Inference and Evidence

Daniël Lakens

Eindhoven University of Technology

@Lakens

Why is this interesting?

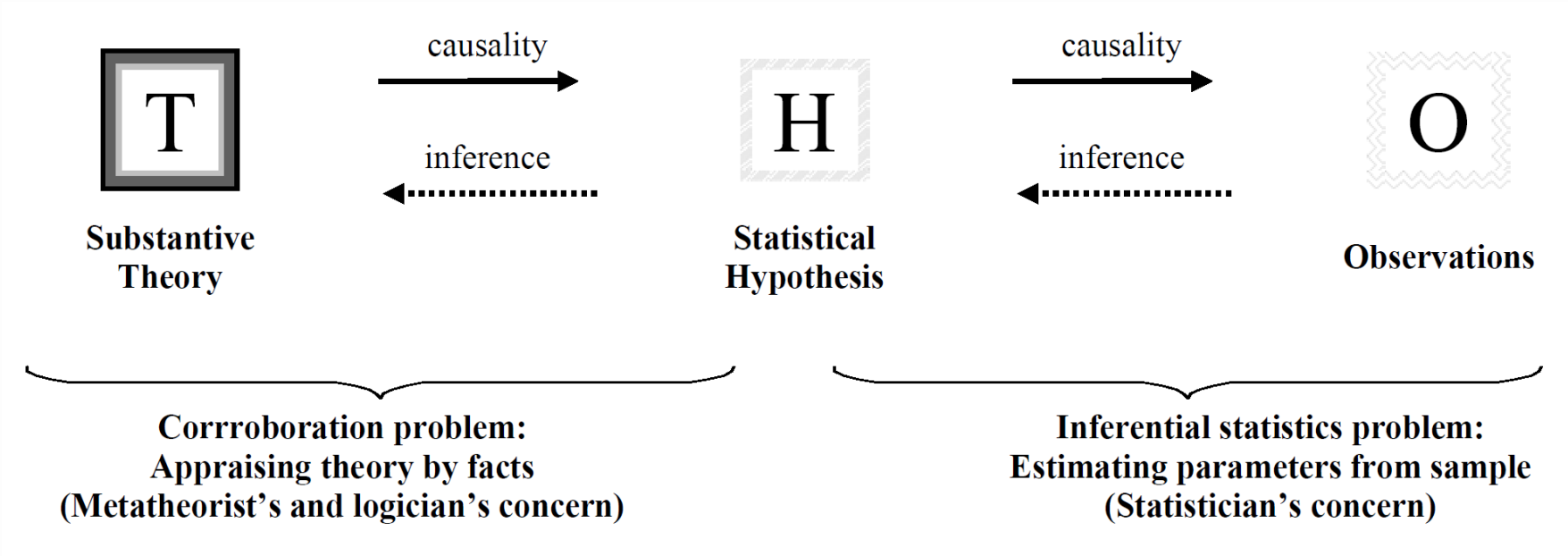
"The first principle is that you must not fool yourself - and *you are the easiest person to fool*"

Feynman, 1974

Increased attention for these topics across science.
Better knowledge of inferences and evidence will:

- improve your own inferences
- increase your contributions to cumulative science
- make your research lines more efficient.

Why is this **un**interesting?



Meehl, 1990

What are we doing?

What is the goal of collecting data, and reporting statistics?

("I don't know, but reviewers seem to like it!")

Three Paths to Salvation

"Truth is One, The Paths are Many"

[Bhagavad Gita]

- The Karma yoga: The path of Action
- The Jnana yoga: The path of Devotion
- The Bhakti yoga: The path of Knowledge

Three Paths to Salvation

"Three Questions One Might Ask"

[Royall, 1997]

- What should I DO? (The path of Action)
- What should I BELIEVE? (The path of Devotion)
- How should I treat data as RELATIVE EVIDENCE? (The path of Knowledge)

The Path of Action

But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$ reject H, if $x \leqq x_0$ accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leqq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false.

[Neyman & Pearson, 1933]

The Path of Action

- Reject the null hypothesis (H0) whenever $p < \alpha$
  - *p = 0.048? p = 0.00001? Potayto, potahto*

- This is a rule to **govern our behavior**.
- It tells us **nothing** about the current test, we can only say '**in the long run, we won't be wrong very often**'

Error Control

- Reject the null hypothesis (H0) whenever $p < \alpha$
  - *p = 0.048? p = 0.00001? Potayto, potahto*

- This is a rule to **govern our behavior**.
- It tells us **nothing** about the current test, we can only say '**in the long run, we won't be wrong very often**'

The history of NHST

- The history of NHST starts with a fierce debate between Fisher's significance test (e.g., Fisher, 1925) and Neyman and Pearson's hypothesis test (e.g., Neyman and Pearson, 1928).

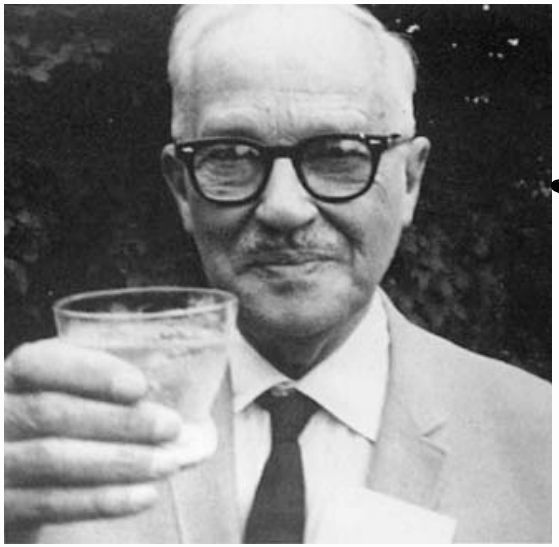
- **NHST is often practiced as a hybrid procedure that combines these two different viewpoints**.

Neyman-Pearson

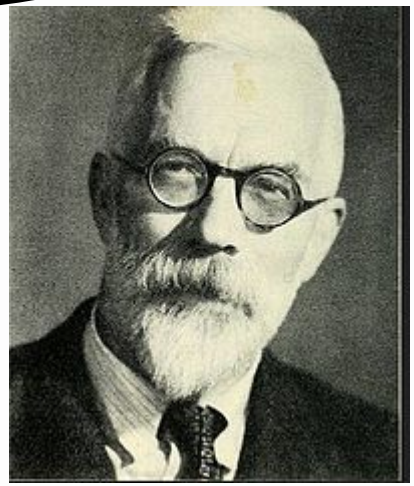
- The Neyman-Pearson approach is the standard **logic** underlying almost all statistics you will see in journals, though few of its users would recognize the name.

- Though researcher often don't **understand** the logic, and many people misuse *p*-values.

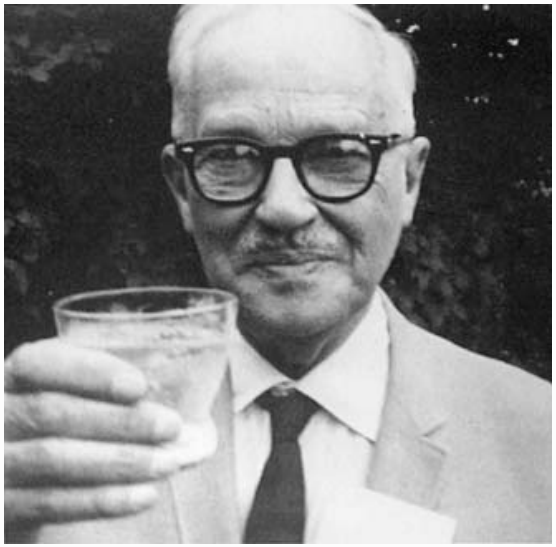
Neyman vs. Fisher



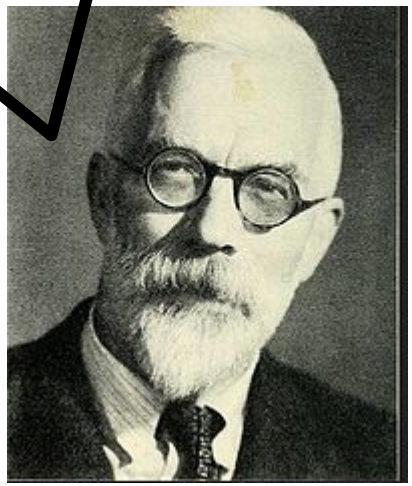
Haha, I've won, my approach to statistics is the underlying logic of almost all statistical tests you see in journals!



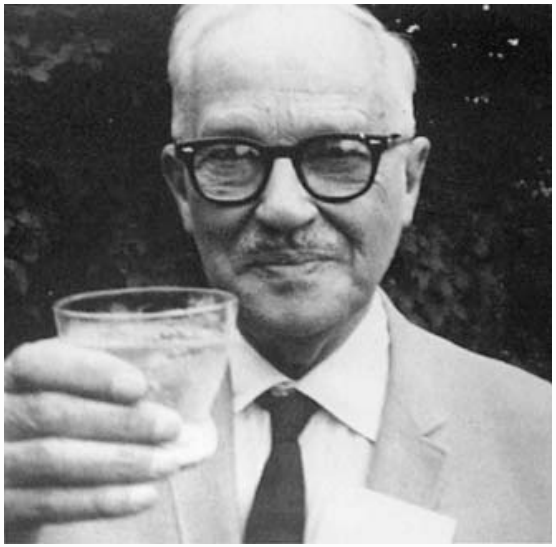
Neyman vs. Fisher



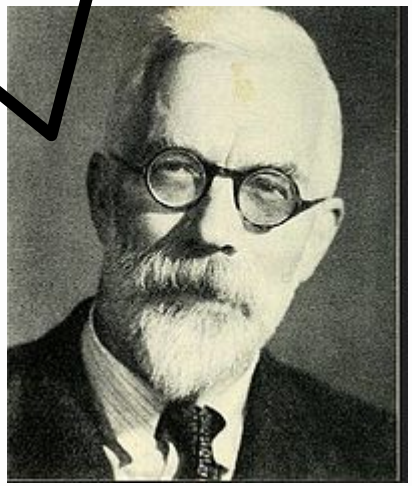
Oh shut up! No one knows your name, and everyone uses *p*-values in the incorrect way I proposed!



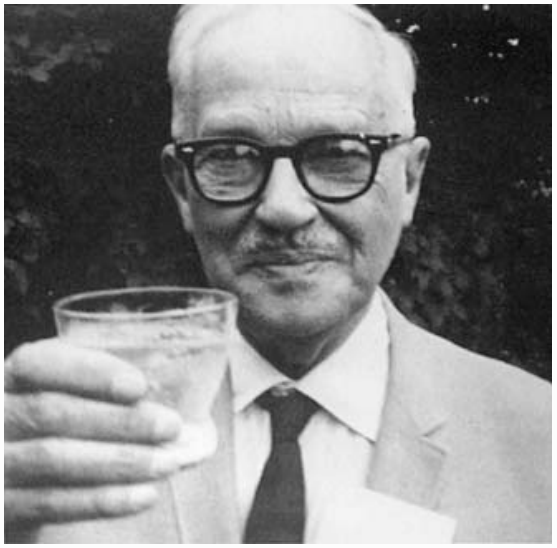
Neyman vs. Fisher



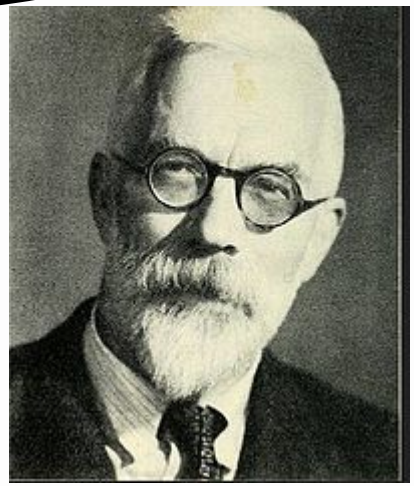
And, in case you didn't know, people love me so much, they named the *F*-distribution in my honor!



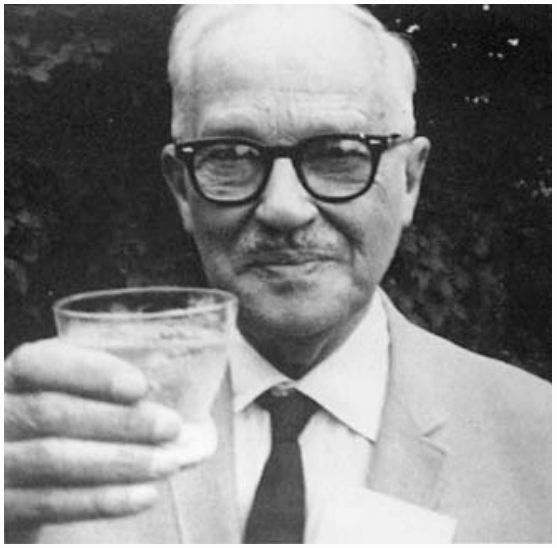
Neyman vs. Fisher



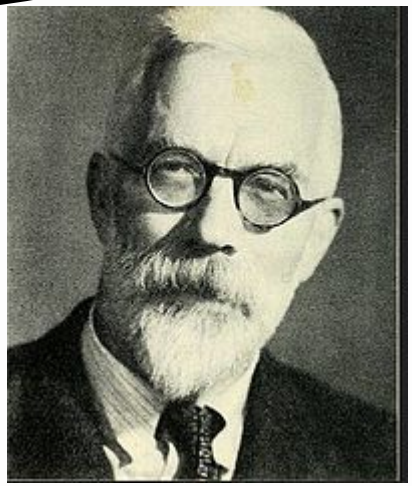
.....



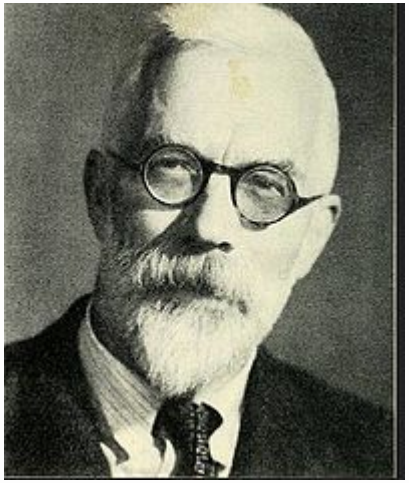
Neyman vs. Fisher



Like, whatever, Mr Eugenicist

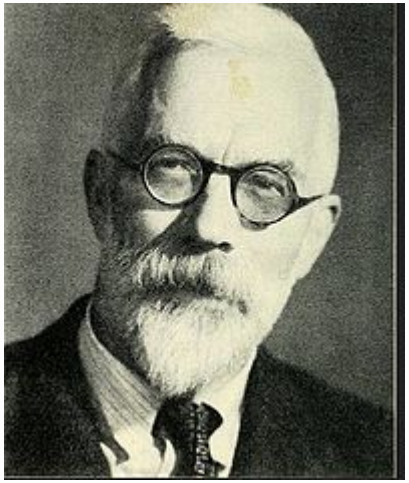


Neyman vs. Fisher vs. Bayes



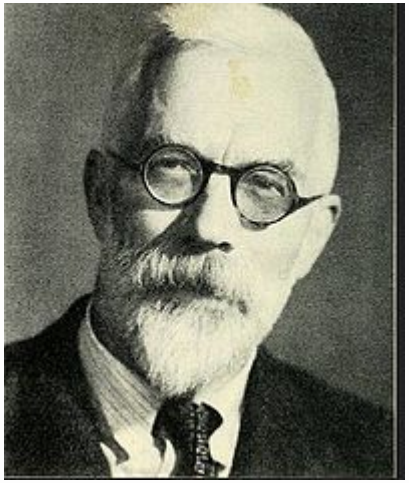
Neyman vs. Fisher vs. Bayes

Gentleman! Calm down! In the future, everyone will use Bayesian statistics anyway! One journal has already banned your silly p-value



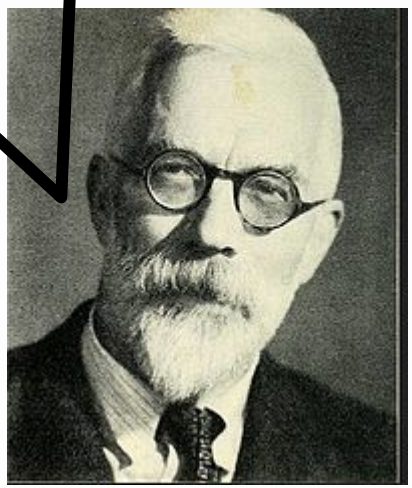
Neyman vs. Fisher vs. Bayes

Yeah, right. My prior on that happening isn't very high, Reverend Bayes.

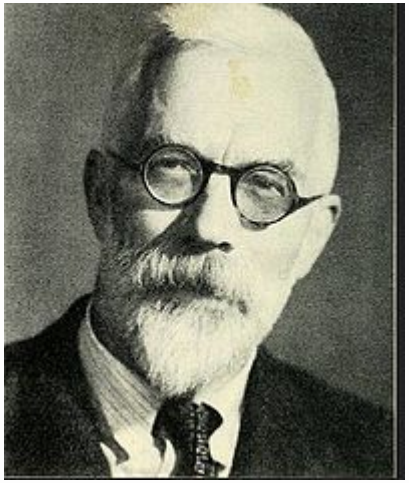


Neyman vs. Fisher vs. Bayes

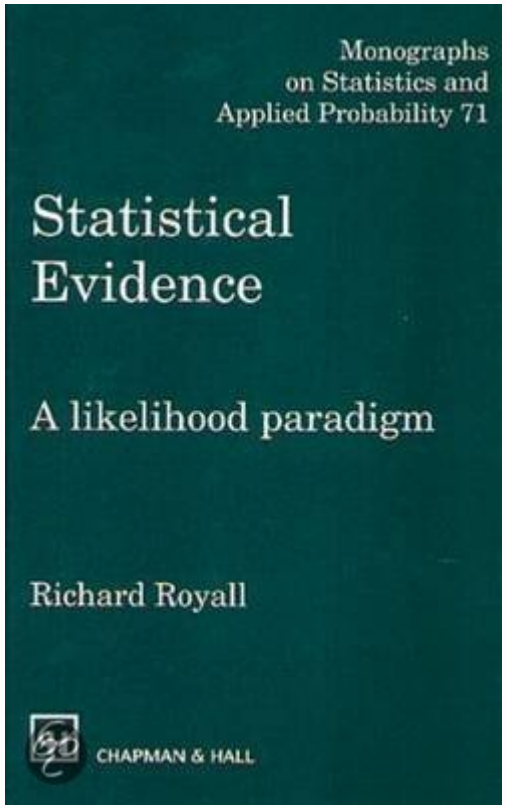
Haha, good one Jerzy, my Frequentist friend. Come, let's go for a long run.



Neyman vs. Fisher vs. Bayes

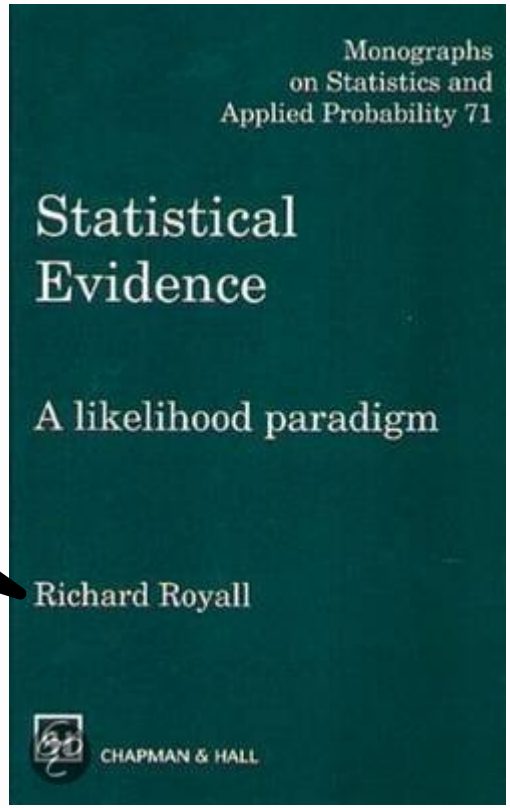


Bayes vs. Royall



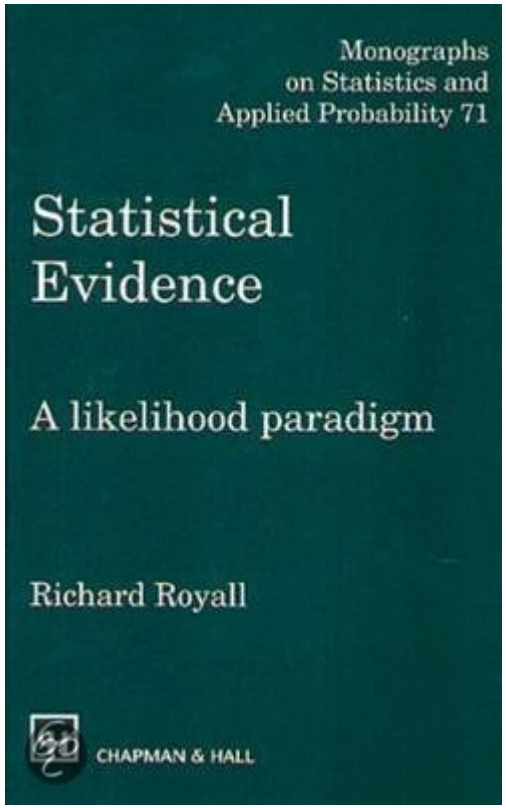
Bayes vs. Royall

No one cares about your subjective opinion, Reverend Bayes. Let's use likelihoods without priors!



Bayes vs. Royall

Who are you? I mean, I can't even find your picture on the internet, dude!



Three Paths to Salvation

"Truth is One, The Paths are Many"

[Bhagavad Gita]

- The Karma yoga: The path of Action
- The Jnana yoga: The path of Devotion
- The Bhakti yoga: The path of Knowledge

Three Paths to Salvation

"Truth is One, The Paths are Many"

[Bhagavad Gita]

- The path of Action (Neyman-Pearson)
- The path of Devotion (Bayes)
- The path of Knowledge (Royall)

What is a *p*-value?

This should be easy, right? Right.

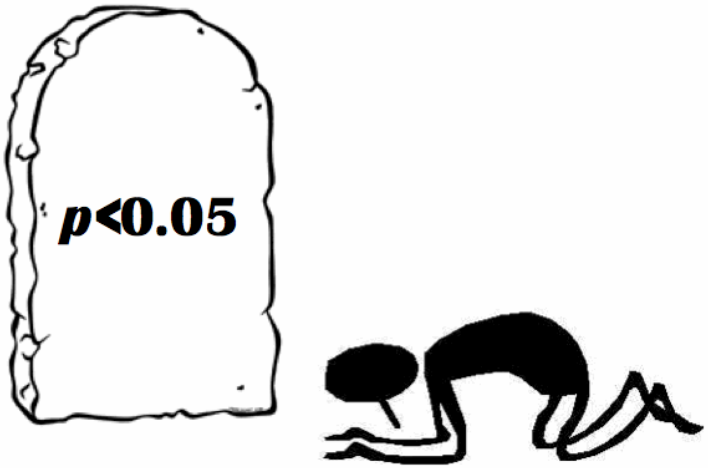
What is a *p*-value?

- P-values are what you use if you don't know Bayesian statistics.

What is a *p*-value?

- Does the use of a cell-phone increase the likelihood of getting into a car accident compared to not using a cell phone?

- This difference is either larger than 0, is not.



What is a *p*-value?

- Name: Null-Hypothesis Significance Testing.
  - But you can call me NHST

- However, effects are not always ‘significant’ (in the common meaning of ‘important’).
- We’ll say: **null-hypothesis testing**
- Observed effects are **statistically different from zero** (even though the ‘null’ does not *need* to be ‘nil’, or 0, it often *is*).

What is a *p*-value?

- If you compare 2 groups on some dependent variable, the difference will not be exactly 0. What if you find people who call while driving get into 0.58 accidents more, on average?
  - A) That means they get into accidents more
  - B) That could just be random variation around 0

What is a *p*-value?

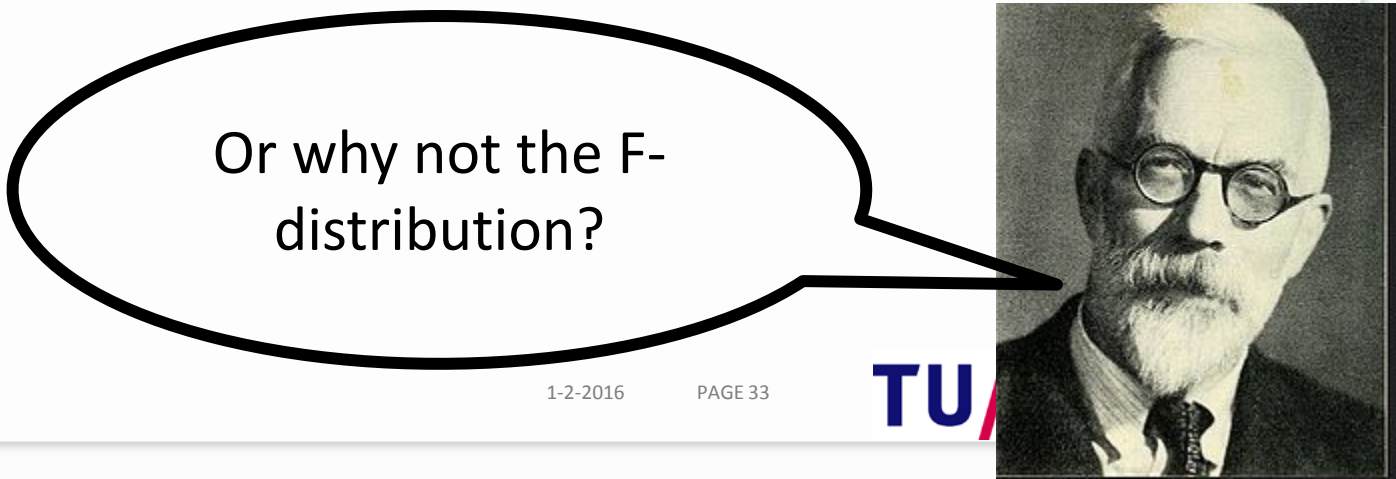
- We need a test statistic to tell us whether this value of 0.58 is surprising or not.

- We compare this test statistic to a distribution (normal distribution, *t* distribution, chi-square)

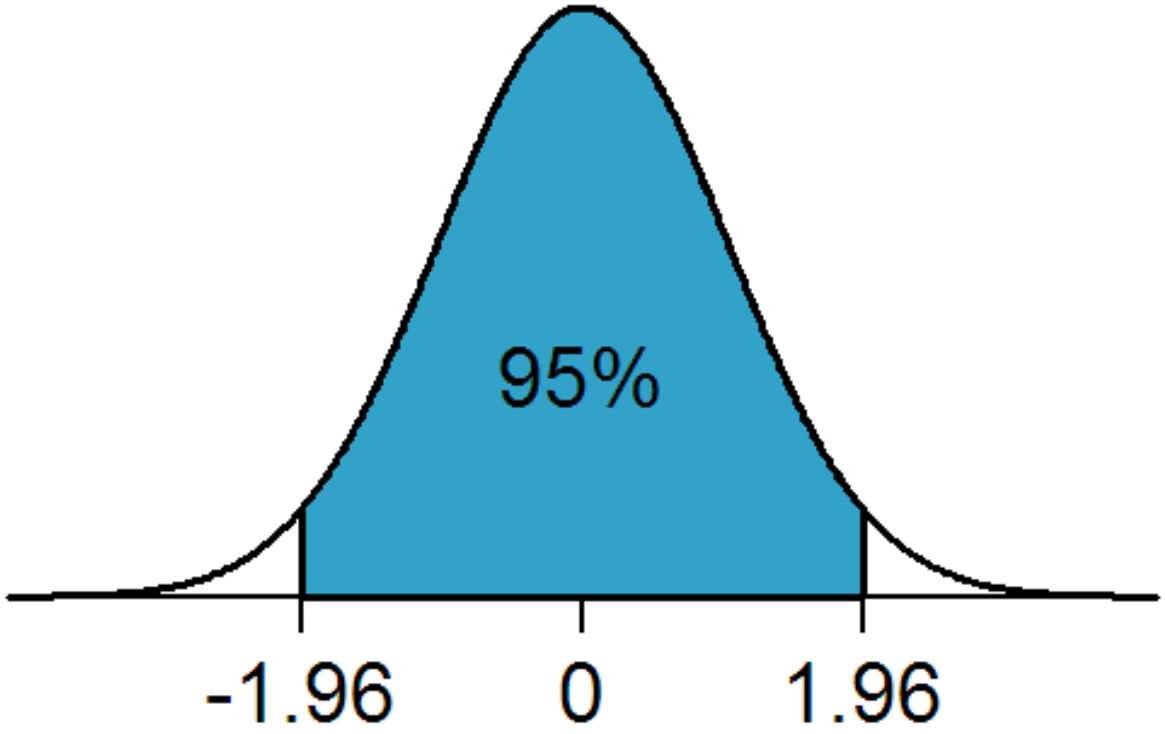
What is a *p*-value?

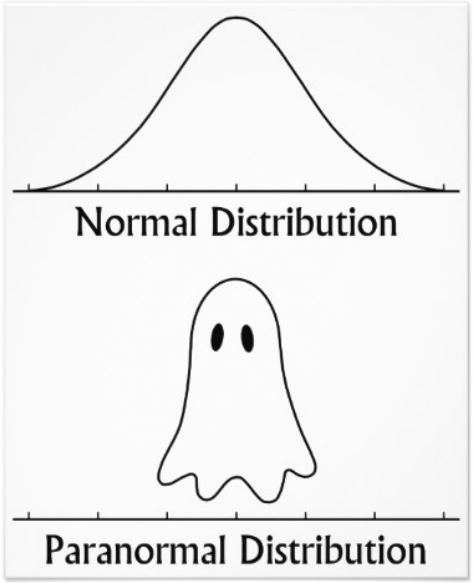
- We need a test statistic to tell us whether this value of 0.58 is surprising or not.

- We compare this test statistic to a distribution (normal distribution, *t* distribution, chi-square)



P-value





P-value

- Now that we have a *p*-value, what does it mean?

- **A *p*-value is the probability of getting the observed or more extreme data, assuming the null hypothesis is true.**

- (see how it's a statement about your data?)

P-value

👎👎We found a *p*-value < 0.05, so our theory

👍👍We found a *p*-value <0.05, so our data

What does a $p<.05$ mean?

- From http://www.popsci.com/race-prove-spooky-quantum-connection-may-have-winner

With only 245 events, statistics dictates a four percent chance that the result was due to chance, meaning that Bell's threshold may not have actually been crossed.

"In other words, there's a 96 percent chance that they won the race," says Paul Kwiat, an experimental quantum physicist who works with photons and is

SCIENCE

THE RACE TO PROVE 'SPOOKY' QUANTUM CONNECTION MAY HAVE A WINNER

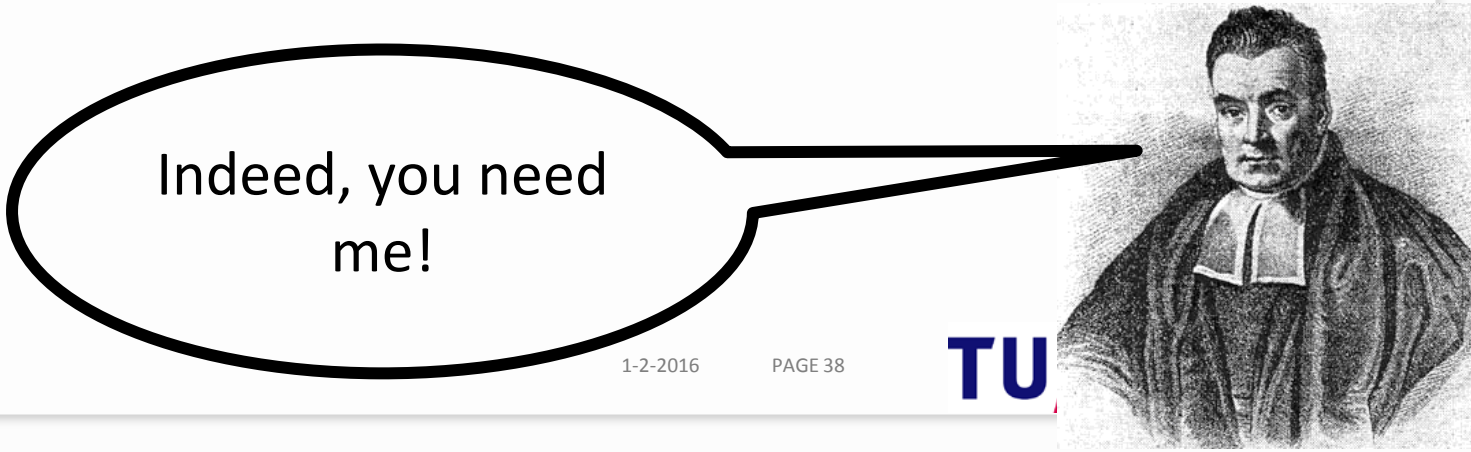
ENTANGLEMENT BREAKTHROUGH COULD LEAD TO UNHACKABLE INTERNET

By Devin Powell Posted August 29, 2015

What does a *p*<.05 mean?

- The data we have observed should therefore be considered **surprising** if H0 would be true.

- A *p*-values does not give the probability that the null-hypothesis is true, given the data (we need Bayesian statistics for this).



What does a *p*<.05 mean?

- We have rejected ('falsified') the null (with a certain error percentage).
  - The Higgs boson used 5 sigma, or $p < 0.0000003$.
  - (Because if you are going to spend $13.25 billion on a scientific finding, you'd better be pretty sure.)

- But Popper is only impressed if you made a **bold** hypothesis. The null is not bold.

What does a *p*<.05 mean?

- You cannot ‘prove’ the alternative hypothesis is true (ever!). You can only ‘corroborate’ it.
- ‘Mere supporting instances are as a rule too cheap to be worth having’ [Popper, 1983]
- One of the ways to introduce Popper’s notion of corroboration is by means of the notion of a **severe test**.

What does a *p*>.05 mean?

- If a *p*-value is **larger** than 0.05, the data we have observed is not surprising. This doesn't imply that the null-hypothesis is true.

- **The *p*-value can only be used as a test to reject the null-hypothesis. It can never be used to accept the null-hypothesis as true**.

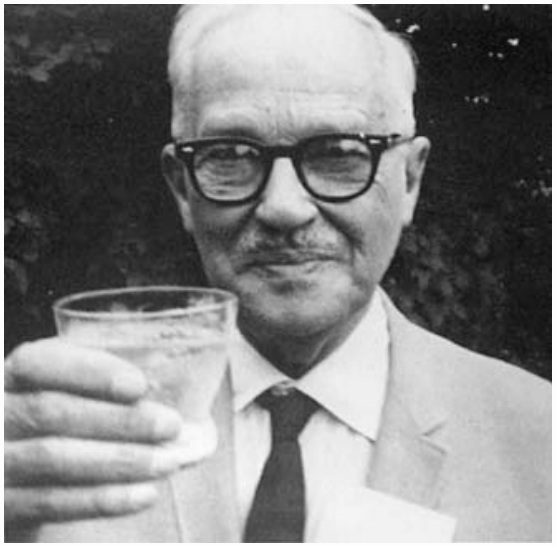
What does a *p*>.05 mean?

- I try to think of it as MU (無)
- A monk asked Joshu, a Chinese Zen master: `Has a dog Buddha-nature or not?'Joshu answered: `Mu.' [Mu is the negative symbol in Chinese, meaning `No-thing' or `Nay'.]
- “Un-asking” the question

What does a *p*>.05 mean?

- I try to think of it as MU (無)
- A monk asked Joshu, a Chinese Zen master: `Has a dog Buddha-nature or not?'Joshu answered: `Mu.' [Mu is the negative symbol in Chinese, meaning `No-thing' or `Nay'.]
- “Un-asking” the question

What does a *p*>.05 mean?



But you can **ACT** as if the null-hypothesis is true!

"Every test of a statistical hypothesis consists in a rule of **rejecting** the hypothesis when a specified character, x, of the sample lies within certain critical limits, and **accepting** it or remaining **in doubt** in all other cases.

What does a *p*>.05 mean?

- Lakatos:
  - Research programmes based on a ‘hard core’ of theoretical assumptions that cannot be abandoned or altered without changing the programme.
  - A ‘protective belt’ around the hard core consists of **auxiliary hypotheses**.
  - Popper had a very negative attitude to such ‘ad-hoc’ theoretical amendments. But Lakatos differentiates between progressive and degenerative research lines.

What does a *p*>.05 mean?

- Progressive research lines:
  - The changes to the theory have increased it's predictive power. It can now explain more than before
- Degenerative research lines:
  - Offering some explanation for troublesome evidence.
- So $p > 0.05$ takes you further into a degenerative research line.
- But before a degenerative research line can be abandoned, we need a *viable alternative*.
  - Sometimes, this alternative is simply: It was a fluke.

What does a *p*>.05 mean?

The lesson of history is that a bold and plausible theory that fills a scientific need is seldom broken by the impact of contrary facts and arguments. Only with an alternative theory can we hope to displace a defective one.

[Stevens, 1957]

Interpreting *p*-values

- *P*-values are correlated with evidential value, but far from perfectly correlated with evidential value (as shown by Bayes Statistics).

- In general, a low *p*-values warrants further research, but is not in itself support for a theory.

Misinterpreting *p*-values

P-values are not:

- The probability a theory is true (you need **Bayesian statistics** for this)
- The probability a finding will replicate (this depends on the **power** of a study)
- The probability you have made a type 1 error (this depends on **probability H0 is true**, only 5% if p(H0)=1)

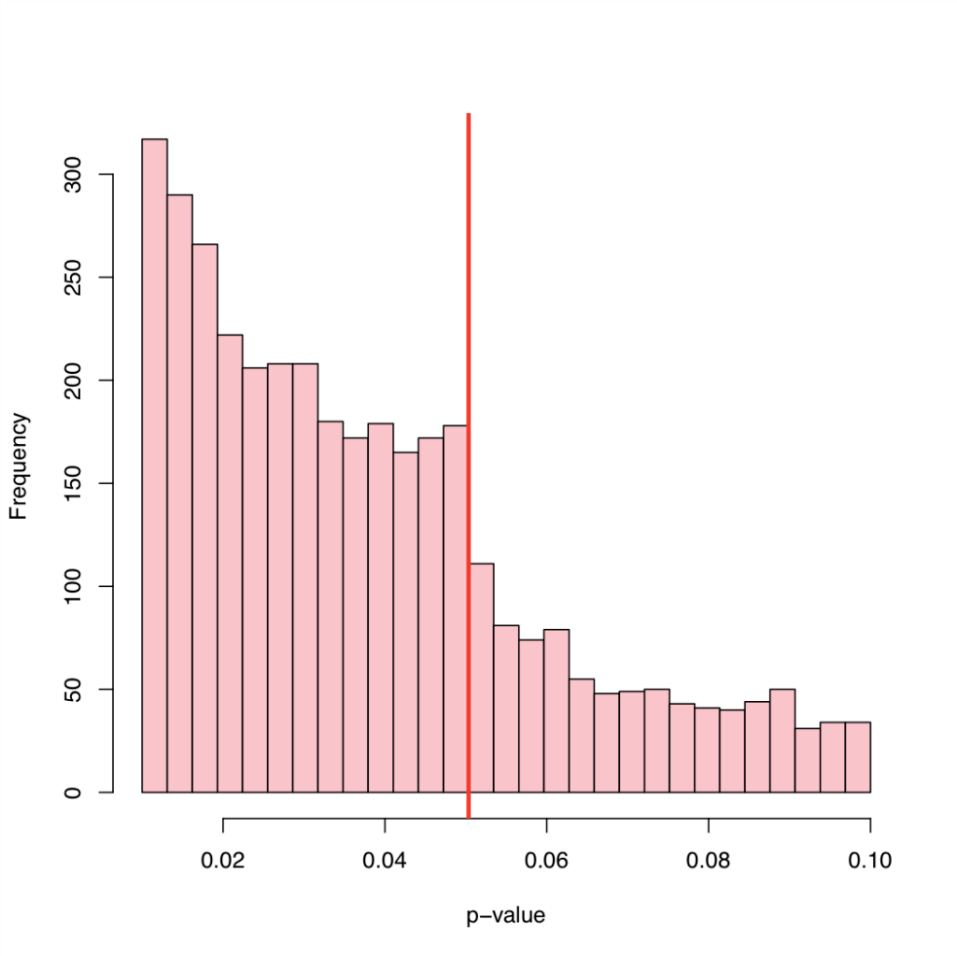
(see Nickerson, 2000 – really, it's very good)

Problems with a focus on *p* <0.05

- One of the biggest problems with the widespread focus on *p*-values is their use as a selection criterion of which findings provide 'support' for a hypothesis and which don't.

- Due to publication bias, tests with p-values below 0.05 are much more likely to be published than those above 0.05.

Problems with a focus on p <0.05



The history of NHST

- when a small *p*-value is observed: … [e]ither an exceptionally rare chance has occurred, or [H0] is not true [i.e., strong evidence against H0]" (Fisher 1956, p. 39)

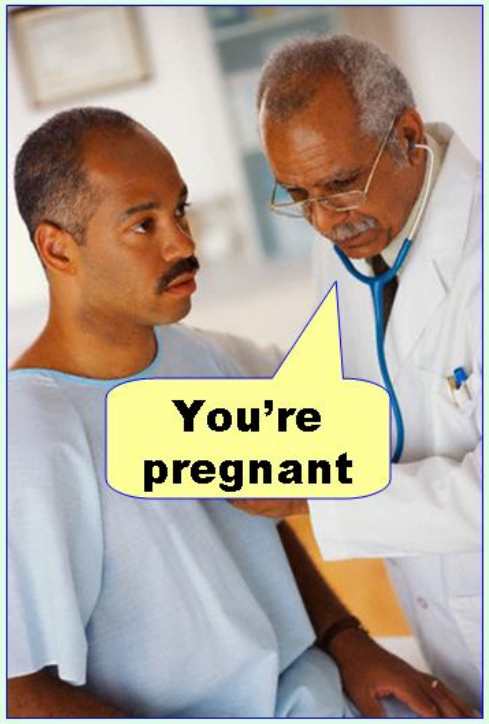
- **a significant *p*-value allows us to reject H0, but a non-significant *p*-value does not allow us to accept H0**

The history of NHST

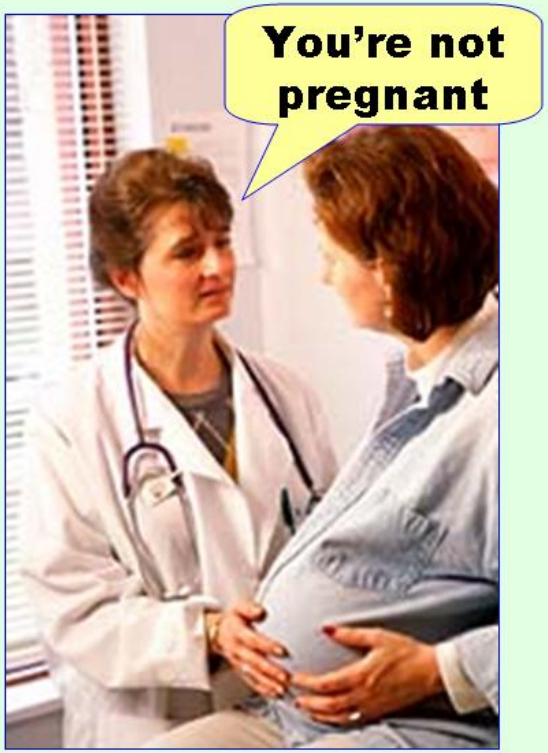
- Something both Fisher and Neyman agreed upon, but which is now often lost in statistical inferences, is that **statistical inferences should be used with "discretion and understanding, and not as instruments which themselves give the final verdict"**

Neyman-Pearson

Type I error
(false positive)



Type II error
(false negative)



4 possible outcomes of a study

	H0 True	H1 True
Significant Finding	False Positive (α)	True Positive (1-β)
Non-Significant Finding	True Negative (1- α)	False Negative (β)

- **The percentage of false positives equals the Type 1 error rate (or α, the significance level)**.
  - This means that if you would perform 1,000 studies, and set the α level to 5% (or 0.05) as is normally done, then you can expect to observe 50 studies that show an effect that is **statistically different from zero in the sample you collected**, even though there is **no real difference in the population**.

Power

- The probability of correctly rejecting the null hypothesis is known as the *power* of a statistical test (Cohen, 1988)

- The statistical power of a study is determined by the size of the effect, the sample size of the study (and the reliability of the sample result), and the significance criterion (typically $\alpha = .05$).

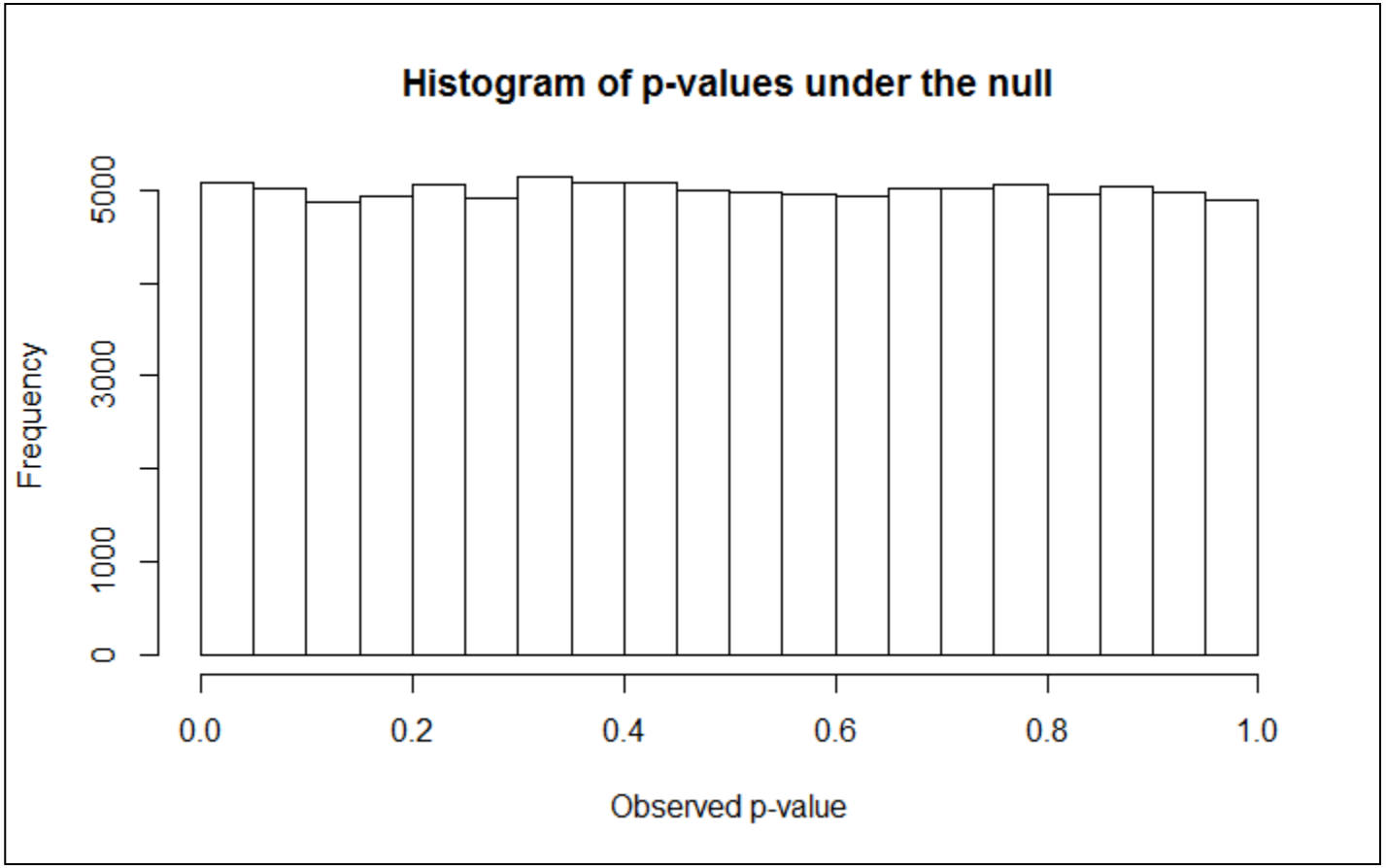
- **The percentage of false negatives (or Type 2 errors, β) equals 1 minus the power of the study.**
  - This means that if your study has 90% power (so a probability of 90% to find an effect that is statistically different from zero, if there really is an effect) then there obviously is a 10% probability of not finding it when it is there, or a 10% Type 2 error rate.

Which *p*-values can you expect?

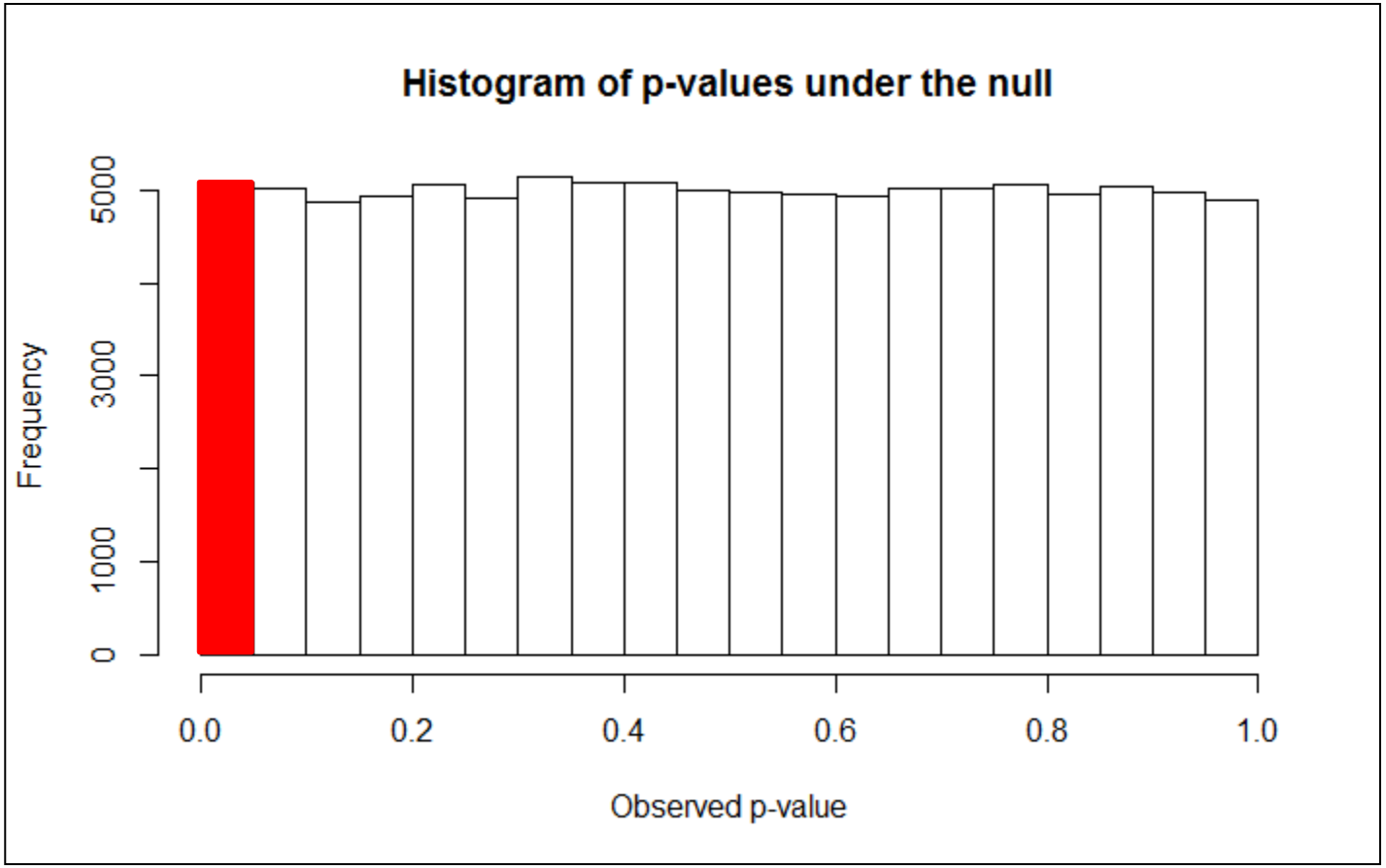
Which *p*-values can you expect?

- **Assuming the null hypothesis is true** (in other words: having 0 power)**, *p*-values are uniformly distributed**. Every *p*-value is equally likely to be observed.

Which *p*-values can you expect?

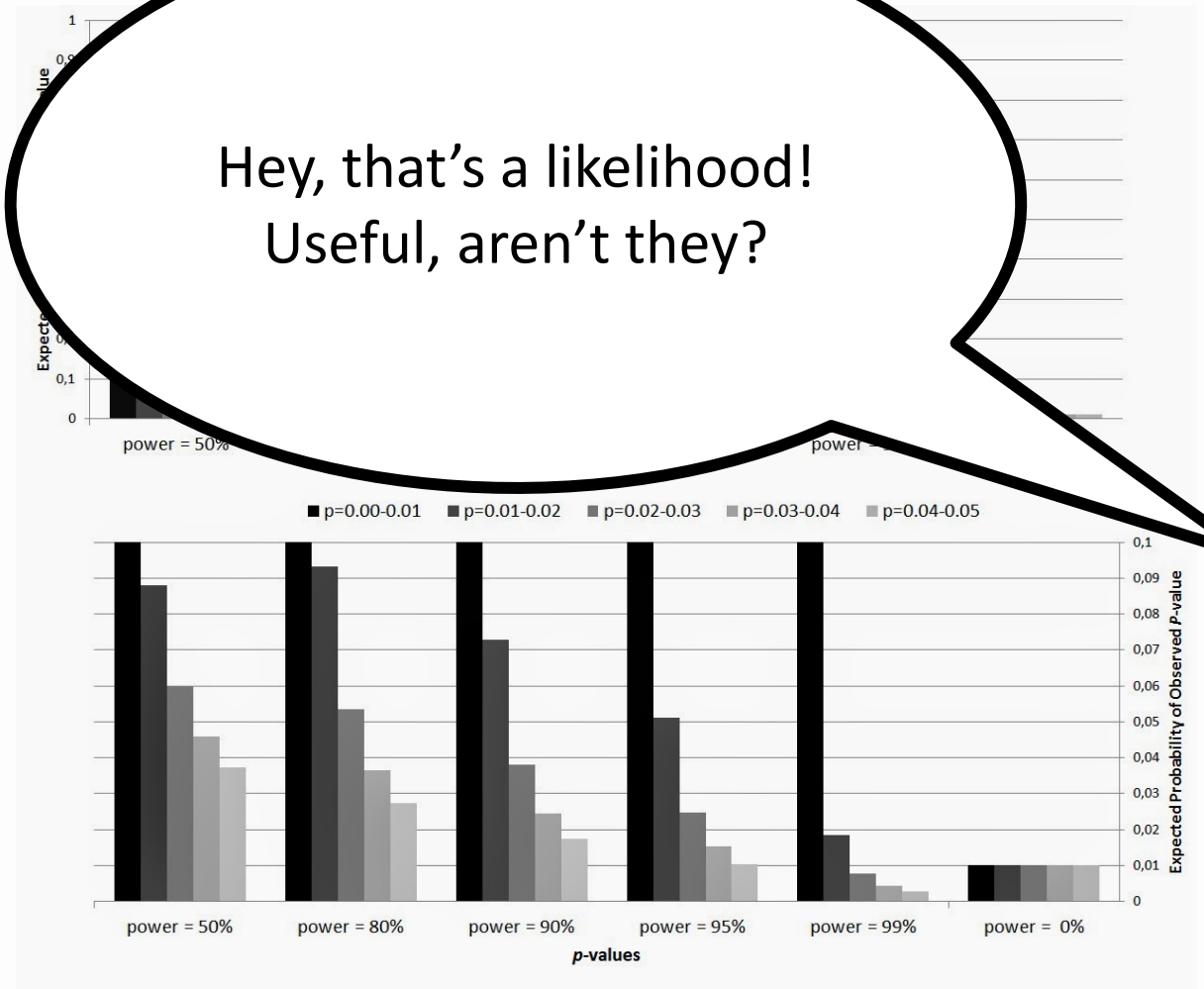


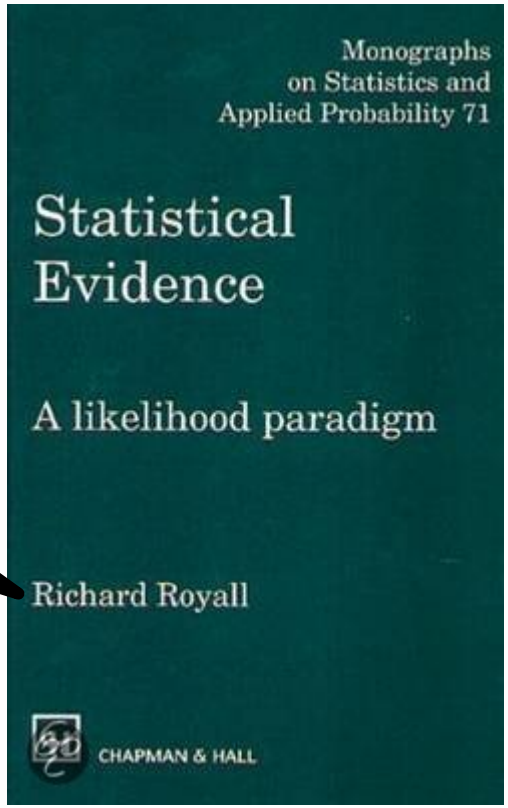
Which *p*-values can you expect?



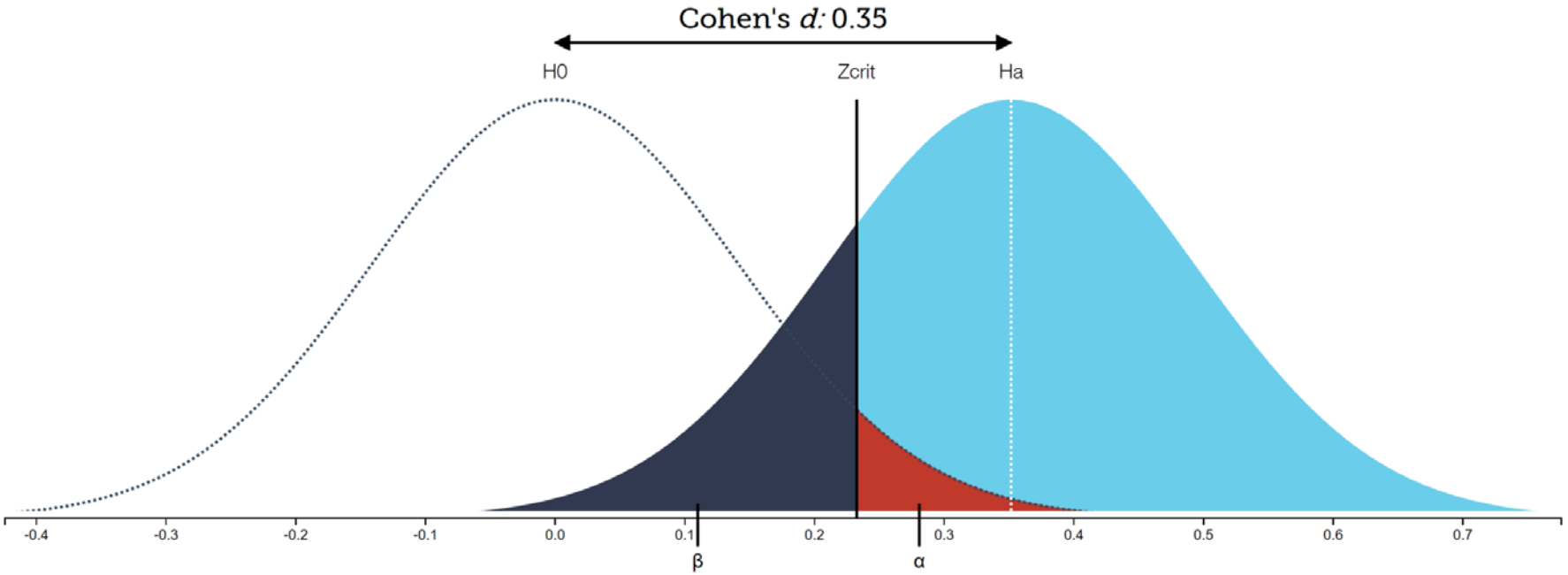
Which *p*-values can you expect?

Hey, that's a likelihood!
Useful, aren't they?





Which *p*-values can you expect?



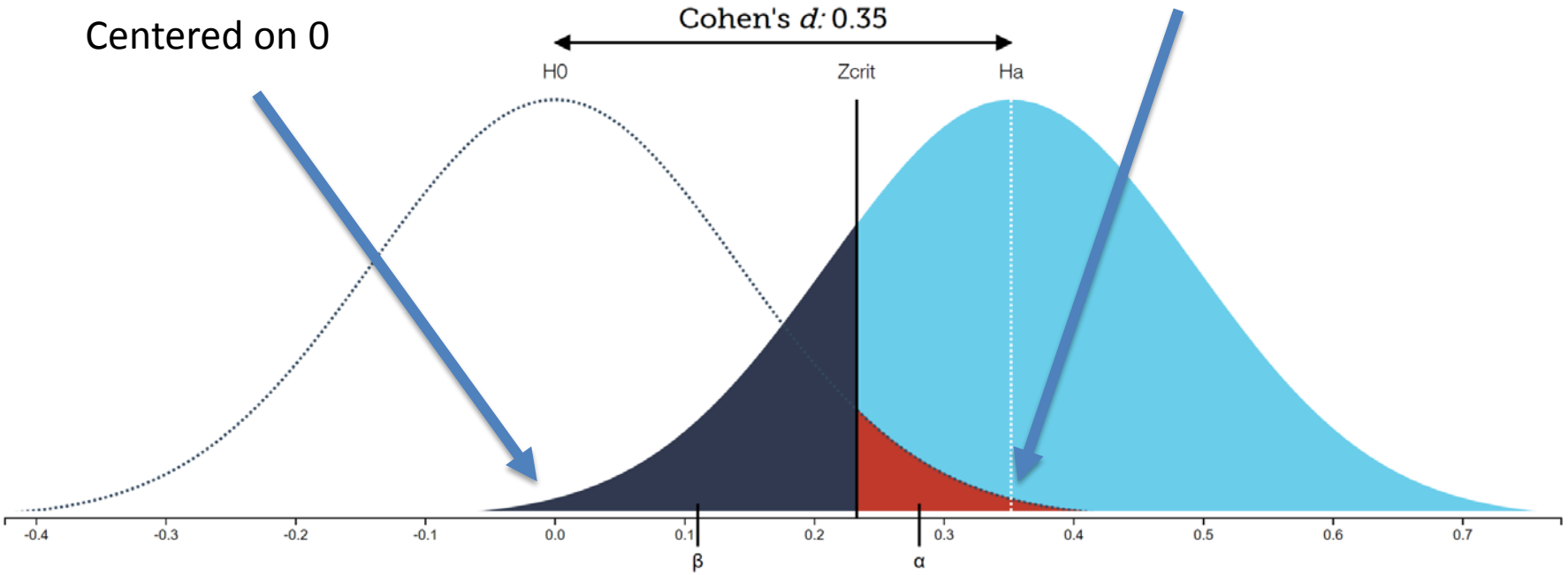
From: http://rpsychologist.com/d3/NHST/

Be sure to visit Kristoffer Magnusson's site!

Which *p*-values can you expect?

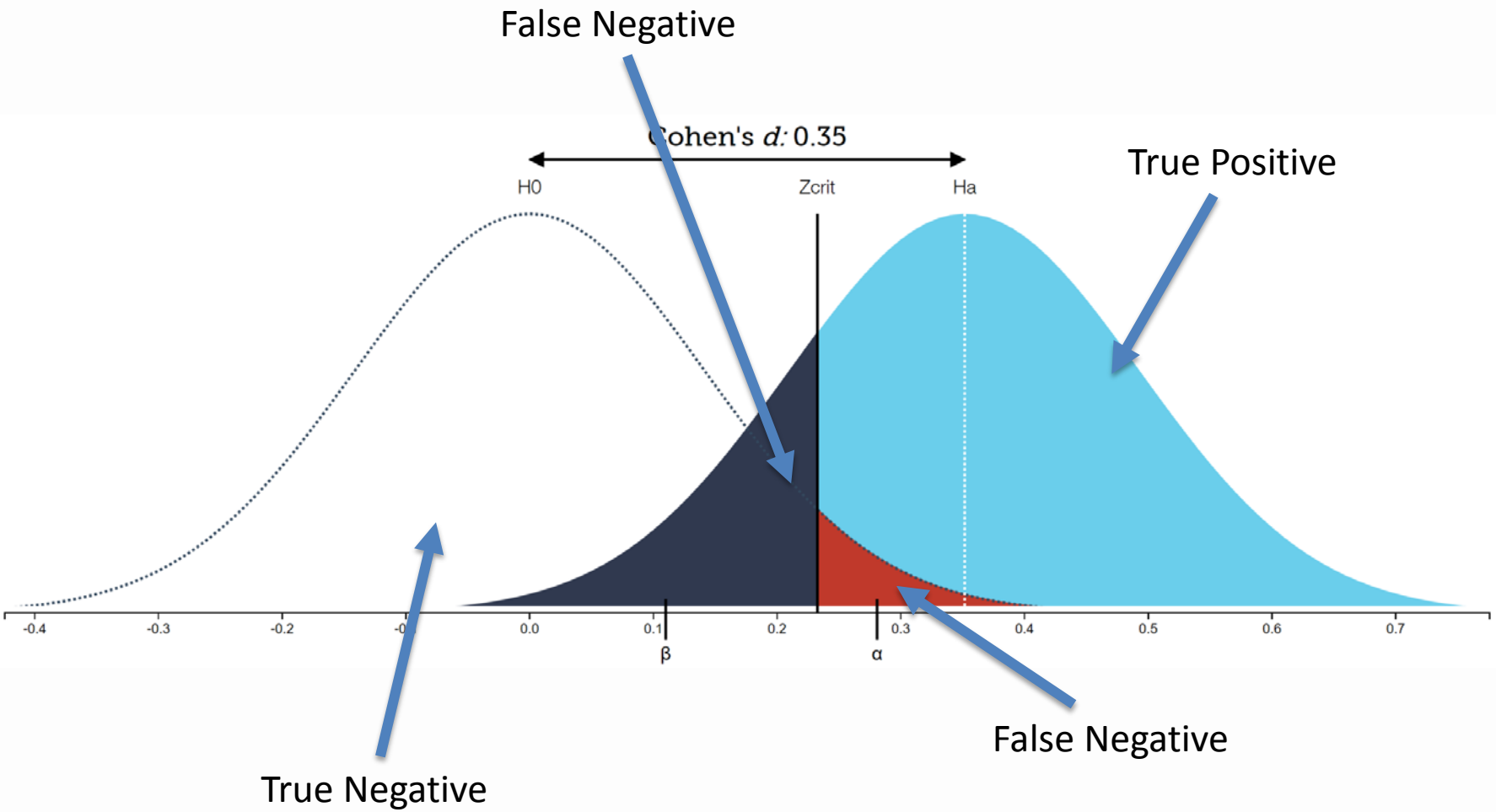
Distribution if H1 is true
Centered on d = 0.35

Distribution if H0 is true
Centered on 0



From: http://rpsychologist.com/d3/NHST/
Be sure to visit Kristoffer Magnusson's site!

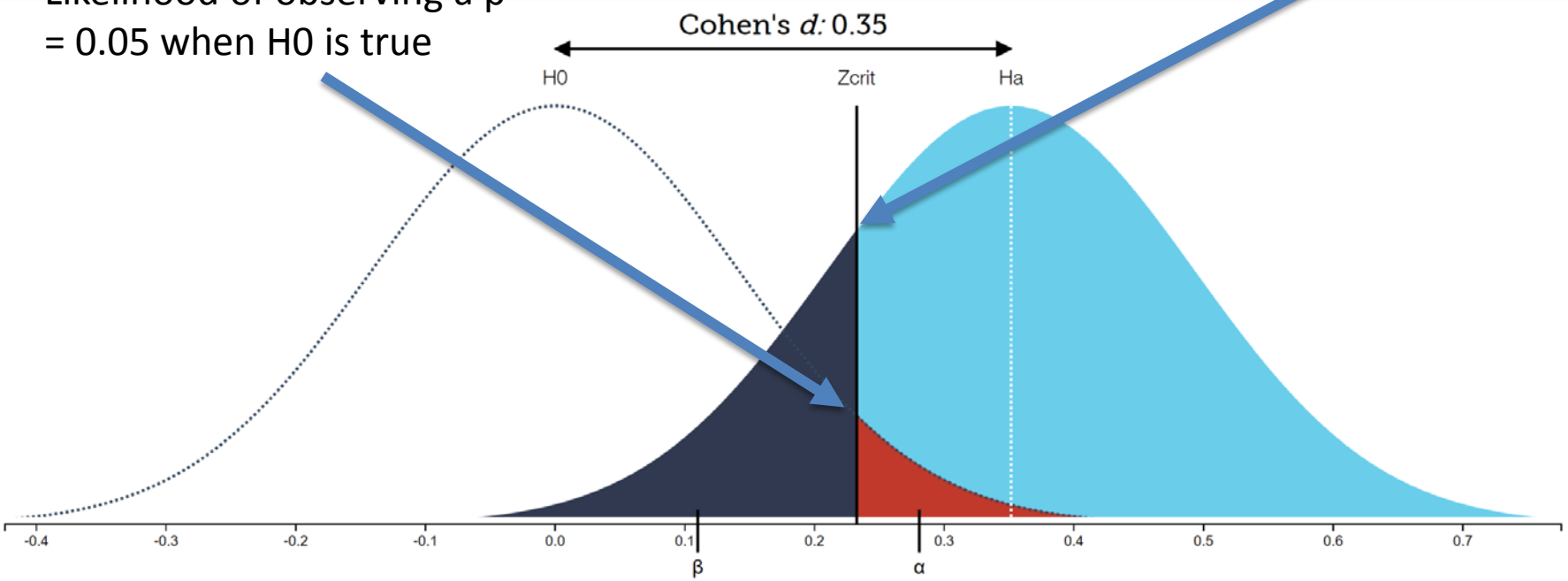
Which *p*-values can you expect?



Which *p*-values can you expect?

Likelihood of observing a p = 0.05 when H0 is true

Likelihood of observing a p = 0.05 when H1 is true

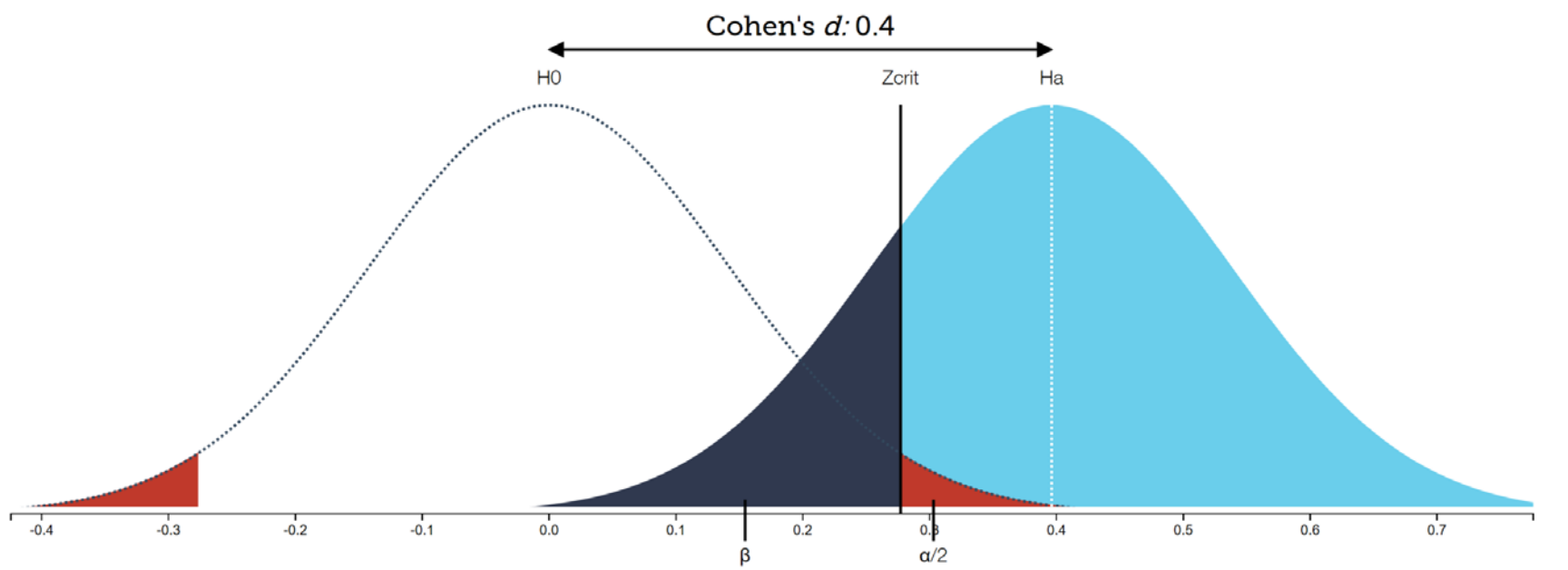


From: http://rpsychologist.com/d3/NHST/

Be sure to visit Kristoffer Magnusson's site!

Which *p*-values can you expect?

Here, we have 80% power



Which *p*-values can you expect?

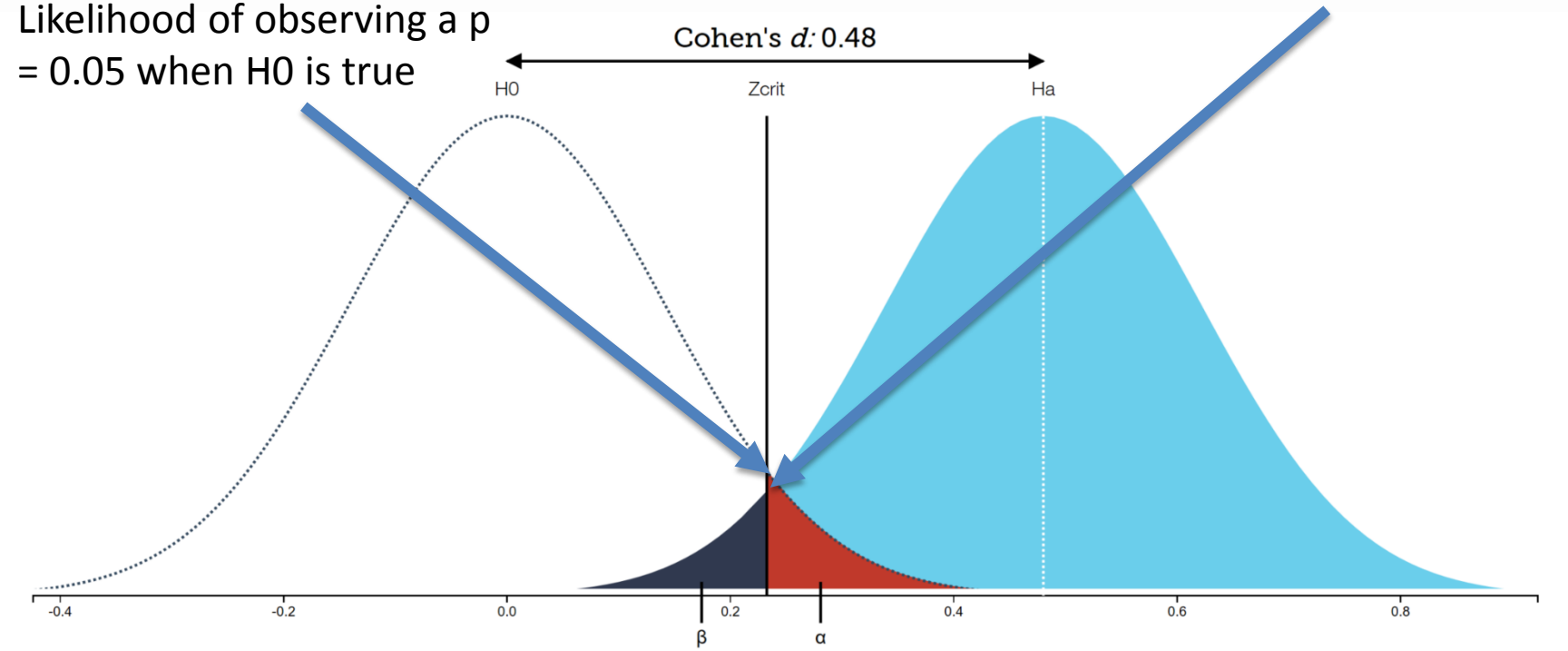
- So, with 80% power, finding a $p = 0.05$ when H1 true is more likely than finding a $p = 0.05$ when H0 is true. The *p*-value is evidence for H1 **relative to** H0.
- This is a likelihood, or the *path of knowledge*. Likelihoods tell us the relative likelyhood of the data under a specific hypothesis.

Which *p*-values can you expect?

Here, we have 96% power

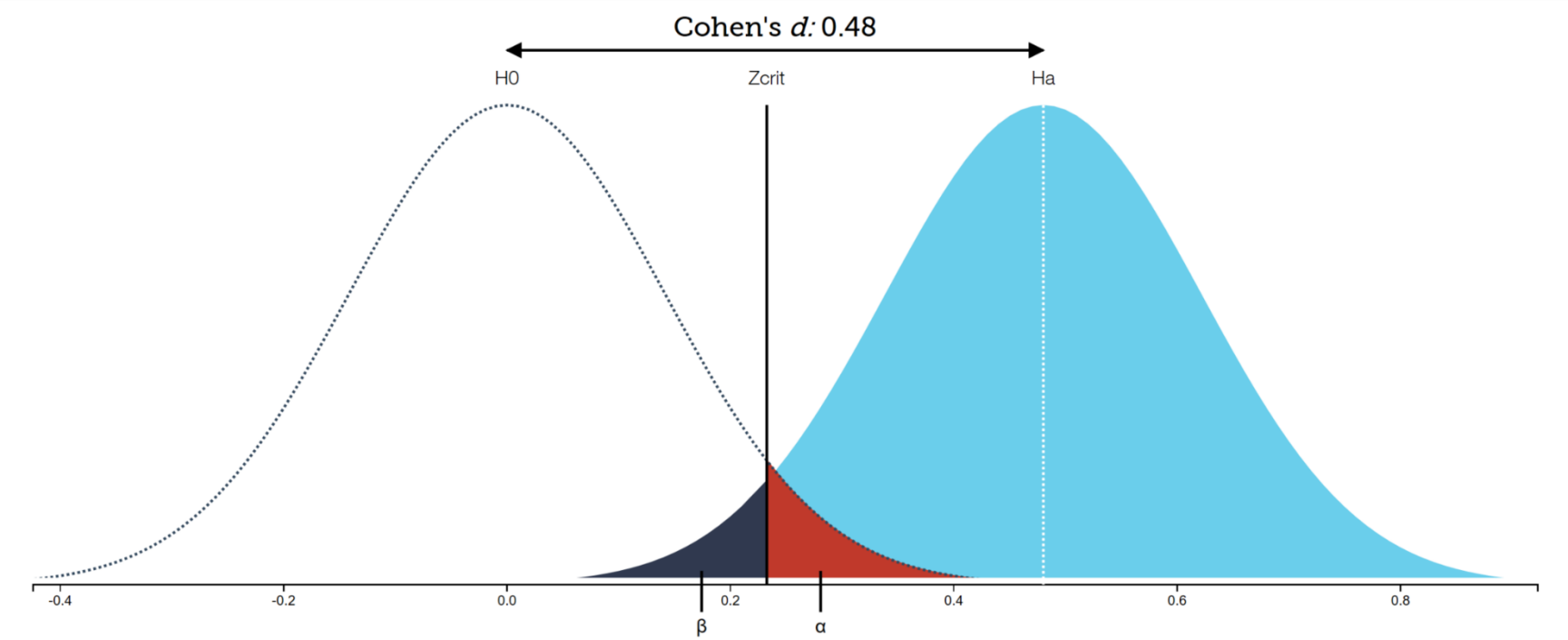
Likelihood of observing a p = 0.05 when H1 is true

Likelihood of observing a p = 0.05 when H0 is true



Which *p*-values can you expect?

Here, we have 96% power



Which *p*-values can you expect?

- So, with 96% power, finding a (one-sided) p = 0.05 when H0 true is more likely than finding a p = 0.05 when H1 is true. The *p*-value is evidence for H0 **relative to** H1.
- But which H1?
  - That's up to you. But you can only have relative evidence, so you always need to compare 2 hypotheses when you want to provide evidence.

Interpreting *p*-values

CrossMark click for updates

Gender contributes to personal research funding success in The Netherlands

Romy van der Lee[1] and Naomi Ellemers[2]

Department of Social and Organizational Psychology, Institute of Psychology, Leiden University, 2300 RB Leiden, The Netherlands

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved August 19, 2015 (received for review May 26, 2015)

Application Evaluations and Success Rates

Of the total population of 2,823 applications examined, 467 were awarded, resulting in an overall success rate of 16.5%. The success rate was systematically lower for female applicants than for male applicants [14.9% vs. 17.7%; $\chi^2(1) = 4.01$, $P = 0.045$, Cramer's $V = 0.04$] (Table S1). Examination

Interpreting *p*-values

"The data reported herein provide **compelling evidence** of gender bias in personal grant applications"

- If you have a sample of 2823 individuals, a p = 0.045 is not 'compelling evidence' for your hypothesis.

(see also http://blog.casperalbers.nl/science/nwo-gender-bias-and-simpsons-paradox/)

Bayesian Statistics

- But given the evidence, how much should you *believe* the result is true?

- For this, you need to weigh your prior belief. Did you get a $p = 0.045$ on a Stroop effect, or in an experiment examining pre-cognition?

Bayesian Statistics

- Major difference is Bayesian statistics expresses the probability a hypothesis is true, based on the data, and a prior.

$$\Omega = \frac{Pr(H_0 \mid \text{data})}{Pr(H_1 \mid \text{data})} = \frac{f(\text{data} \mid H_0)}{f(\text{data} \mid H_1)} \times \frac{Pr(H_0)}{Pr(H_1)}$$

Posterior probability = Likelihood x prior probability

Bayesian Statistics

- The difference between *Pr*(D|H0) and *Pr*(H0|D) might not be immediately clear, but the two probabilities can be completely different.

Probability of being **Dead**, given that your **Head** is bitten of by a shark:

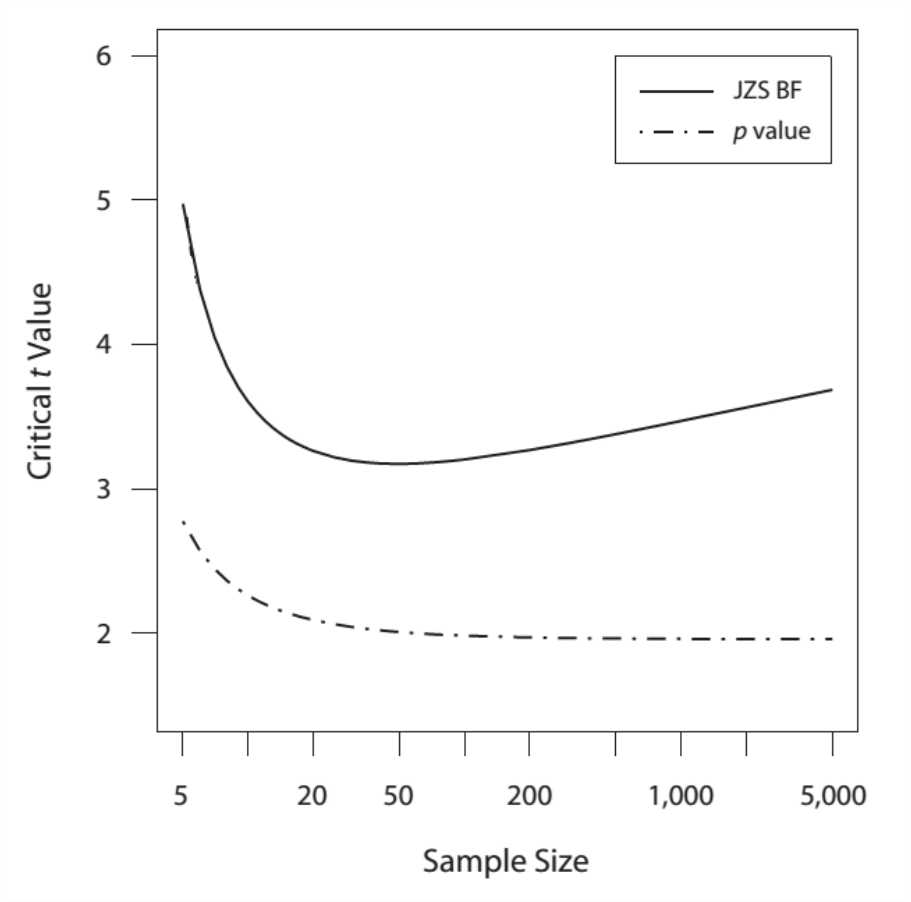
P(D|H) = 0.9999999

Probability of your **Head** being bitten off by a shark, given that you are **Dead**:

P(H|D) = 0.0000002

(It also works is D means Data, and H means Hypothesis)

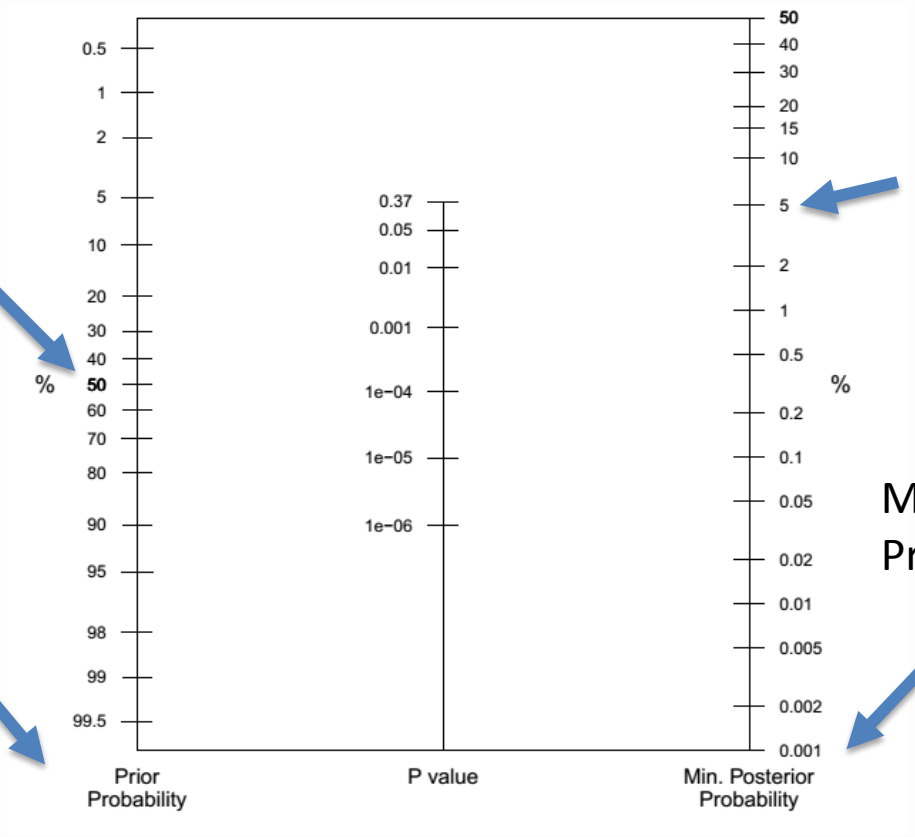
Bayes Factors vs. *p*-values



Bayes Factors vs. *p*-values

E.g., 50%: Before collecting data, you believe there is 50% probability H0 is true (it's like flipping a coin)

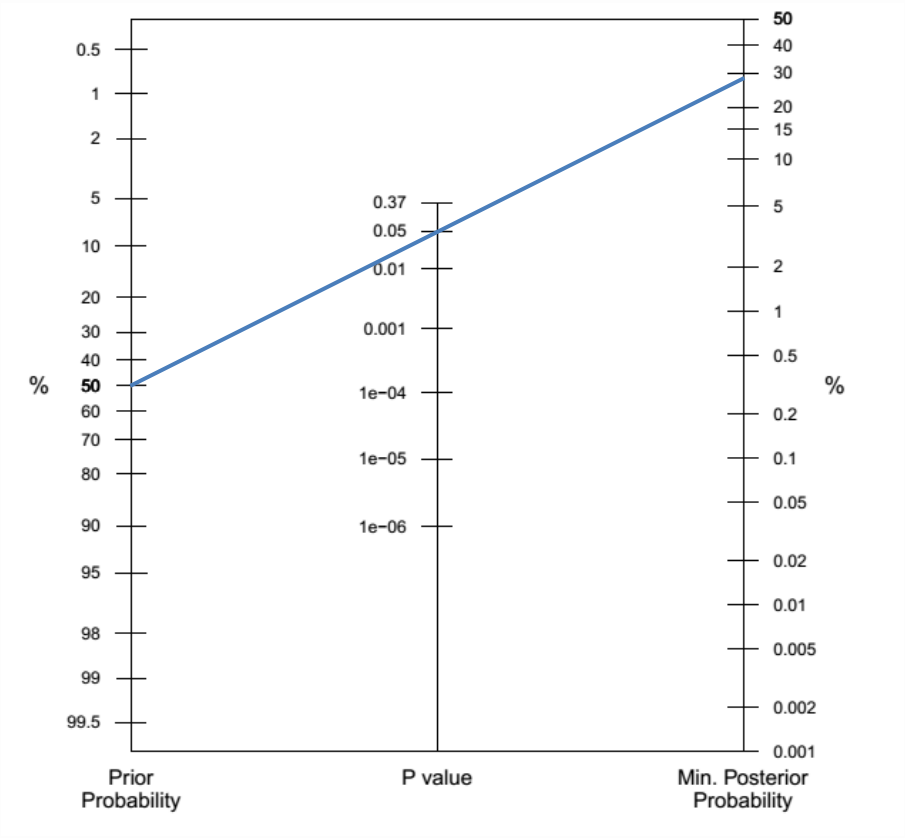
Prior Probability H0 is true



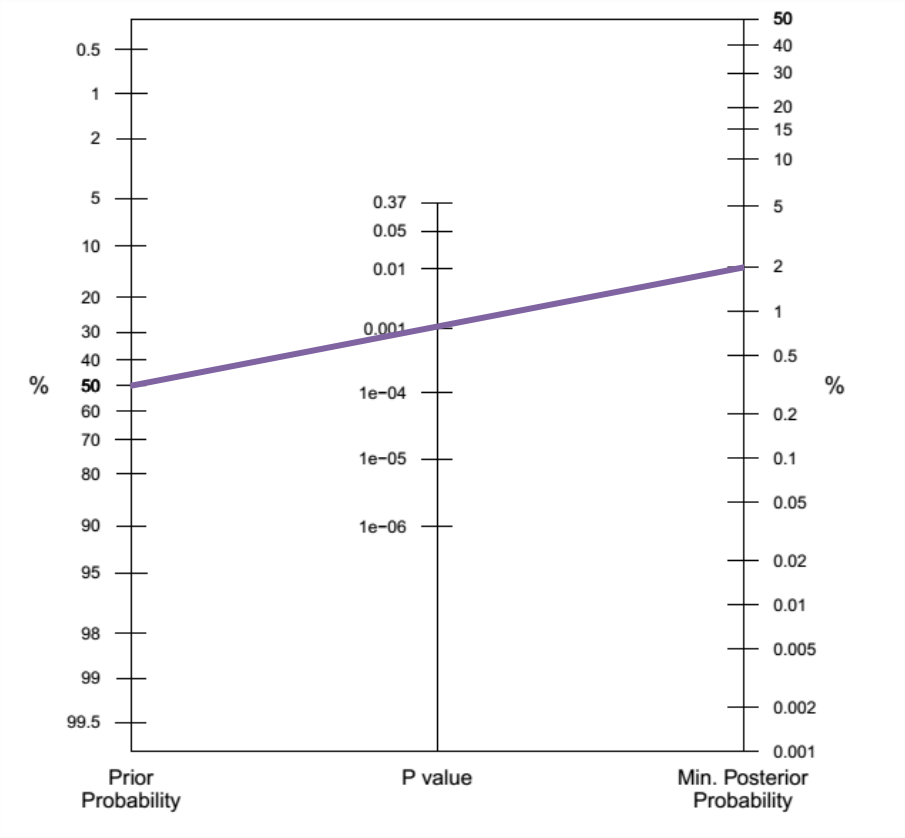
E.g., 5%: After collecting data, the probability H0 is true is only 5%

Minimal Posterior Probability H0 is true

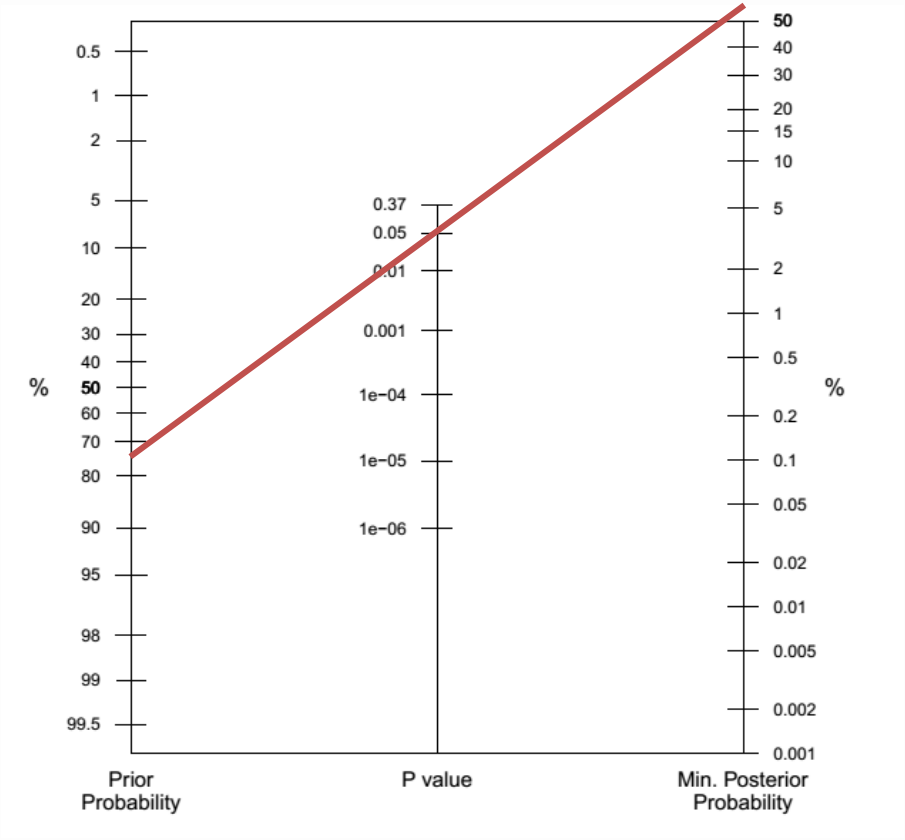
Bayes Factors vs. *p*-values



Bayes Factors vs. *p*-values



Bayes Factors vs. *p*-values

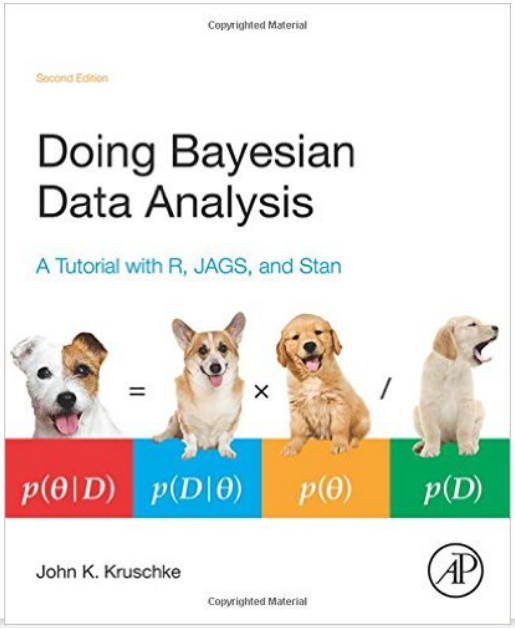


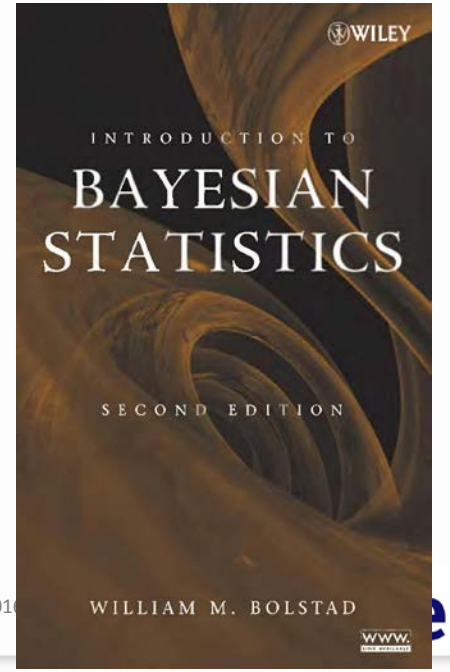
Interpreting *p*-values

- What to do if you have observed a $p = 0.045$?
  - If possible, replicate the study (typically with a larger sample).
  - If not possible, acknowledge the relatively low evidential value of the data.
  - Is the effect surprising? Then it might not be real. Is it predicted a-priori based on solid theory and earlier results? Then it might be real.

Bayesian Statistics

- It's important to understand Bayesian statistics. Subjective beliefs should be important to you, and quantifying it (instead of 'feeling' it) is useful.
- I'm not the person to teach you Bayesian stats, but I can recommend these books:





Bayesian Statistics

- Remember: It's not either-or.
- You can use both Bayesian stats and Frequentist stats – they will lead to similar inferences, most of the time, especially with sufficient data.