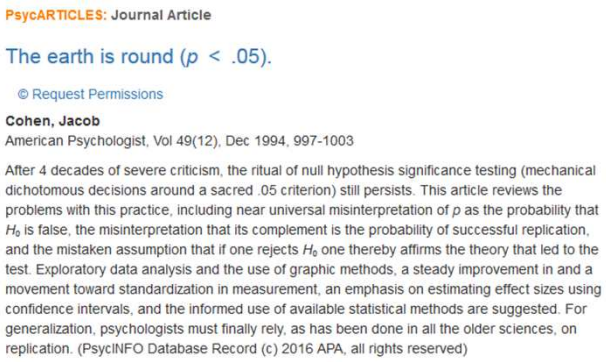
- **1 - p** ist nicht die Wahrscheinlichkeit für das Zutreffen von H_1 (gegeben die Daten), **p** ist nicht Wahrscheinlichkeit für H_0
- **1 - p** ist kein Maß für die **Replizierbarkeit** von Studienergebnissen
- **p** misst nicht die Größe/Bedeutsamkeit eines Zusammenhangs
- u.u.u.:

Problem 3: „dichotomia“

- Reduziere Ergebnisse ohne Notwendigkeit auf **binäre Entscheidung,** p < .05?

- **p** sollte eigentlich gar nicht kategorisiert werden

(willkürlich; .05 ist bloß eine soziale Konvention, die mechanisch angewendet wird; s.u.)



PsycARTICLES: Journal Article

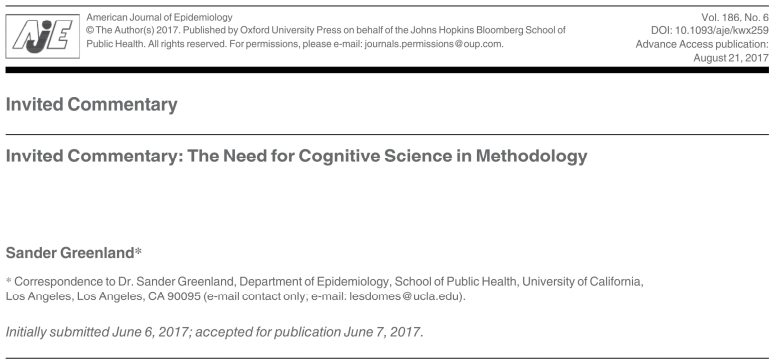
The earth is round ($p < .05$).

© Request Permissions

Cohen, Jacob

American Psychologist, Vol 49(12), Dec 1994, 997-1003

After 4 decades of severe criticism, the ritual of null hypothesis significance testing (mechanical dichotomous decisions around a sacred .05 criterion) still persists. This article reviews the problems with this practice, including near universal misinterpretation of p as the probability that H_0 is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects H_0 one thereby affirms the theory that led to the test. Exploratory data analysis and the use of graphic methods, a steady improvement in and a movement toward standardization in measurement, an emphasis on estimating effect sizes using confidence intervals, and the informed use of available statistical methods are suggested. For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication. (PsycINFO Database Record (c) 2016 APA, all rights reserved)



American Journal of Epidemiology
© The Author(s) 2017. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 186, No. 6
DOI: 10.1093/aje/kwx259
Advance Access publication: August 21, 2017

Invited Commentary

Invited Commentary: The Need for Cognitive Science in Methodology

Sander Greenland*

* Correspondence to Dr. Sander Greenland, Department of Epidemiology, School of Public Health, University of California, Los Angeles, Los Angeles, CA 90095 (e-mail contact only, e-mail: lesdomes@ucla.edu).

Initially submitted June 6, 2017; accepted for publication June 7, 2017.

“[…] degrading continuous measures of evidence into decisive conclusions, feeding the strong cognitive bias of dichotomania: the compulsion to replace quantities with dichotomies (“black-and-white thinking”), even when such dichotomization is unnecessary and misleading for inference.”

Problem 4: Falschinterpretation von Testergebnissen je nach Stichprobengröße

- **Kleine Stichproben, p ≥. 05**: Behauptung, Nullhypothese treffe zu, obwohl geringe Power (Sekundäranalysen!)

- **Große Stichproben, p < .05**: Behauptung, es gebe einen bedeutsamen Unterschied, dabei kann bereits kleiner Unterschied zur Ablehnung von H_0 führen; falsche Praxis von *,**,**-Signifikanz (hat keine methodische Basis; Vermischung der stat. Schulen von K. Pearson und R.A. Fisher; wahrnehmungspsychologisch sehr problematisch;).

Problem 5: Statistische Tests sind asymmetrisch

Die Verteilung von Teststatistiken ist oft nur berechenbar, wenn H_0 zutrifft (z.B. χ^2-Unabhängigkeitstest) → man kann besser auf H_1 schließen.

p < vorgegebenes α, z.B. α = .05 → **H_0 verwerfen**

p ≥ vorgegebenes α, z.B. α = .05 → **H_0 beibehalten**

Korrekt: Es gibt keine Evidenz/keinen Hinweis dafür, dass H_0 verletzt ist.
Falsch: H_0 trifft zu.

Problem 6: Statistisches Modell kann falsch sein

Falsche Ergebnisse, weil z.B. die Annahmen eines Modells (ausgedrückt in Likelihood-Funktion) verletzt sind („analytical bias“; eher Regel als Ausnahme).

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers

Andy P. Field [a, *], Rand R. Wilcox [b]

[a] School of Psychology, University of Sussex, Falmer, Brighton, BN1 9QH, UK
[b] Department of Psychology, University of Southern California, 618 Seeley Mudd Building, University Park Campus, Los Angeles, CA 90089-1061, USA

Problem 7: Annahmen sind intransparent

- Annahmen im Gedankenexperiment
- Bevor Daten erhoben, sind alle Parameterwerte (z.B. Effekt = -2, 1, 0, +1 , +2) „gleichwahrscheinlich“, „Gleichverteilung“ (korrekter: alle gleichbreiten Intervalle haben Wahrscheinlichkeit, z.B. [-2, -1], [-1, 0], [0, 1] …], stetige Verteilung, Wahrscheinlichkeit für jeden Punktwert = 0)
- Mit 100% Wahrscheinlichkeit keinerlei Bias in einer Schätzung, z.B. durch gemeinsame Ursachen von **X** und **Y**, Messfehler in **X** und **Y,** Selektion, s.u.

Problem 8: instabile Schätzungen

- Kleine Datensätze, zu viele Einflussvariablen
- Sehr große Zusammenhänge/Effekte (Regressionskoeffizienten β) aber oft unwahrscheinlich
- In frequentistischer Statistik variieren Schätzungen der β jedoch völlig frei (zwischen - ∞ und + ∞).

Research Methods & Reporting

Sparse data bias: a problem hiding in plain sight

BMJ 2016 ; 352 doi: https://doi.org/10.1136/bmj.i1981 (Published 27 April 2016)
Cite this as: *BMJ* 2016;352:i1981

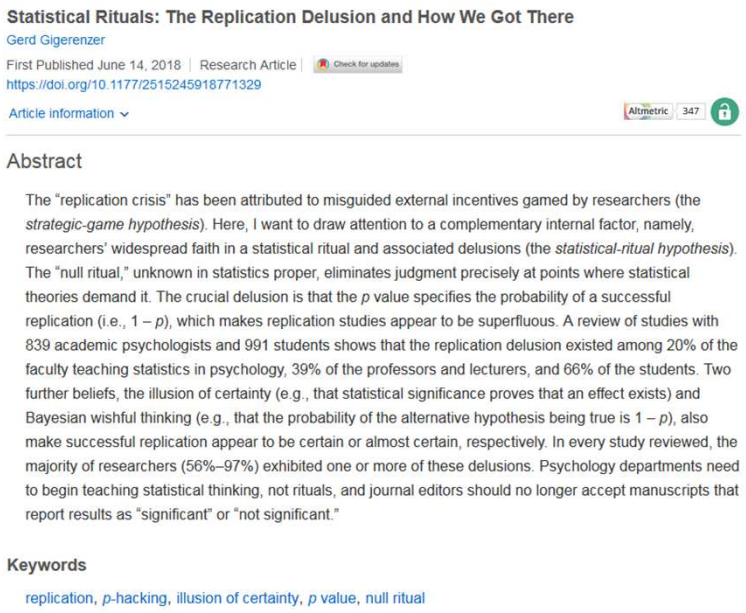
Sander Greenland, professor of epidemiology and statistics [1], *Mohammad Ali Mansournia, assistant professor of epidemiology* [2], *Douglas G Altman, professor of statistics in medicine* [3]

[1] *Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA*
[2] *Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, PO box 14155-6446, Tehran, Iran*
[3] *Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK*

Correspondence to: M A Mansournia mansournia_ma@yahoo.com
Accepted 7 March 2016

Problem 9: grundsätzlich falscher Gebrauch von Statistik

- Statistik wird gedankenlos, mechanisch und als bloßes Ritual verwendet.
- p-hacking: Missbrauch der Variation in den Ergebnissen je nach Auswertungsmethode, um best. Ergebnis zu erhalten



Statistical Rituals: The Replication Delusion and How We Got There

Gerd Gigerenzer

First Published June 14, 2018 | Research Article | Check for updates

https://doi.org/10.1177/2515245918771329

Article information

Altmetric 347

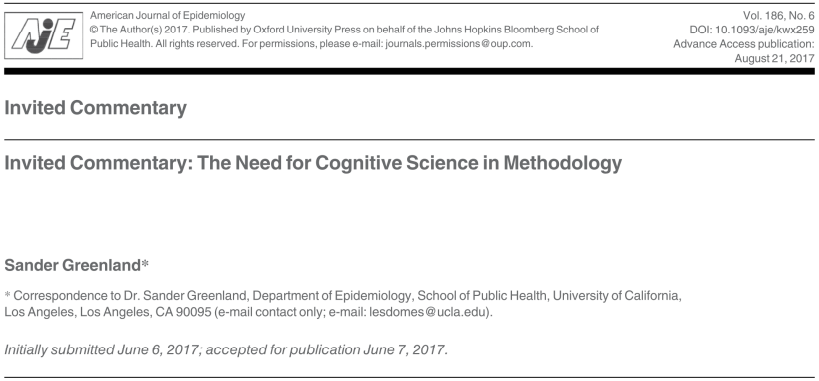
Abstract

The "replication crisis" has been attributed to misguided external incentives gamed by researchers (the *strategic-game hypothesis*). Here, I want to draw attention to a complementary internal factor, namely, researchers' widespread faith in a statistical ritual and associated delusions (the *statistical-ritual hypothesis*). The "null ritual," unknown in statistics proper, eliminates judgment precisely at points where statistical theories demand it. The crucial delusion is that the *p* value specifies the probability of a successful replication (i.e., 1 – *p*), which makes replication studies appear to be superfluous. A review of studies with 839 academic psychologists and 991 students shows that the replication delusion existed among 20% of the faculty teaching statistics in psychology, 39% of the professors and lecturers, and 66% of the students. Two further beliefs, the illusion of certainty (e.g., that statistical significance proves that an effect exists) and Bayesian wishful thinking (e.g., that the probability of the alternative hypothesis being true is 1 – *p*), also make successful replication appear to be certain or almost certain, respectively. In every study reviewed, the majority of researchers (56%–97%) exhibited one or more of these delusions. Psychology departments need to begin teaching statistical thinking, not rituals, and journal editors should no longer accept manuscripts that report results as "significant" or "not significant."

Keywords

replication, *p*-hacking, illusion of certainty, *p* value, null ritual

- Testergebnisse werden als Wahrheitsaussage statt als Verhaltensregel verstanden. Ein signifikantes Ergebnis bedeutet nur, dass es eine Auffälligkeit gibt, die erklärt werden muss.



American Journal of Epidemiology

Vol. 186, No. 6
DOI: 10.1093/aje/kwx259
Advance Access publication:
August 21, 2017

Invited Commentary

Invited Commentary: The Need for Cognitive Science in Methodology

Sander Greenland*

* Correspondence to Dr. Sander Greenland, Department of Epidemiology, School of Public Health, University of California, Los Angeles, Los Angeles, CA 90095 (e-mail contact only; e-mail: lesdomes@ucla.edu).

Initially submitted June 6, 2017, accepted for publication June 7, 2017.

[…] both confidence intervals and α-level tests were conceived as **decision rules for behavior** but were rapidly misinterpreted as **rules for belief**, and thus fed the false notion that a single study can by itself tell us whether an effect is present or absent.”

Bayesianische Statistik

Frequentistische Statistik: P(Daten| **θ**)

Wie erhält man daraus das interpretierbare P(**θ** |Daten)?

Wahrscheinlichkeit z.B. für **θ =** Effekt > 0,
gegeben die Daten (und das Modell dazu/Likelihood)?

(Man kann auch analog wie im Folgenden vorgehen, nur bedingt darauf, dass eine **H_1 über θ zutrifft**; „prior“ ist dann die Wahrscheinlichkeit für das Zutreffen von H_1, bevor man die Daten sieht:
P(Daten| H_1) → P(H_1 |Daten)?

Zweiseitige Hypothesen über **θ,** z.B. H_0: **θ** = 0 lassen sich damit untersuchen; aber nicht, wenn man (wie auf den folgenden Folien) eine stetige Verteilung („prior“) über **θ** zugrundelegt, dann würde gelten: P(H_0) = 0.)

Theorem von Bayes:

Prior:

Wahrscheinlichkeitsverteilung der Größe des Effekts, bevor man Daten untersucht

Posterior:

$$P(Effekt|Daten) = \frac{P(Daten|Effekt) \times P(Effekt)}{Konstante *}$$

Falls prior = **Gleichverteilung** („flat prior“) = jeder Wert gleichwahrscheinlich

$$= \frac{P(Daten|Effekt)}{Konstante *}$$

Prior → „Basisraten-problem“ in der Psycho-logie des Entscheidens:

Bild: weltbild.com

= Ergebnis frequentistischer Statistik, aber bessere Interpretation mit **transparenter** Annahme (man weiß nichts über den Effekt, bevor man die Daten sieht)

* Konstante dient nur dazu, dass das Ergebnis als Verteilung interpretiert werden kann; sorgt dafür, dass „Wahrscheinlichkeit über alle möglichen β-Werte“ = 1

(V.a. für kausale Inferenz erweitert um **Bias-Parameter***
und Verteilungen, die die Unsicherheit über diese modellieren)

J. R. Statist. Soc. A (2005)
168, *Part* 2, *pp.* 267–306

Multiple-bias modelling for analysis of observational data

Sander Greenland
University of California, Los Angeles, USA

Theorem von Bayes, erweitert:

Verteilungen von Bias-Parametern*, die Unsicherheiten über diese widerspiegeln

prior

$$P(Effekt|Daten, Annahmen) = \frac{P(Daten|Effekt)\ P(Annahmen)\ P(Effekt)}{Konstante}$$

Wahrscheinlichkeit für überhaupt kein Bias = 1!

prior = Gleichverteilung

$$= \frac{P(Daten \mid Effekt)}{Konstante}$$

= Ergebnis frequentis-tischer Statistik

* Z.B. Effekte unberücksichtigten Confounders auf **X** und **Y**, Selektionswahrscheinlichkeiten, Messfehler.

Posterior = Wahrscheinlichkeitsverteilung von **θ**

(z.B. Ausmaß eines Effekts), gegeben Daten und prior
(und Modell/Likelihood)

Vorteil: daraus lässt sich alles bestimmen, was man wissen möchte, z.B. **Wahrscheinlichkeit** für

H_1 = Effekt > 0

Effekt > **s**

Effekt liegt in beliebigem Intervall [$\mathbf{s_1}$, $\mathbf{s_2}$].

s = minimale Größe für inhaltlich relevanten Effekts, z.B. via „klinische Signifikanz“

Umsetzung: Wie bestimmt man die Posterior?

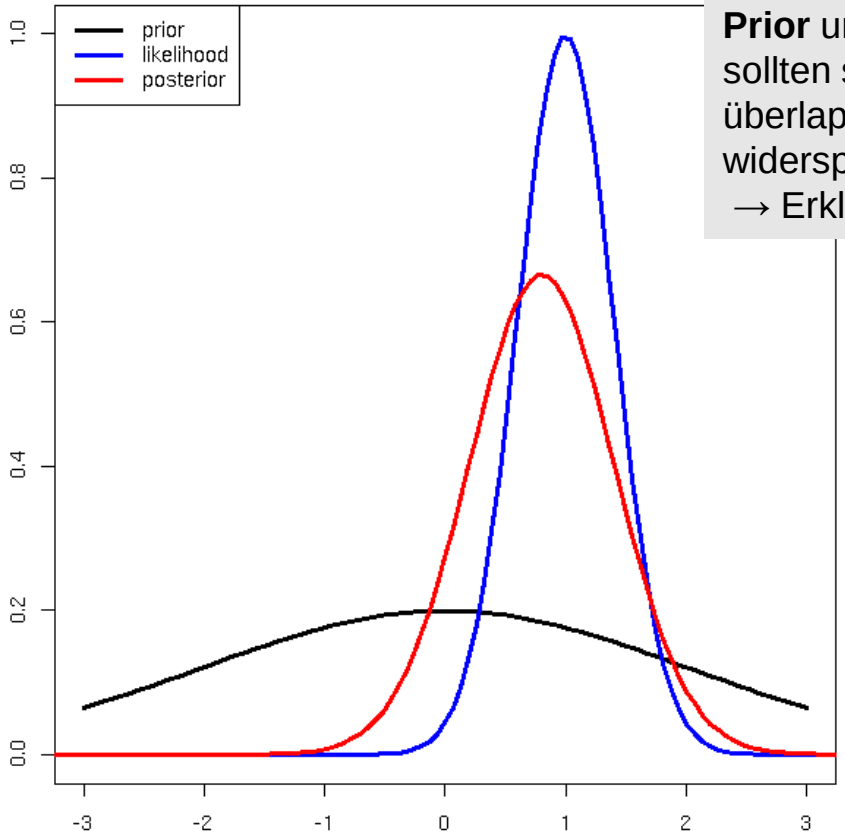
Simuliert man oft mit MCMC = „Monte-Carlo-Markov-Chain“

https://www.youtube.com/watch?v=OTO1DygELpY&feature=youtu.be

Beispiel für **prior**, **likelihood** und **posterior**

Prior: Normalverteilung um Erwartungswert 0 = kein Effekt; weder positiven, noch negativen Werten Präferenz gegeben; große Varianz, aber große positive wie negative Effekte unwahrscheinlicher

Likelihood: Normalverteilung um aus Daten geschätzten Wert von 1; hier viel weniger Varianz
(Schätzung von Regressionskoeff. β ist normalverteilt, zum. In großen Stichproben)



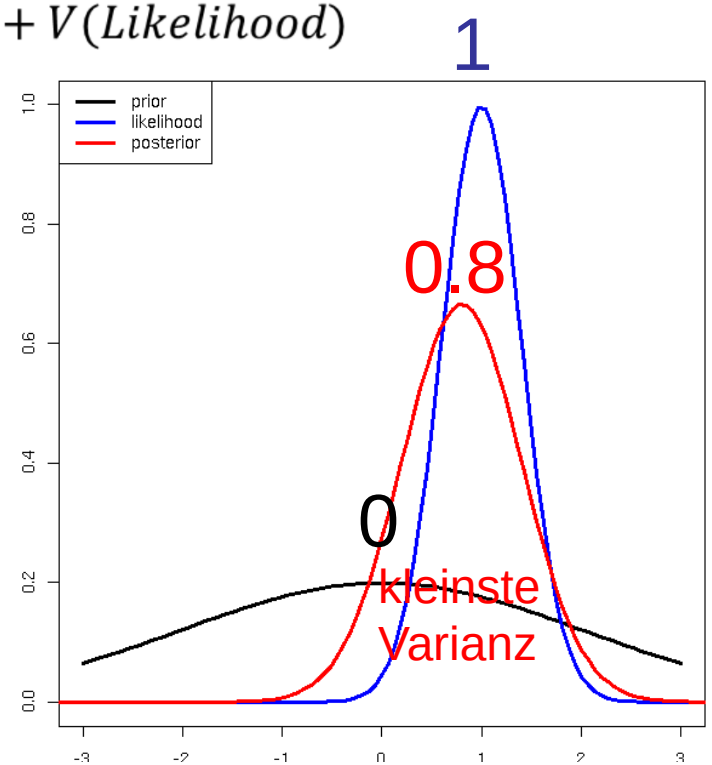
Prior und **likelihood** sollten sich gut überlappen, sonst widersprüchlich, → Erklärungsbedarf)

Beispiel **Normalverteilung**

Erwartungswert (E) der Posterior = mit Varianz (V) gewichtetes Mittel von Prior und Likelihood ("inverse variance-weighted"):

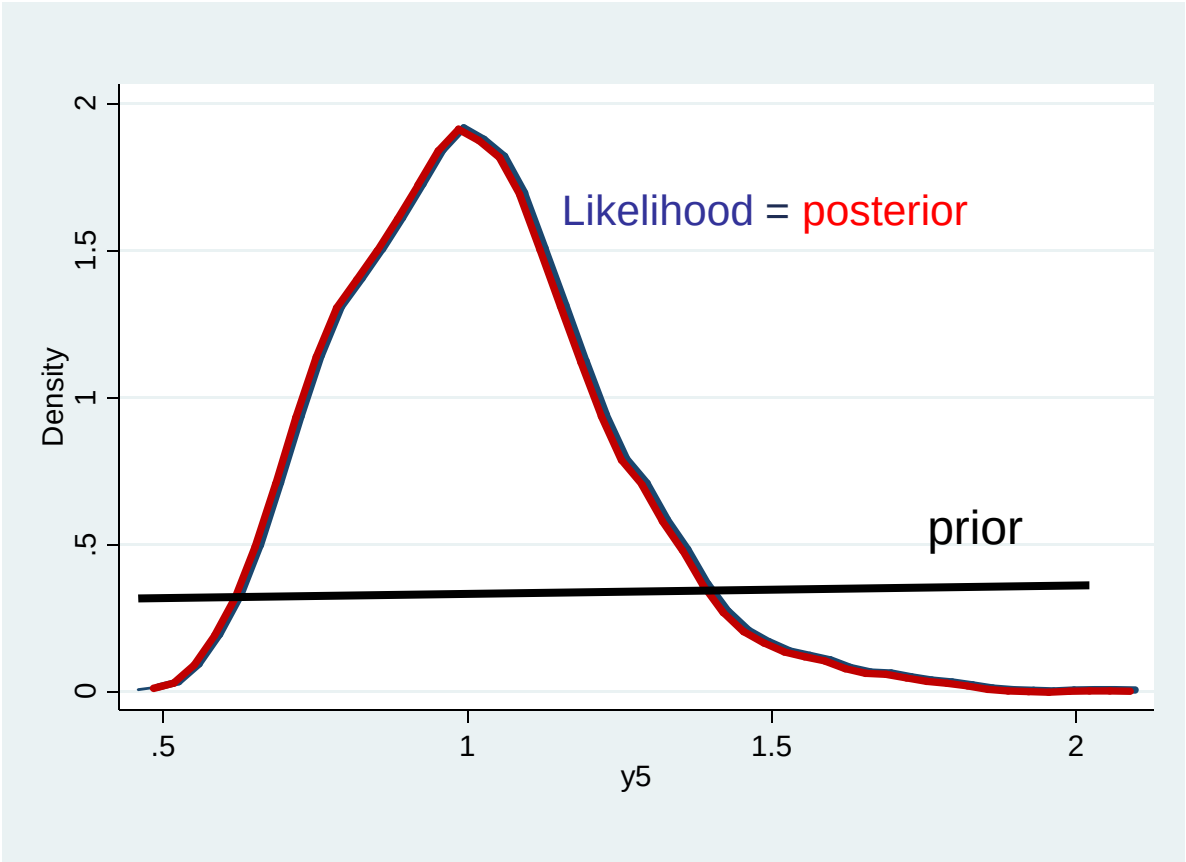
$$E(Posterior) = \frac{\frac{E(Prior)}{V(Prior)} + \frac{E(Likelihood)}{V(Likelihood)}}{V(Prior) + V(Likelihood)}$$

Posterior hat kleinste Varianz; je kleiner die Varianz der **Prior**, desto größer ihr Einfluss auf die Posterior.

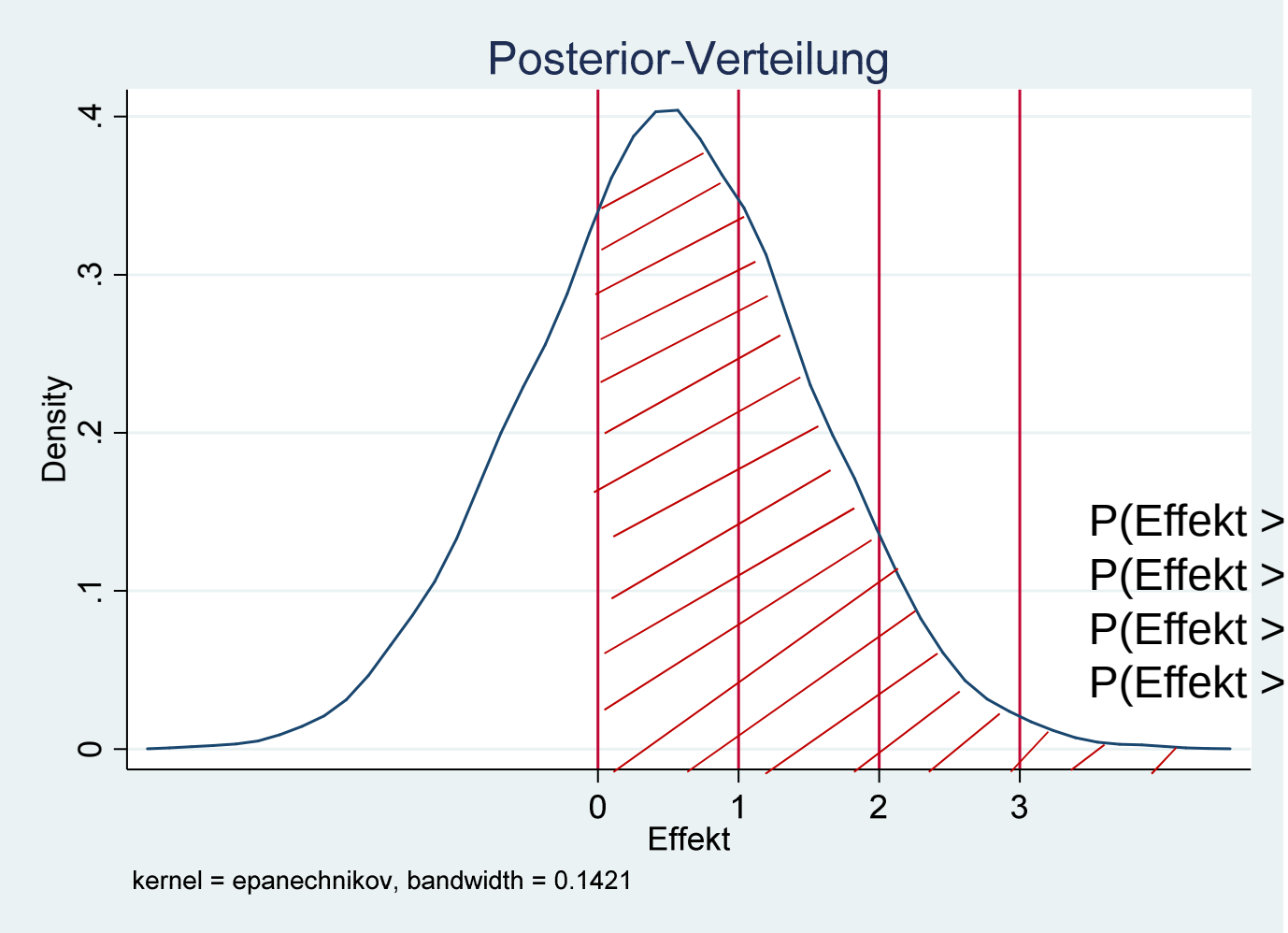


Frequentische Statistik bayesianisch betrachtet: Gleichverteilung

(„flat prior“, „uninformative prior“) ⇒ **likelihood** = **posterior**



Allgemein: Aus der **Posterior** ist jede Intervall-Wahrscheinlichkeit berechenbar



Fiktives Beispiel

In einem Experiment habe man zwei Gruppen à 50 Personen randomisiert in **X** = 0 und **X** = 1 eingeteilt.

X = 1: Übungen zum Training der Intelligenz durchgeführt.
X = 0: Magazin zum Lesen gegeben, z.B. „Beef!“.

Y = Wert in IQ-Test einen Tag später.

Frequentistische Auswertung

```
. regress y x

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(1, 98)        =
       Model |  916.334659         1  916.334659   Prob > F        =
    Residual |  25274.2761        98  257.900776   R-square        =
-------------+----------------------------------   Adj R-sq        =
       Total |  26190.6107        99  264.551624   Root MSE        =

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    6.054204   3.211858     1.88   0.062    -.3196234    12.42803
       _cons |    99.18632   2.271126    43.67   0.000     94.67934    103.6933
------------------------------------------------------------------------------
```

Unterschied zweiseitig nicht signifikant (p=.062), aber unter **X** = 1 könnte IQ um bis zu 12.4 höher sein!

Einseitiger p-Wert für H_1: Effekt > 0 = 0.062/2 = 0.031 (erlaubt, falls Schätzung von β das erwartete Vorzeichen hat (hier positiv), sonst p > 0.5)).

Bayesianische Auswertung

a. „flat prior“: gleiches Ergebnis
(bis auf zufällige Abweichungen — durch Simulationsmethode in der Berechnung):

```
. bayes , rseed(64674376) prior({ y:x }, normal(0, 1000000))   mcmcsize(100000) burnin(5000) thinning(2) savin
> g(post) : regress y x
note: discarding every other sample observation; using observations 1,3,5,...

Burn-in ...
Simulation ...

file post.dta saved

Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ regress(xb_y,{sigma2})

Priors:
      {y:x} ~ normal(0,1000000)                                       (1)
  {y:_cons} ~ normal(0,10000)                                         (1)
   {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.

Bayesian linear regression                      MCMC iterations  =    204,999
Random-walk Metropolis-Hastings sampling        Burn-in          =      5,000
                                                MCMC sample size =    100,000
                                                Number of obs    =        100
                                                Acceptance rate  =      .4402
                                                Efficiency:  min =      .1509
                                                             avg =      .2263
Log marginal likelihood = -434.50671                         max =      .3762

------------------------------------------------------------------------------
                                                              Equal-tailed
             |      Mean   Std. Dev.     MCSE     Median  [95% Cred. Interval]
-------------+----------------------------------------------------------------
y            |
           x |  6.109508   3.255388   .026502    6.12915  -.3341289   12.51243
       _cons |  99.12405    2.30417   .018705   99.13456   94.61286   103.6284
-------------+----------------------------------------------------------------
      sigma2 |  263.2974   38.48576   .198435   259.4171   198.5666   349.0749
------------------------------------------------------------------------------
Note: Default priors are used for some model parameters.
```

technische Details

Punktschätzung

SE

Bayesianisches Schätzintervall = „credibility interval“

Anmerkung: Es gibt hier keine „p-Werte“, aber man kann zum Hypothesenprüfen vorgehen wie auf der folgenden Folie.

Daraus berechnen, **wie wahrscheinlich** Differenz > 0 ist:

```
. bayestest interval {y:x}, lower(0) upper(100)

Interval tests     MCMC sample size =    10,000

       prob1 : 0 < {y:x} < 100
```

	Mean	Std. Dev.	MCSE
prob1	.9721	0.16469	.0038601

Gegeben die Daten und die Prior beträgt diese Wahrscheinlichkeit 97.2%!

Theoretisch = 1 - einseitiger p-Wert (.031) vorhin

.. Oder wie wahrscheinlich praktisch relevante Differenz von mindestens 10 (= 2/3*SD)

```
. bayestest interval {y:x}, lower(10) upper(100)

Interval tests     MCMC sample size =    10,000

       prob1 : 10 < {y:x} < 100
```

	Mean	Std. Dev.	MCSE
prob1	.116	0.32024	.008951

Gegeben die Daten und die Prior beträgt diese Wahrscheinlichkeit 11.6%!

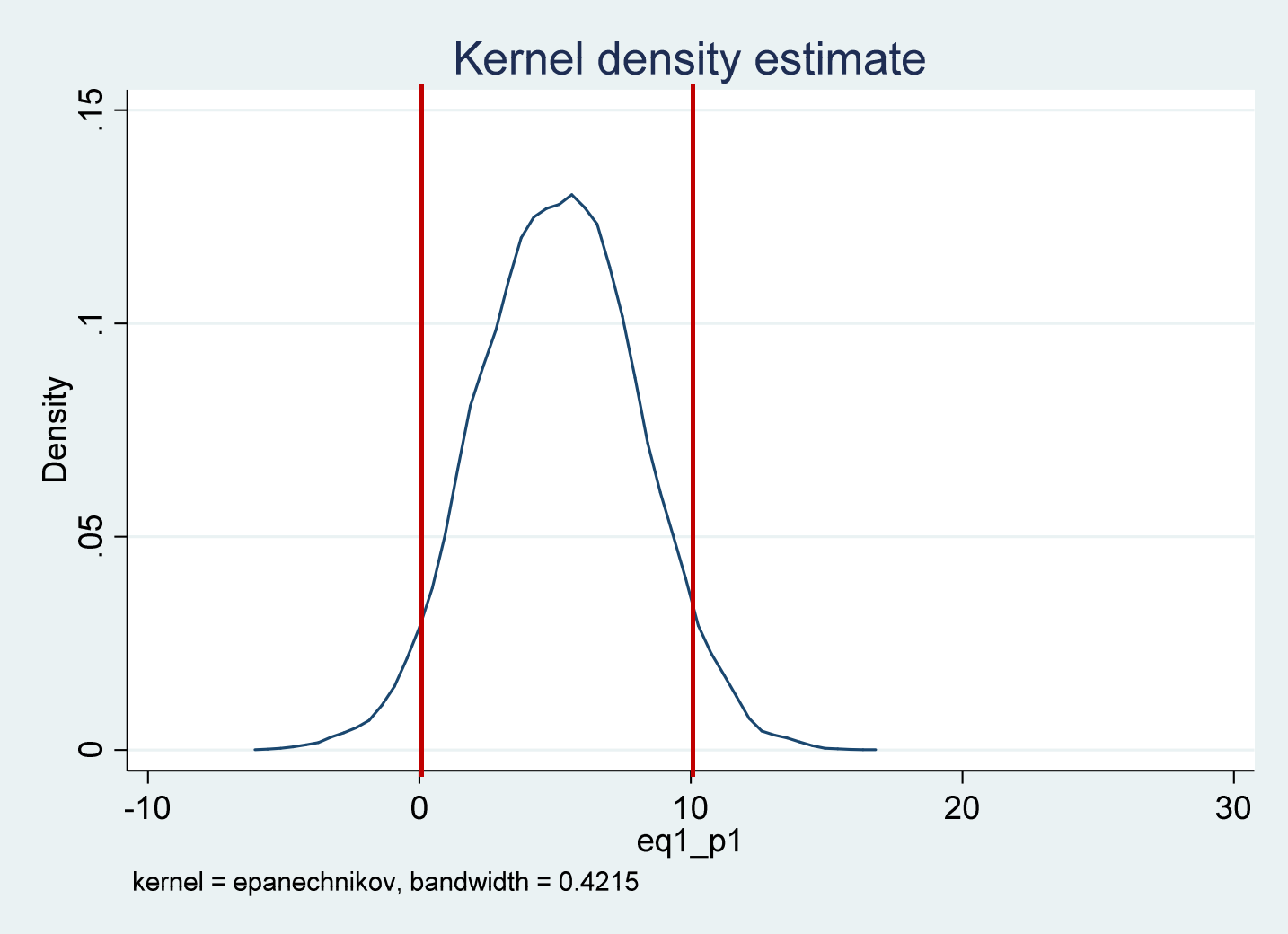
Andere Prior („informative“): N(0, 56.25), keine Präferenz für eine der beiden Bedingungen (Erwartungswert = 0), aber großer Effekt in beide Richtungen unwahrscheinlich (SD = √56=7.5: jeweils 2.5% W.keit, dass eine der beiden Bedingungen um mehr als 15 (SD der IQ-Verteilung) besser ist; bei Normalverteilung liegt 95% W.keitsmasse zwischen Erwartungswert und +/- 2*SD).

```
. bayes  , rseed(64674270) prior({y:x} , normal(0,56.25)) mcmcsize(10000) burnin(5000) thinning(2) : regress y x
note: discarding every other sample observation; using observations 1,3,5,...

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ regress(xb_y,{sigma2})

Priors:
      {y:x} ~ normal(0,56.25)                                       (1)
  {y:_cons} ~ normal(0,10000)                                       (1)
   {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.

Bayesian linear regression                      MCMC iterations  =     24,999
Random-walk Metropolis-Hastings sampling        Burn-in          =      5,000
                                                MCMC sample size =     10,000
                                                Number of obs    =        100
                                                Acceptance rate  =      .4444
                                                Efficiency:  min =      .1677
                                                             avg =      .2434
Log marginal likelihood = -429.99019                         max =      .3865

------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. Dev.     MCSE     Median  [95% Cred. Interval]
-------------+----------------------------------------------------------------
y            |
           x |   5.153889   2.933098   .069891   5.170395   -.4623962   10.96649
       _cons |   99.81853   2.221009   .054232   99.56019    95.38456   104.0596
-------------+----------------------------------------------------------------
      sigma2 |   263.2707    38.6592   .621865   259.0131     199.058    350.758
------------------------------------------------------------------------------
Note: Default priors are used for some model parameters.
```

Punktschätzung zu 0 hin verschoben (a-priori Erwartungswert = 0)

Schmaleres Schätzintervall als bei non-informative prior

Posterior



Effekt von Intelligenztraining auf IQ-Wert

Aus posterior ausrechnen, **wie wahrscheinlich** Differenz > 0 und Differenz > 10 sind:

```
. bayestest interval {y:x} , lower(0) upper(100)

Interval tests     MCMC sample size =    10,000

       prob1 : 0 < {y:x} < 100
```

	Mean	Std. Dev.	MCSE
prob1	.9625	0.18999	.0032323

Gegeben die Daten und die prior beträgt diese Wahrscheinlichkeit hier 96% (kaum geringer).

Oder ein praktisch relevanter Unterschied von mindestens 10 IQ-Punkten:

```
. bayestest interval {y:x} , lower(10) upper(100)

Interval tests     MCMC sample size =    10,000

       prob1 : 10 < {y:x} < 100
```

	Mean	Std. Dev.	MCSE
prob1	.0488	0.21546	.0036812

Gegeben die Daten und die prior beträgt diese Wahrscheinlichkeit hier nur 5%!

Gretchenfrage: Wie bestimmt man die Prior? „educated guess“

(falls man sie nicht aus anderen Daten bestimmen kann)

Es gibt viele mögliche Verteilungstypen einer Prior, auch für schiefe Verteilungen, z.B. Gamma-Verteilung.

Als ob man wetten würde

- Regressionskoeffizient β (Assoziation, Effekt) normalverteilt = viele, unabhängige Einflüsse auf Vorwissen
- Wert **e**, für den gilt: genauso viel darauf wetten, dass β > **e** wie dass β < **e** → Erwartungswert = **e**
- Bestimme Wert **sd**, sodass du 2:1 darauf wetten würdest, dass wahres β zwischen **e** - **sd** und **e** + **sd** liegt
  (bei Normalverteilung liegen ca. 2/3 Wahrscheinlichkeitsmasse zwischen **e** – Standardabweichung und **e** + Standardabweichung)
- Varianz = $\mathbf{sd}^2$
- Ergibt Prior = N(**e**, $\mathbf{sd}^2$)

Mittels „data equivalents“

(ausführliches Beispiel über Assoziation im Anhang)

- Beispiel Schätzung unbekannter Prävalenz für Depression
- Nehme an, Vorwissen entspreche einer **fiktiven Studie**, die bei **n** = 100 p = 10/100 = 10% Depression geschätzt hat
- Entspricht „Betaverteilung“* beta(a,b) = beta(p* (n-2) , (p-2)* (1-p))

* Verteilungsmodell für Wahrscheinlichkeiten

Int J Epidemiol. 2006 Jun;35(3):765-75. Epub 2006 Jan 30.

Bayesian perspectives for epidemiological research: I. Foundations and basic methods.

Greenland S[1].

Spezielle Bayes-Methoden:

- **„Empirical (objective) Bayes“:** Prior wird aus Daten geschätzt. Formel von Bayes dient einfach dazu, Wissen aktuell zu halten (z.B. neue Probanden machen ein Paradigma):

$$P(Effekt|neue\ Daten) = \frac{P(neue\ Daten|Effekt) \times P(Effekt|bisherige\ Daten)}{Konstante}$$

- **„Bayes-Faktor“:** Faktor, um den H_1 wahrscheinlicher wird, nachdem man die Daten analysiert hat

Kritik: die absolute Wahrscheinlichkeit für H_1 ist interessanter. Der Bayes-Faktor wird heuristisch eingeteilt in Stufen der Evidenz → dichotomia

Spezielle Bayes-Methoden:

- **„Semi-Bayes“:** nur Priors für manche Parameter spezifizieren, z.B. für ein β_1 = Zusammenhang von **X** und **Y**, aber nicht für β_0 = Grundniveau von **Y** (hierfür implizit flat prior verwendet; die Beispiele hier sind semi-bayesian)

- **„Reverse Bayes“: Wie unwahrscheinlich darf H_1 gerade noch sein**, damit man, gegeben die Daten, den Schluss ziehen kann, dass H_1 zutrifft?

$$P(H_1|neue\ Daten) = \frac{P(neue\ Daten|H_1) \times P(H_1|bisherige\ Daten)}{Konstante}$$

Welche Probleme löst die bayesianische Statistik nun?

1. Gedankenexperiment ✓

Entfällt, da man die gewünschte Wahrscheinlichkeit aus der Posterior enthält

2. Falsche Interpretation des p-Werts ✓

Bayesianische Wahrscheinlichkeiten haben die gewünschte Interpretation.

3. Dichotomia ✓—

Solange man Wahrscheinlichkeit aus Posterior nicht wieder dichotomisiert (kategorisiert).

4. Fehlinterpretation je nach Stichprobengröße —

$P(H_1|Daten)$ hängt (über Likelihood) von Stichprobengröße ab (jedoch weniger, umso informativer (geringere Varianz) die Prior ist).

5. Asymmetrie statistischer Tests ✓

H_0 und H_1 werden symmetrisch behandelt

$P(H_1|Daten)= 1- P(H_0|Daten)$

6. Statistisches Modell kann falsch sein —

Analytical bias (in der Likelihood) geht in die Posterior ein, erhält allerdings durch die Prior geringeres Gewicht.

7. Annahmen intransparent ✓

Annahmen stecken explizit in Prior, dadurch sollte diese Gegenstand der Diskussion über die Ergebnisse werden.

8. Instabile Schätzungen ✓

Die Posterior hat höchstens so große Varianz wie die Likelihood. Bei entsprechendem Vorwissen kann man Schätzungen stark stabilisieren (siehe Anhang).

9. Falscher Gebrauch von Statistik —

Durch die Wahl der Prior noch eine Ebene mehr, auf der man Variation in den Ergebnissen erzeugen kann.
→ Prior unbedingt vorher registrieren!

Pro und kontra Bayes

- **Studienergebnisse nicht vergleichbar**, wenn sie unterschiedliche Priors benutzen
  Ausweg: immer auch Likelihood-Verteilung angeben. Eine Leserin kann diese dann mit einer Prior ihrer Wahl kombinieren und damit ihre eigene Posterior erhalten.

- Bayes ist durch die Prior **subjektiv**, frequentistische Statistik objektiv
  Die **scheinbare Objektivität frequ. Statistik** basiert auf absurden Annahmen:
  1. kein Vorwissen — ohne dem würde man bestimmten Effekt gar nicht untersuchen; Forschung beginnt nie bei null; wenn man trotzdem bei null anfängt, verschwendet man Ressourcen (da unnötig große Stichproben).
  2. keinerlei Bias mit 100%iger Wahrscheinlichkeit

  Subjektivität lässt sich verringern, z.B. durch data equivalents.

Literatur

Baldwin SA, Larson MJ. An introduction to using Bayesian linear regression with clinical data. *Behaviour Research and Therapy* 2017; 98:58-75

Etz A, Vandekerckhove J. Introduction to Bayesian Inference for Psychology *Psychon Bull Rev* 2018; 25:5–34

Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol* 2006; 35:765-75

Greenland S. Putting background information about relative risks into conjugate priors. *Biometrics* 2001;57:663-70

Quintana DS, Williams DR: Bayesian alternatives for common null hypothesis significance tests in psychiatry: a non-technical guide using JASP. *BMC Psychiatry* 2018; 18:178

Anhang: **Beispiel für data equivalents und Stabilisierung von Schätzungen**

- kleine Stichprobe mit großem Zufallsfehler in der Schätzung eines **θ**
- Beinspiel für binäres **X**, binäres **Y**
- Numerisch sehr instabiles, schlecht zu interpretierendes, aber häufig benutztes Zusammenhangsmaß: odds ratio (OR)

		Faktor X		
		0	**1**	
Outcome Y	**0**	49 (98.9%)	14 (90%)	
	1	1 (1.1%)	6 (10%)	
		50 (100%)	20 (100%)	70

- OR geschätzt als 21 (2.3 – 189.3), unrealistisch groß!

```
. logistic Y X [fweight=freq]

Logistic regression                             Number of obs     =         70
                                                LR chi2(1)        =      11.27
                                                Prob > chi2       =     0.0008
Log likelihood = -17.119242                     Pseudo R2         =     0.2477

------------------------------------------------------------------------------
           Y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |          21   23.55844     2.71   0.007     2.329783     189.288
       _cons |   .0204082   .0206154    -3.85   0.000     .0028181    .1477908
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Research Methods & Reporting

Sparse data bias: a problem hiding in plain sight

BMJ 2016 ; 352 doi: https://doi.org/10.1136/bmj.i1981 (Published 27 April 2016)
Cite this as: *BMJ* 2016;352:i1981

Sander Greenland, professor of epidemiology and statistics [1], *Mohammad Ali Mansournia, assistant professor of epidemiology* [2], *Douglas G Altman, professor of statistics in medicine* [3]

[1] *Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA*
[2] *Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, PO box 14155-6446, Tehran, Iran*
[3] *Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK*

Correspondence to: M A Mansournia mansournia_ma@yahoo.com
Accepted 7 March 2016

- Stabilisieren mit Prior, z.B. ln(OR) ~ N(0,0.5)
- Entspricht OR = 1.0, 95%-CI = (0.25, 4)
- = wette 2/1 darauf, dass OR zwischen 0.5 und 2 liegt
  (halbe Intervalllänge = SD)

```
.  penlogit Y X [fweight=freq] if ind==0 , nprior(X  ln(1) 0.5) or


Penalized logistic regression                         No. of obs  =         70

Normal prior for X: exact prior median OR (95% PL): 1.00 (0.25, 4.00)
Data approx. equivalent to prior: cases=4.54  noncases=4.54  exp(offset)=1

Penalized log likelihood = -24.38717

------------------------------------------------------------------------------
           Y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   3.182524   1.695816     2.17   0.030     1.119972      9.0435
       _cons |   .0709895   .0342561    -5.48   0.000     .0275705     .182786
------------------------------------------------------------------------------
```

- OR nun viel stabiler geschätzt: 3.2 (1.1 – 9.0).

Beobachtet

Prior entspricht diesem Auffüllen der Daten

Nachgerechnet, man muss nach Indikator `ind` adjustieren:

		Häufigkeit in Zelle	Indikatorvariable für aufgefüllte Daten
X	Y	freq	ind
0	0	49	0
0	1	1	0
1	0	14	0
1	1	6	0
0	0	100000	1
0	1	4.5	1
1	0	100000	1
1	1	4.5	1

```
.  logistic  Y  factor ind [iweight=freq]

Logistic regression                             Number of obs     =          8
                                                LR chi2(2)        =      94.86
                                                Prob > chi2       =     0.0000
Log likelihood = -118.56639                     Pseudo R2         =     0.2857

------------------------------------------------------------------------------
           Y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      factor |   3.174619   1.78973     2.05   0.040     1.051518    9.584441
         ind |      .0003   .000164   -14.83   0.000     .0001027    .0008761
       _cons |   .0710765   .0348868    -5.39   0.000     .0271599    .1860051
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Int J Epidemiol. 2006 Jun;35(3):765-75. Epub 2006 Jan 30.

Bayesian perspectives for epidemiological research: I. Foundations and basic methods.

Greenland S[1].