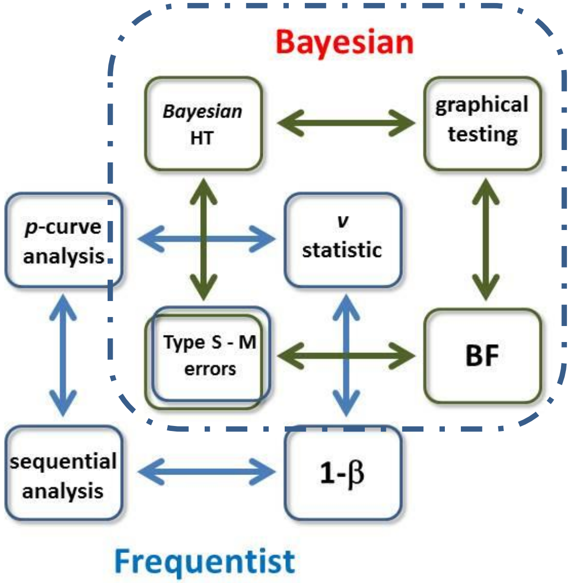
Methodological workshop

Frequentist and Bayesian approaches to improving your statistical inferences

Luigi Lombardi

Dept. of Psychology and Cognitive Science, University of Trento



Part 2

UNIVERSITÀ DEGLI STUDI DI TRENTO

1 Problems with the null hypothesis (N-H) testing approach

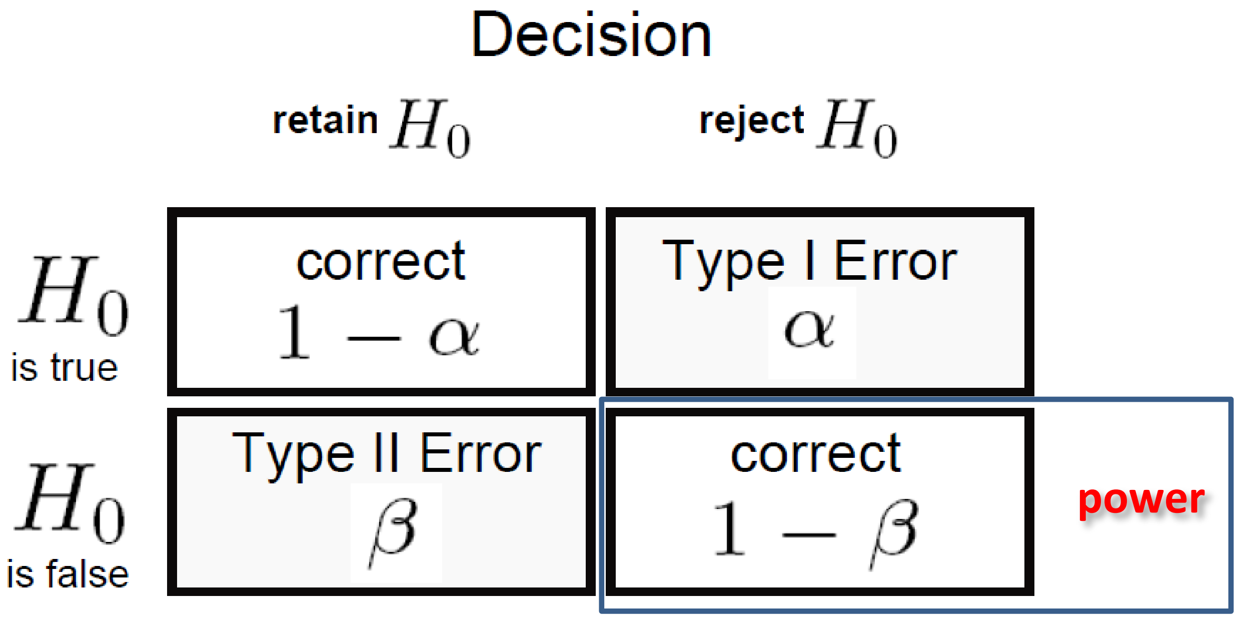
The Neyman-Pearson paradigm (N-H)

- In the Null Hypothesis (N-H) approach, the probability distributions are grouped into two aggregates:
  - H_0: the **null hypothesis**
  - H_A: the **alternative hypothesis**

  (there are several common variations on this notation; the alternative hypothesis, for example, is sometimes denoted as H_1 or even K.)

- The alternative hypothesis H_A is the **logical negation** of the null hypothesis H_0, and *vice versa*.

1 Problems with the null hypothesis (N-H) testing approach



	Decision retain H_0	 reject H_0
H_0 is true	correct $1 - \alpha$	Type I Error α
H_0 is false	Type II Error β	correct $1 - \beta$

power

The N-H table

Probabilistic interpretation

- **Type I error** α: $P(\text{reject } H_0 | H_0 \text{ is true})$
- **Type II error** β: $P(\text{retain } H_0 | H_0 \text{ is false})$
- **Power** $1 - \beta$: $P(\text{reject } H_0 | H_0 \text{ is false}) = 1 - P(\text{retain } H_0 | H_0 \text{ is false})$
- $1 - \alpha$: $P(\text{retain } H_0 | H_0 \text{ is true}) = 1 - P(\text{reject } H_0 | H_0 \text{ is true})$

Note: these are conditional probabilities!! The p-**value** is

$$P(T \text{ at least as extreme as } v^* | H_0 \text{ is true})$$

with v^* being the value of the observed statistic $T(\mathbf{x})$.

1 Problems with the null hypothesis (N-H) testing approach

Graphical interpretation

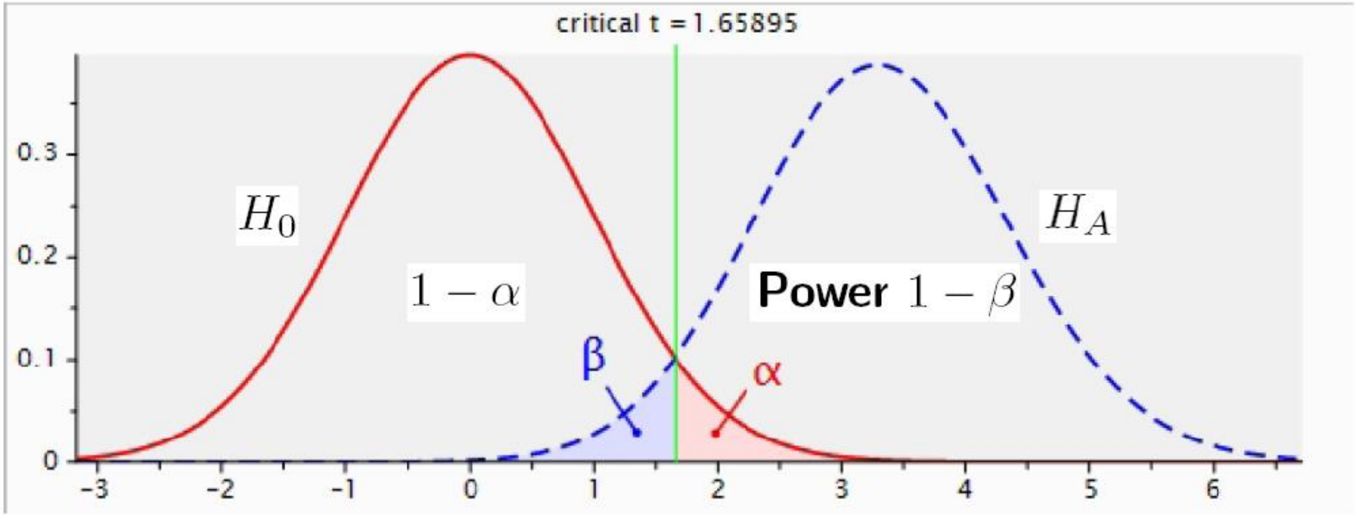


Figure 3: Type I error and type II error for a t statistic.

Note that in an ideal situation the test T would have $\alpha = \beta = 0$, but this is not feasible in practice. For real data, it is always the case that, for a fixed sample size N, **in order to decrease α, the probability β must be increased**, and vice versa.

Decision rules (one tailed)

- **Decision rule** Ψ (based on the critical value and the observed statistic):

$$\Psi(v_c, v^*) = \begin{cases} \text{retain } H_0 & \text{if} \quad v^* \leq v_c \\ \text{reject } H_0 & \text{if} \quad v^* > v_c \end{cases} \qquad (1)$$

- **Decision rule** Φ (based on α and the p-value):

$$\Phi(\alpha, \text{p-value}) = \begin{cases} \text{retain } H_0 & \text{if} \quad p\text{-value} \geq \alpha \\ \text{reject } H_0 & \text{if} \quad p\text{-value} < \alpha \end{cases} \qquad (2)$$

1 Problems with the null hypothesis (N-H) testing approach

Connection between Ψ e Φ (one tailed)

$$\begin{tikzcd}
\boxed{\alpha} \arrow[r, phantom, ">"] \arrow[d, Leftrightarrow] & \boxed{\text{p-value}} \arrow[r, Rightarrow] \arrow[d, Leftrightarrow] & H_A \text{ (on the basis of } \Phi) \\
\boxed{v_c} \arrow[r, phantom, "<"] & \boxed{v^*} \arrow[r, Rightarrow] & H_A \text{ (on the basis of } \Psi)
\end{tikzcd}$$

$$\begin{tikzcd}
\boxed{\alpha} \arrow[r, phantom, "\leq"] \arrow[d, Leftrightarrow] & \boxed{\text{p-value}} \arrow[r, Rightarrow] \arrow[d, Leftrightarrow] & H_0 \text{ (on the basis of } \Phi) \\
\boxed{v_c} \arrow[r, phantom, "\geq"] & \boxed{v^*} \arrow[r, Rightarrow] & H_0 \text{ (on the basis of } \Psi)
\end{tikzcd}$$

Replicability problem

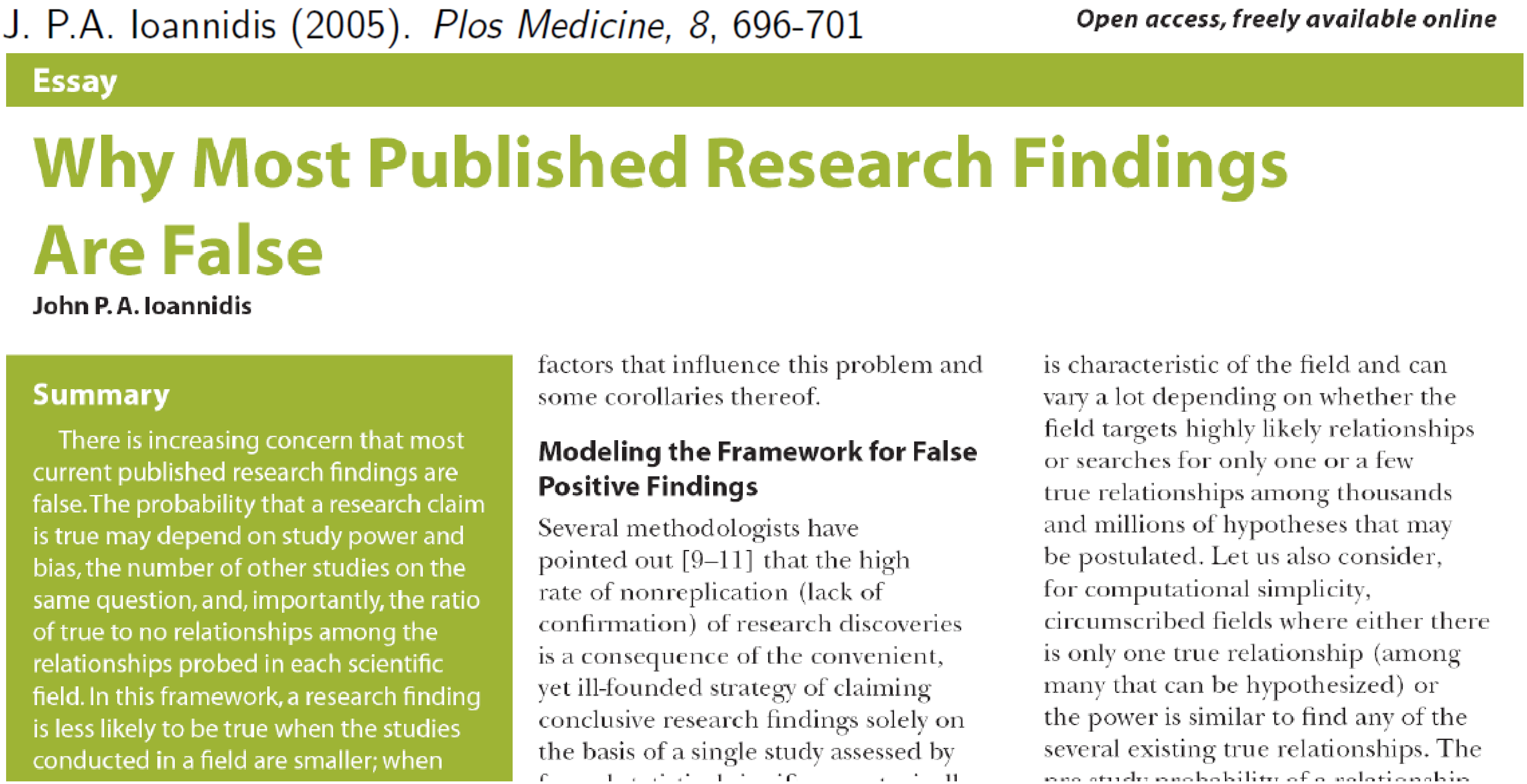
There is increasing concern that most current published research findings are suffering from high rate of nonreplication (lack of confirmation) of their results.

According to some researchers, this is a consequence of applying standard statistical paradigms to derive research findings (adoption of formal statistical significance, e.g., p-value less than 0.05).

empirical researches plagued by *false positive findings*

Replicability problem

J. P.A. Ioannidis (2005). *Plos Medicine*, *8*, 696-701



Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The

Replicability problem

According to Ioannidis (2005)

"high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05" (p. 696).

See the special issue

Perspective in Psychological Science, 2014, Vol 9(1)

1 Problems with the null hypothesis (N-H) testing approach

Replicability problem

The basic problem is that we are usually interested in the posterior conditional probability

$$P(\underline{H_A}|\text{reject } H_0) \equiv P(H_0 \text{ is false}|\text{reject } H_0)$$

Alternative hypothesis is true

1 Problems with the null hypothesis (N-H) testing approach

This posterior probability represents the positive predictive value (PPV) of a true finding. That is to say

$$PPV = P(H_A|\text{reject } H_0).$$

Note the difference:

- **Type I error** α: $P(\text{reject } H_0|H_0 \text{ is true})$
- **Type II error** β: $P(\text{retain } H_0|H_0 \text{ is false})$
- **Power** $1-\beta$: $P(\text{reject } H_0|H_0 \text{ is false}) = 1 - P(\text{retain } H_0|H_0 \text{ is false})$
- $1-\alpha$: $P(\text{retain } H_0|H_0 \text{ is true}) = 1 - P(\text{reject } H_0|H_0 \text{ is true})$

1 Problems with the null hypothesis (N-H) testing approach

$$P(H_A|\text{reject } H_0) = \frac{P(H_A)P(\text{reject } H_0|H_A)}{P(H_A)P(\text{reject } H_0|H_A) + P(H_0)P(\text{reject } H_0|H_0)}$$

1 Problems with the null hypothesis (N-H) testing approach

$$\begin{aligned} P(H_A|\text{reject } H_0) &= \frac{P(H_A)P(\text{reject } H_0|H_A)}{P(H_A)P(\text{reject } H_0|H_A) + P(H_0)P(\text{reject } H_0|H_0)} \\ &= \frac{P(H_A)(1-\beta)}{P(H_A)(1-\beta) + P(H_0)\alpha} \end{aligned}$$

1 Problems with the null hypothesis (N-H) testing approach

prior probability of the alternative hypothesis $P(H_A)$

$$P(H_A|\text{reject } H_0) = \frac{P(H_A)P(\text{reject } H_0|H_A)}{P(H_A)P(\text{reject } H_0|H_A) + P(H_0)P(\text{reject } H_0|H_0)}$$

$$= \frac{P(H_A)(1-\beta)}{P(H_A)(1-\beta) + P(H_0)\alpha} = PPV$$

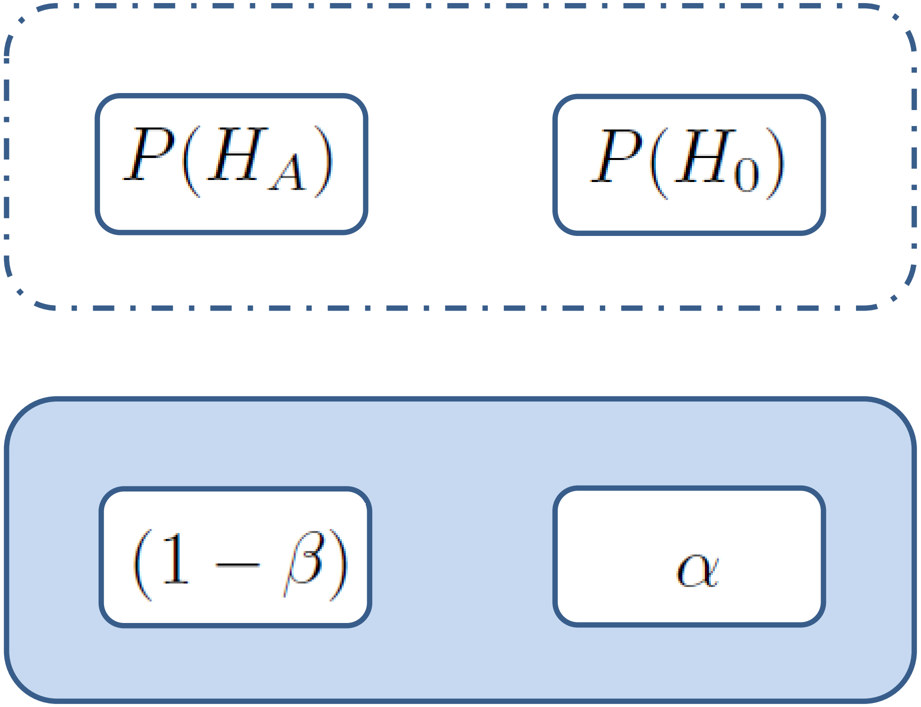
prior probability of the null hypothesis $P(H_0)$

1 Problems with the null hypothesis (N-H) testing approach

By using a similar representation we can also derive the **negative predictive value** $P(H_0|\text{reject } H_0)$:

$$NPV \;=\; 1 - PPV$$

1 Problems with the null hypothesis (N-H) testing approach



$P(H_A)$ $P(H_0)$

$(1-\beta)$ α

How do we compute/estimate these values?

Probability terms
In the PPV

(usually)
Theorical values

1 Problems with the null hypothesis (N-H) testing approach

$P(H_A)$ $P(H_0)$

How do we compute/estimate these values?

Ioannidis reported some procedures to compute the prior probability H0 on the basis of prior information, empirically based meta-analytic information, case scenario analysis, and expecially the so called potential bias

J. P.A. Ioannidis (2005). *Plos Medicine, 8*, 696-701

1 Problems with the null hypothesis (N-H) testing approach

According to Ioannidis (2005), a **bias** is the combination of various design, data, analysis, and presentation factors that tend to produce research findings when *they should not be produced*.

$$\boxed{P(H_A) = \frac{u\left(\frac{n_A}{n_0}\right)}{u\left(\frac{n_A}{n_0}\right) + 1}}$$

(*Pre-Study Odds*)

$$n_A, n_0 > 0$$

Let $u \in [0,1]$ be the proportion of probed analyses that would not have been research findings (negative results), but nevertheless end up presented and reported as positive ones, because of bias.

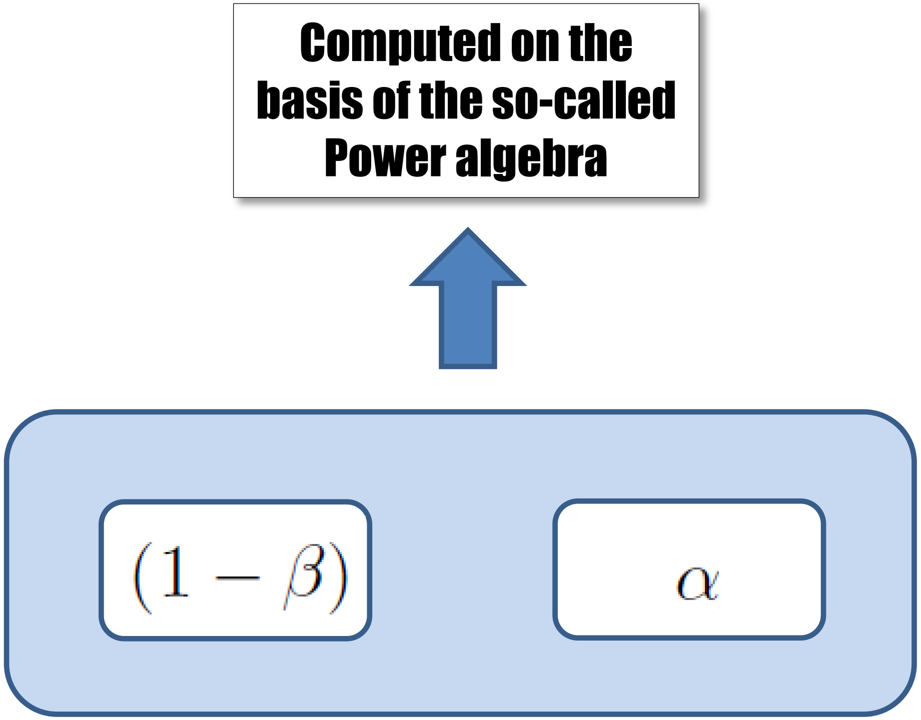
The six corollaries

- **Corollary 1:** "The smaller the studies conducted in a scientific field, the less likely the research findings are to be true."
- **Corollary 2:** "The smaller the effect sizes in a scientific field, the less likely the research findings are to be true."
- **Corollary 3:** "The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true."

The six corollaries

- **Corollary 4:** "The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true."
- **Corollary 5*:** "The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true."
- **Corollary 6*:** "The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true."

1 Problems with the null hypothesis (N-H) testing approach



Computed on the basis of the so-called Power algebra

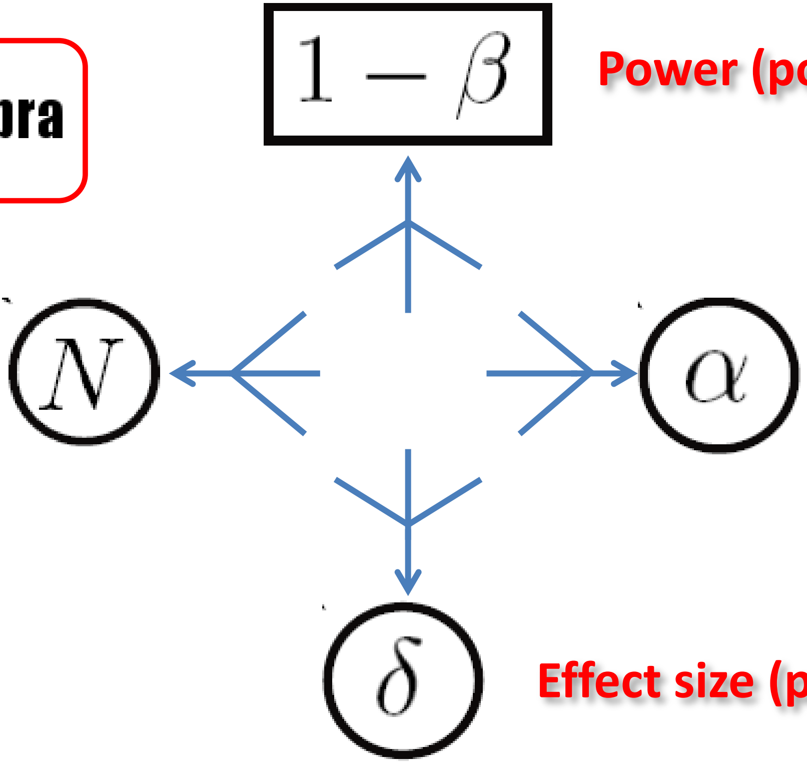
$(1-\beta)$

α

1 Problems with the null hypothesis (N-H) testing approach

Power analysis is based on four different parameters:

The power algebra



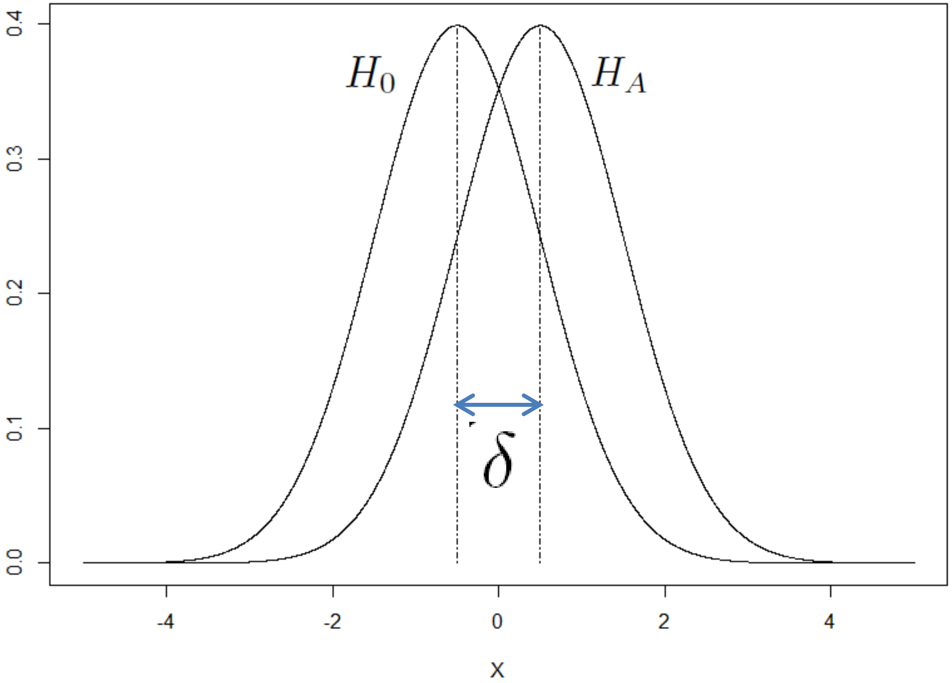
Power (population level)

Hypothetical Sample size

Type I error (population level)

Effect size (population level)

1 Problems with the null hypothesis (N-H) testing approach

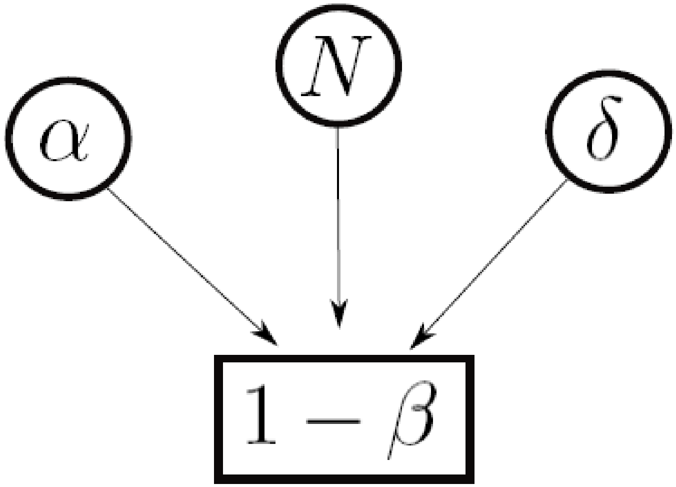


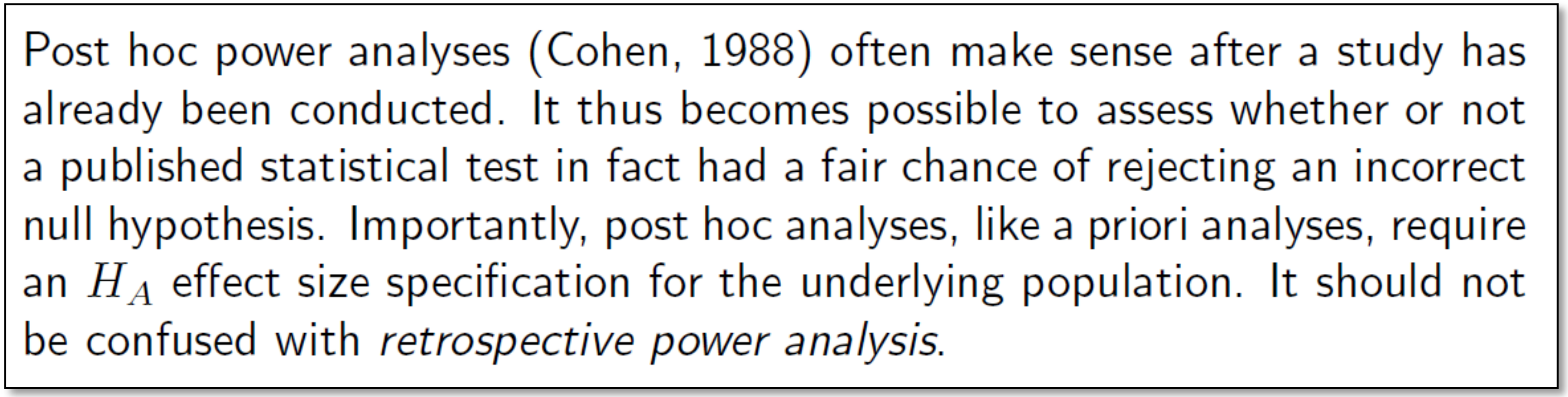
Effect size parameter defining H_A; it represents the degree of deviation from H_0 in the underlying population

δ

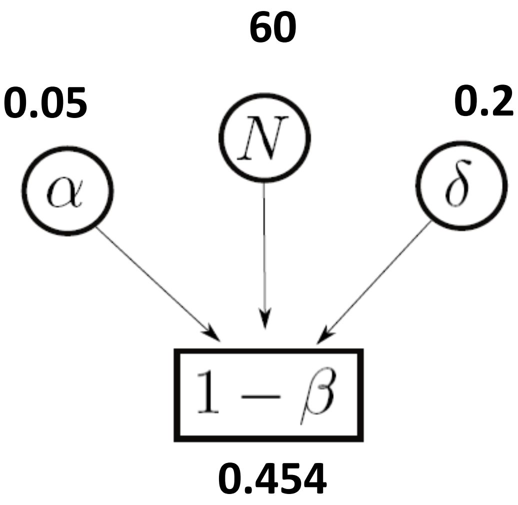
Effect size (population level)

Post hoc power analysis




Post hoc power analyses (Cohen, 1988) often make sense after a study has already been conducted. It thus becomes possible to assess whether or not a published statistical test in fact had a fair chance of rejecting an incorrect null hypothesis. Importantly, post hoc analyses, like a priori analyses, require an H_A effect size specification for the underlying population. It should not be confused with *retrospective power analysis*.

Post hoc power analysis: an example using the pwr package



One-sample t-test: H0 $\mu \leq 0$

```
pwr.t.test(d=0.2,n=60,sig.level=0.05,power=NULL,type=
"one.sample",alternative="greater")
```

R syntax

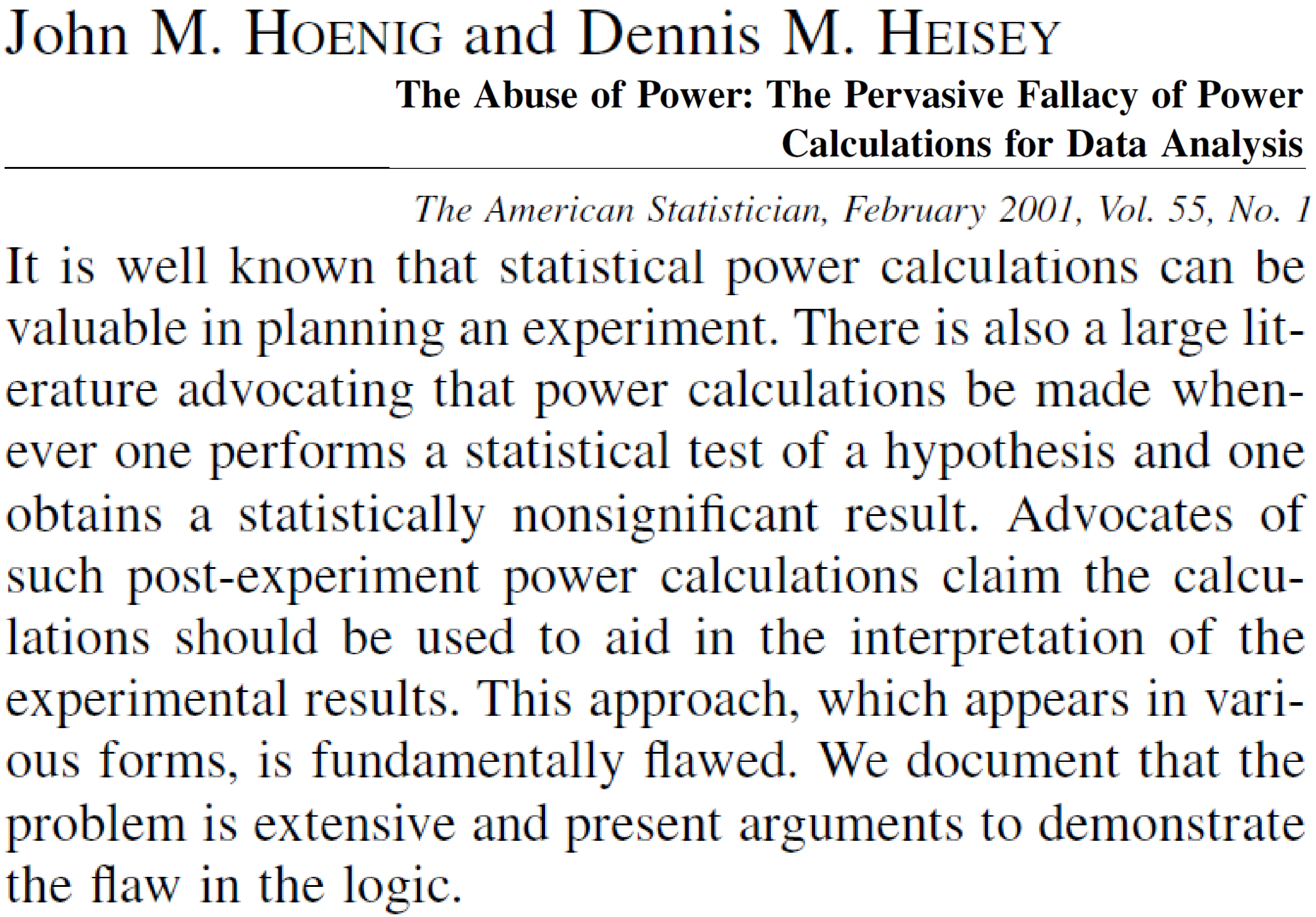
```
   One-sample t test power calculation

              n = 60
              d = 0.2
      sig.level = 0.05
          power = 0.4548365
    alternative = greater
```

R output

1 Problems with the null hypothesis (N-H) testing approach

The power fallacy



John M. Hoenig and Dennis M. Heisey

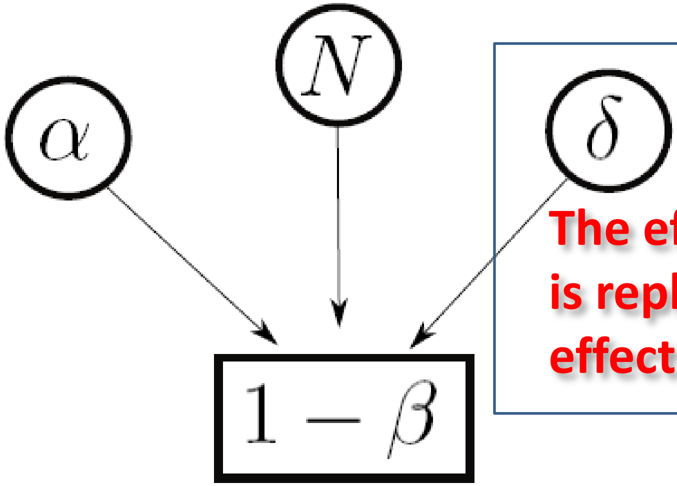
The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis

The American Statistician, February 2001, Vol. 55, No. 1

It is well known that statistical power calculations can be valuable in planning an experiment. There is also a large literature advocating that power calculations be made whenever one performs a statistical test of a hypothesis and one obtains a statistically nonsignificant result. Advocates of such post-experiment power calculations claim the calculations should be used to aid in the interpretation of the experimental results. This approach, which appears in various forms, is fundamentally flawed. We document that the problem is extensive and present arguments to demonstrate the flaw in the logic.

Observed power analysis

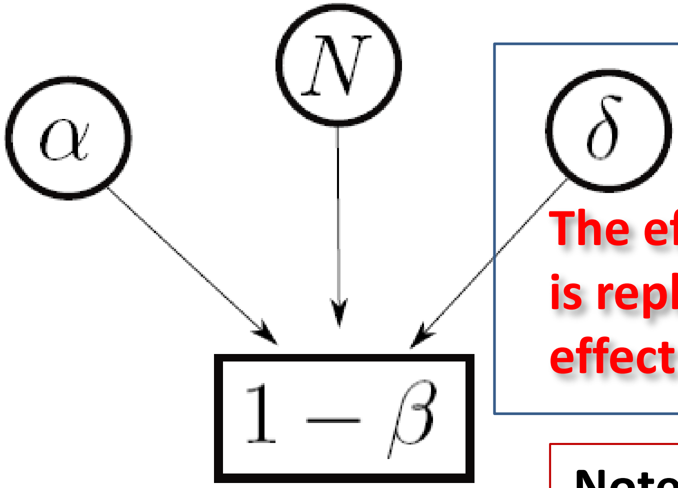
The basic idea of observed power analysis is that there is evidence for the null hypothesis being true if $p > \alpha$ and the computed power is high at the observed effect size *d*



The effect size (at population level) is replaced with the observed effect size *d* (at the sample level)

1 Problems with the null hypothesis (N-H) testing aproach

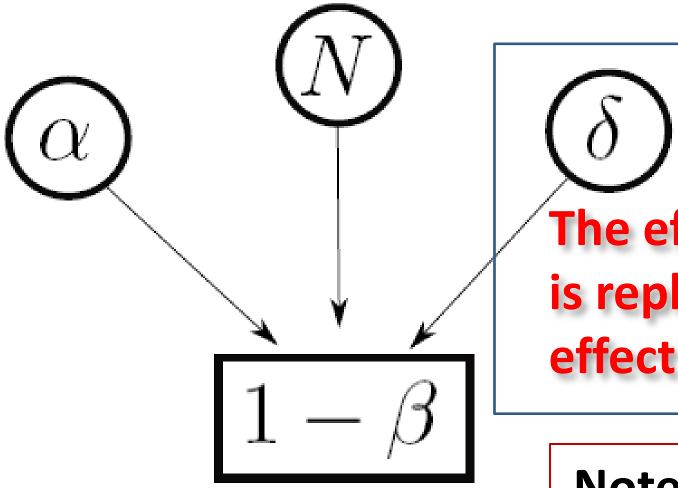
Observed power analysis



The effect size (at population level) is replaced with the observed effect size *d* (at the sample level)

Note *d* is not a theoretical value (hypothetical value)

Observed power analysis

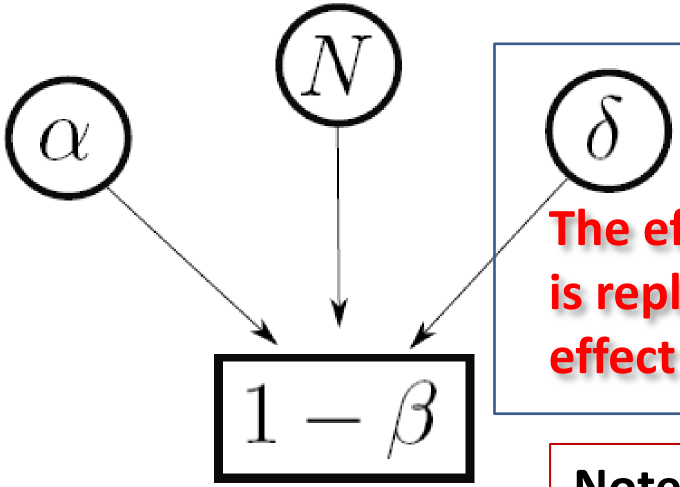


The effect size (at population level) is replaced with the observed effect size *d* (at the sample level)

Note *d* is not a theoretical value (hypothetical value)

It is estimated from the sample according to the theoretical model for the null hypothesis

Observed power analysis



The effect size (at population level) is replaced with the observed effect size *d* (at the sample level)

Note *d* is not a theoretical value (hypothetical value)

It is estimated from the sample according to the theoretical model for the null hypothesis

It is biased!!!

Observed power analysis – hypothetical derivations

Basic power analysis claim:

($p > \alpha$) AND (power is high) entails «evidence for H_0 is high»

Some 'derivations':

NOT [($p > \alpha$) AND (power is high)] iff

NOT($p > \alpha$) OR NOT(power is high)

Some 'derivations':

1. NOT($p > \alpha$) AND (power is high) entails ??
2. ($p > \alpha$) AND NOT(power is high) entails ??
3. NOT($p > \alpha$) AND NOT(power is high) entails ??

1 Problems with the null hypothesis (N-H) testing approach

Observed power analysis – hypothetical derivations

Some interpretations:

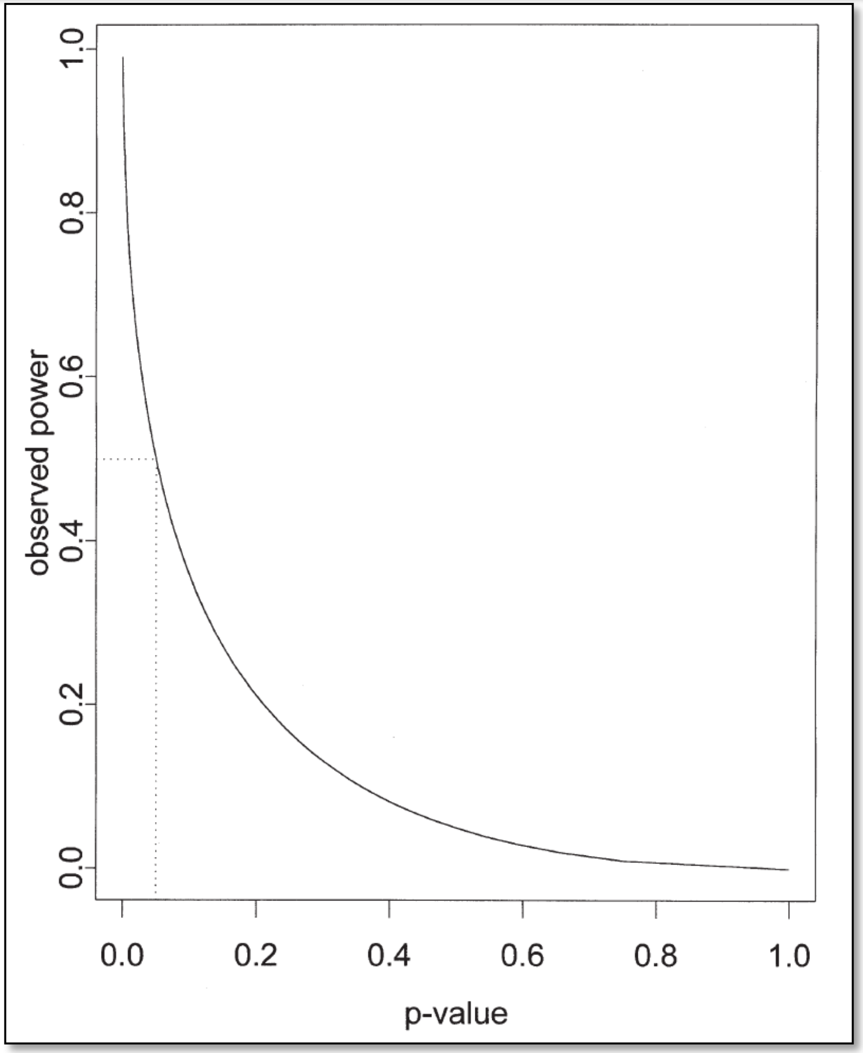
($p > \alpha$) AND NOT(power is high) entails «evidence for H_0 is weak»

The underlying idea is: if we increase the sample size, then we raise the power, and probably we can reject H_0!

However some of these interpretations lead us to the a paradox!

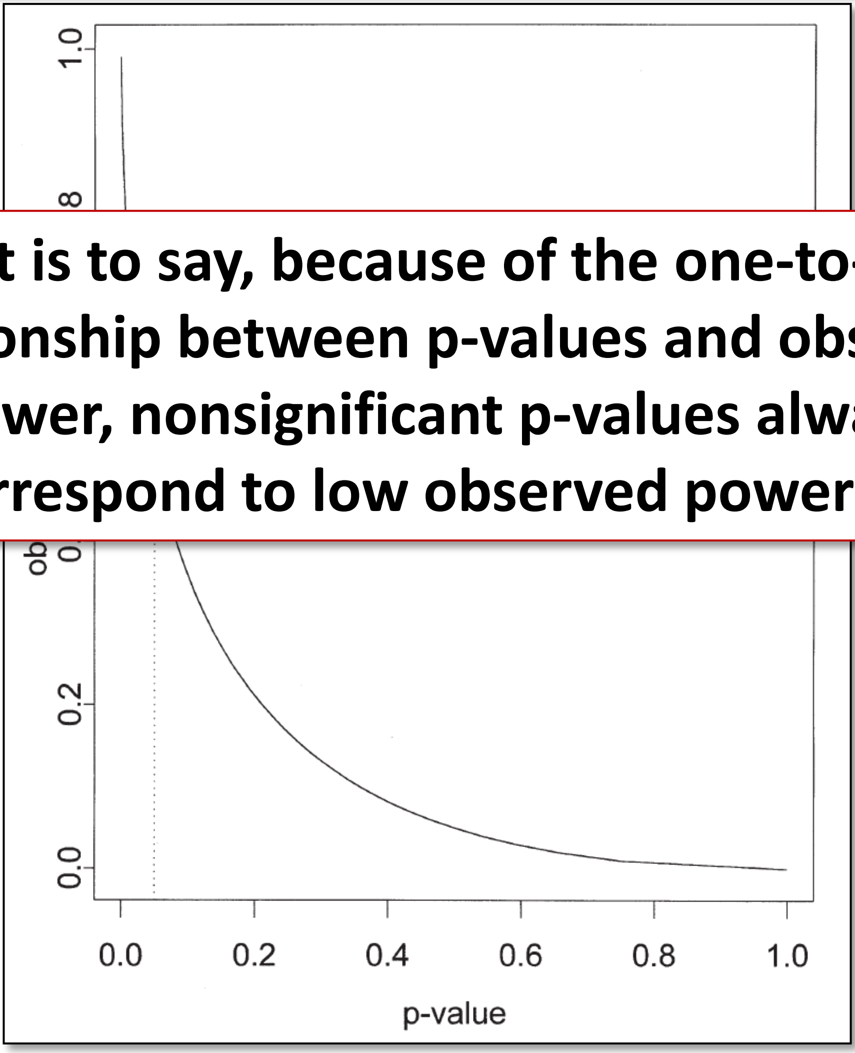
The power approach paradox (PAP)

1 Problems with the null hypothesis (N-H) testing approach



There is a negative monotonic relationship between observed power and p-value!

1 Problems with the null hypothesis (N-H) testing approach

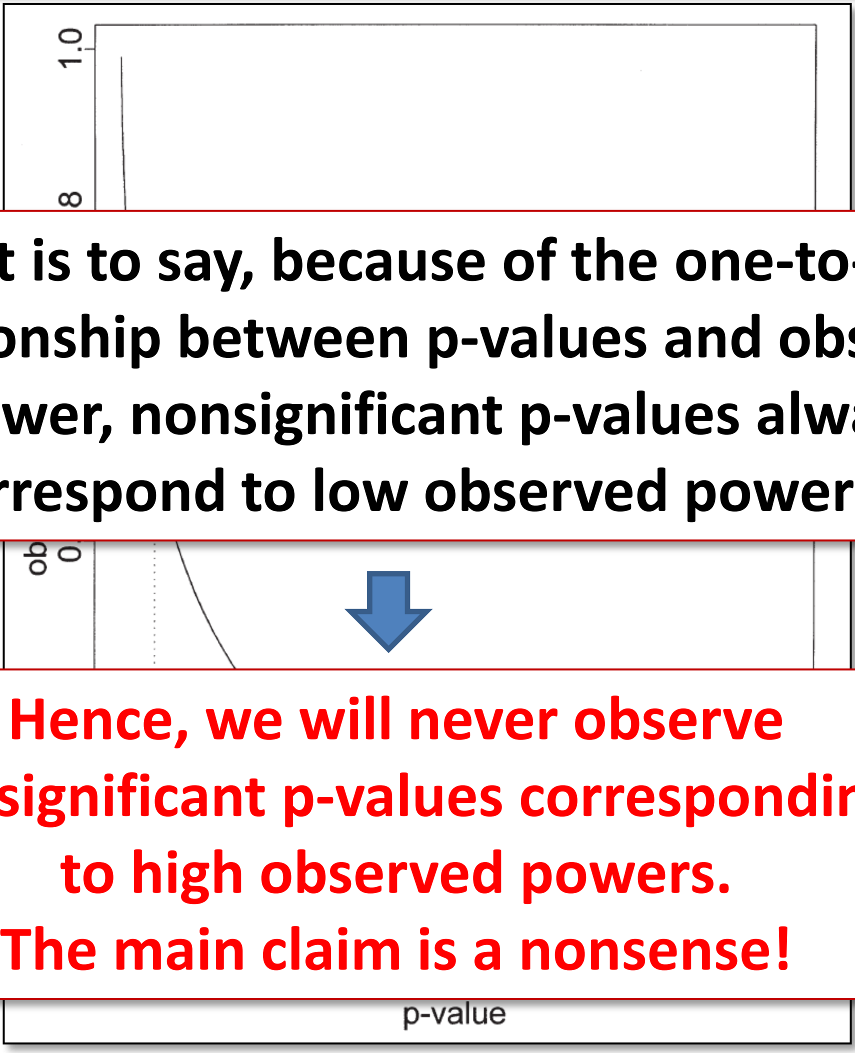


That is to say, because of the one-to-one relationship between p-values and observed power, nonsignificant p-values always correspond to low observed powers!!!

re is a gative otonic ionship between observed power and p-value!

1

Problems with the null hypothesis (N-H) testing approach



That is to say, because of the one-to-one relationship between p-values and observed power, nonsignificant p-values always correspond to low observed powers!!!

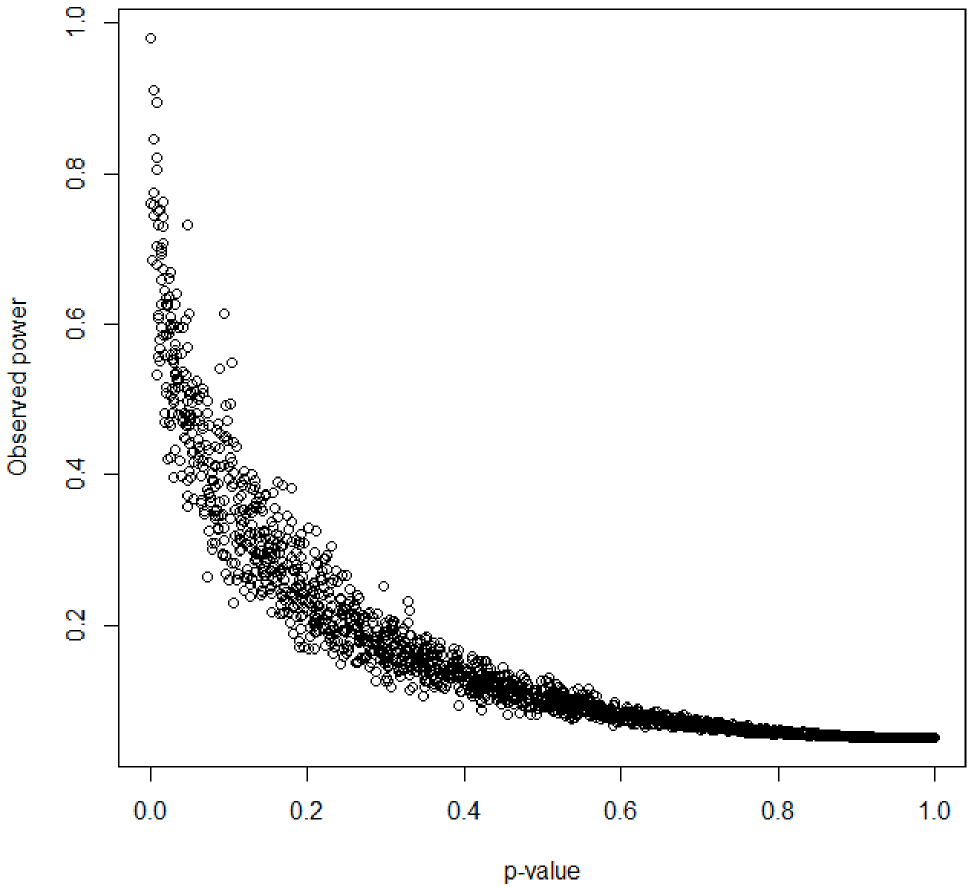
re is a gative otonic ionship between observed power and p-value!

Hence, we will never observe nonsignificant p-values corresponding to high observed powers.

The main claim is a nonsense!

1 Problems with the null hypothesis (N-H) testing approach

relationship between observed power and p-value – simulation study



1 Problems with the null hypothesis (N-H) testing approach

One-sample t-test: H0 $\mu_1 = 0$ (simulation study)

```
n <- 50
mu0 <- 0
sd <- 1
B <- 2000
simPv <- rep(0,B)
simPw <- rep(0,B)

for (b in 1:B) {

  X <- rnorm(n,mu0,sd)
  dobs <- (mean(X))/sqrt(((n-1)*sd^2)/(n-1))
  simPv[b] <- t.test(X)$p.value
  simPw[b] <- pwr.t.test(d=dobs,n=n,sig.level=0.05,power=NULL,
  type="one.sample",alternative="two.sided")$power

}

plot(simPv,simPw,ylab="Observed power", xlab="p-value")
```

R syntax

2 Beyond power calculations

2

Beyond power calculations

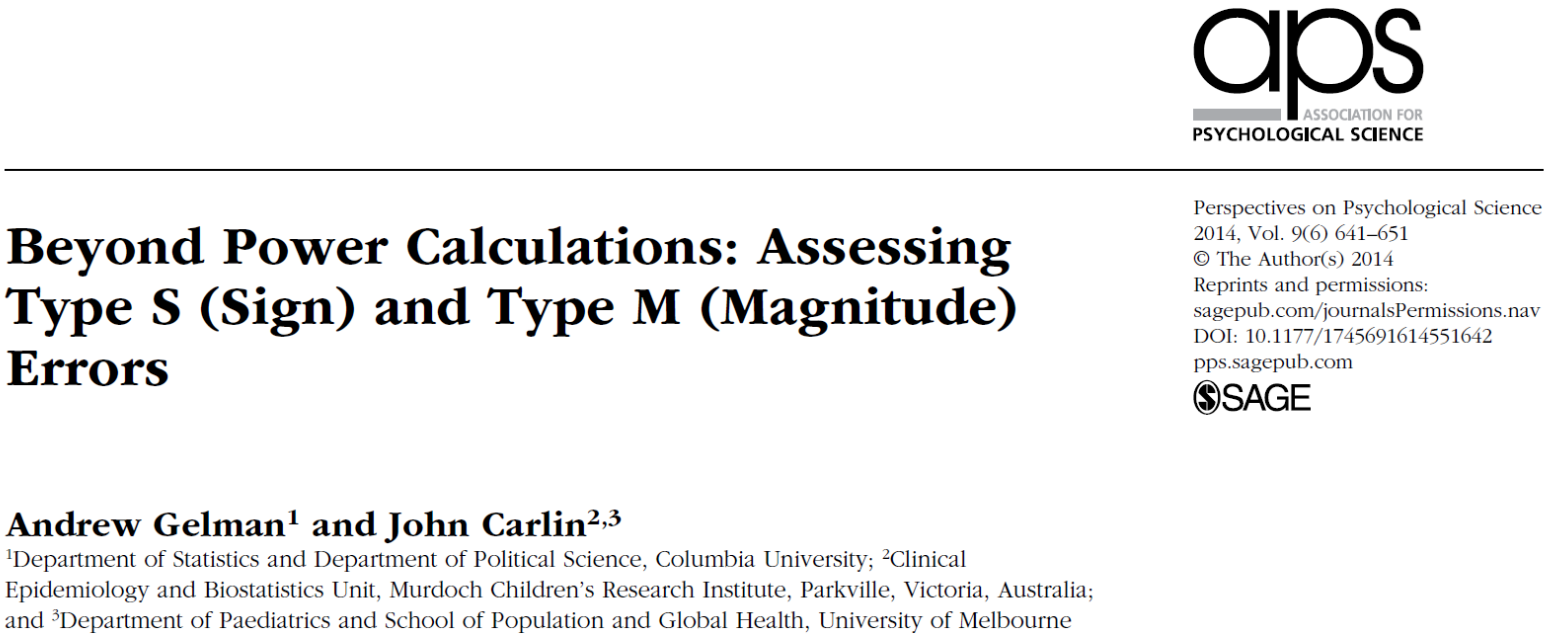
One of the main problems of standard power analysis is that it puts a narrow emphasis on statistical significance which is the primary focus of many study designs. However, in noisy, small-sample settings, statistically significant results can often be misleading. This is particularly true when observed power analysis is used to evaluate the statistical results.

2 Beyond power calculations

A better approach would be

Design Analysis (DA): a set of statistical calculations about **what could happen under hypothetical replications of a study** (that focuses on estimates and uncertainties rather than on statistical significance)

2 Beyond power calculations



Perspectives on Psychological Science
2014, Vol. 9(6) 641–651
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614551642
pps.sagepub.com
SAGE

Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors

Andrew Gelman[1] and John Carlin[2,3]
[1]Department of Statistics and Department of Political Science, Columbia University; [2]Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria, Australia; and [3]Department of Paediatrics and School of Population and Global Health, University of Melbourne

Somehow this work represents a kind of conceptual «bridge» linking the Frequentist approach with a more Bayesian oriented perspective

DA main tokens

$d \in \mathbb{R}$ The observed effect

$D \in \mathbb{R}$ The true population effect

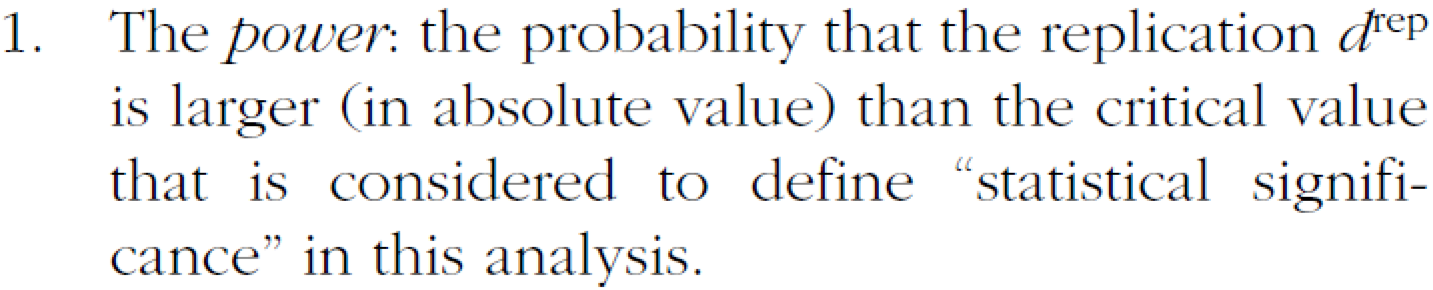
$s \in \mathbb{R}^+$ The standard error (SE) of the observed effect

$\alpha = 0.05$ The Type I error

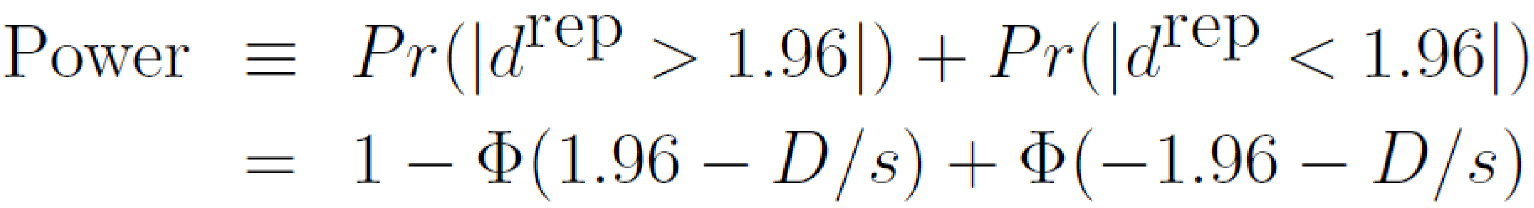
$d^{\text{rep}} \sim N(D, s)$ A hypothetical normally distributed random variable with parameters D and s (note this constitutes a conceptual leap)

DA main tokens

The main goals are to compute:



1. The *power*: the probability that the replication d^{rep} is larger (in absolute value) than the critical value that is considered to define "statistical significance" in this analysis.



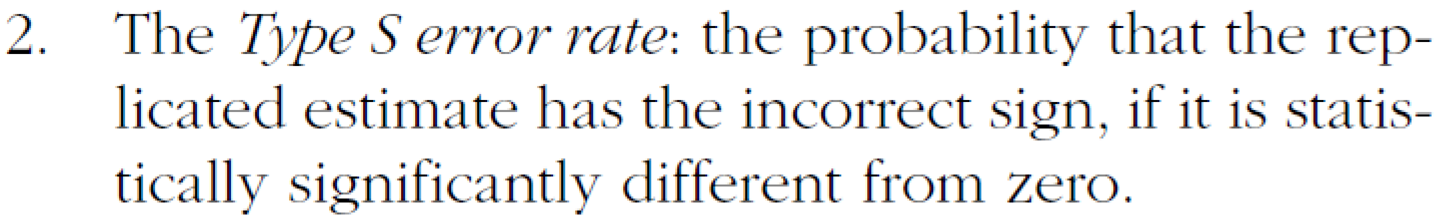
$$\begin{aligned}\text{Power} &\equiv Pr(|d^{\text{rep}} > 1.96|) + Pr(|d^{\text{rep}} < 1.96|)\\ &= 1 - \Phi(1.96 - D/s) + \Phi(-1.96 - D/s)\end{aligned}$$

Φ being the cumulative standard normal distribution

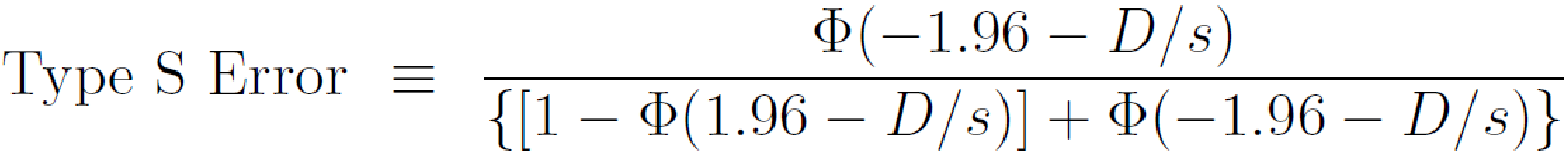
2 Beyond power calculations

DA main tokens

The main goals are to compute:



2. The *Type S error rate*: the probability that the replicated estimate has the incorrect sign, if it is statistically significantly different from zero.



$$\text{Type S Error} \equiv \frac{\Phi(-1.96 - D/s)}{\{[1 - \Phi(1.96 - D/s)] + \Phi(-1.96 - D/s)\}}$$

DA main tokens

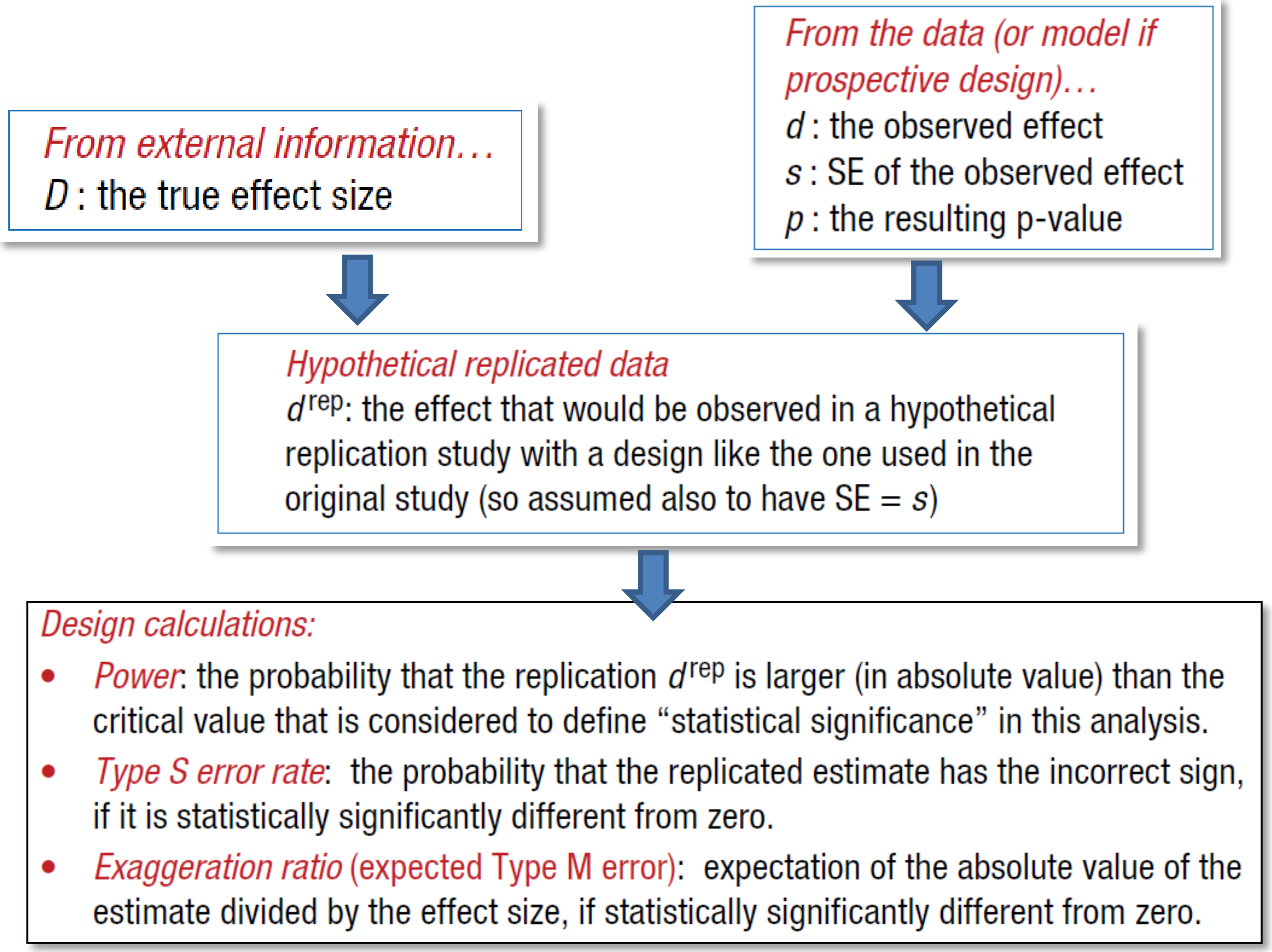
The main goals are to compute:

3. The *exaggeration ratio* (expected Type M error): the expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero.

$$\text{Type M Error} \equiv \frac{\mathbb{E}[d_+^{\text{rep}} | d_+^{\text{rep}} > 1.96]}{D}$$

$$d_+^{\text{rep}} = |d^{\text{rep}}|$$

2 Beyond power calculations



From external information…
D : the true effect size

From the data (or model if prospective design)…
d : the observed effect
s : SE of the observed effect
p : the resulting p-value

Hypothetical replicated data
d^{rep}: the effect that would be observed in a hypothetical replication study with a design like the one used in the original study (so assumed also to have SE = s)

Design calculations:

- *Power*: the probability that the replication d^{rep} is larger (in absolute value) than the critical value that is considered to define “statistical significance” in this analysis.
- *Type S error rate*: the probability that the replicated estimate has the incorrect sign, if it is statistically significantly different from zero.
- *Exaggeration ratio* (expected Type M error): expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero.

Gelman & Carlin (2014), p. 644

2

Beyond power calculations

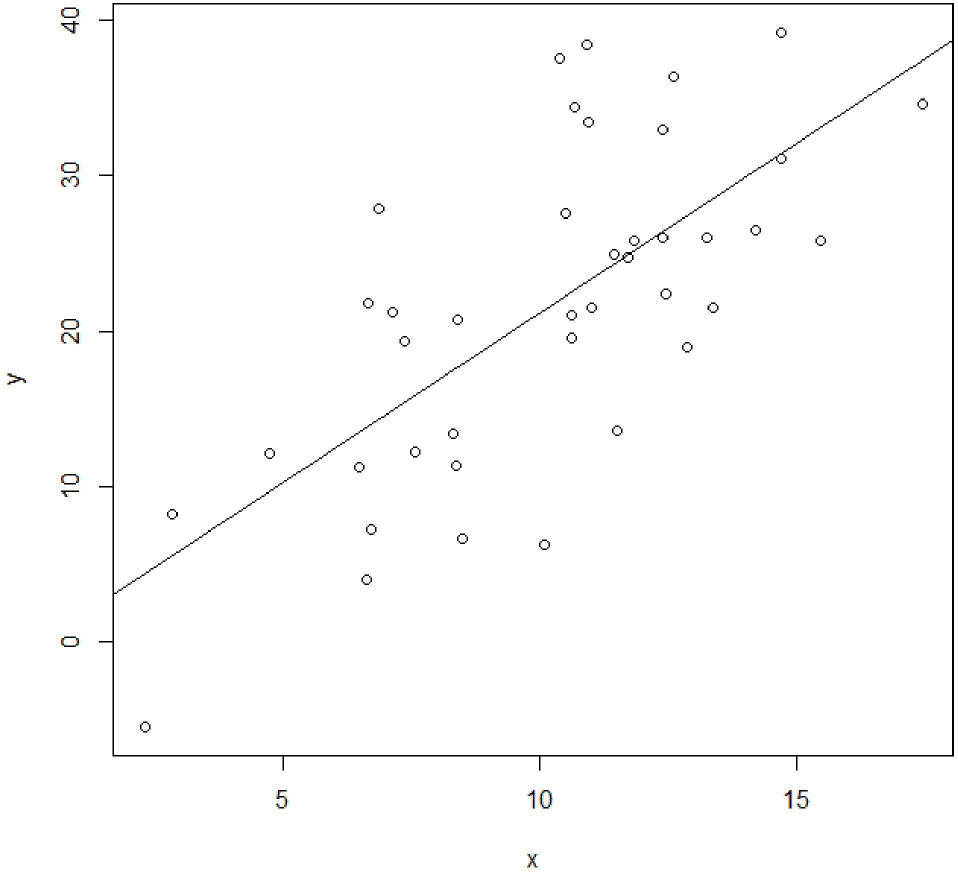
```
retrodesign <- function(A, s, alpha=.05, df=Inf, n.sims=10000){
   z <- qt(1-alpha/2, df)
   p.hi <- 1 - pt(z-A/s, df)
   p.lo <- pt(-z-A/s, df)
   power <- p.hi + p.lo
   typeS <- p.lo/power
   estimate <- A + s*rt(n.sims,df)
   significant <- abs(estimate) > s*z
   exaggeration <- mean(abs(estimate)[significant])/A
   return(list(power=power,typeS=typeS,exaggeration=exaggeration))
}
```

R function: Gelman & Carlin (2014), p. 644

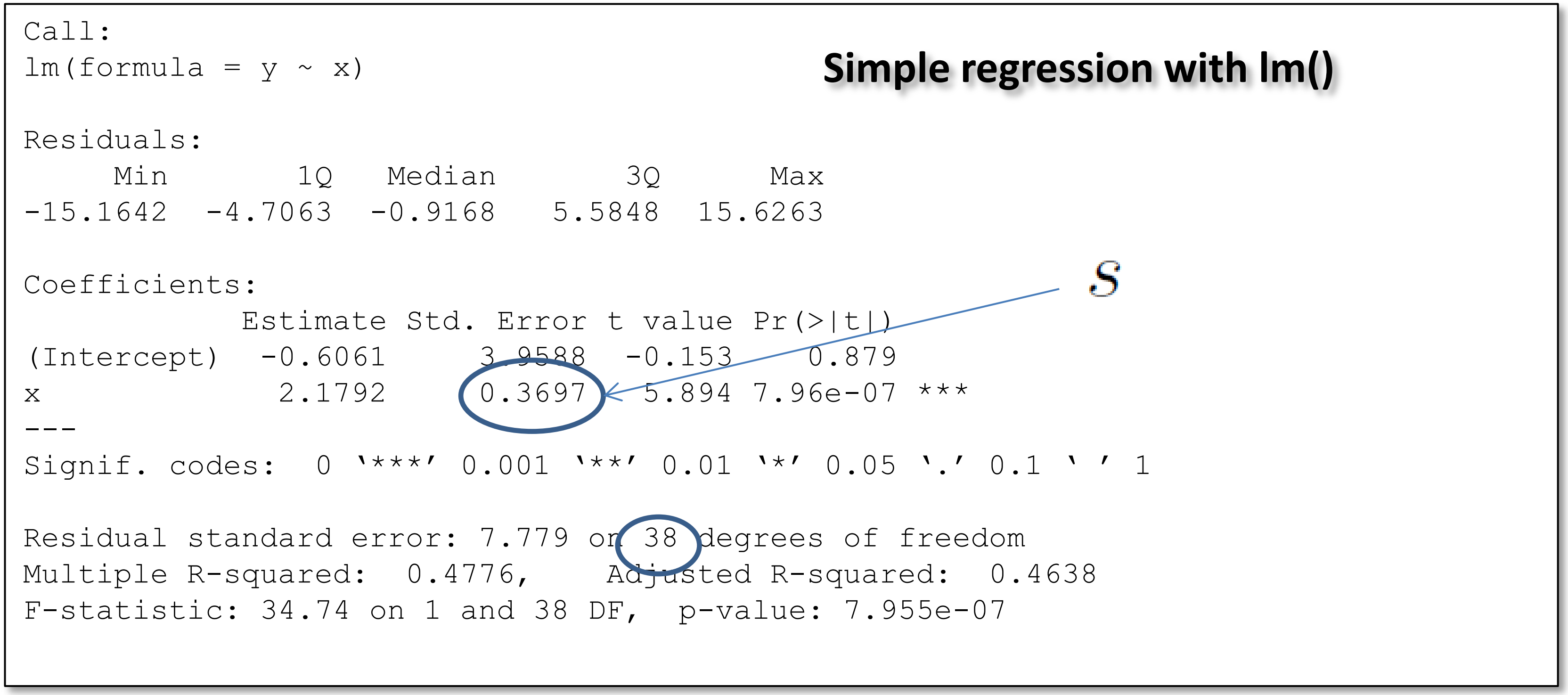
2

Beyond power calculations

A simple example: linear regression



2 Beyond power calculations



```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.1642  -4.7063  -0.9168   5.5848  15.6263 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.6061     3.9588  -0.153    0.879    
x             2.1792     0.3697   5.894 7.96e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.779 on 38 degrees of freedom
Multiple R-squared:  0.4776,    Adjusted R-squared:  0.4638 
F-statistic: 34.74 on 1 and 38 DF,  p-value: 7.955e-07
```

Simple regression with lm()

s

R syntax

2 Beyond power calculations

```
> retrodesign(1, 0.3697, df=38)
$power
[1] 0.7498592

$typeS
[1] 2.054527e-05

$exaggeration
[1] 1.161278
```

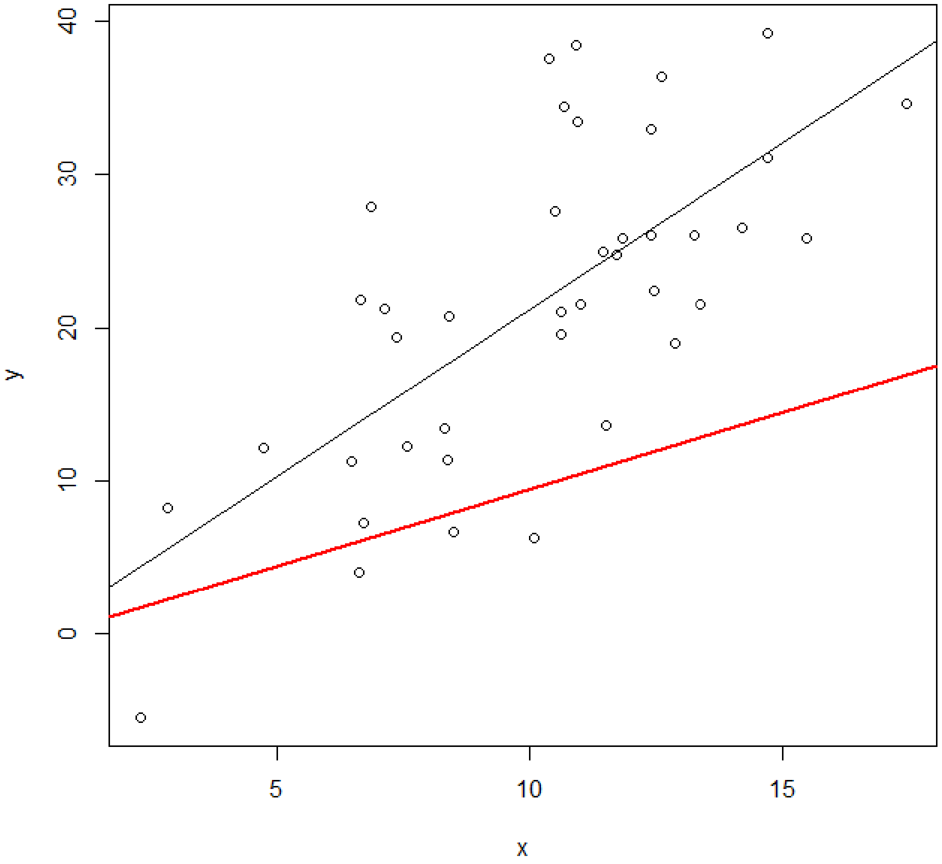
Design Analysis

$$D = 1$$

True population effect

R syntax

2 Beyond power calculations



$D = 1$

2 Beyond power calculations

```
> retrodesign(0.5, 0.3697, df=38)
$power
[1] 0.2536931

$typeS
[1] 0.003356801

$exaggeration
[1] 1.962419
```

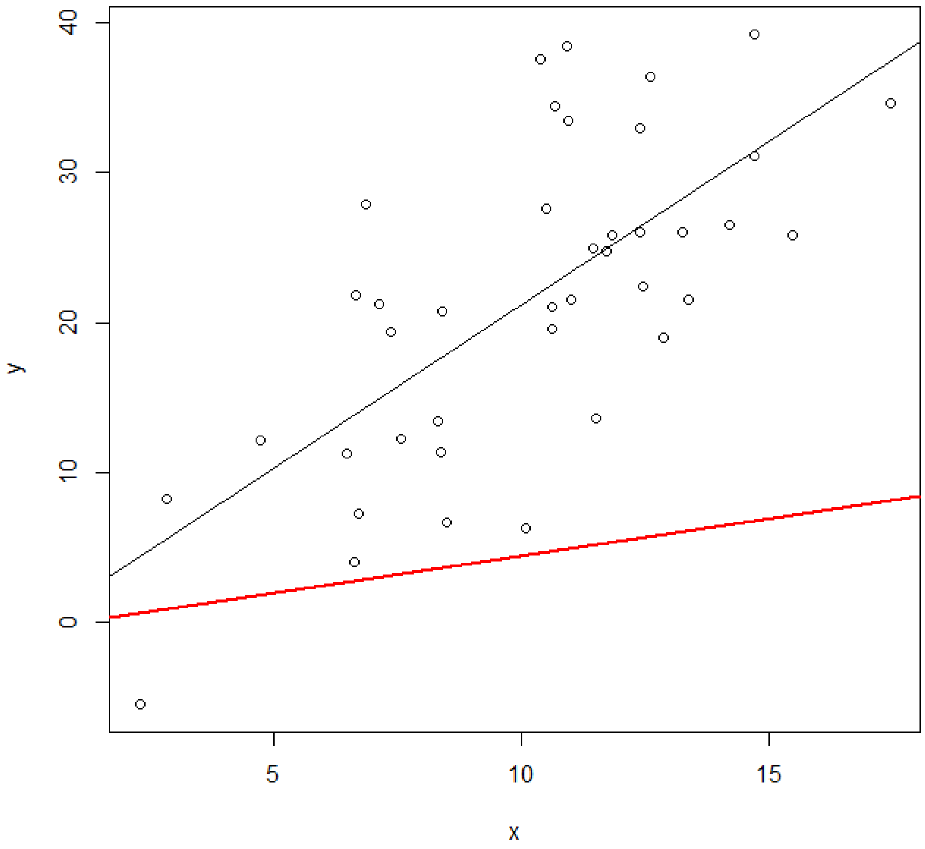
Design Analysis

$$D = 0.5$$

True population effect

R syntax

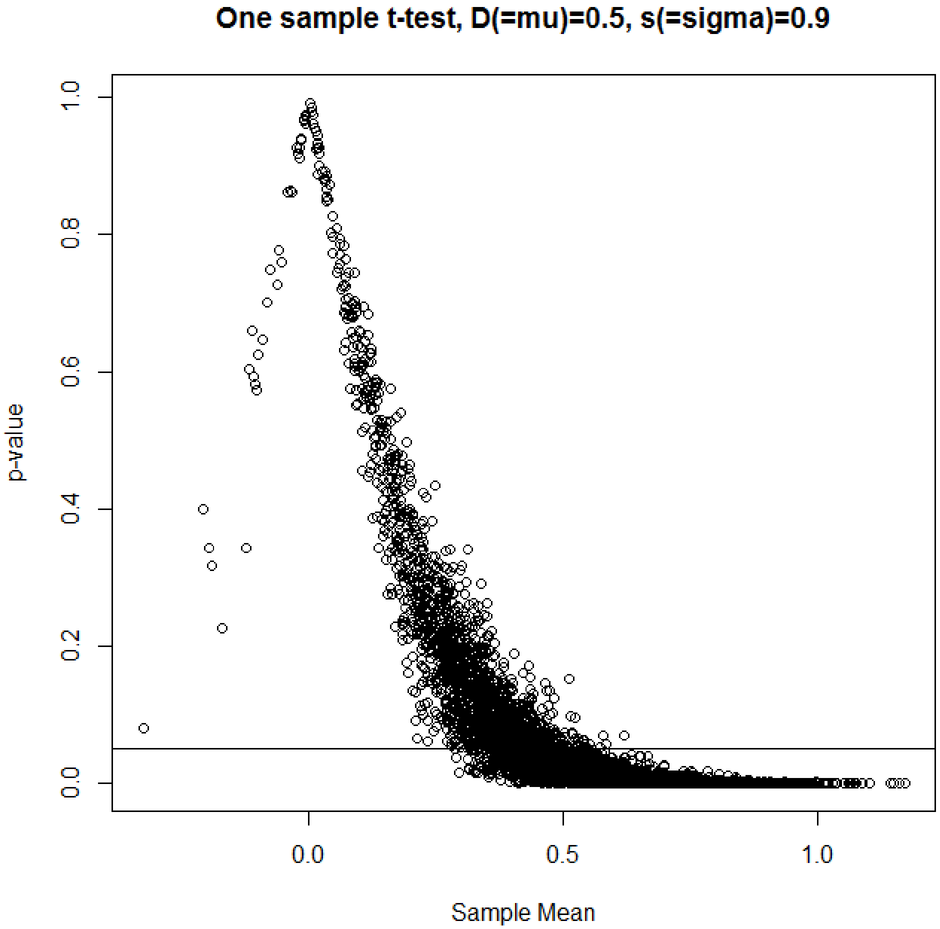
2 Beyond power calculations



$D = 0.5$

2

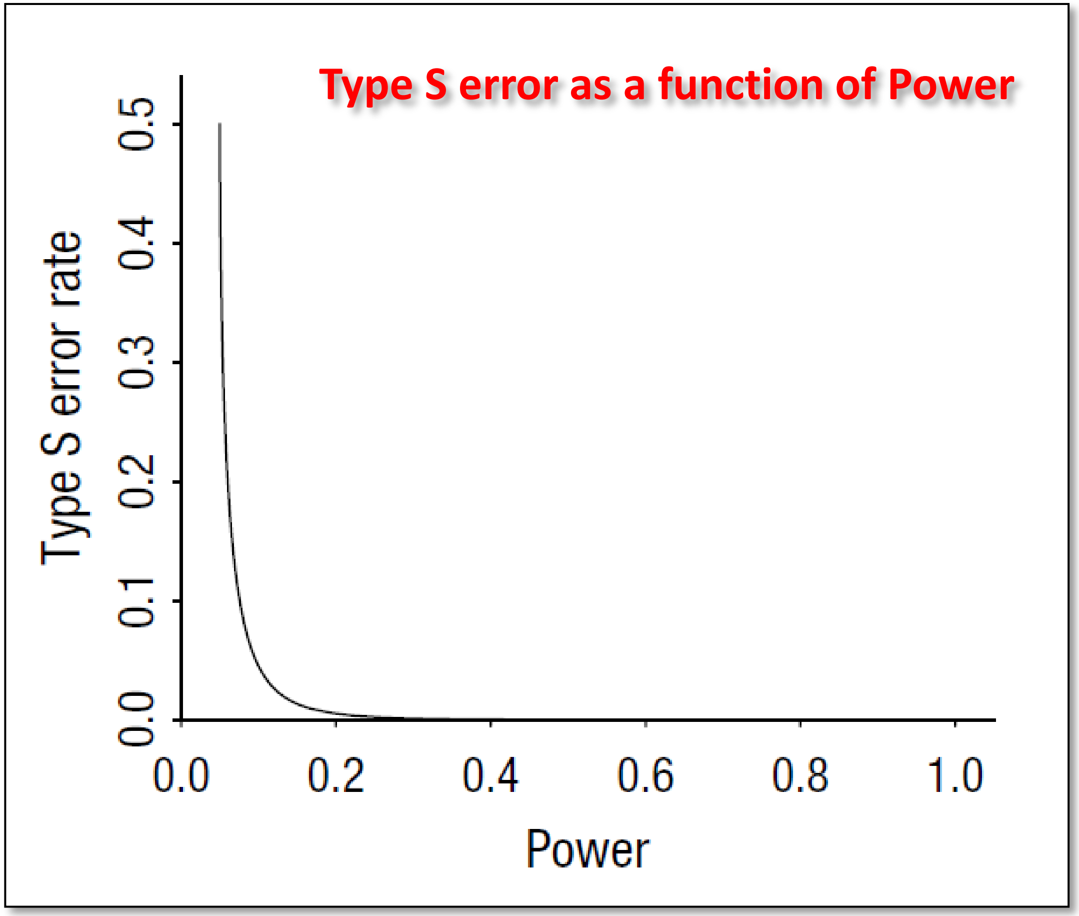
Beyond power calculations



5000 simulated
samples with 20 observations each
from a normal distribution with
parameters μ = 0.5; s = 0.9

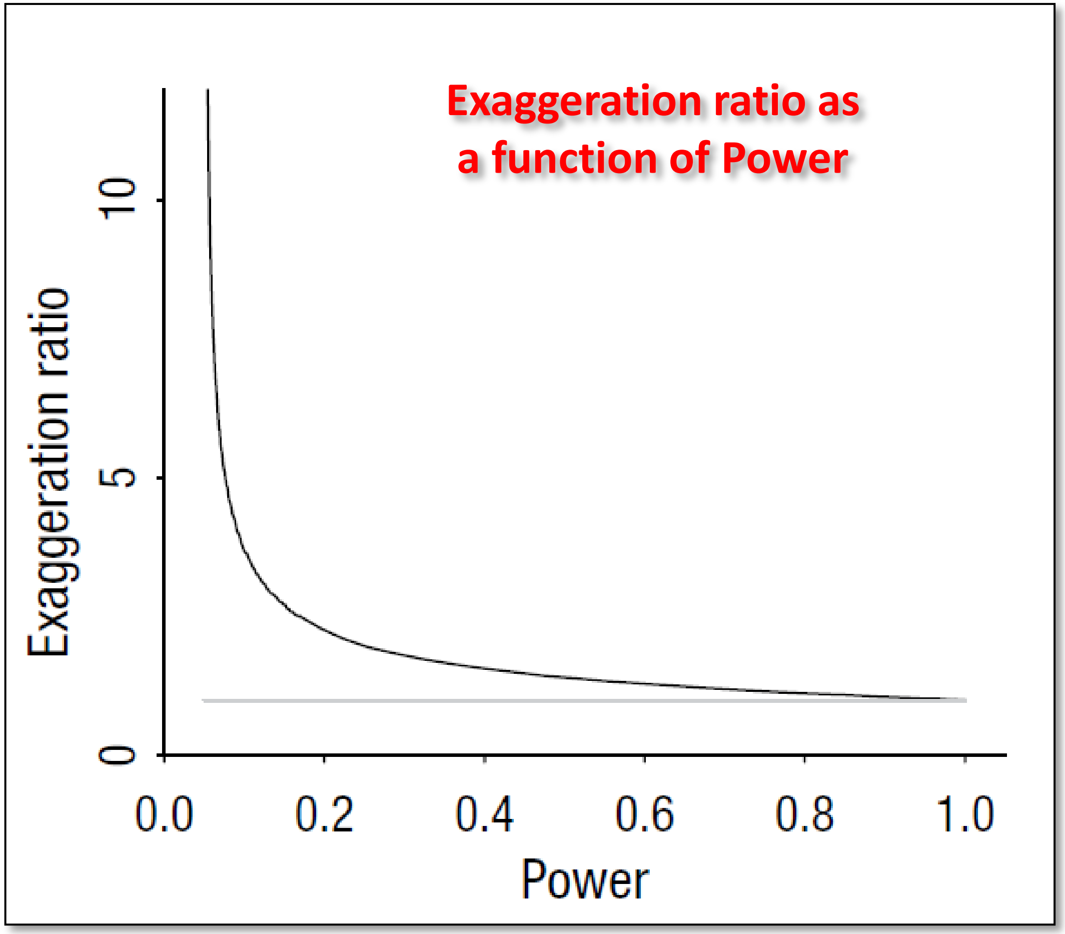
% of significant results (≠ 0) : 39.7
% of sample means > D(=μ) : 32.3

2 Beyond power calculations



Gelman & Carlin (2014), p. 644

2 Beyond power calculations



Gelman & Carlin (2014), p. 644

Practical implications:

Design Analysis strongly suggests larger sample sizes than those that are commonly used in psychology. In particular, if sample size is too small, in relation to the true effect size, then what appears to be a win (statistical significance) may really be a loss (in the form of a claim that does not replicate).

For a more formal presentation of the DA approach see Gelman A. & Tuerlinckx F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15, 373–390.

3 Pros and cons of the Bayes factor (BF)

3

Pros and cons of the Bayes factor

Recall

Positive predictive value (PPV)

$$PPV = P(H_A | \text{reject } H_0)$$

Negative predictive value (PPV)

$$NPV = 1 - PPV$$

We need the Bayes theorem to derive these posterior probabilities for the contrasting hypotheses

3 Pros and cons of the Bayes factor

Recall

Positive predictive value (PPV)

$$PPV = P(H_A | \text{reject } H_0)$$

Negative predictive value (PPV)

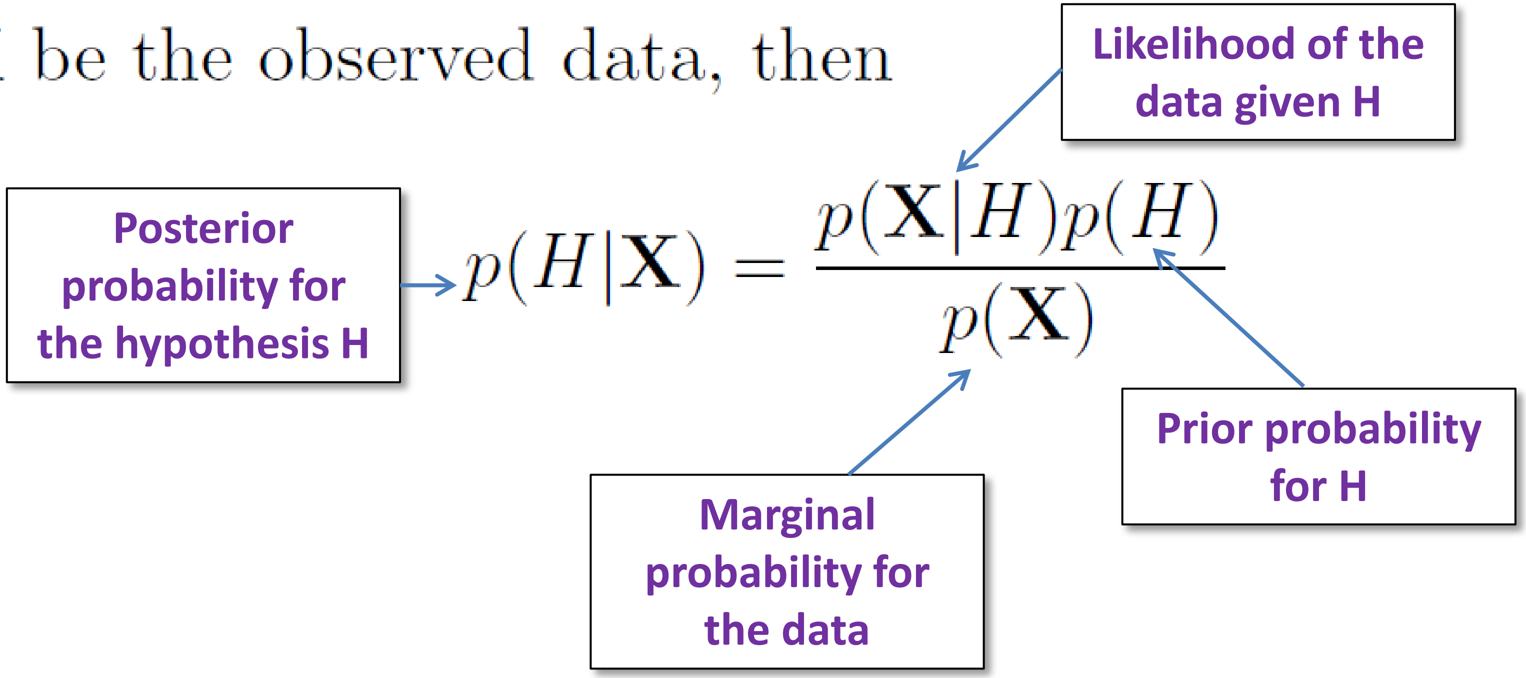
$$NPV = 1 - PPV$$

We need the Bayes theorem to derive these posterior probabilities for the contrasting hypotheses

The same applies if we want to compute the posterior probabilities explicitely given the observed data

3 Pros and cons of the Bayes factor

Let $\mathbf{X}$ be the observed data, then



$$p(H|\mathbf{X}) = \frac{p(\mathbf{X}|H)p(H)}{p(\mathbf{X})}$$

Posterior probability for the hypothesis H

Likelihood of the data given H

Prior probability for H

Marginal probability for the data

3

Pros and cons of the Bayes factor

The relative posterior probability of the null and alternative hypotheses →

$$\frac{p(H_0|\mathbf{X})}{p(H_A|\mathbf{X})} = \frac{\frac{p(\mathbf{X}|H_0)p(H_0)}{p(\mathbf{X})}}{\frac{p(\mathbf{X}|H_A)p(H_A)}{p(\mathbf{X})}}$$

$$= \frac{p(\mathbf{X}|H_0)p(H_0)}{p(\mathbf{X}|H_A)p(H_A)}$$

Pros and cons of the Bayes factor

In general it is assumed that $p(H_A) = p(H_0)$, then

Bayes Factor (BF) →
$$\frac{p(H_0|\mathbf{X})}{p(H_A|\mathbf{X})} = \frac{p(\mathbf{X}|H_0)}{p(\mathbf{X}|H_A)}$$

The analytic derivation of BF can be very difficult (see, for example, Kass & Raftery, 1995)

3 Pros and cons of the Bayes factor

A possible way out is to approximate the BF by means of some function of the Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = -2\ln(L) + k\ln(n)$$

L

Maximum likelihood of the data

k

Number of free parameters In the model

n

Number of independent observations

3

Pros and cons of the Bayes factor

The BF can be approximated according to the following equation

$$\mathrm{BF} = \frac{p(\mathbf{X}|H_0)}{p(\mathbf{X}|H_A)} \approx \mathrm{e}^{(\Delta\mathrm{BIC})/2}$$

Exponential function

where $\Delta\mathrm{BIC} = \mathrm{BIC}(H_A) - \mathrm{BIC}(H_0)$

3

Pros and cons of the Bayes factor

The BF can be approximated according to the following equation

$$\mathrm{BF} = \frac{p(\mathbf{X}|H_0)}{p(\mathbf{X}|H_A)} \approx \mathrm{e}^{(\Delta \mathrm{BIC})/2}$$

Warning: This represents a very basic approximation only!

Please see, for example, Kass & Raftery (1995), Wagenmakers (2007), and Bollen, Ray, Zavisca, & Harden (2012) for more rigorous derivations.

3

Pros and cons of the Bayes factor

Finally, the posterior probability of H0 is

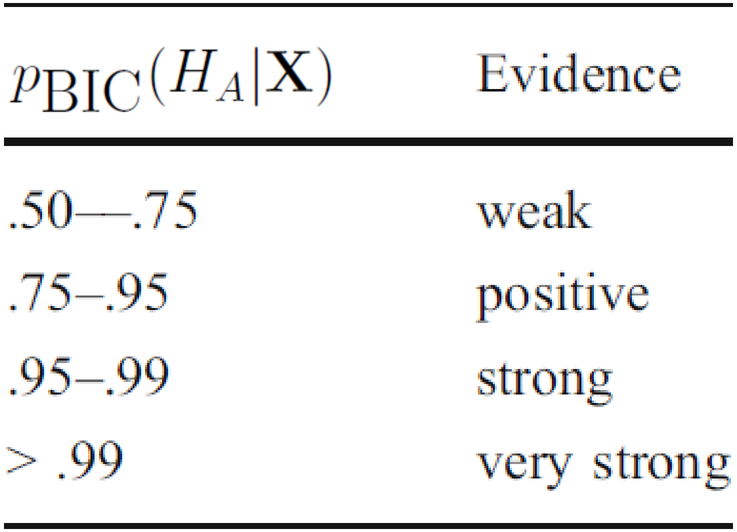
$$p_{\mathrm{BIC}}(H_0|\mathbf{X}) = \frac{\mathrm{BF}}{\mathrm{BF}+1}$$

consequently, the posterior probability of HA is

$$p_{\mathrm{BIC}}(H_A|\mathbf{X}) = 1 - p_{\mathrm{BIC}}(H_0|\mathbf{X})$$

3 Pros and cons of the Bayes factor

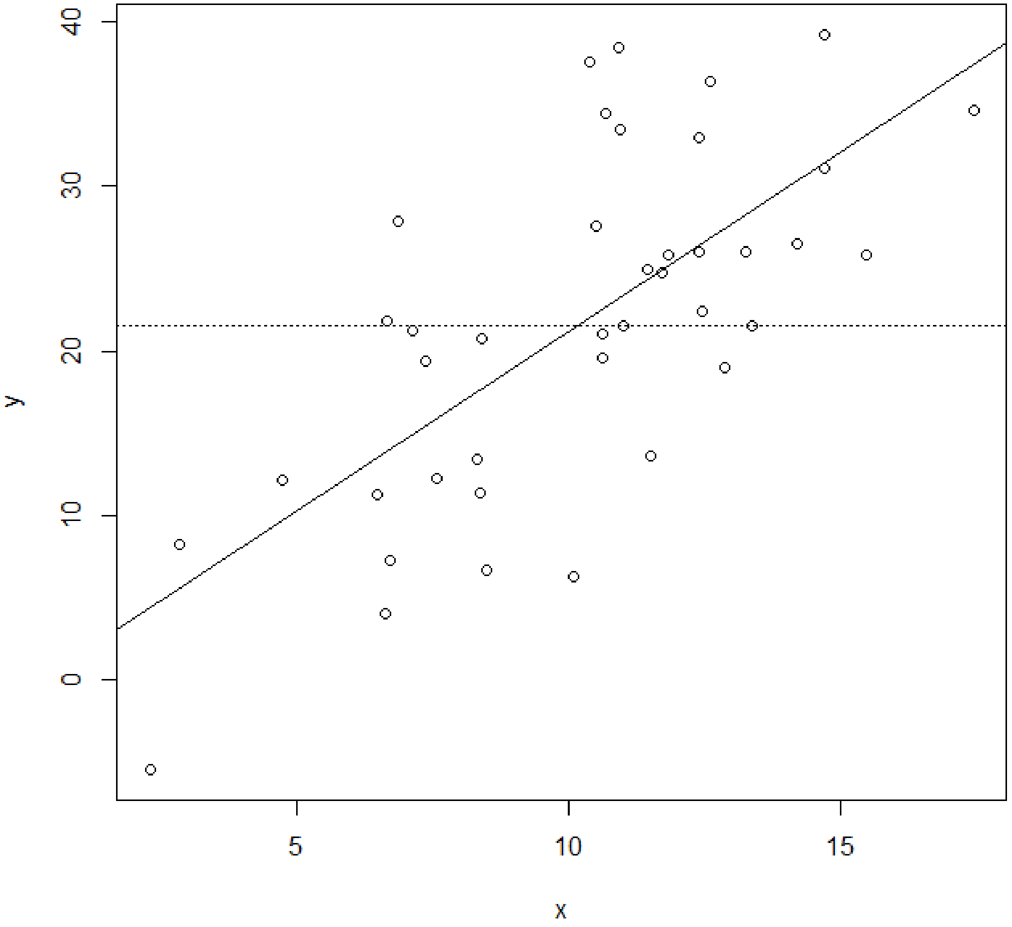
Raftery (1995) suggests the following substantive interpretations for the posterior probability



$p_{\mathrm{BIC}}(H_A \mid \mathbf{X})$	Evidence
.50—.75	weak
.75–.95	positive
.95–.99	strong
$>$ .99	very strong

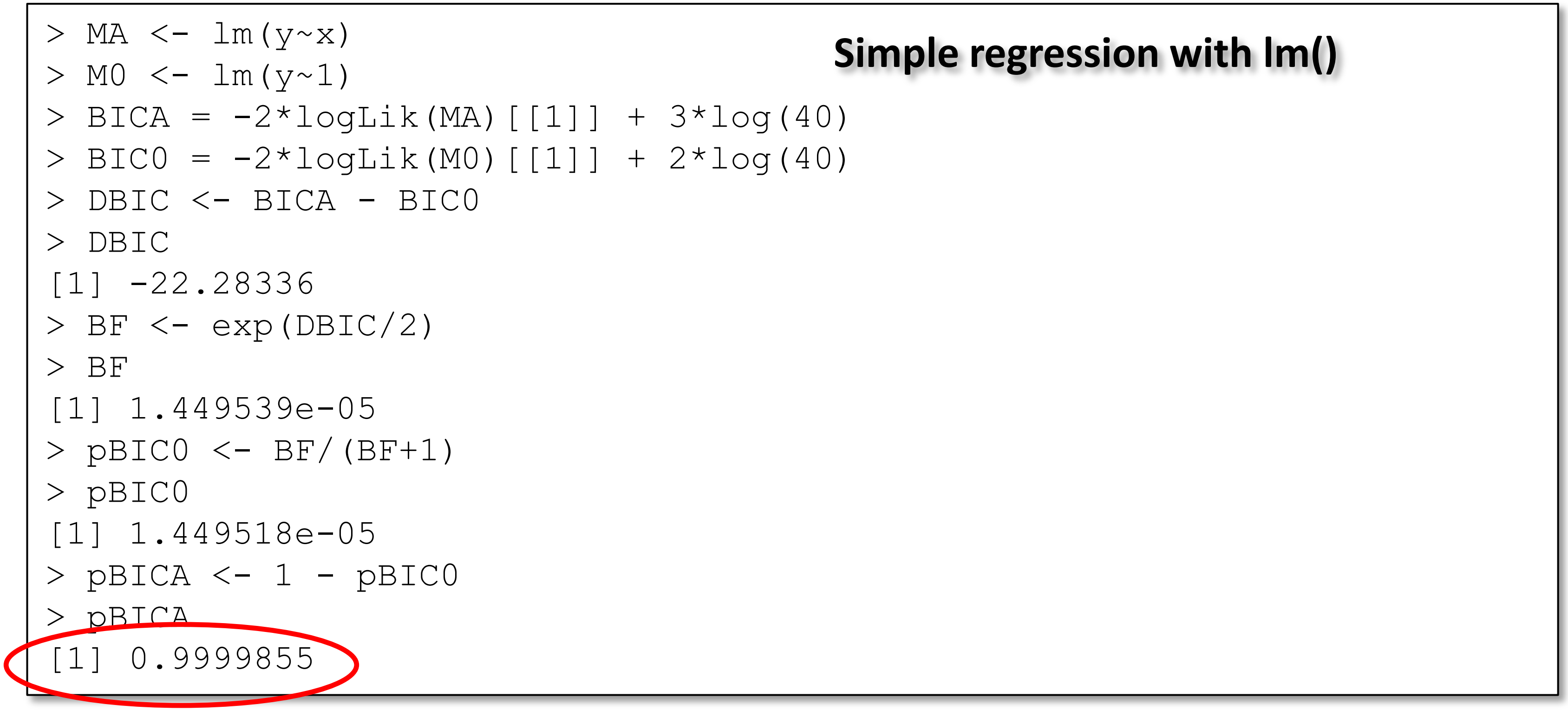
Pros and cons of the Bayes factor

A simple example: linear regression



March 23-24, 2015

3 Pros and cons of the Bayes factor



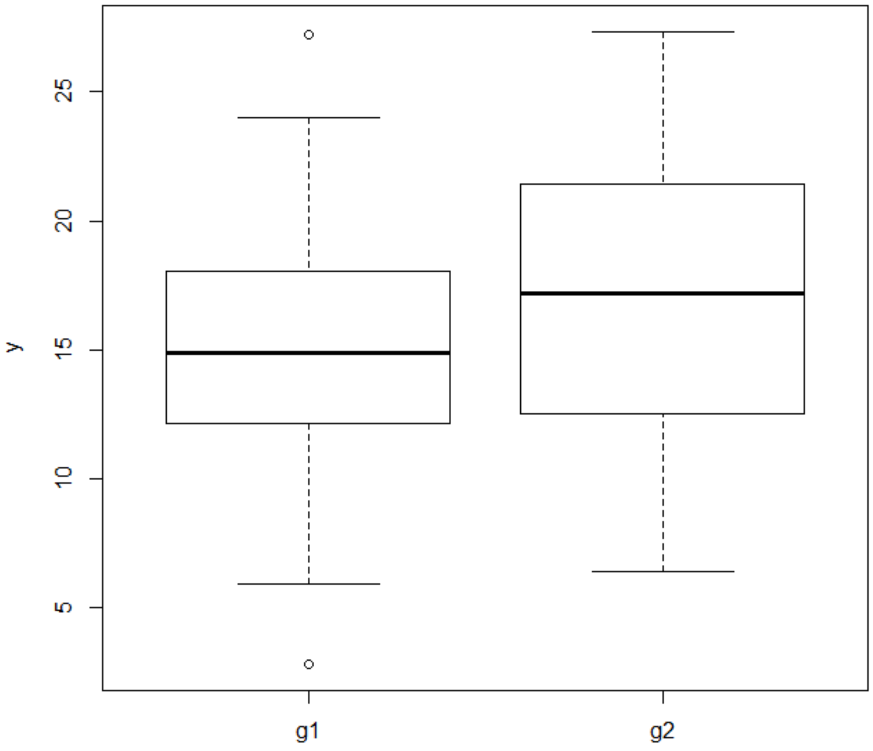
```
> MA <- lm(y~x)
> M0 <- lm(y~1)
> BICA = -2*logLik(MA)[[1]] + 3*log(40)
> BIC0 = -2*logLik(M0)[[1]] + 2*log(40)
> DBIC <- BICA - BIC0
> DBIC
[1] -22.28336
> BF <- exp(DBIC/2)
> BF
[1] 1.449539e-05
> pBIC0 <- BF/(BF+1)
> pBIC0
[1] 1.449518e-05
> pBICA <- 1 - pBIC0
> pBICA
[1] 0.9999855
```

Simple regression with lm()

R syntax

Pros and cons of the Bayes factor

A simple example: linear regression with categorical predictor



Pros and cons of the Bayes factor

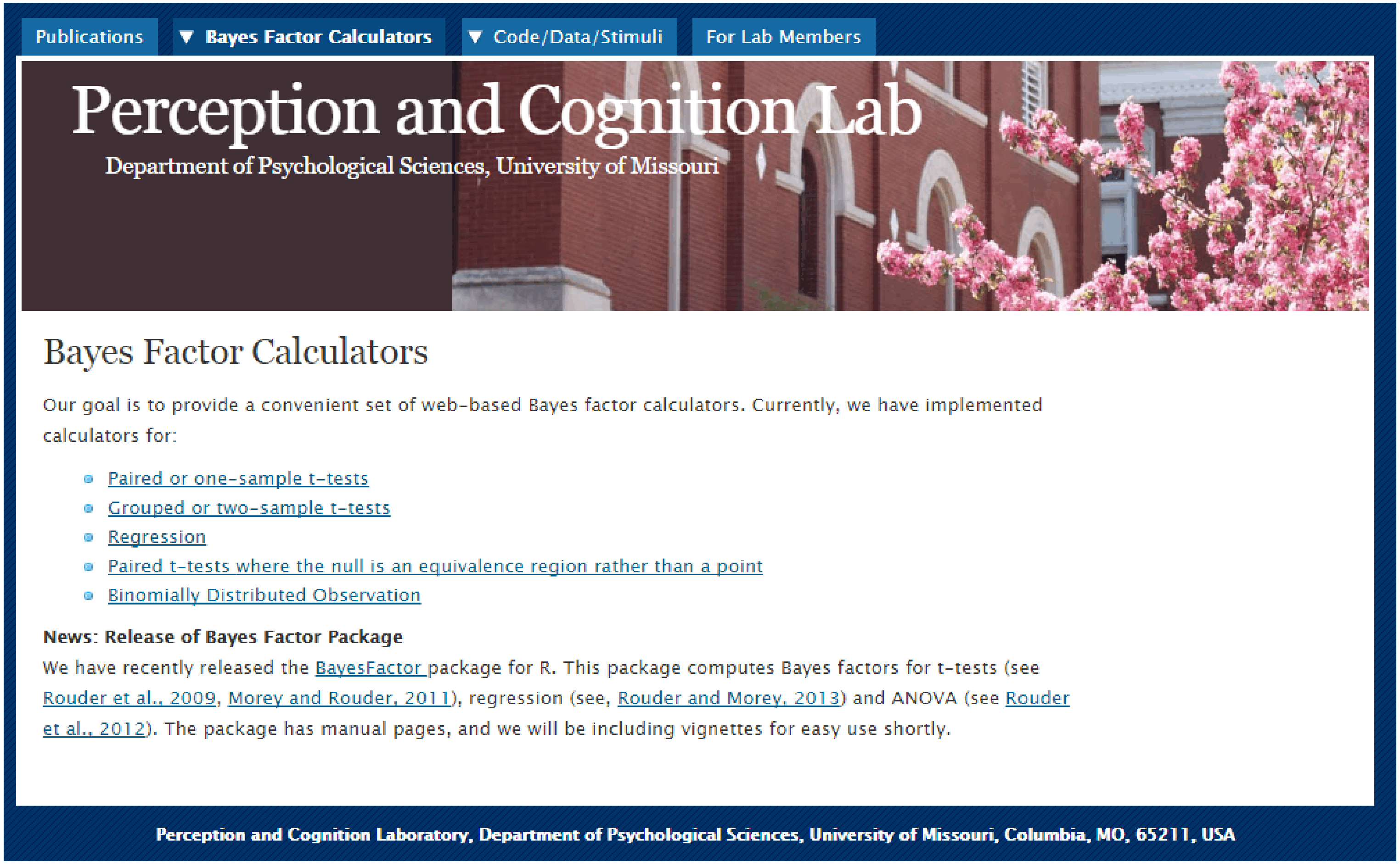
Simple regression with lm()

```
> x1 <- rnorm(25,15,6)
> x2 <- rnorm(25,15.5,6)
> boxplot(x1,x2,names=c("g1","g2"),ylab="y")
> G1 <- rep("g1",25)
> G2 <- rep("g2",25)
> G <- c(G1,G2)
> y <- c(x1,x2)
> MA <- lm(y~G)
> M0 <- lm(y~1)
> BICA = -2*logLik(MA)[[1]] + 3*log(50)
> BIC0 = -2*logLik(M0)[[1]] + 2*log(50)
> DBIC <- BICA - BIC0
> DBIC
[1] 1.17938
> BF <- exp(DBIC/2)
> BF
[1] 1.803429
> pBIC0 <- BF/(BF+1)
> pBIC0
[1] 0.643294
> pBICA <- 1 - pBIC0
> pBICA
[1] 0.356706
```

R syntax

3 Pros and cons of the Bayes factor

Different resources for computing BF according to other approaches (es. http://pcl.missouri.edu/bayesfactor)



Perception and Cognition Lab

Department of Psychological Sciences, University of Missouri

Bayes Factor Calculators

Our goal is to provide a convenient set of web-based Bayes factor calculators. Currently, we have implemented calculators for:

- Paired or one-sample t-tests
- Grouped or two-sample t-tests
- Regression
- Paired t-tests where the null is an equivalence region rather than a point
- Binomially Distributed Observation

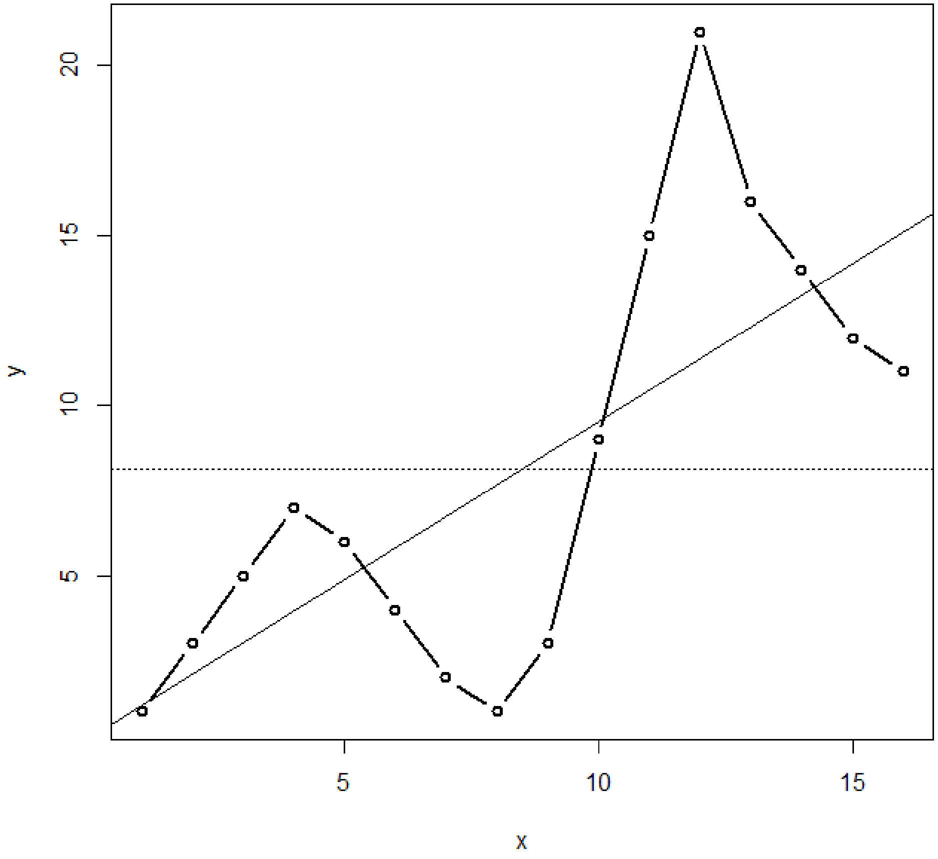
News: Release of Bayes Factor Package

We have recently released the BayesFactor package for R. This package computes Bayes factors for t-tests (see Rouder et al., 2009, Morey and Rouder, 2011), regression (see, Rouder and Morey, 2013) and ANOVA (see Rouder et al., 2012). The package has manual pages, and we will be including vignettes for easy use shortly.

Perception and Cognition Laboratory, Department of Psychological Sciences, University of Missouri, Columbia, MO, 65211, USA

The main problem of the BF

Let us consider the following graphical representation



3

Pros and cons of the Bayes factor

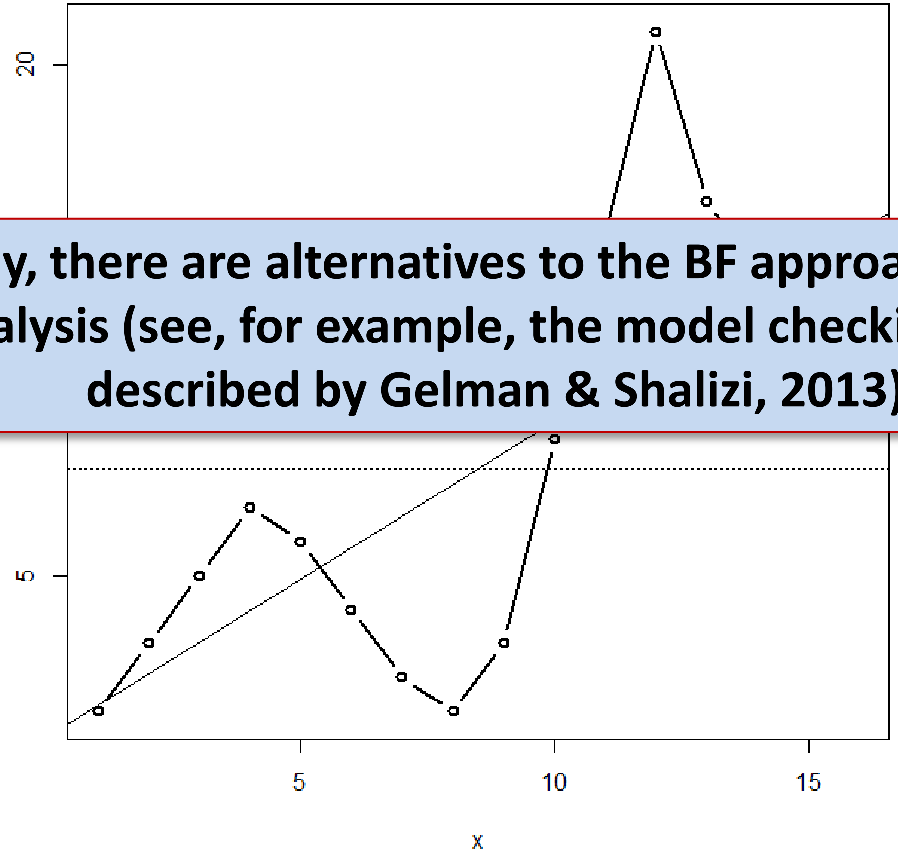
```
> x <- c(1:16)
> y <- c(c(1,3,5,7,6,4,2,1),3*c(1,3,5,7,6,4,2,1))
> plot(x,y,type="b",lwd=2)
> x <- c(1:16)
> y <- c(c(1,3,5,7,6,4,2,1),3*c(1,3,5,7),10+c(6,4,2,1))
> plot(x,y,type="b",lwd=2)
> MA <- lm(y~x)
> M0 <- lm(y~1)
> abline(MA)
> abline(M0,lty=3)
> BICA = -2*logLik(MA)[[1]] + 3*log(16)
> BIC0 = -2*logLik(M0)[[1]] + 2*log(16)
> DBIC <- BICA - BIC0
> DBIC
[1] -9.079352
> BF <- exp(DBIC/2)
> BF
[1] 0.01067687
> pBIC0 <- BF/(BF+1)
> pBIC0
[1] 0.01056407
> pBICA <- 1 - pBIC0
> pBICA
[1] 0.9894359
```

R syntax

3

Pros and cons of the Bayes factor

The BF cannot recognize that both the models are bad models (the problem of relative comparisons)



Fortunately, there are alternatives to the BF approach in Bayesian data analysis (see, for example, the model checking proposal described by Gelman & Shalizi, 2013)

Thank you for your attention!