A little number and a big controversy: p-Values

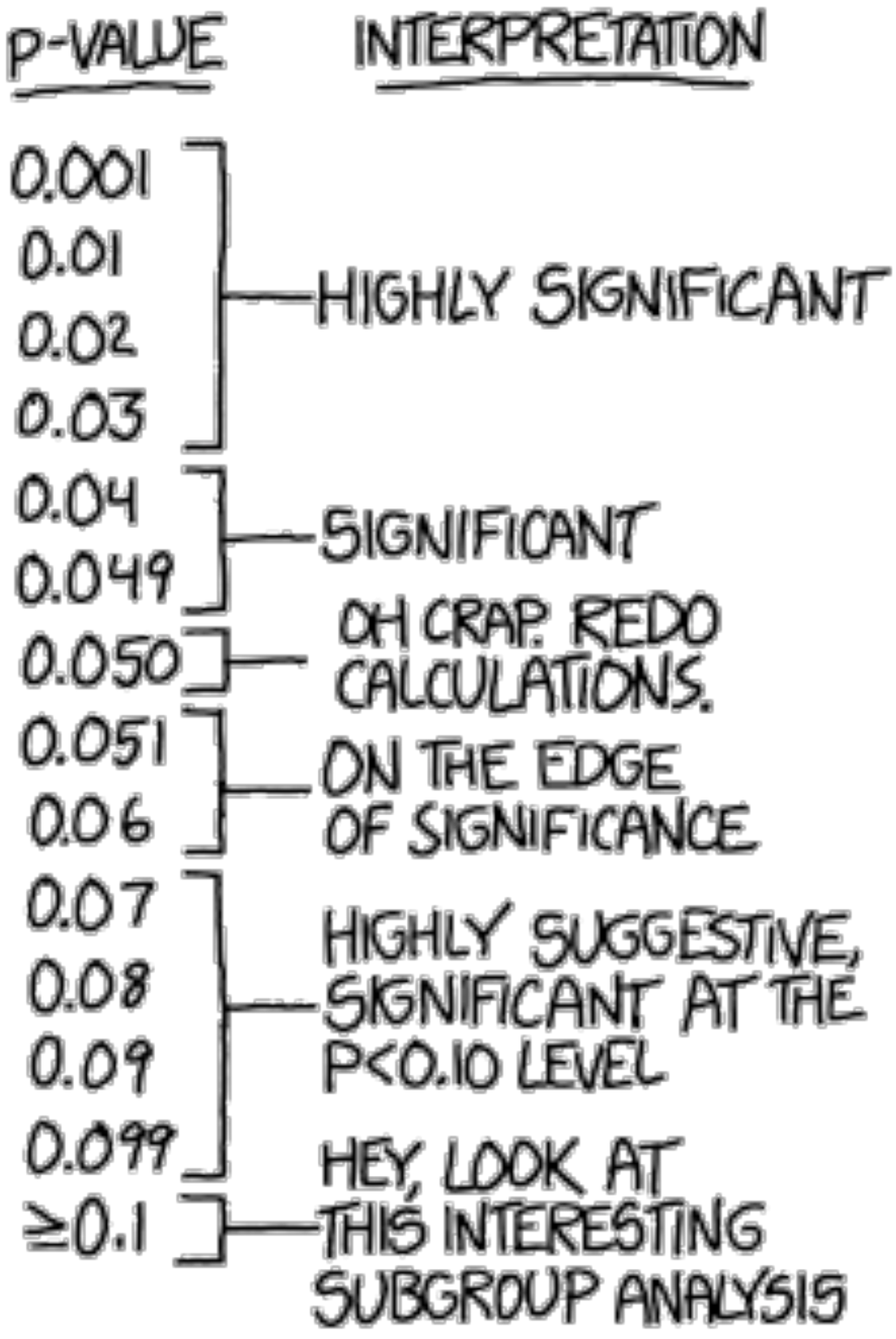
Andrew Q. Philips
Texas A&M University
Feb. 2, 2017
IPSA-USP Summer School 2017

- Definitions
- Advantages of p-values
- Disadvantages
- What else to use?
- Conclusion

https://xkcd.com/882/

Given 20 independent tests, a 5% significance level, probability of false positive:

$$1 - (1 - .05)^{20} = 64\%$$


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.099
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1

ASA Feb. 2014 discussion forum (Wasserstein and Lazar 2016):

"why do so many colleges and grad schools teach p = 0.05?"

"because that's still what the scientific community and journal editors use"

"why do so many people still use p = 0.05?"

"because that's what they were taught in college or grad school"

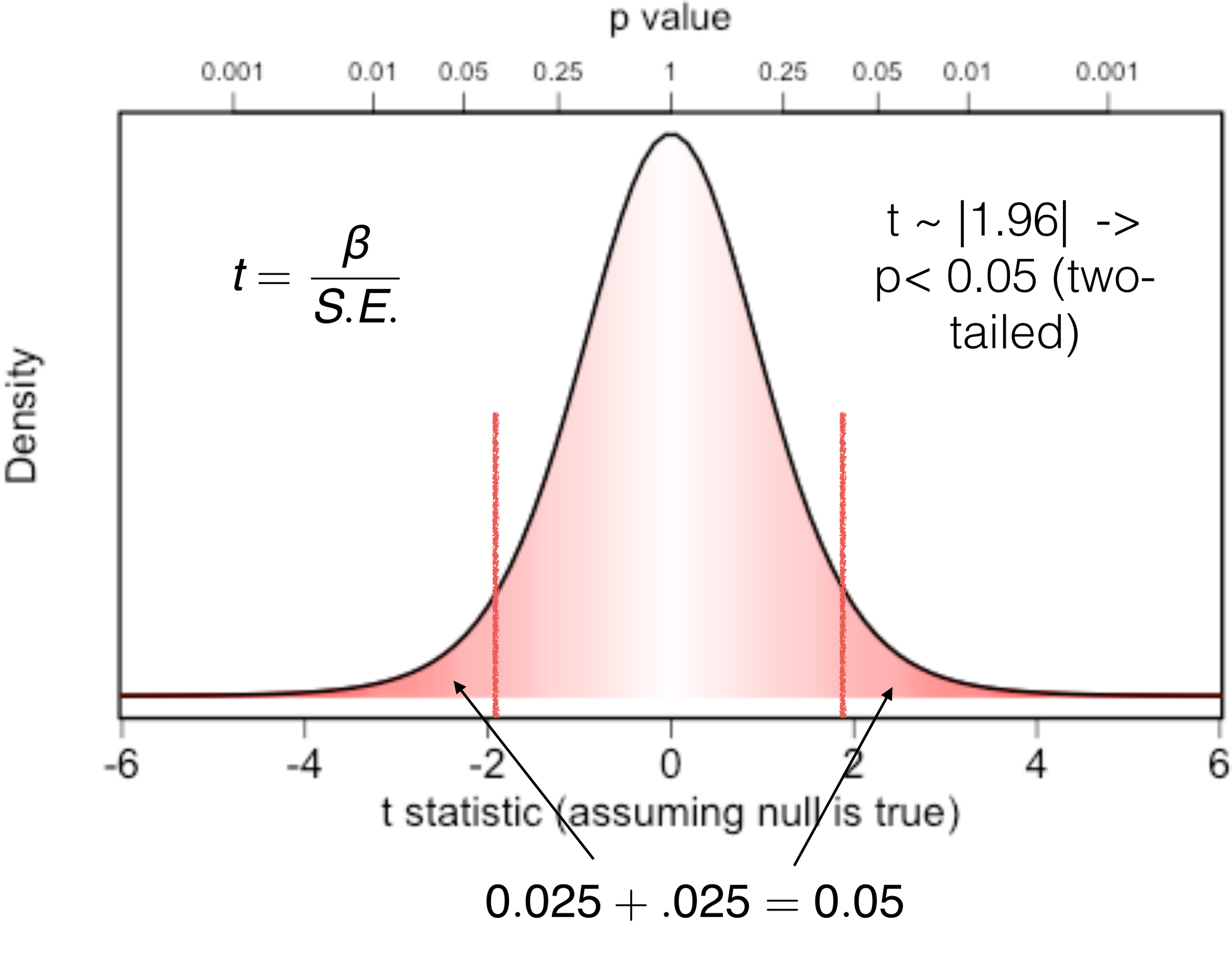
Definition:

"how frequently would I observe a result at least as extreme as the one obtained if Ho were true?" (Jackman)

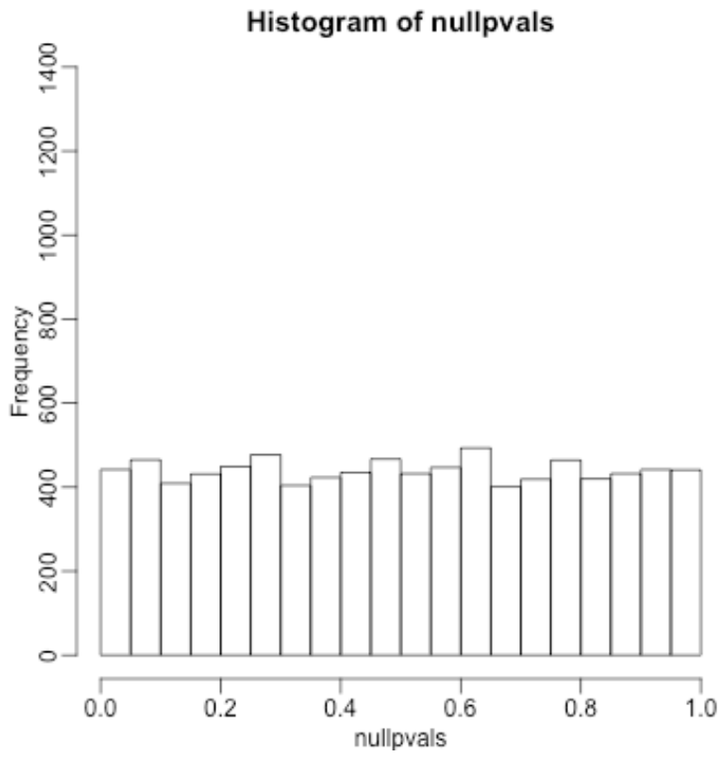
"strength of evidence against the null hypothesis" (Wagenmakers)

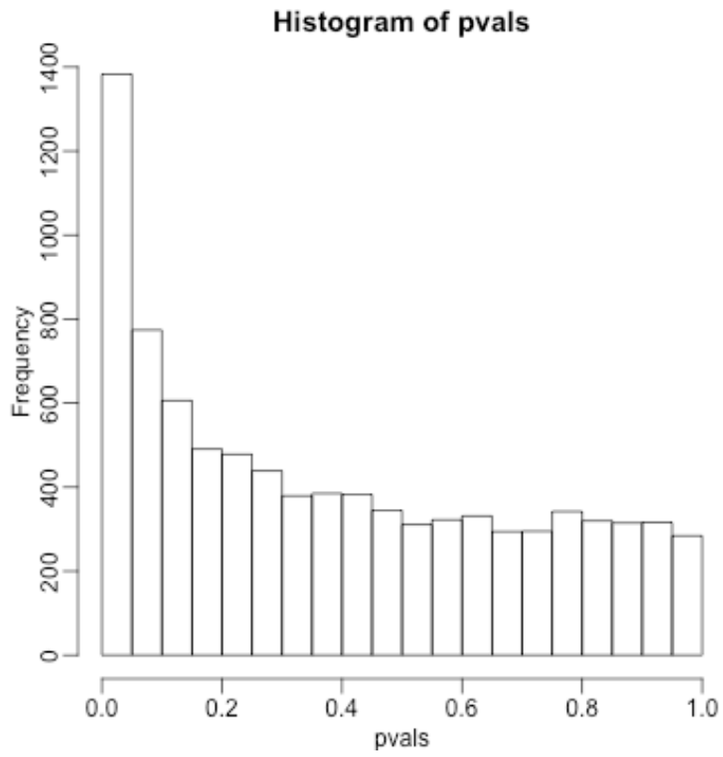
Used to assess statistical significance of a finding

Null-Hypothesis Significance Testing (NHST)


p value
0.001
0.01
0.05
0.25
1
0.25
0.05
0.01
0.001
$t = \frac{\beta}{S.E.}$
t ~ |1.96| -> p< 0.05 (two-tailed)
Density
-6
-4
-2
0
2
4
6
t statistic (assuming null is true)
$0.025 + .025 = 0.05$

If the null hypothesis is true, the probability distribution of p is uniform [0,1]





If the alternative hypothesis is true, the distribution of p depends on sample size and the true value of the parameter of interest

e.g. two-tailed test that 5 flips of a coin (T T T T T) is likely:

$$2 \cdot (\frac{1}{2})^5 = 0.0625$$

History

Ronald Fisher (1920s)…though Pearson and Laplace discussed p-values

Differs from Neyman-Pearson framework (power, Type I, Type II error)

Unlike Fisher, NP approach involves explicitly specifying Ha

Advantages of p-value

Only need to specify null hypothesis (i.e. proposed model used to summarize incompatibility with the data)

$$H_0 : \beta = 0$$

Smaller p-values correspond with greater incompatibility between the (null) model and the data.

Evidence against the null hypothesis

p-values can be looked up using relevant t/z statistics

Disadvantages

p-values do not tell us whether the null hypothesis (or the alternative) is true

p-values do not tell us the probability that random chance produced the data observed

0.05 threshold is not a dichotomous threshold between “true” effects and “false” effects.

“p-hacking” leads to faulty scientific progress (large increases in Type I error)

“My p-value is 0.01…phew; there’s only a 1% chance that the results I’m seeing are not real”

We never know the odds that the effect existed in the first place….the “plausibility of the hypothesis”

“my p-value is 0.04…the alternative hypothesis is true and the null hypothesis is false”

We never know if the null hypothesis (of no effect) is true or false.

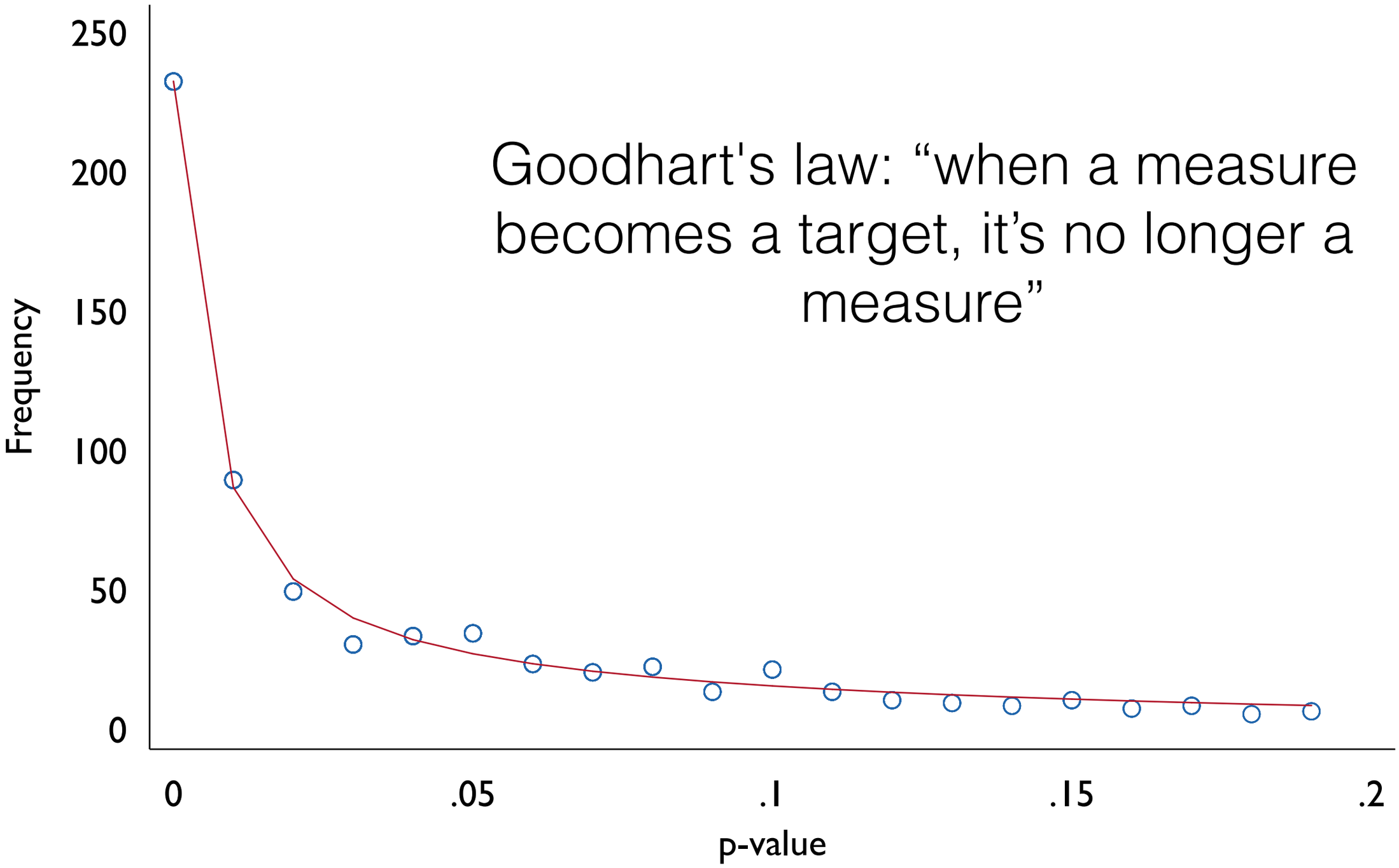
The p-value simply provides the probability that the data are unlikely to have been generated if the null was true, given the data we’re seeing.

“When I include x1 in the model, its p-value is 0.05…but z’s is 0.06. Only x1 is affecting y”

p-value of 0.05 by convention is an arbitrary cut-off point. Z is simply less compatible with the data, given the null of no effect

p-hacking

Evidence of Publication Bias in the PBC Literature



Goodhart's law: “when a measure becomes a target, it’s no longer a measure”

Data from Philips (2016). 622 study-model obs.

Solutions?

Alternatives to p-values?

Basic and Applied Social Psychology bans p-values

ASA statement on statistical significance and p-values

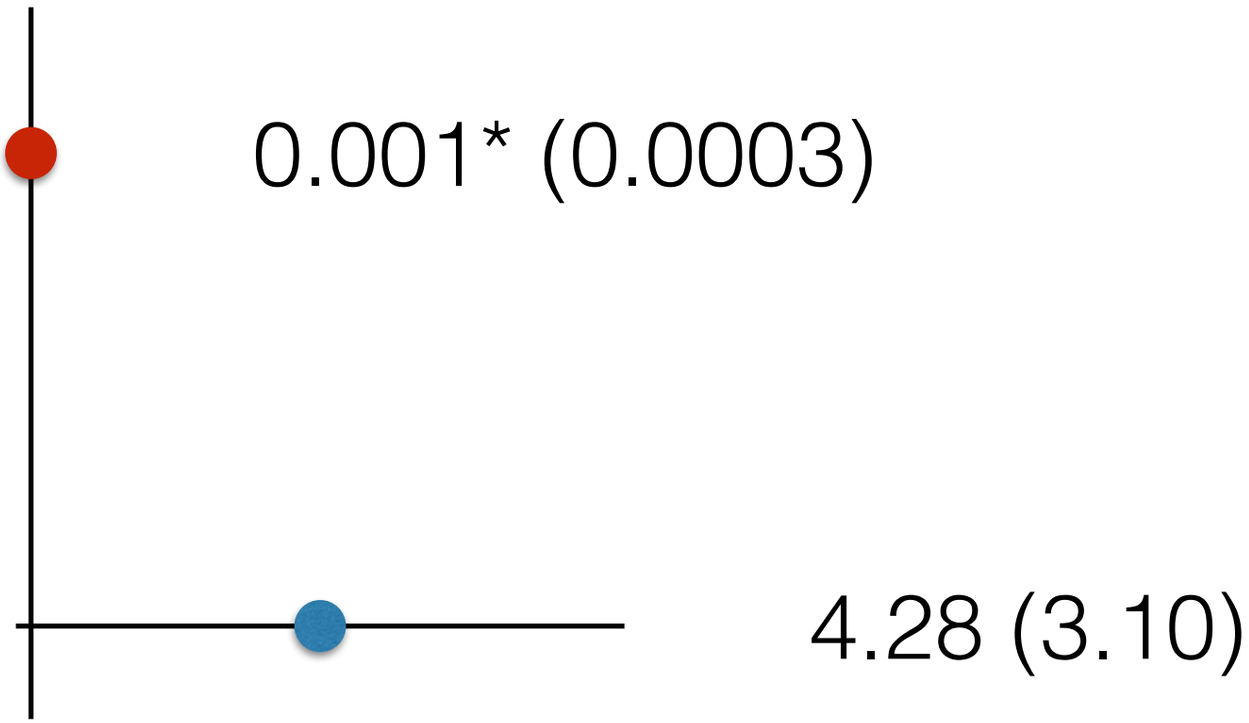
Pre-specification, clear methods, data access and transparency, robustness…

Substantive Significance

p-values say nothing about the substantive effect

As sample size increases, test power goes to 1

Which effect matters more?



Confidence intervals, predicted/expected values, substantive quantities of interest probably better test the substantive results

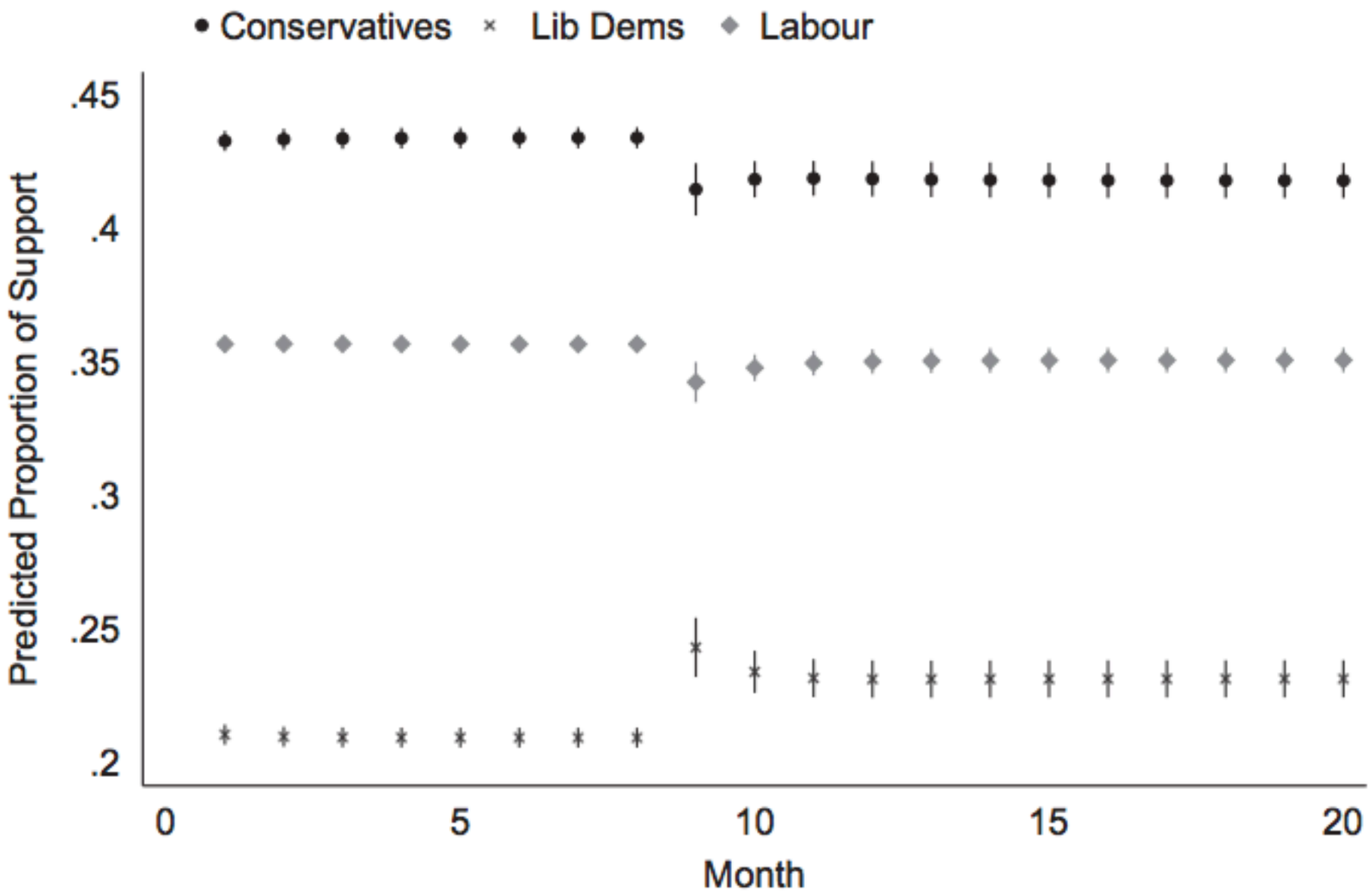
Confidence Interval

In repeated samples, we would expect the true value of the coefficient to lie within this interval “x”% of the time

Less sharp cutoff, more substantive feel

Bayesian: 95% posterior intervals

FIGURE 3 Dynamic Simulation of an Increase in the Average Evaluation of the Liberal Democratic Leader



Note: The 95% confidence intervals are shown.

(Philips, Rutherford, and Whitten 2016)

Others

Likelihood ratios

How much more likely are the data generated from model M1 vs. model M2?

fully Bayesian

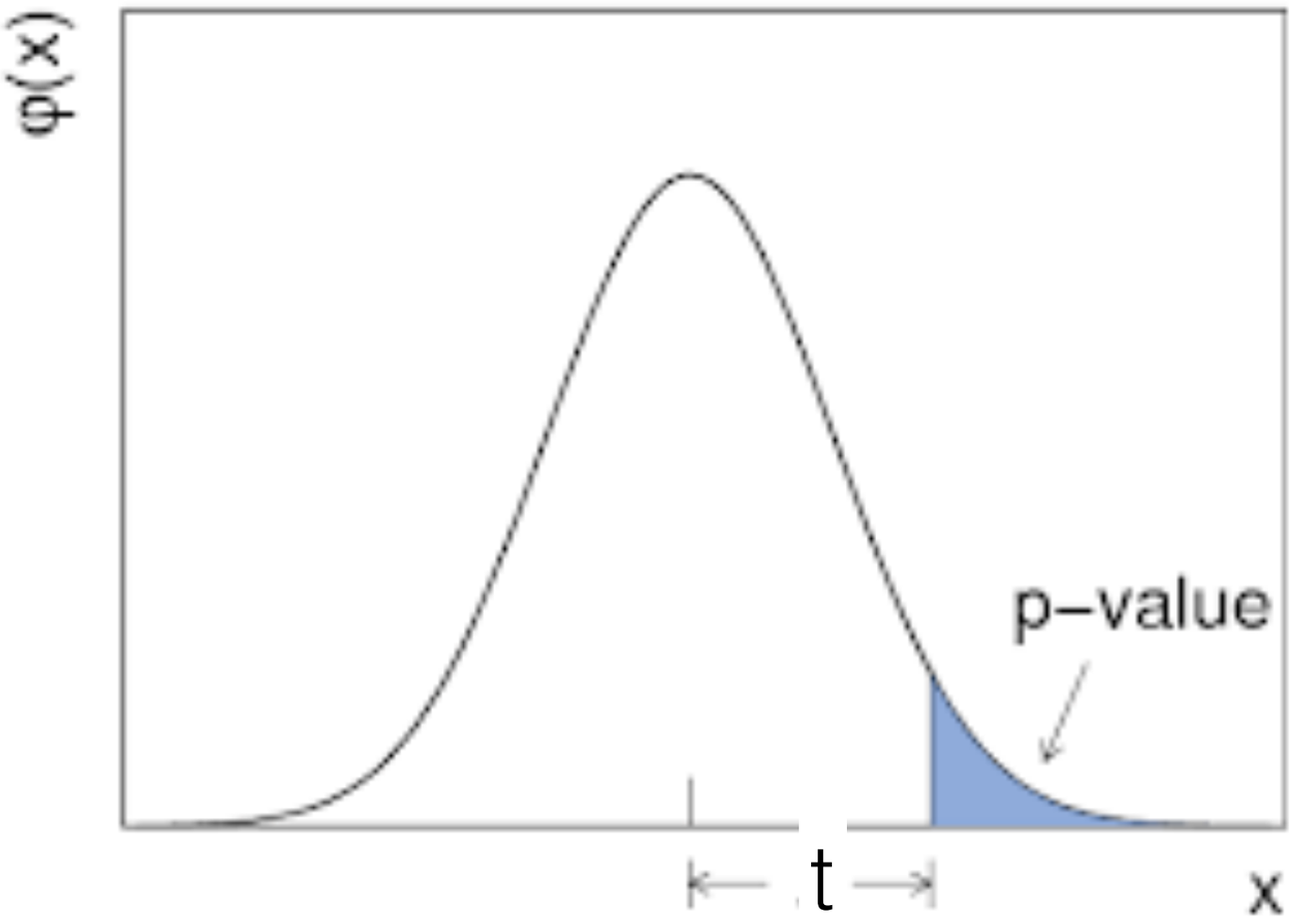
Bootstrapping (~Bayesian with uninformative priors)

Bayes factors

Relative odds of the null hypothesis vs. the alternative

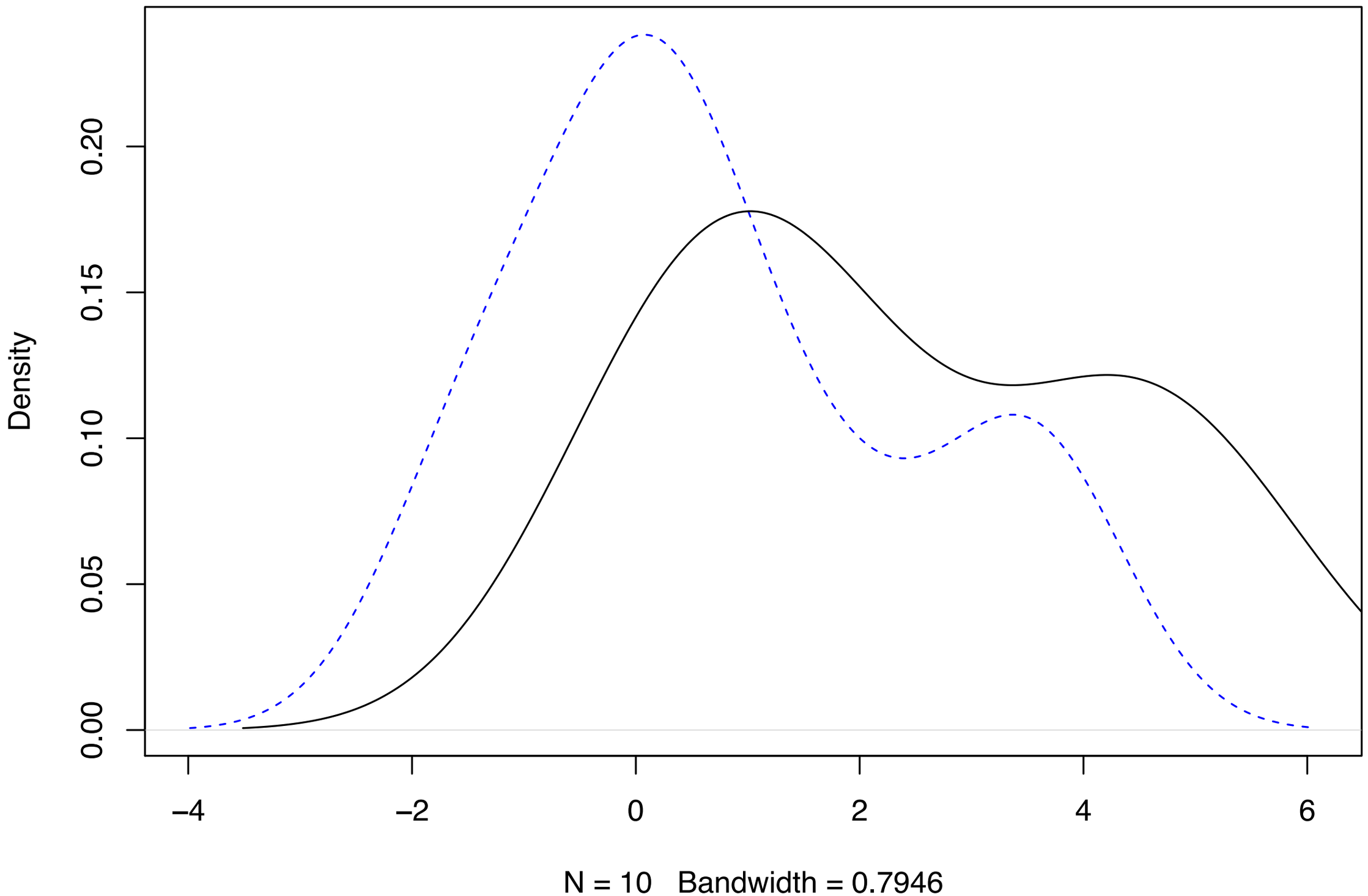
Bayes Factors

Does a patient's sleep improve before vs. after taking a drug?

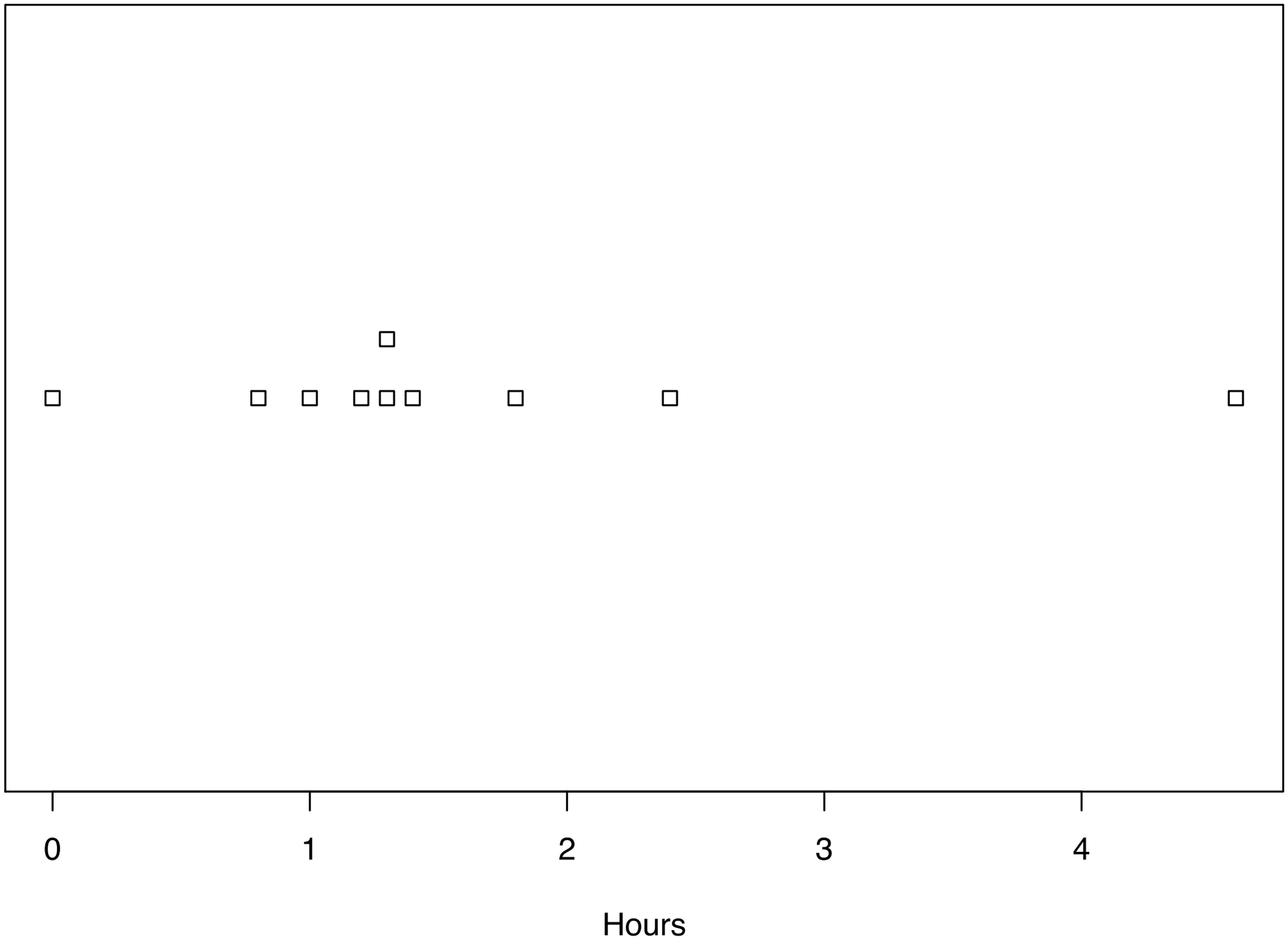


One-sample t-test

t = 4.0621, df = 9, p-value = 0.002833 (2-sided)

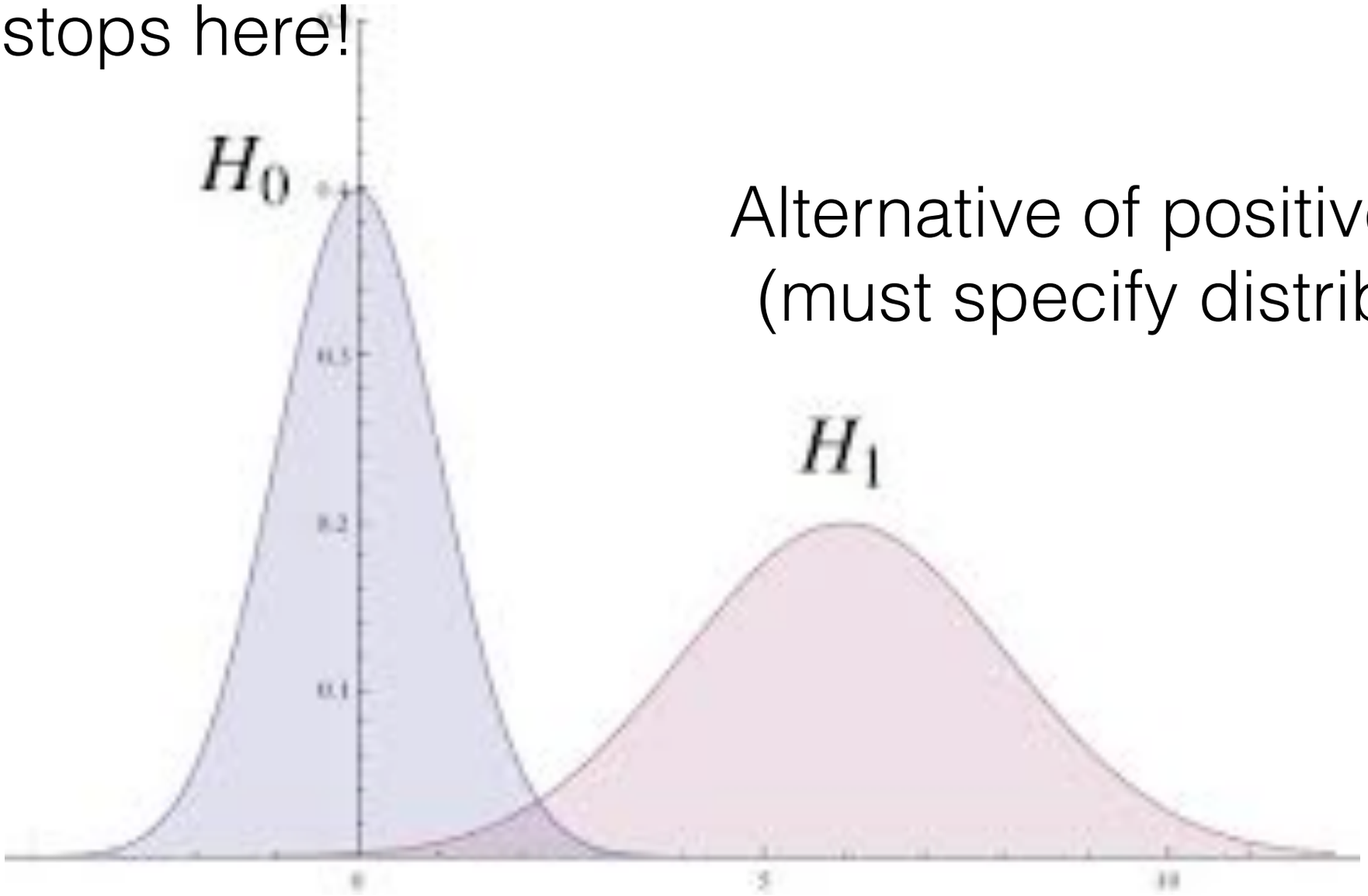

Does A Drug Increase Sleep?
Density
0.00
0.05
0.10
0.15
0.20
−4
−2
0
2
4
6
N = 10 Bandwidth = 0.7946

Increase in Patient's Sleep after Receiving Drug



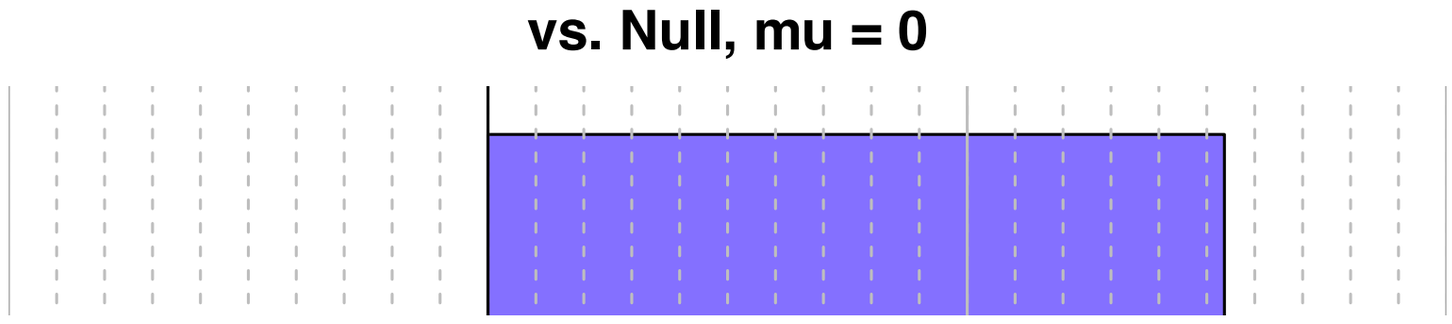
Null of no effect of drug on sleep

p-value stops here!

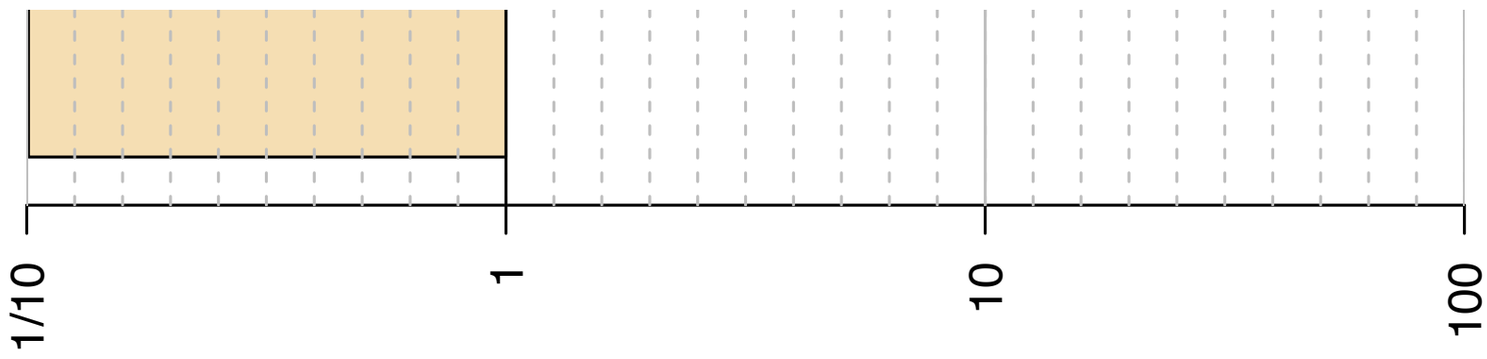


Alternative of positive effect (must specify distribution)

BF: Is the data (relatively) more consistent with Ha than Ho?



2 ln *K*	*K*	Strength of evidence
0 to 2	1 to 3	not worth more than a bare mention
2 to 6	3 to 20	positive
6 to 10	20 to 150	strong
>10	>150	very strong



Conclusions

p-values are not going anywhere

Useful, but often misinterpreted

Use in conjunction with other approaches