

Bayesian hypothesis testing

PSM2 2019/2020

Bennett Kleinberg

17 March 2020

Bayesian hypothesis testing

Screening terrorists

Problem 1: A secret government agency has developed a scanner which determines whether a person is a terrorist. The scanner is fairly reliable; 95% of all scanned terrorists are identified as terrorists, and 95% of all upstanding citizens are identified as such. An informant tells the agency that exactly one passenger of 100 aboard an aeroplane in which you are seated is a terrorist. The agency decide to scan each passenger and the shifty looking man sitting next to you is tested as “TERRORIST”. What are the chances that this man *is* a terrorist? Show your work!

Your turn

What are the chances that this man is a terrorist?

Problem 1: A secret government agency has developed a scanner which determines whether a person is a terrorist. The scanner is fairly reliable; 95% of all scanned terrorists are identified as terrorists, and 95% of all upstanding citizens are identified as such. An informant tells the agency that exactly one passenger of 100 aboard an aeroplane in which you are seated is a terrorist. The agency decide to scan each passenger and the shifty looking man sitting next to you is tested as “TERRORIST”. What are the chances that this man *is* a terrorist? Show your work!

Formalising the problem

CONDITIONAL Probability:

Probability of TERRORIST **given** that there is an ALARM

Looking for: $P(\text{terrorist} \text{ GIVEN } \text{alarm})$

Formal: $P(\text{terrorist} | \text{alarm})$

Solving the problem (method 1)

	Terrorist	Passenger	
Terrorist	950	50	1,000
Passenger	4,950	94,050	99,000
	5,900	94,100	100,000

$$P(\text{terrorist} | \text{alarm}) = 950 / 5900 = 16.10\%$$

Method 2: Bayes' rule

Setting the stage:

- $P(T)$ -> probability of terrorist
- $P(A)$ -> probability of alarm

We want:

- $P(T|A)$

We know:

- accuracy = $P(A|T) = 0.95$
- baserate = $P(T) = 0.01$

Bayes' rule

```
accuracy = 0.95 #P(A|T)  
baserate = 0.01 #P(T)
```

Bayes' rule: $P(T|A) = (P(A|T) * P(T)) / P(A)$

$P(A)$ -> probability of any alarm???

$$P(A) = P(A|T) * P(T) + P(A|\text{not}T) * P(\text{not}T)$$

```
(Prob_notT = 1 - baserate) #P(notT) = 1 - P(T)
```

```
## [1] 0.99
```

```
(Prob_A_given_notT = 1 - accuracy) #P(A|notT) = 1 - P(A|T)
```

```
## [1] 0.05
```

Bayes' rule (cont'd)

Putting it together:

```
#Bayes' rule:  
Prob_A = accuracy * baserate + Prob_A_given_notT * Prob_notT #P(A) = P(A  
Prob_A
```

```
## [1] 0.059
```

```
Prob_T_given_A = (accuracy * baserate) / Prob_A #P(T|A) = ( P(A|T) * P(T  
Prob_T_given_A
```

```
## [1] 0.1610169
```

Bigger picture: why all this?

Suppose:

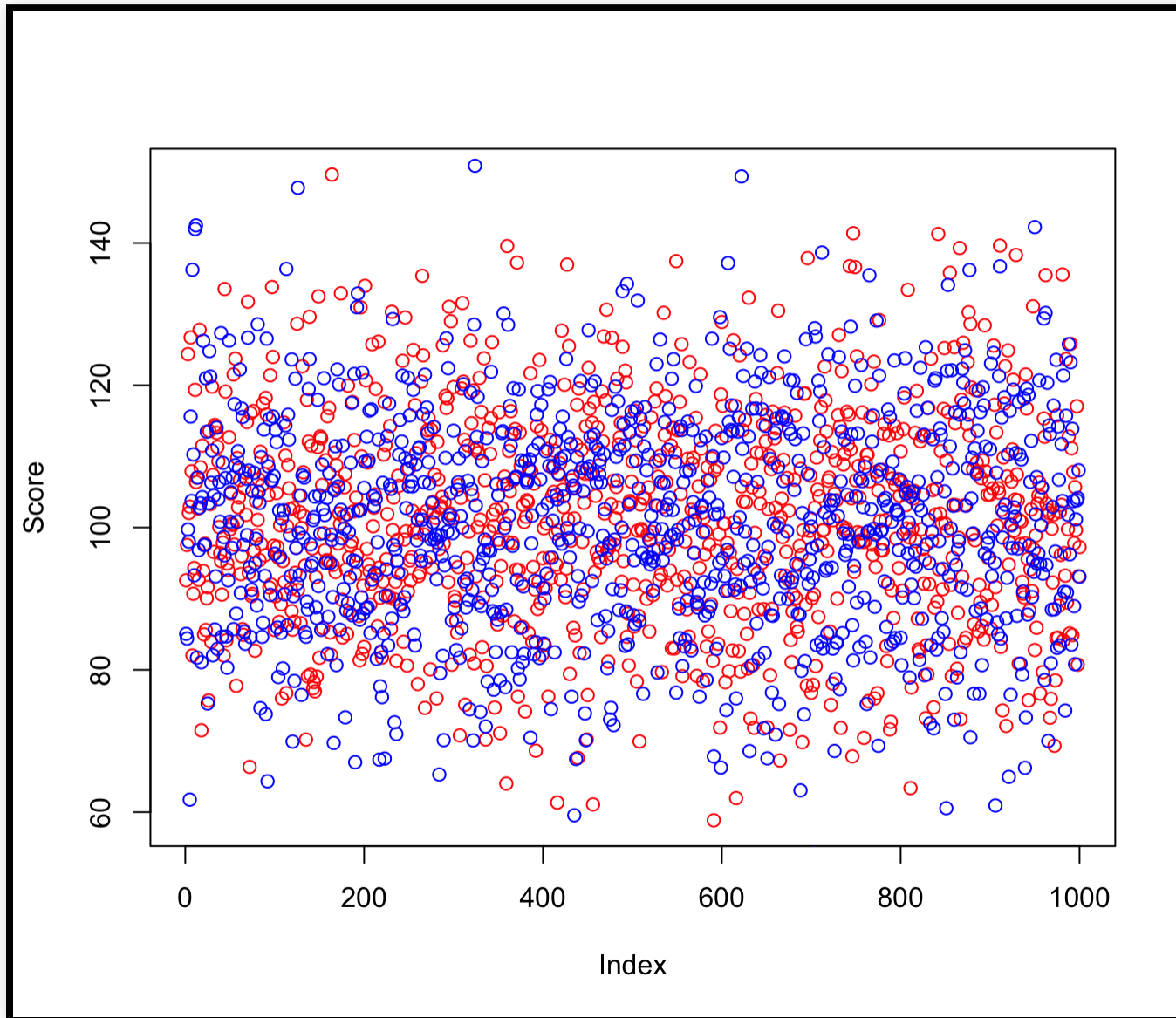
- you have two groups (left-handed vs right-handed)
- you take their IQ score
- you want to test if one group has a higher IQ score than the other

How would you do it?

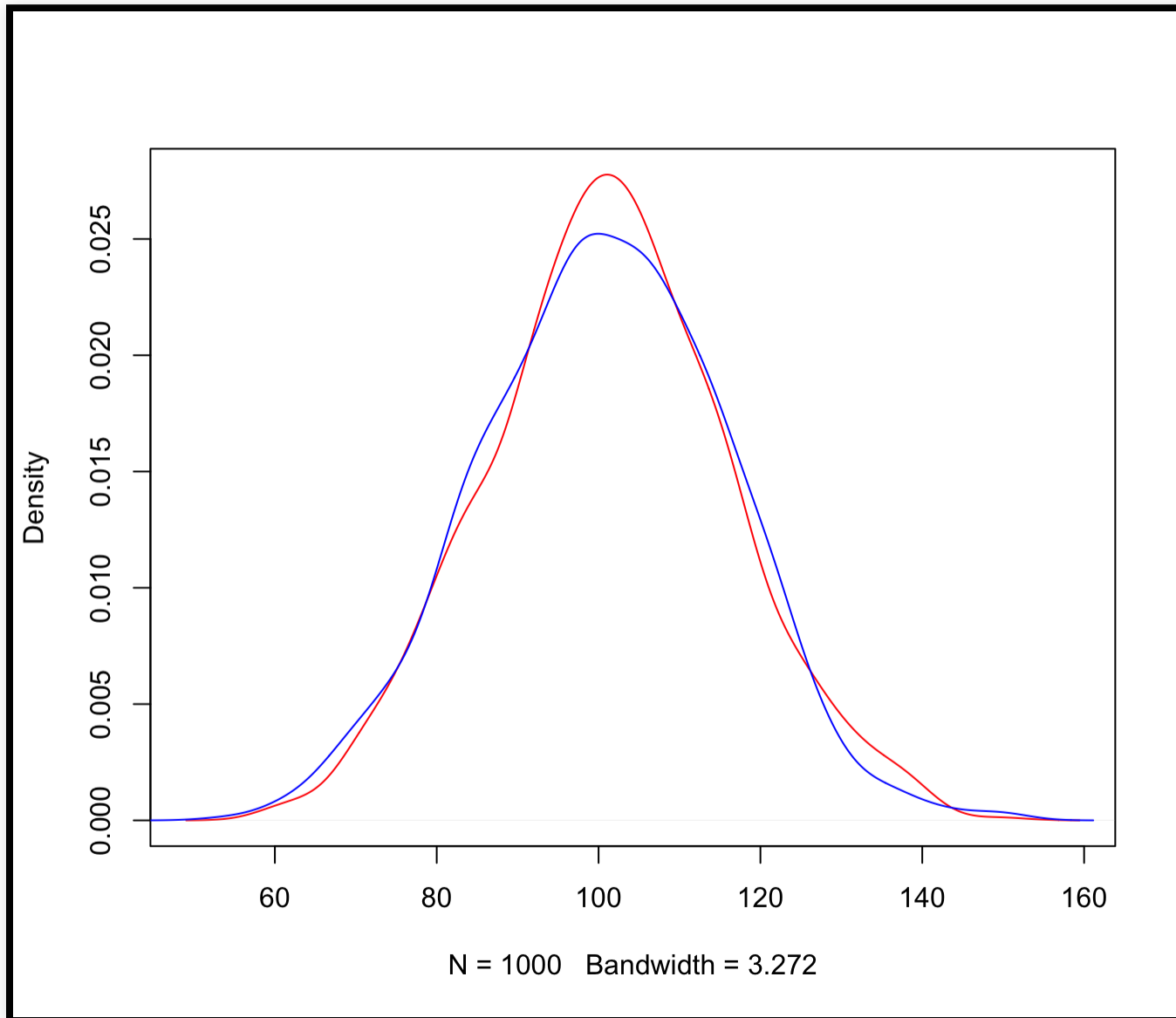
Rewind

Null hypothesis significance testing (NHST)

Hypothesis testing the old way



Hypothesis testing the old way



Hypothesis testing the old way

NULL hypothesis testing

- $H_0 : M_A \approx M_B$
 - there is no difference in the means between Group A and Group B
- $H_A : M_A \neq M_B$
 - there is a difference in the means between Group A and Group B
 - Directed hypotheses:
 - $H_A : M_A > M_B$
 - $H_A : M_A < M_B$

Hypothesis testing the old way

Purpose:

- test whether the data allow us to reject H_0
- remember: rejecting $H_0 \neq$ accepting H_A
- remember: not rejecting $H_0 \neq M_A == M_B$
- obsession with the p -value

In fact: *all we can ever say* is whether H_0 was rejected or not!

Today

Bayesian statistics

- What is it?
- How does it differ from NHST?
- What can it solve?
- Why should I care?
- How do I do it?

There are more problems

- we're bad at interpreting NHST results (e.g. p-values, CIs)
- strong assumptions about the data
- no stopping rule (increase n and everything becomes significant)

Quite problematic

The misunderstandings surrounding p-values and CIs are particularly unfortunate because they constitute the main tools by which ~~psychologists~~ crime scientists draw conclusions from data.

Robust misinterpretation of confidence intervals

Rink Hoekstra • Richard D. Morey • Jeffrey N. Rouder •
Eric-Jan Wagenmakers

Why do we need hypothesis testing anyway?

- core of inference testing
- core of scientific endeavour
 - think of the 'reproducibility crisis'
 - we want to avoid fishing expeditions

So: we desperately need hypotheses, but NHST is weak

Enter

A photograph of a blue neon sign mounted on a dark, textured wall. The sign displays the formula for conditional probability: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. The neon is bright blue and the background is dark, making the formula stand out. The sign is slightly tilted and has some faint, illegible markings on the wall behind it.

Two ideas of probability:
Frequentist vs Bayesian

Laymen's explanation

I have misplaced my phone somewhere in the home. I can use the phone locator on the base of the instrument to locate the phone and when I press the phone locator the phone starts beeping.

Problem: Which area of my home should I search?

Adapted from [this SO post](#)

Frequentist Reasoning

I can hear the phone beeping. I also have a mental model which helps me identify the area from which the sound is coming. Therefore, upon hearing the beep, I infer the area of my home I must search to locate the phone.

Bayesian Reasoning

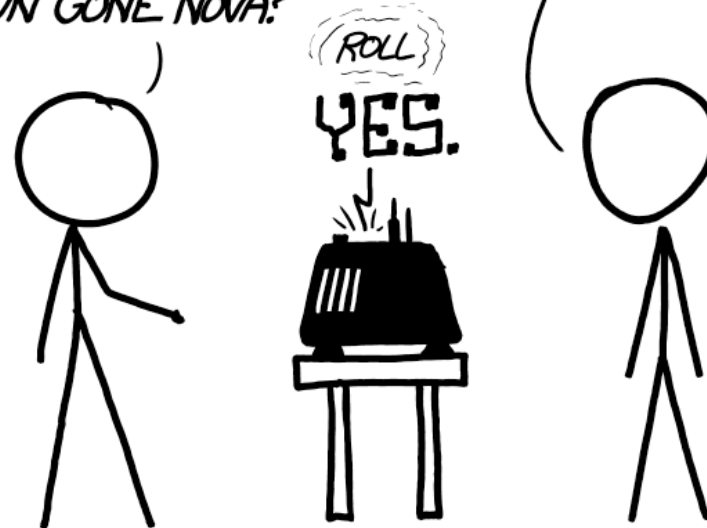
I can hear the phone beeping. Now, apart from a mental model which helps me identify the area from which the sound is coming from, I also know the locations where I have misplaced the phone in the past. So, I combine my inferences using the beeps and my prior information about the locations I have misplaced the phone in the past to identify an area I must search to locate the phone.

DID THE SUN JUST EXPLODE? (IT'S NIGHT, SO WE'RE NOT SURE.)

THIS NEUTRINO DETECTOR MEASURES
WHETHER THE SUN HAS GONE NOVA.

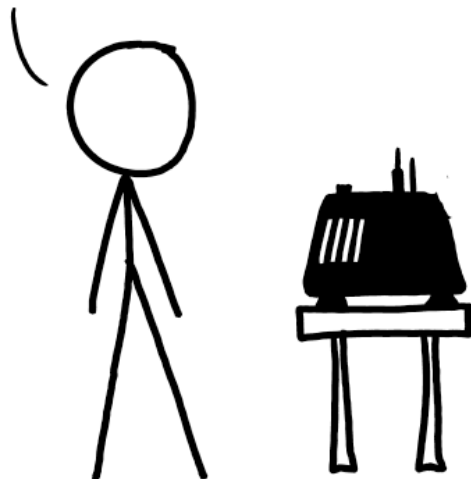
THEN, IT ROLLS TWO DICE. IF THEY
BOTH COME UP SIX, IT LIES TO US.
OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY.
DETECTOR! HAS THE
SUN GONE NOVA?



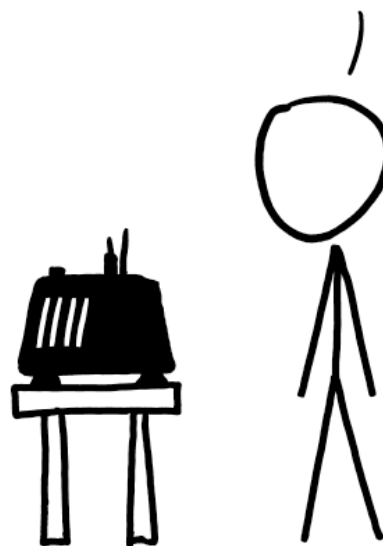
FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT
HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE
THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU \$50
IT HASN'T.



What is all this?

Remember?

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

$$P(terrorist|alarm) = \frac{P(alarm|terrorist)*P(terrorist)}{P(alarm)}$$

Translated to hypothesis testing

- $P(H)$: prob. of hypothesis H **prior** to have seen the data
- $P(D)$: marginal prob. of the data (same for all hyp.)
- $P(D|H)$: compatibility of the data with the hyp.
(**likelihood**)

We want to know:

$P(H|D)$: prob. of the hyp. given the data (**posterior**)

$$P(H|D) = \frac{P(D|H)*P(H)}{P(D)}$$

$$\textit{posterior} = \frac{\textit{likelihood}*\textit{prior}}{\textit{marginal}}$$

Formally

Since: $P(D)$ does not involve the hypothesis, ...

$$P(H|D) \propto P(D|H) * P(H)$$

Conceptually

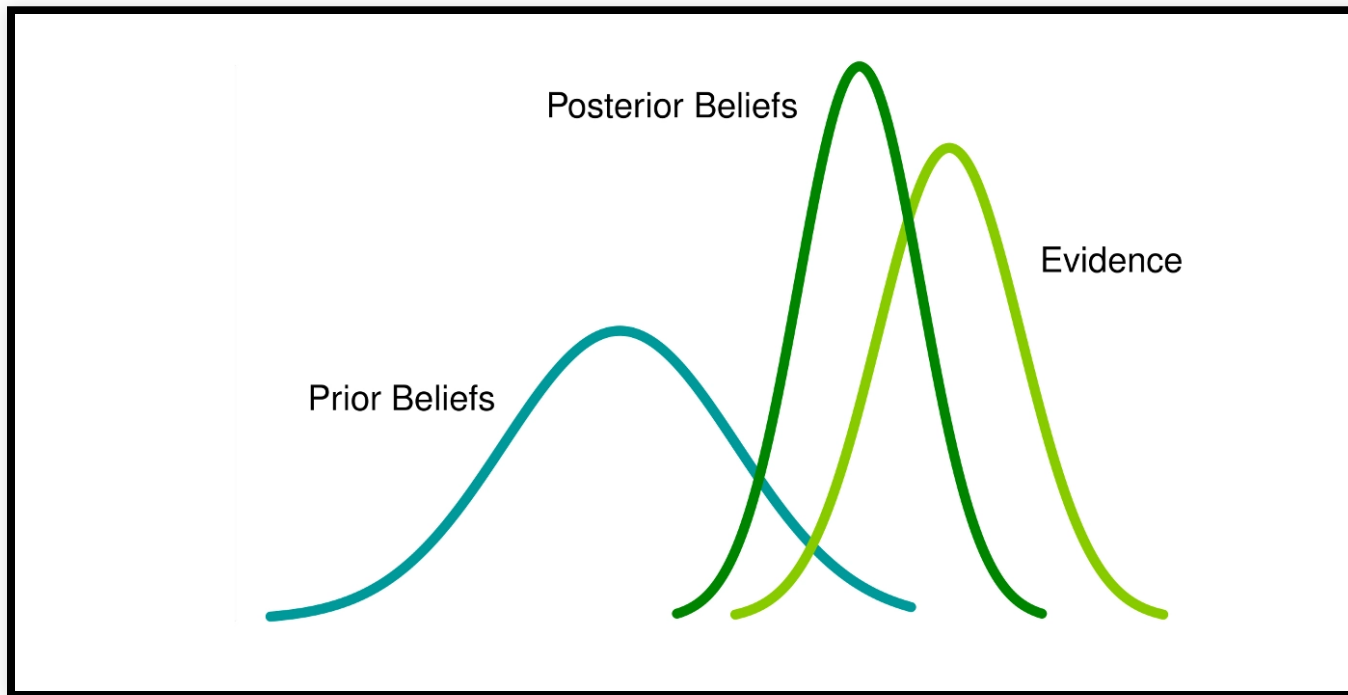
$$\textit{posterior} \propto \textit{likelihood} * \textit{prior}$$

- posterior: what we know after having seen the data (i.e. **what we learned from the data**)
- prior: our *prior* beliefs
- likelihood: observation

Think for a second

- this means that evidence can/must convince
- if you know that the sun is unlikely to have exploded, the evidence must be very, very strong to convince you otherwise

Bayesian inference is about updating beliefs with the data.



Bayesian hypothesis testing

If for any H :

$$P(H|D) \propto P(D|H) * P(H)$$

... then maybe we can compare the evidence $P(H_0|D)$ with the evidence $P(H_A|D)$?

Bayesian hypothesis testing

Suppose we have not seen the data, then:

$$odds_{0A} = \frac{P(H_0)}{P(H_A)}$$

or:

$$odds_{prior} = \frac{prior_{H_0}}{prior_{H_A}}$$

Important: no special status for H_0 !

Bayesian hypothesis testing

What we need for two hypotheses H_0 and H_A is:

- $P(D|H_A)$: compatibility of the data with H_A
- ... versus ...
- $P(D|H_0)$: compatibility of the data with H_0

The Bayes factor

$$\frac{P(H_A|D)}{P(H_0|D)} = \frac{P(D|H_A)}{P(D|H_0)} * \frac{P(H_A)}{P(H_0)}$$

How much more likely the data are under H_A compared to H_0 .

Called the Bayes Factor BF_{A0}

The evidence in the data favors one hypothesis, relative to another, exactly to the degree that the hypothesis predicts the observed data better than the other.

What is a Bayes factor? (Morey, 2014)

Stepwise example

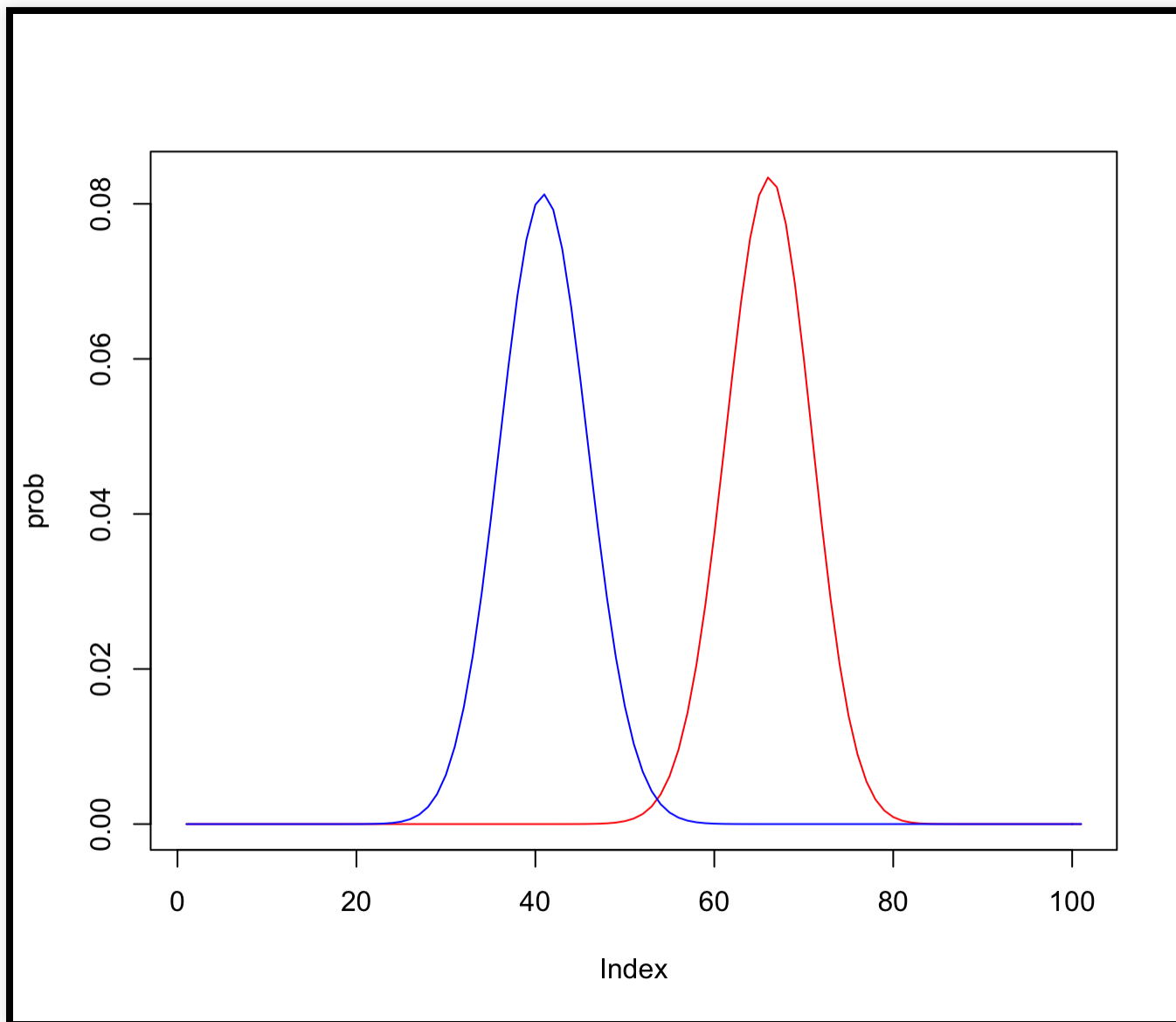
Suppose we have two lines of thought re. successful replication of crime science research:

- Optimists
- Skeptics

Optimists say that 65% of research replicates; skeptics say it's 40%.

Data: 100 replications and their outcome (successful vs fail)

- $H_{optimists} = 0.65$
- $H_{skeptics} = 0.40$



Now the data come in

- 100 replication studies
- 58 successful
- 42 failures
- $58/100 = 0.58$

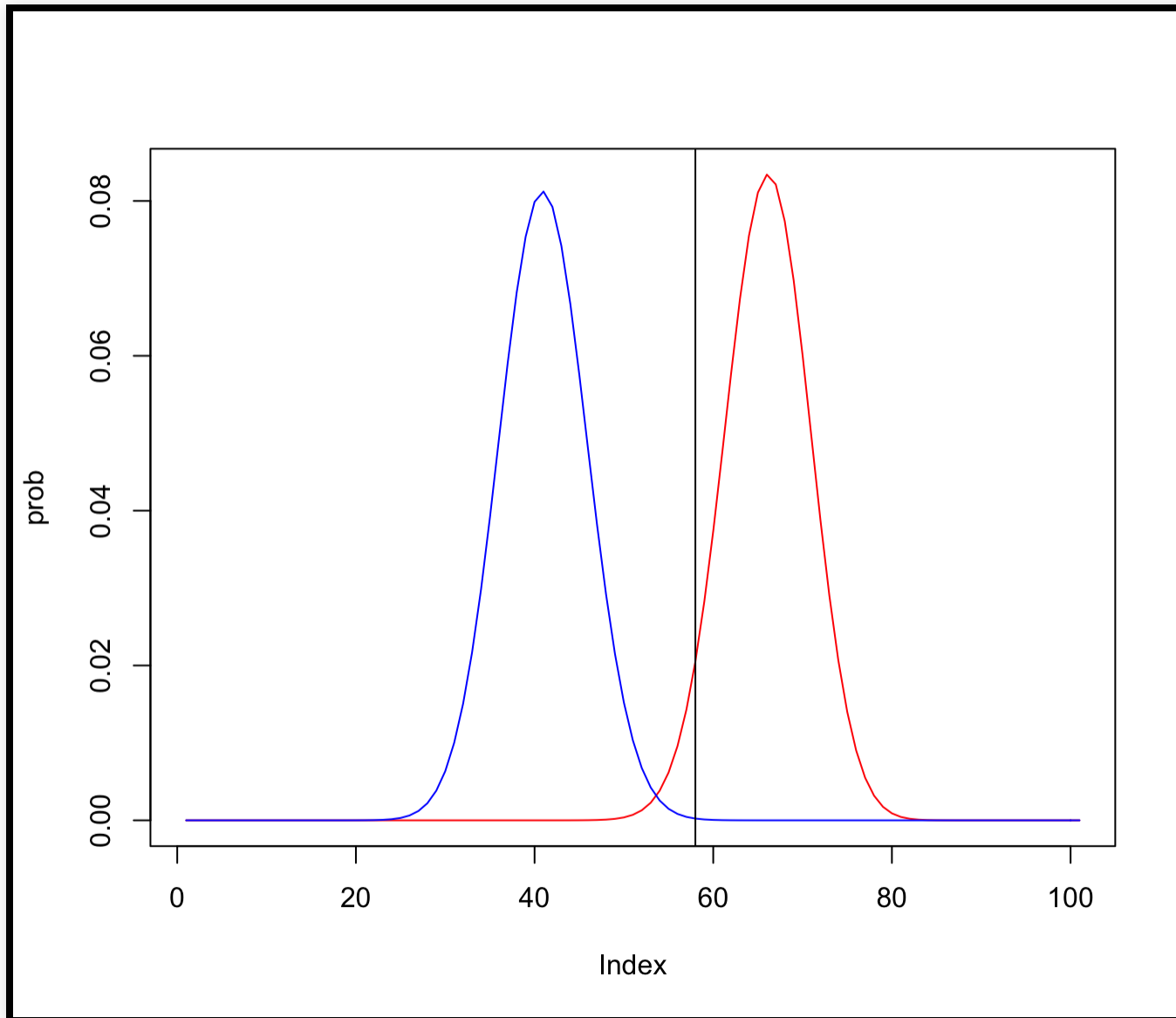
Closer to the optimists, but how much?

Relative weight of evidence

How much does the evidence update our beliefs?

Plausibility of the hypotheses $H_{opt.} = 0.65$ and $H_{skept.} = 0.40$ changes according to Bayes' rule!

Probability of observations



Probability of observations

- 58 successes:
- for $H_{opt.} = 0.65$: $P(D|H_{opt.}) = 0.0284$
- for $H_{skept.} = 0.40$: $P(D|H_{skept.}) = 0.0001$

$$\text{So: } \frac{P(D|H_{opt.})}{P(D|H_{skept.})} = \frac{0.0284}{0.0001} = 250.03$$

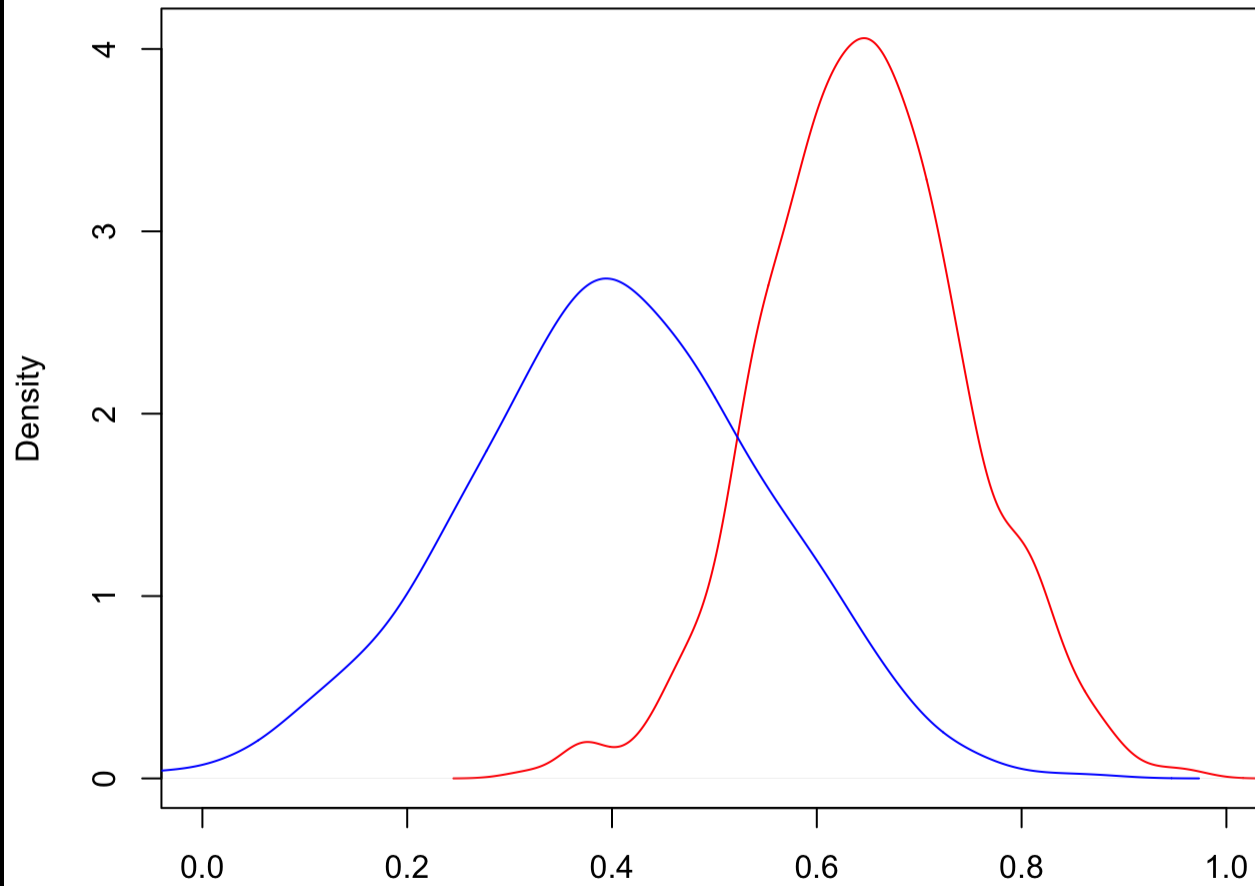
Bayes factor

$$BF = \frac{P(D|H_{opt.})}{P(D|H_{skept.})} = 250.03$$

The data are 250 times more likely under $H_{opt.}$ than under $H_{skept.}$

But what about uncertain priors?

Uncertainty in priors



N = 1000 Bandwidth = 0.02175

Prior beliefs as distributions

- rather than specific point estimates, we use distributions
- $H_{optimists}$ becomes a distribution (here normal distr.)
- $H_{skeptics}$ becomes a distribution (here normal distr.)

Bayesian estimation can handle this.

What can it solve?

What can it solve?

- all hypothesis testing questions!
- those with uncertainty
- aaaaand

What can it solve?

It can solve the H_0 problem!!!!

Now we can quantify relative evidence:

$$BF_{01} = \frac{P(H_0|D)}{P(H_1|D)}$$

Relative evidence of H_0 over H_1

Why should I care?

Why should I care?

Original Articles

Why Isn't Everyone a Bayesian?

B. Efron

Pages 1-5 | Received 01 Jul 1985, Published online: 27 Feb 2012

Efron, 1985

Why should I care?

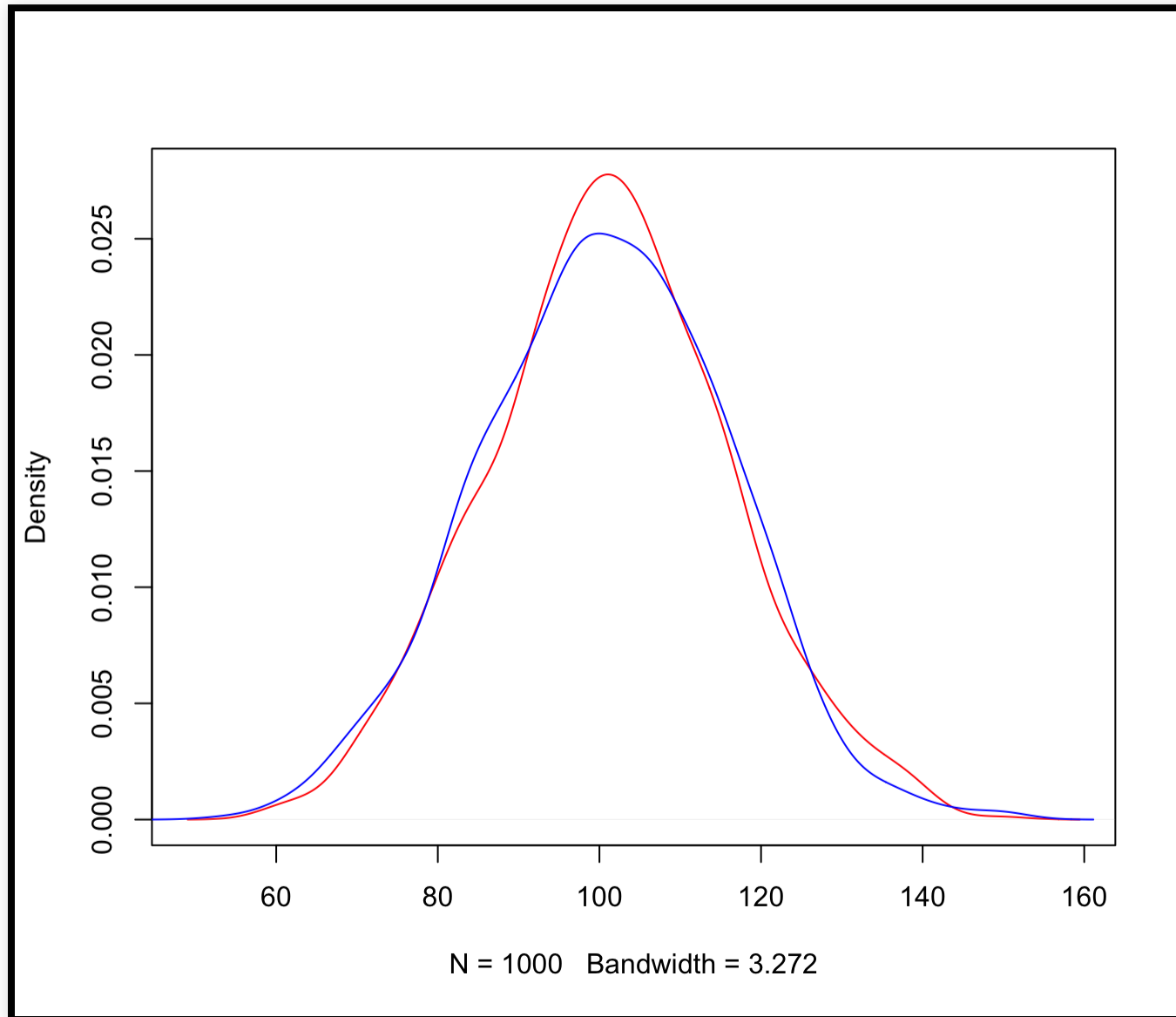
- Bayesian framework widely considered superior
- Bayesian logic fits with “science” better than NHST
- The “tools problem” is overcome
- Will become standard in the future

How to do it?

Two approaches:

- the `BayesFactor` R package
- JASP

How do I do it?



Is there a difference?

```
tapply(mydata$score, mydata$group, mean)
```

```
##           A           B  
## 101.2419 100.6370
```

- $H_0 : M_A \approx M_B$
- $H_1 : M_A \neq M_B$

Old school NHST

```
t.test(score ~ group
       , data = mydata
       , var.eq=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  score by group
## t = 0.90114, df = 1998, p-value = 0.3676
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7115953  1.9214737
## sample estimates:
## mean in group A mean in group B
##      101.2419      100.6370
```

Cohen's d effect size

```
d = 0.90*(sqrt(1/1000 + 1/1000))  
d
```

```
## [1] 0.04024922
```

NHST conclusion: small non-sign. difference.

BayesFactor R

```
library(BayesFactor)
ttestBF(formula = score ~ group
        , data = mydata)
```

```
## Bayes factor analysis
## -----
## [1] Alt., r=0.707 : 0.07520633 ±0%
##
## Against denominator:
##   Null, mu1-mu2 = 0
## ---
## Bayes factor type: BFindepSample, JZS
```

Package reference

BayesFactor R

$BF_{10} = 0.075$, which equals:

$$BF_{01} = 1/0.075 = 13.33$$

→ Evidence quantified for both hypotheses!

Interpreting BFs

Bayes factor	Evidence category
> 100	Extreme evidence for \mathcal{H}_1
$30 - 100$	Very strong evidence for \mathcal{H}_1
$10 - 30$	Strong evidence for \mathcal{H}_1
$3 - 10$	Moderate evidence for \mathcal{H}_1
$1 - 3$	Anecdotal evidence for \mathcal{H}_1
1	No evidence
$1/3 - 1$	Anecdotal evidence for \mathcal{H}_0
$1/10 - 1/3$	Moderate evidence for \mathcal{H}_0
$1/30 - 1/10$	Strong evidence for \mathcal{H}_0
$1/100 - 1/30$	Very strong evidence for \mathcal{H}_0
$< 1/100$	Extreme evidence for \mathcal{H}_0

Interpreting BFs

$$BF_{01} = 1/0.075 = 13.33$$

The data are 13.33 times more likely under H_0 than under H_A . There is strong evidence for H_0

How do I do it?



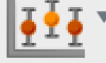








JASP



| JASP

A Fresh Way to
Do Statistics

JASP

<div><div> Descriptives</div><div> T-Tests</div><div> ANOVA</div><div> Regression</div><div> Frequencies</div><div> Factor</div></div>				
	 V1	 score	 group	
1	1	92.5929	A	
2	2	97.5473	A	
3	3	124.381	A	
4	4	102.058	A	
5	5	102.939	A	
6	6	126.726	A	
7	7	107.914	A	
8	8	82.0241	A	
9	9	90.6972	A	
10	10	94.3151	A	
11	11	119.361	A	
12	12	106.397	A	
13	13	107.012	A	

JASP

Independent Samples T-Test

Independent Samples T-Test

	t	df	p	Cohen's d
score	0.901	1997	0.368	0.040

Note. Welch's t-test.

JASP

Bayesian Independent Samples T-Test ▼

Bayesian Independent Samples T-Test

	BF_{01}	error %
score	13.29	$1.078e - 5$

[Psychonomic Bulletin & Review](#)

February 2018, Volume 25, [Issue 1](#), pp 219–234 | [Cite as](#)

How to become a Bayesian in eight easy steps: An annotated reading list

Authors

[Authors and affiliations](#)

Alexander Etz, Quentin F. Gronau, Fabian Dablander, Peter A. Edelsbrunner, Beth Baribault 

Etz et al. 2018

END