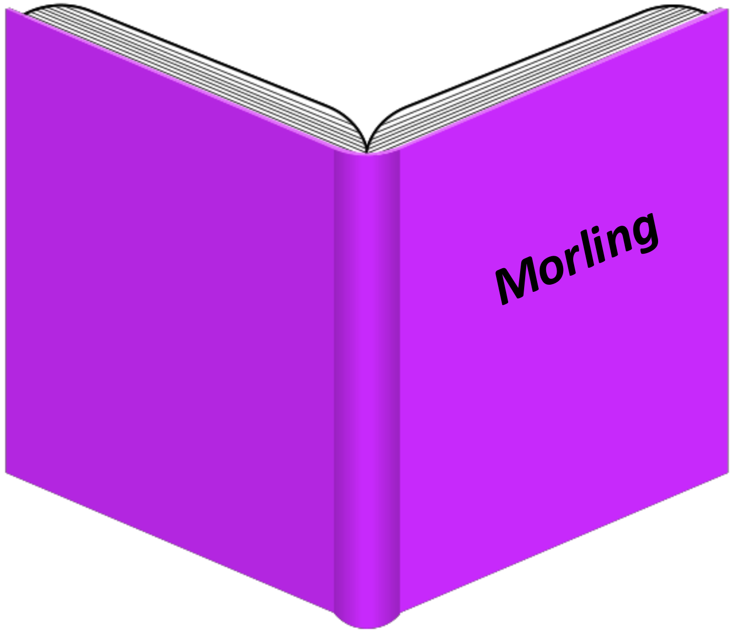
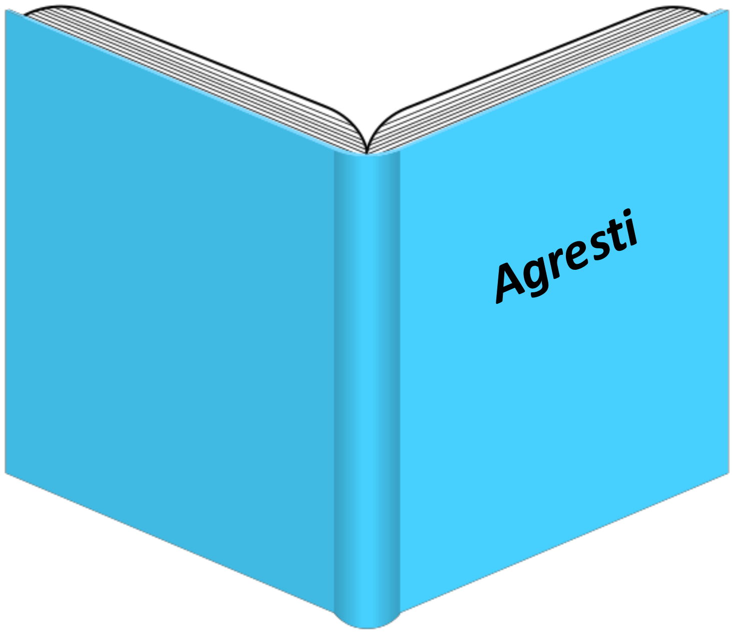




Research Methods and Statistics

Lecture 24: More Bayes…

Johnny van Doorn

Pictures source: pixabay

Today

- **Recap of last week**
- Bayesian Inference
  - Correlation
  - T-test
- Bayes & Frequentism
- Statistics in the Wild
- Recap
  - Example exam question

Bayesian Inference

$$\underbrace{p(\theta \mid \text{data})}_{\text{Posterior beliefs about the world}} = \underbrace{p(\theta)}_{\text{Prior beliefs about the world}} \times \underbrace{\frac{p(\text{data} \mid \theta)}{p(\text{data})}}_{\text{Predictive updating factor}}$$

This is on the level of the ***parameter***, *within a single model*

We compare how well each specific value of θ in the model predicted the data, compared to the other values of θ

Bayesian Hypothesis Testing

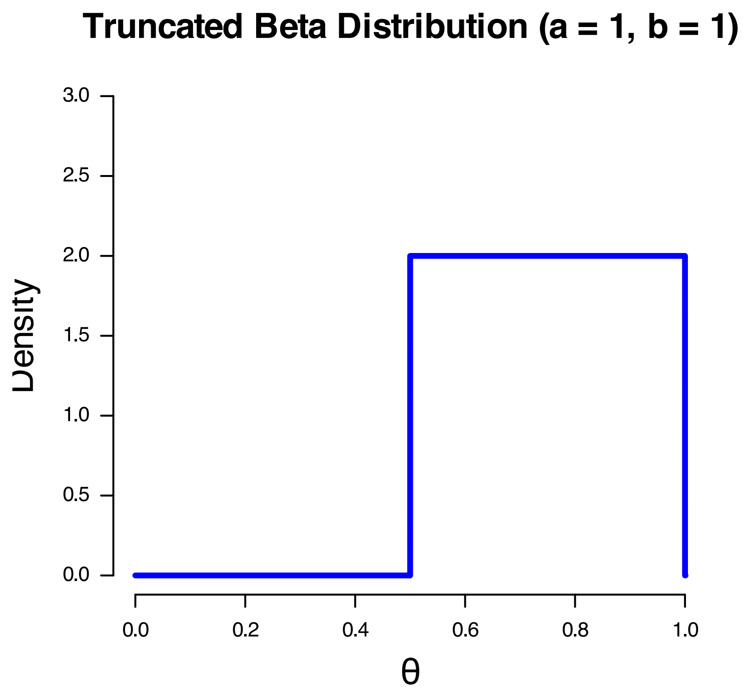
$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{Posterior beliefs} \\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{Prior beliefs} \\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\substack{\text{Predictive} \\ \text{updating factor}}}$$

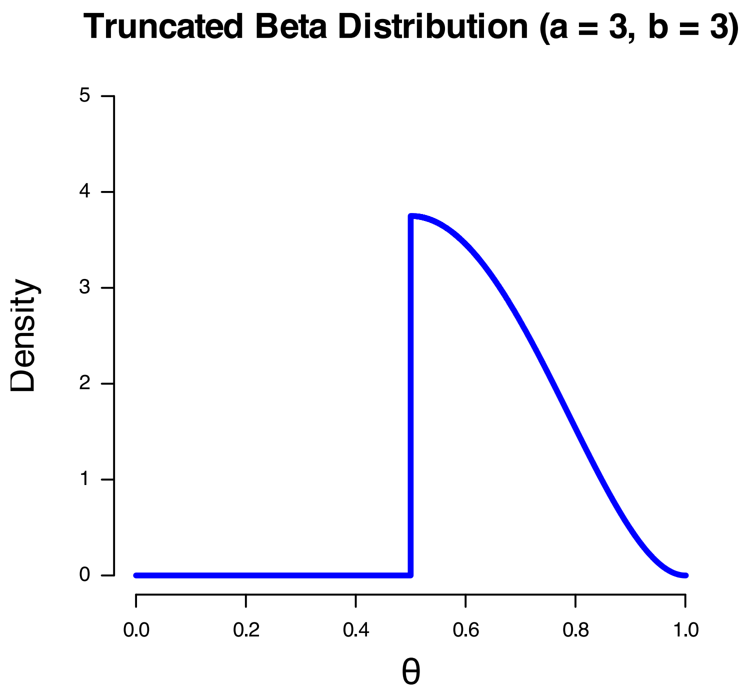
This is on the level of the ***hypothesis***

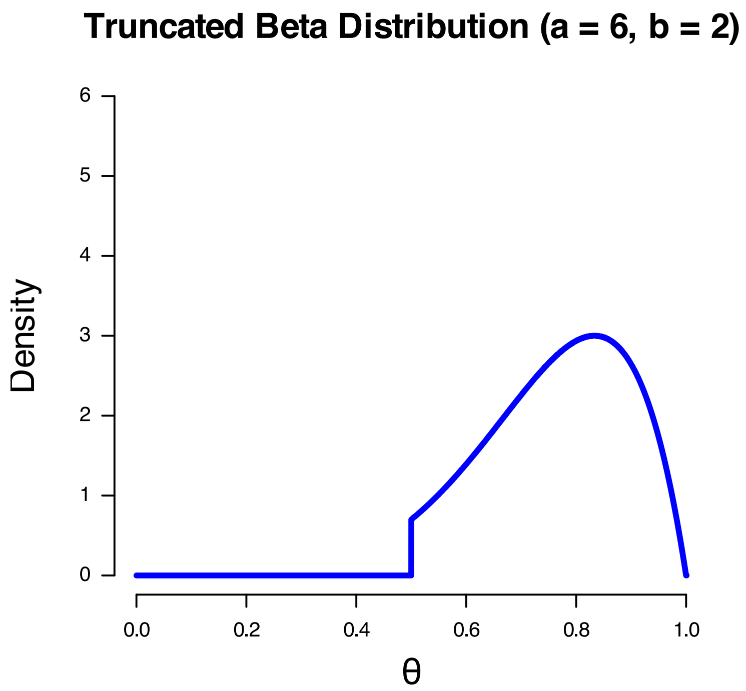
We compare how well the hypotheses predicted the data: we look at how well all values postulated by each hypothesis predicted the data

What Does the Alternative Hypothesis Predict?

"If you had 100€ to bet on likely values of θ, how would you divide it?"







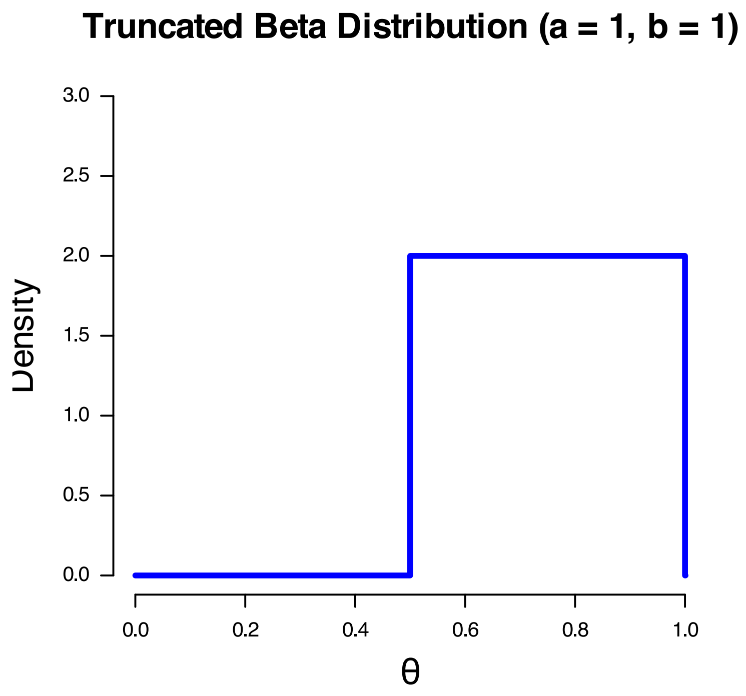
Here, it predicts that only values of θ between 0.5 and 1 are possible, and that all values of θ between 0.5 and 1 are equally likely.

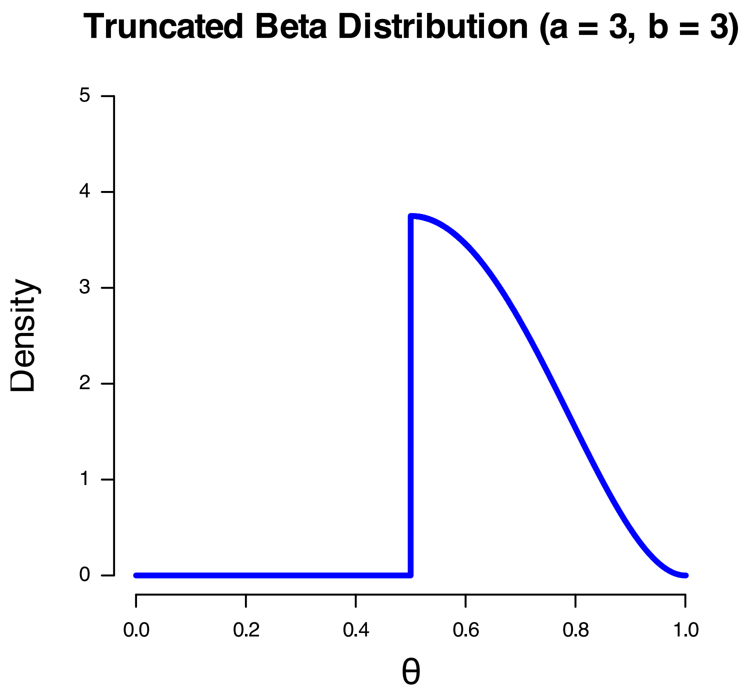
Here, it predicts that only values of θ between 0.5 and 1 are possible, and that values of θ closer to 0.5 are more likely than values close to 1.

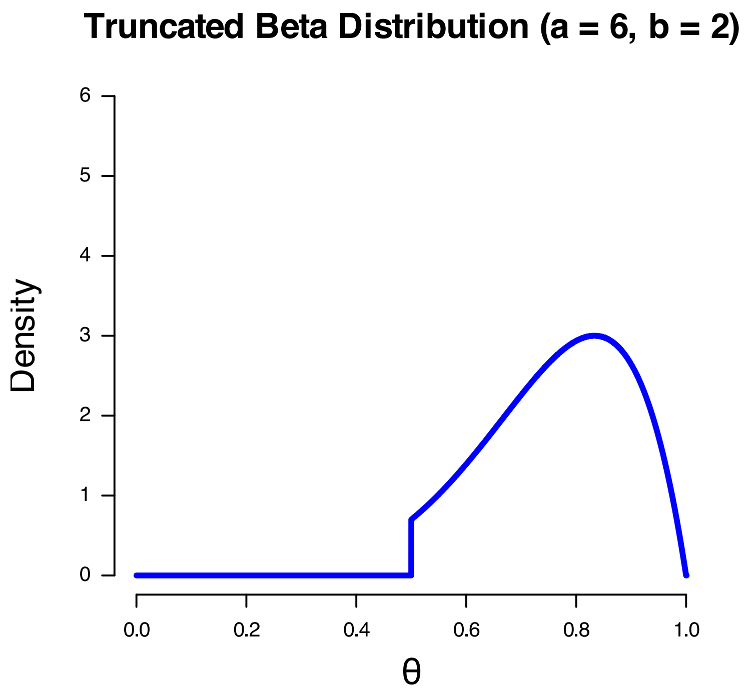
Here, it predicts that all values of θ between 0.5 and 1 are possible, and that values of θ closer to 0.5 are more likely than values close to 1.

What Does the Alternative Hypothesis Predict?

"If you had 100€ to bet on likely values of θ, how would you divide it?"







If the observed proportion (i.e., the data) equals 0.55, it will win **as much** money as if the observed proportion equals 0.8

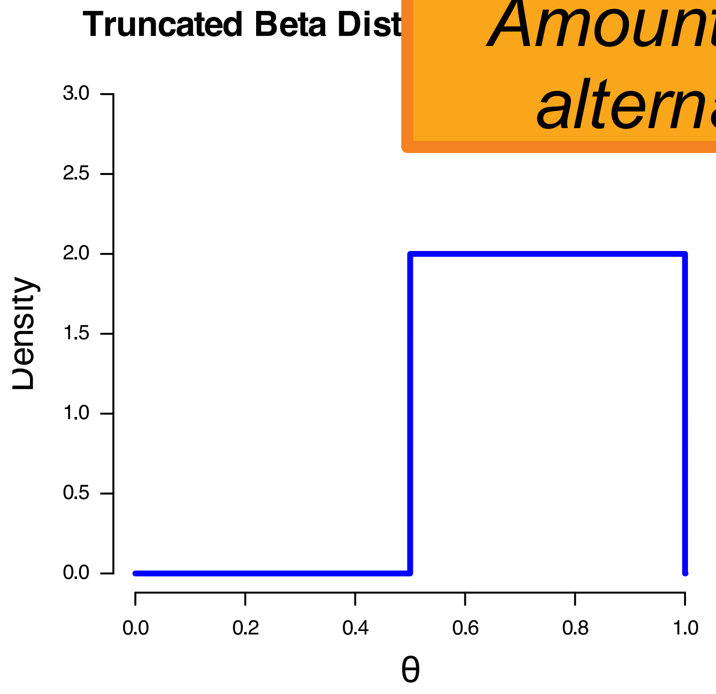
If the observed proportion (i.e., the data) equals 0.55, it will win **more** money than if the observed proportion equals 0.8

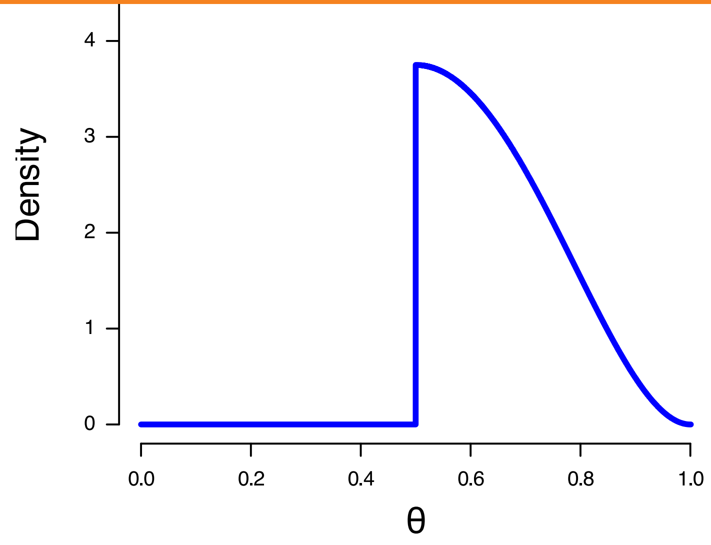
If the observed proportion (i.e., the data) equals 0.55, it will win **less** money than if the observed proportion equals 0.8

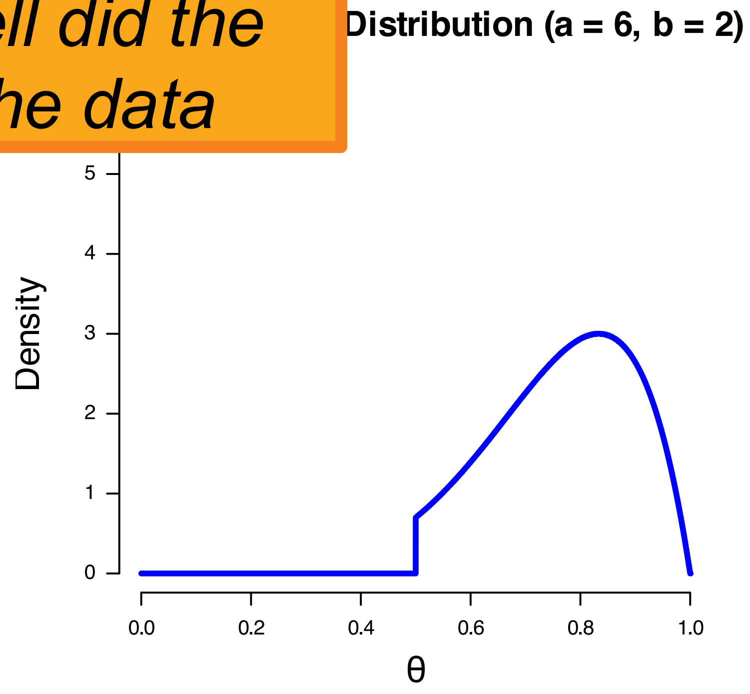
What Does the Alternative Hypothesis Predict?

“If you had 100€ to bet on likely values of θ, how would you divide it?”

Amount of money won = how well did the alternative hypothesis predict the data







If the observed proportion (i.e., the data) equals 0.55, it will win as much money as if the observed proportion equals 0.8

If the observed proportion (i.e., the data) equals 0.55, it will win more money than if the observed proportion equals 0.8

If the observed proportion (i.e., the data) equals 0.55, it will win less money than if the observed proportion equals 0.8

Bayesian Hypothesis Testing: Bayes Factor

The average likelihood across all values predicted by H_1
(i.e., the marginal likelihood)

$$\underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\substack{\text{Predictive} \\ \text{updating factor}}}$$

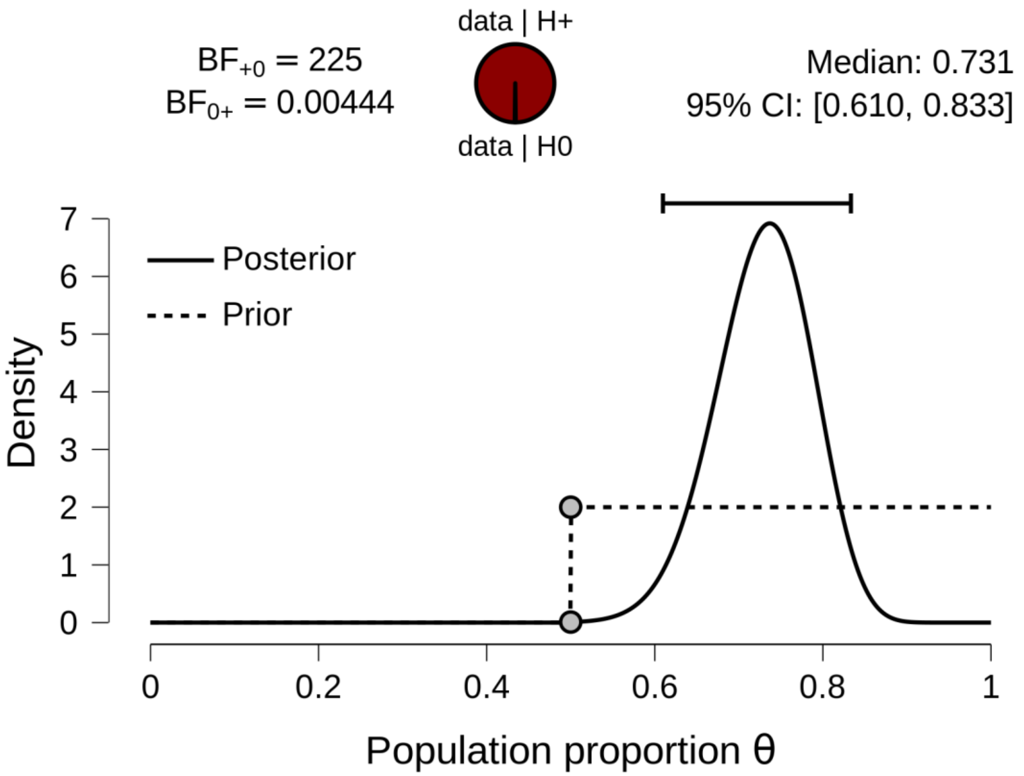
The average likelihood across all values predicted by H_0 (i.e., the likelihood of the data, given the test value (in this case, θ = 0.5)

A computational shortcut for calculating this factor is the Savage-Dickey density ratio:
- Take the prior density at the point of testing (θ = 0.5)
- Take the posterior density at the point of testing (θ = 0.5)

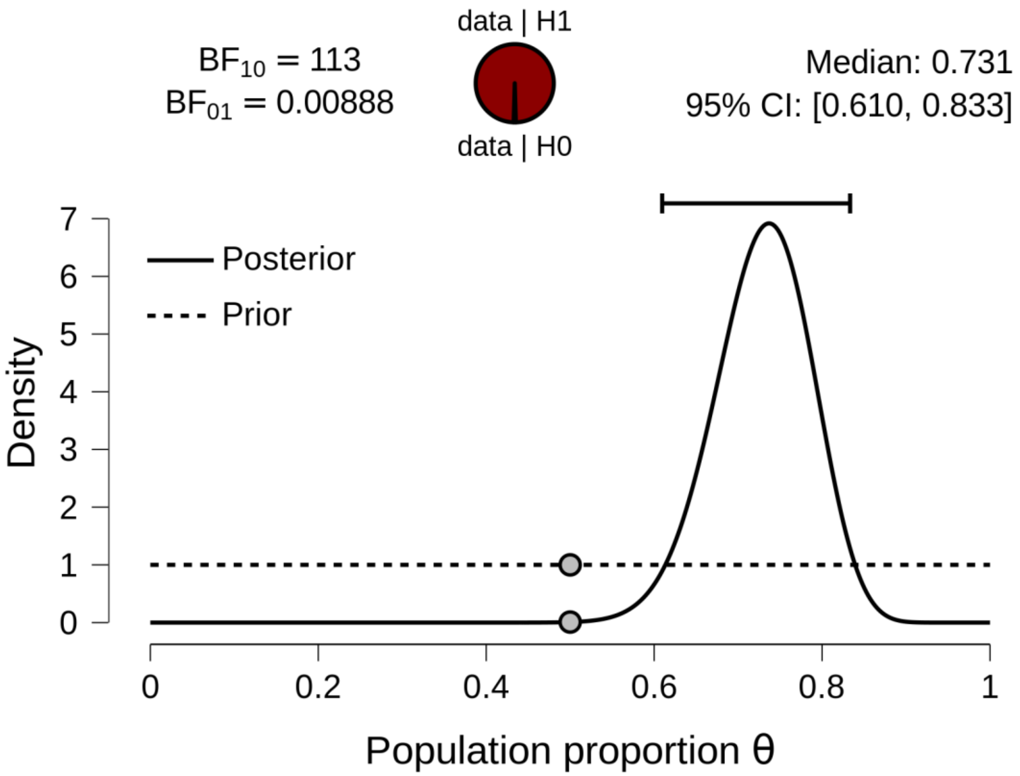
The Bayes factor equals the ratio of those numbers!

Two-Sided Testing Vs. One-Sided Testing

Parsimony: More specific predictions, when still accurate, get rewarded!



$$\mathcal{H}_1 : \theta > 0.5$$



$$\mathcal{H}_1 : \theta \neq 0.5$$

Posterior distribution is beta distribution with

a = prior a + successes

b = prior b + failures

The only calculations you need to know

Bayes factor is ratio of marginal likelihoods

BF_{10} = P(data | H_1) / P(data | H_0)

BF_{01} = P(data | H_0) / P(data | H_1)

Today

- Recap of last week
- **Bayesian Inference**
  - Correlation
  - T-test
- Bayes & Frequentism
- Statistics in the Wild
- Recap
  - Example exam question

Choosing the Prior Distribution

- Before the experiment / looking at the data (so there are no calculations)

- Informed by previous experiments/knowledge or kept uninformative
  (e.g., uniform prior on the population proportion)

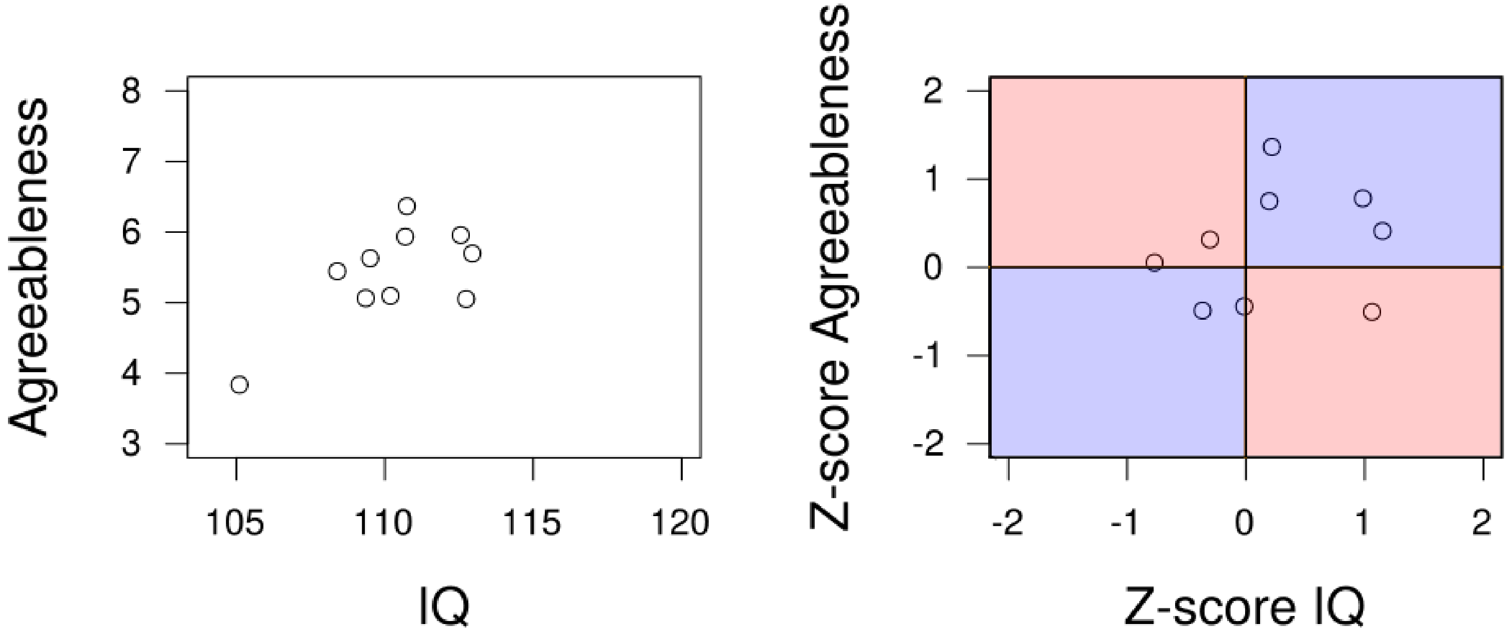
- One-sided / two-sided determined by what you want to test
  (e.g., do you want to know if there is a difference, or a positive difference?)

Choosing the Prior Distribution

- The prior distribution is on the same *domain* as the parameter of interest:
  - For a proportion: [0, 1]
  - For a correlation: [-1, 1]
  - For a difference in means: [-∞, ∞]

Bayesian Correlation

Instead of inference on the population proportion θ, we can estimate/test the **population correlation ρ ("rho")**



Values in the blue quadrant: $Z_x Z_y > 0$
Contribute to *positive* correlation

Values in the red quadrant: $Z_x Z_y < 0$
Contribute to *negative* correlation

Bayesian Correlation

$$P(\rho \mid \text{data}) = P(\rho) \ \frac{P(\text{data} \mid \rho)}{P(\text{data})}$$

Bayesian Correlation: Hypothesis

$$\mathcal{H}_0 : \rho = 0$$

$$\mathcal{H}_1 : \rho \neq 0$$

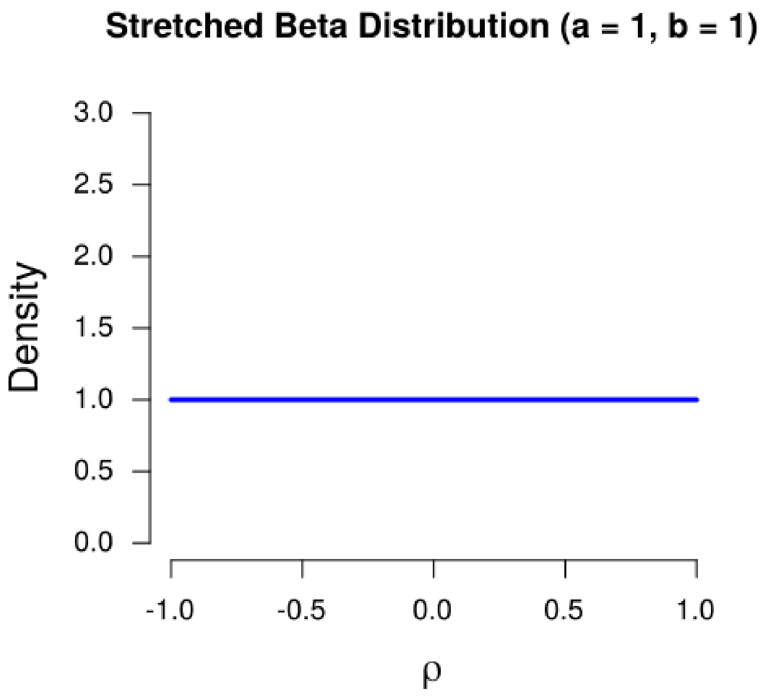
$$\mathcal{H}_+ : \rho > 0$$

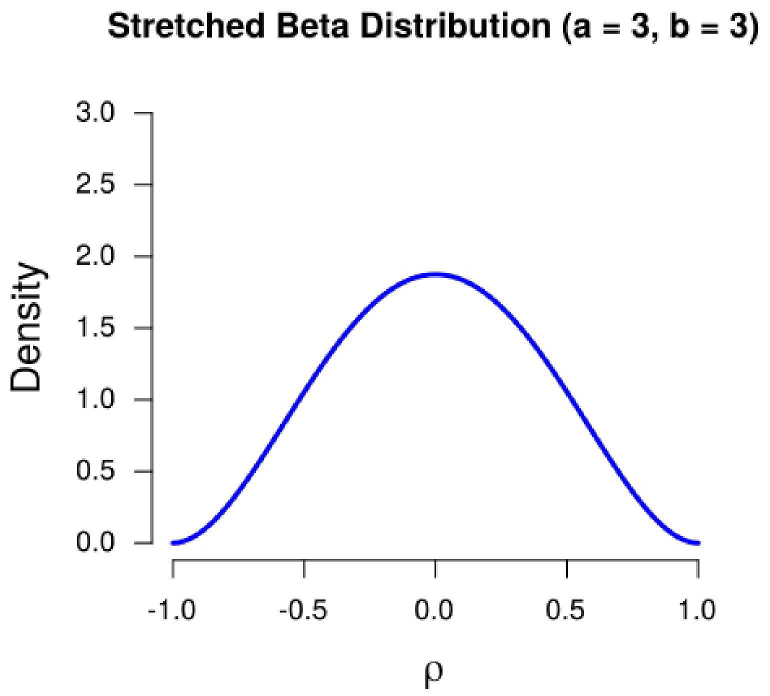
$$\mathcal{H}_- : \rho < 0$$

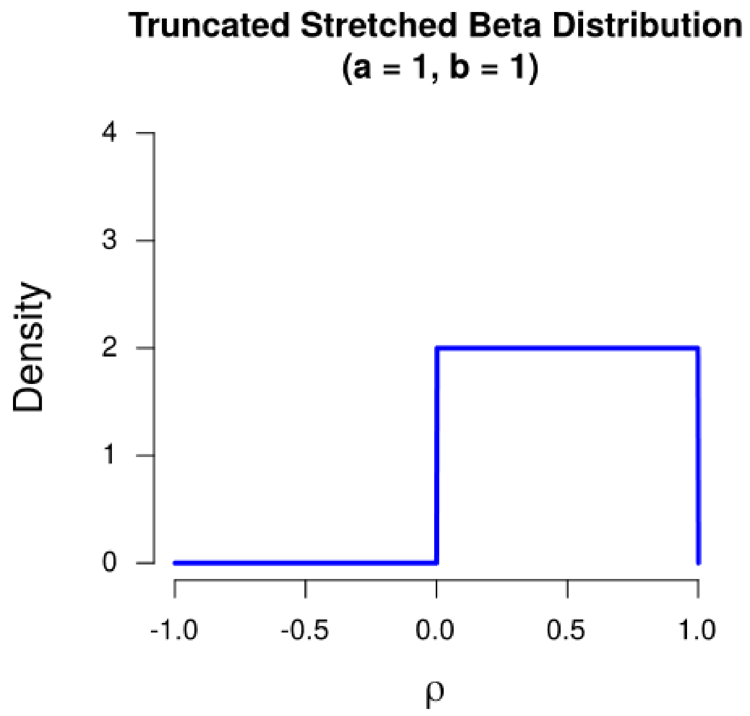
"Alternative Hypothesis"

Bayesian Correlation: Prior Distribution

The prior distribution is on the same domain as the parameter of interest: we can take the **stretched Beta distribution!**
It is the same as Beta, but then stretched to the domain [-1, 1]

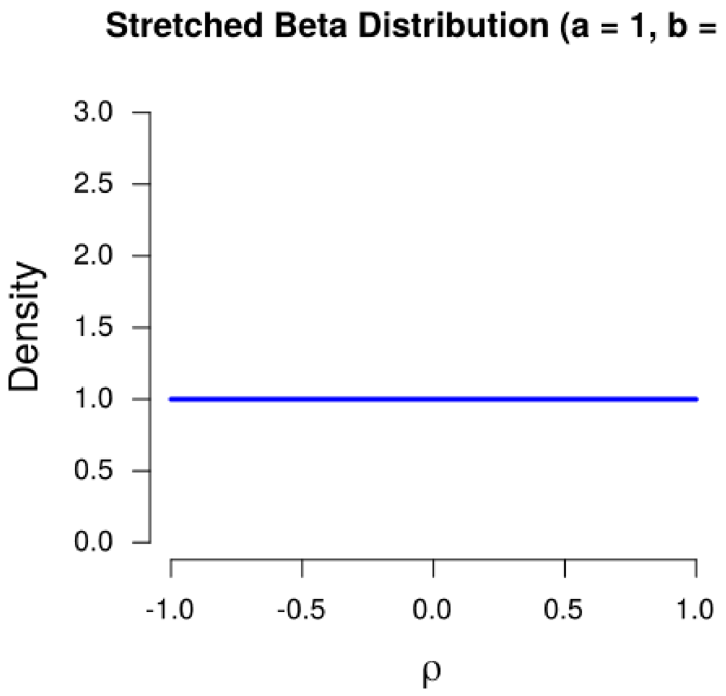






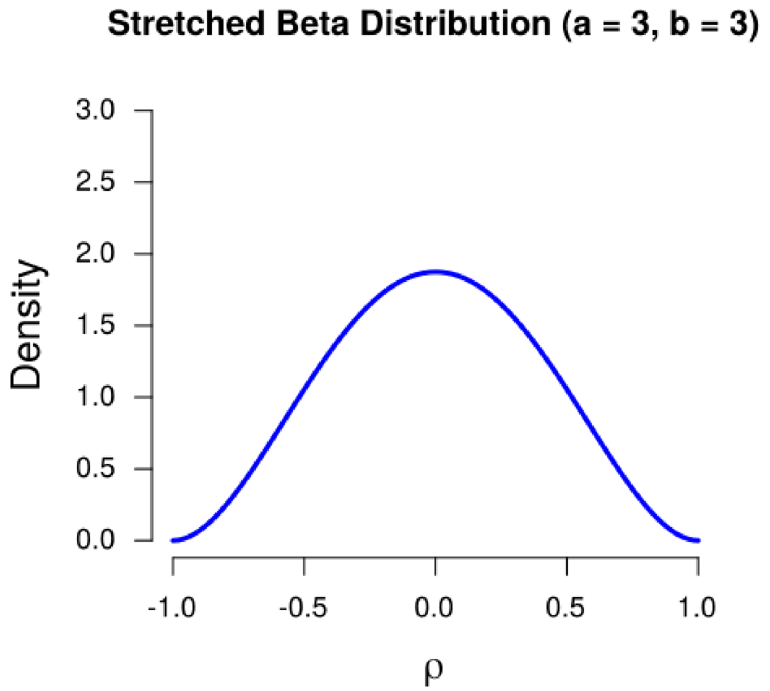
Bayesian Correlation: Prior Distribution

A prior distribution that reflects the belief that all values of ρ are equally plausible, **a priori**, we call this an *uninformative prior*



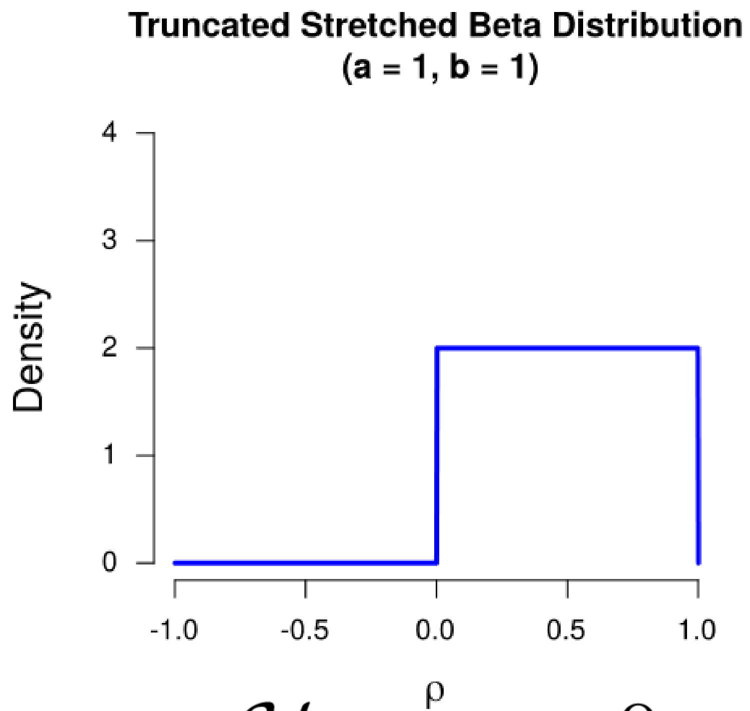
$$\mathcal{H}_1 : \rho \neq 0$$

A prior distribution that reflects the belief that values of ρ close to 0 are more plausible (i.e., the correlation will be around 0 or not be very strong), **a priori**



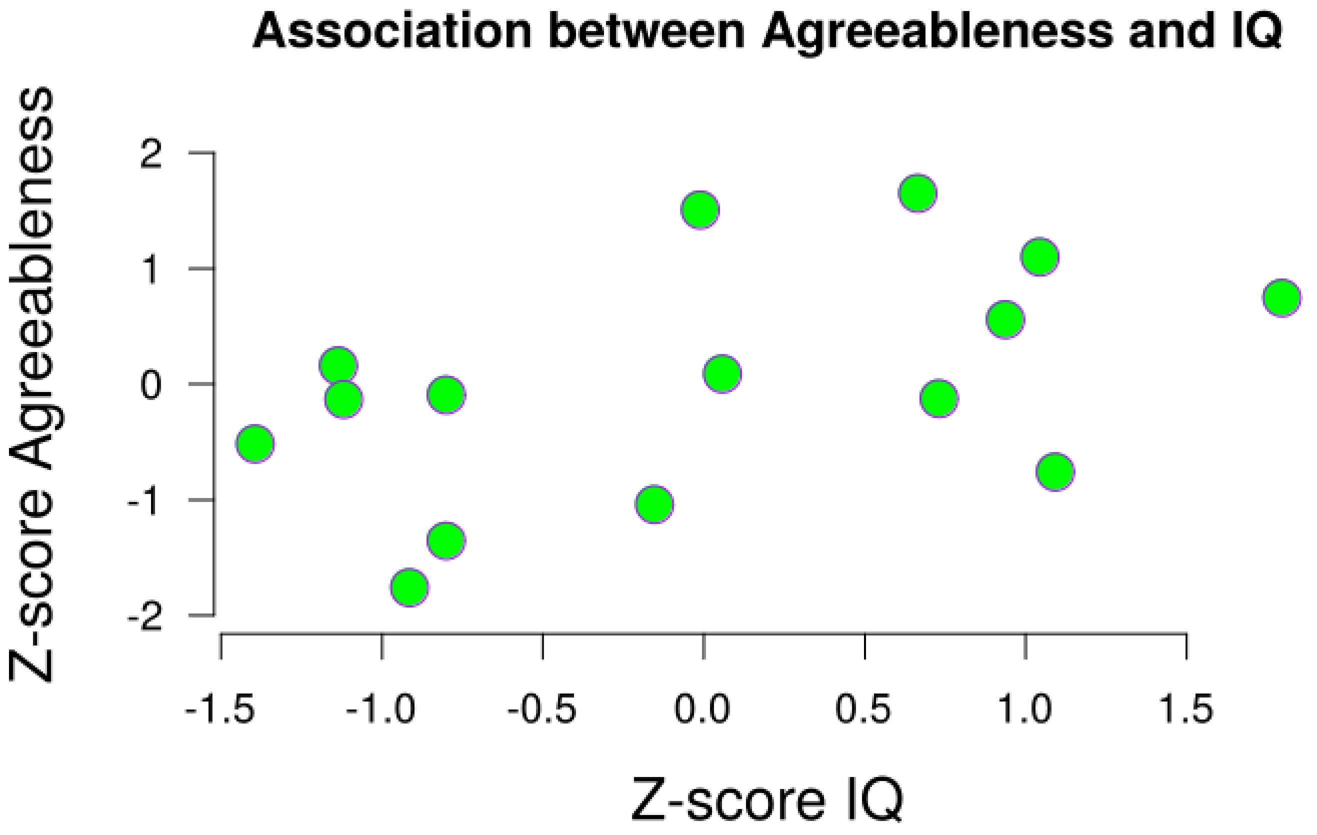
$$\mathcal{H}_1 : \rho \neq 0$$

A prior distribution that reflects the belief that only positive values of ρ are possible and that those values of ρ are equally plausible (i.e., there will be a positive correlation), **a priori**

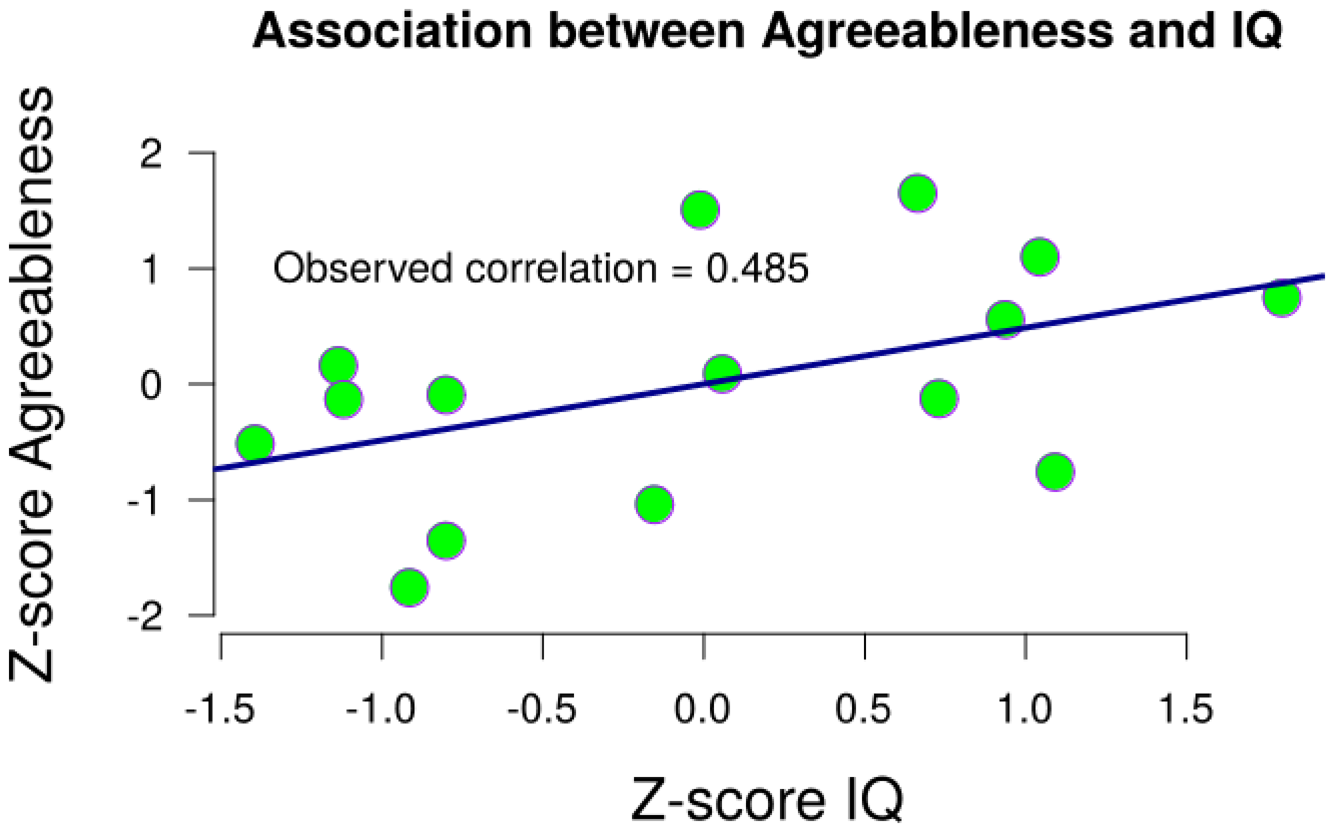


$$\mathcal{H}_+ : \rho > 0$$

Bayesian Correlation: Data

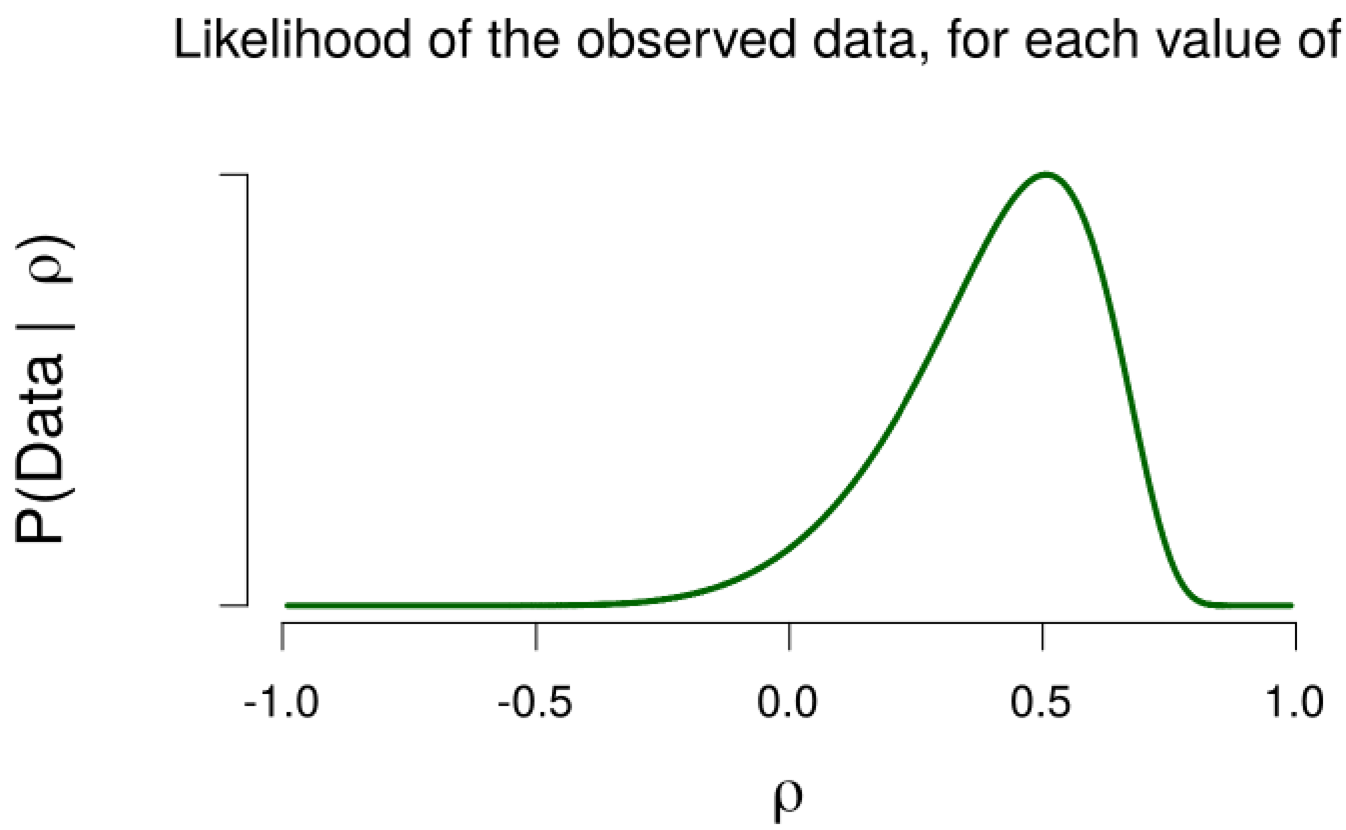


Bayesian Correlation: Data



Bayesian Correlation: Likelihood

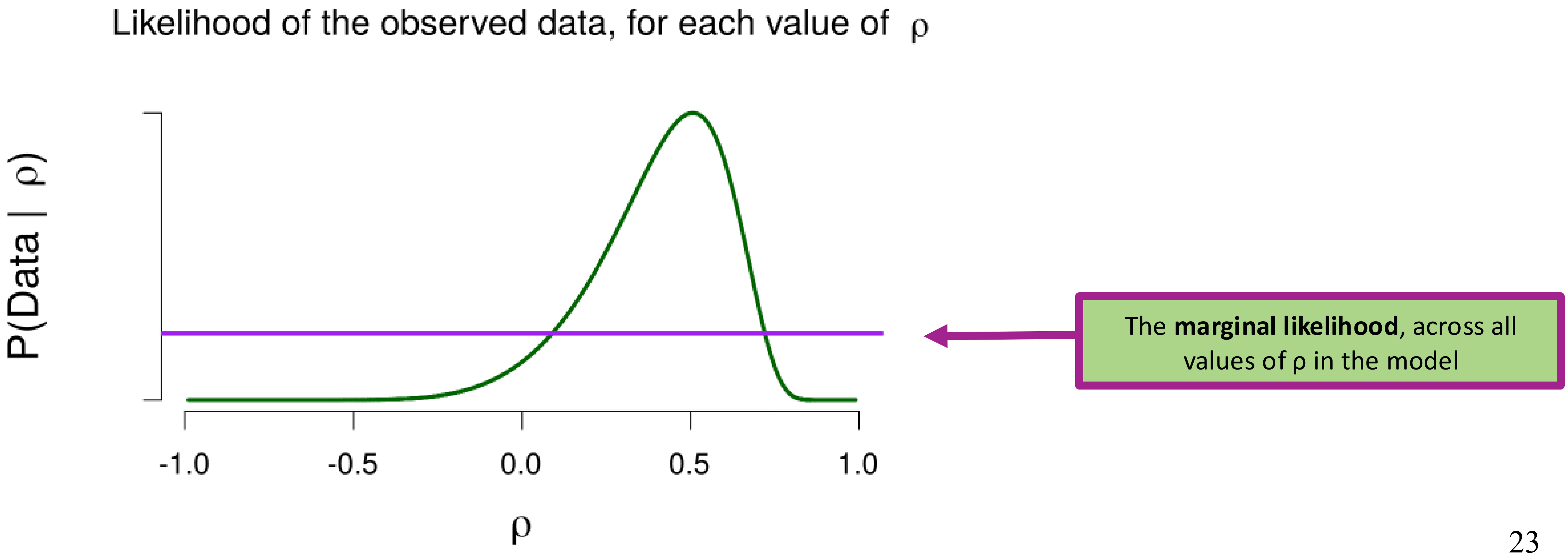
$$P(\text{data} \mid \rho)$$



We see that the data are likely for values of ρ close to 0.5.
This makes sense, because the observed correlation (i.e., the data) is equal to 0.485!

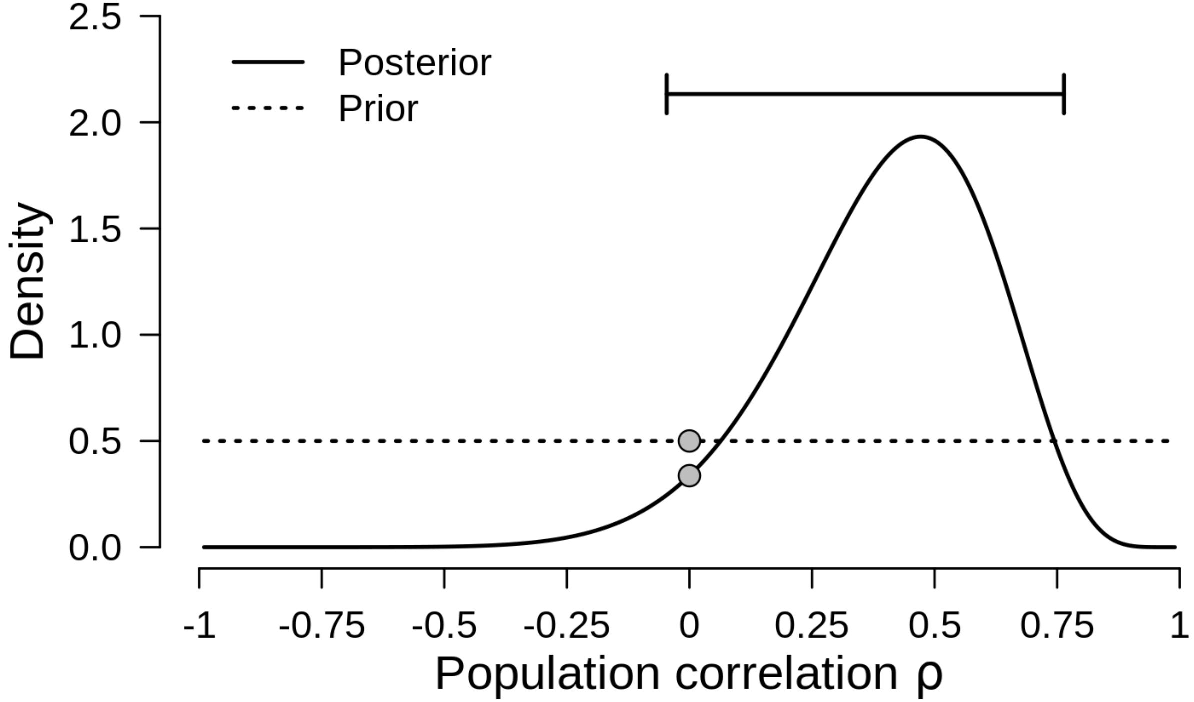
Bayesian Correlation: Likelihood

$$P(\text{data} \mid \rho)$$



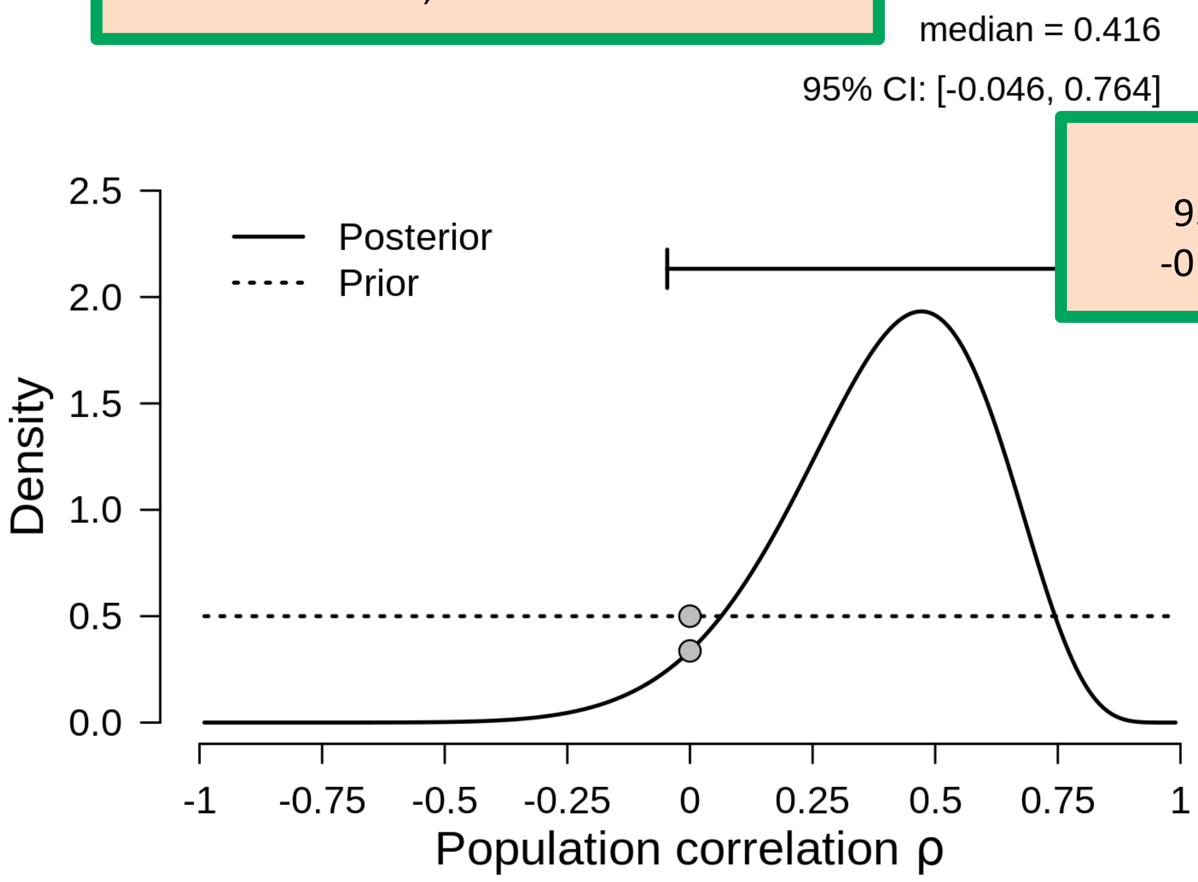
Bayesian Correlation: Posterior Distribution

$$P(\rho \mid \text{data})$$



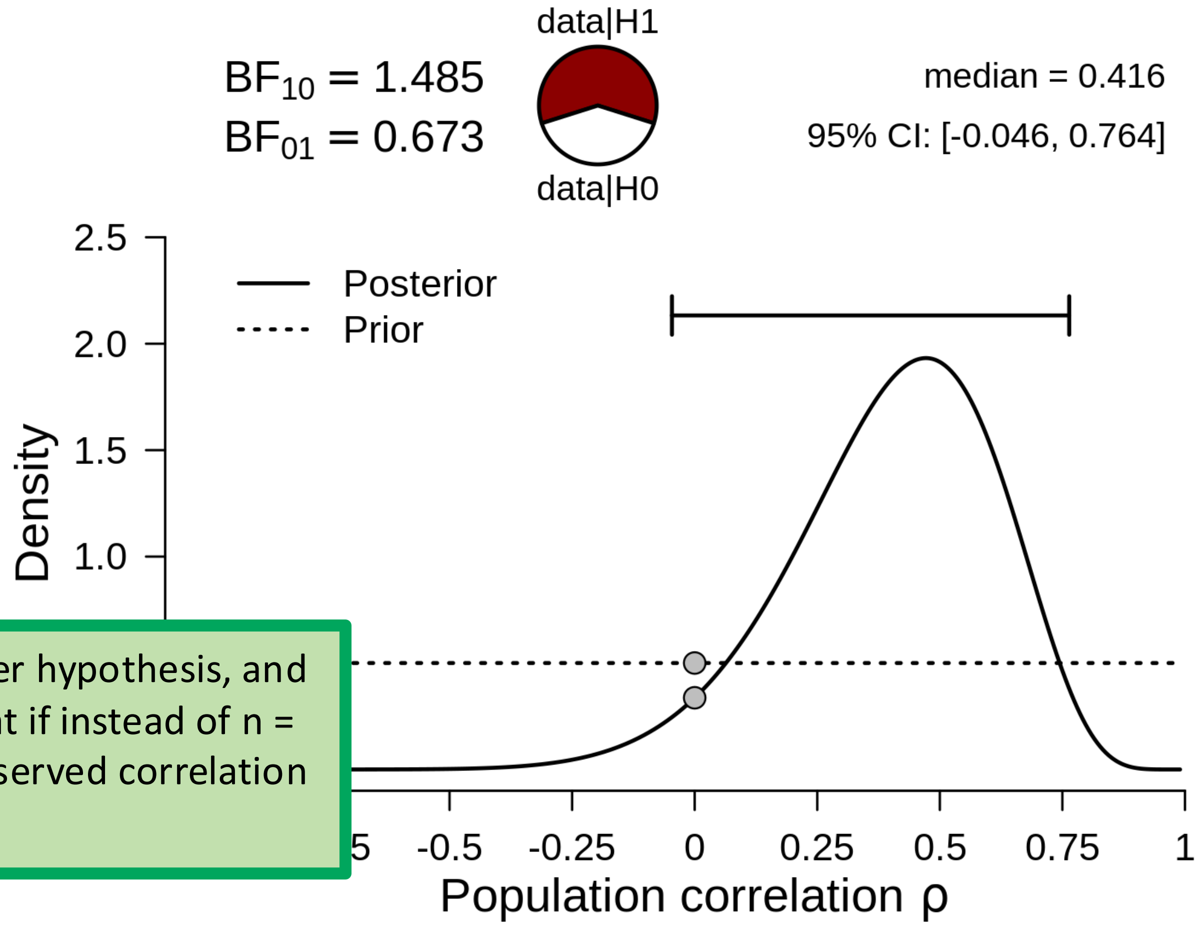
Bayesian Correlation: Posterior Distribution

Median:
50% probability that ρ is equal to or lower than 0.416, under this model



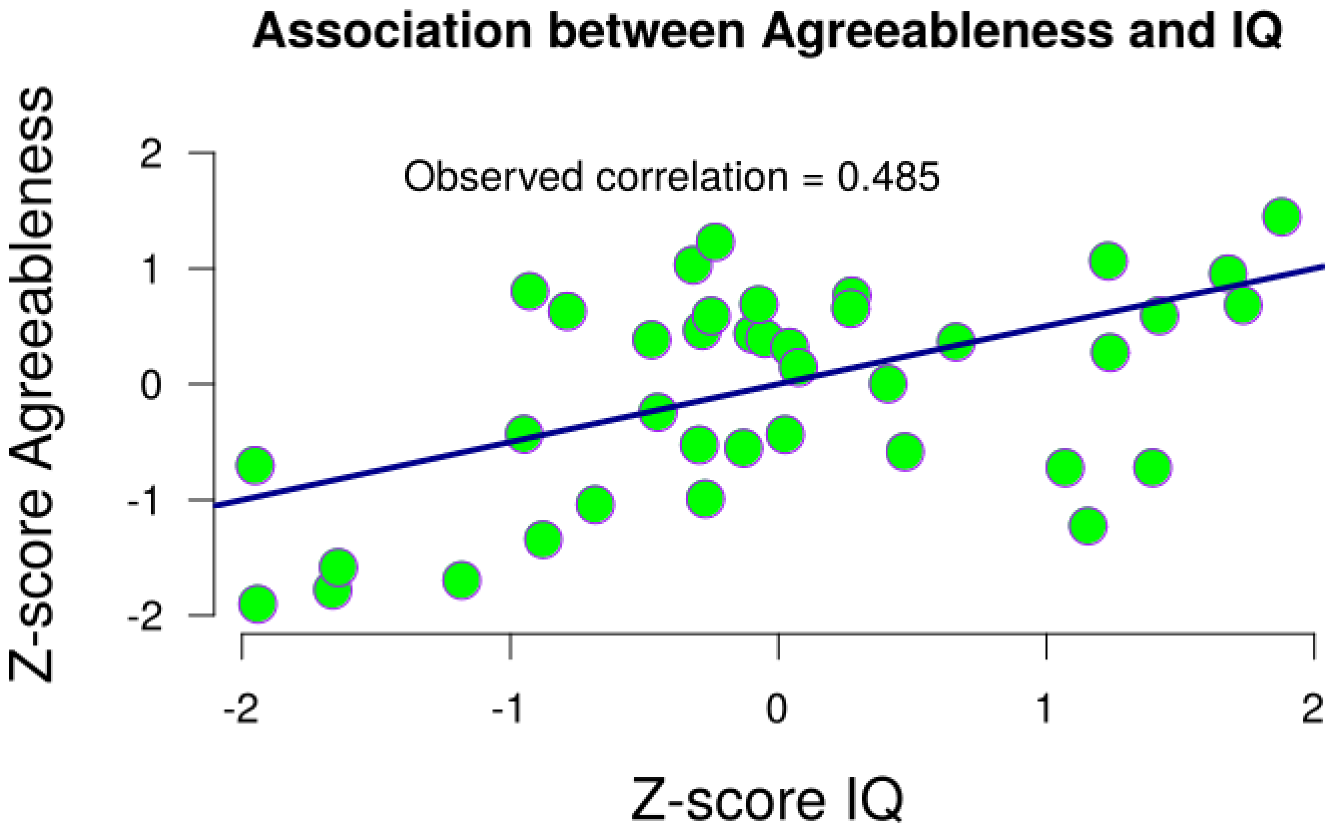
95% Credible interval:
95% probability that ρ is between -0.046 and 0.764, under this model

Bayesian Correlation: Bayes Factor



Very weak evidence in favor of either hypothesis, and a very wide credible interval… What if instead of n = 15, we had n = 40, but the same observed correlation (0.485)?

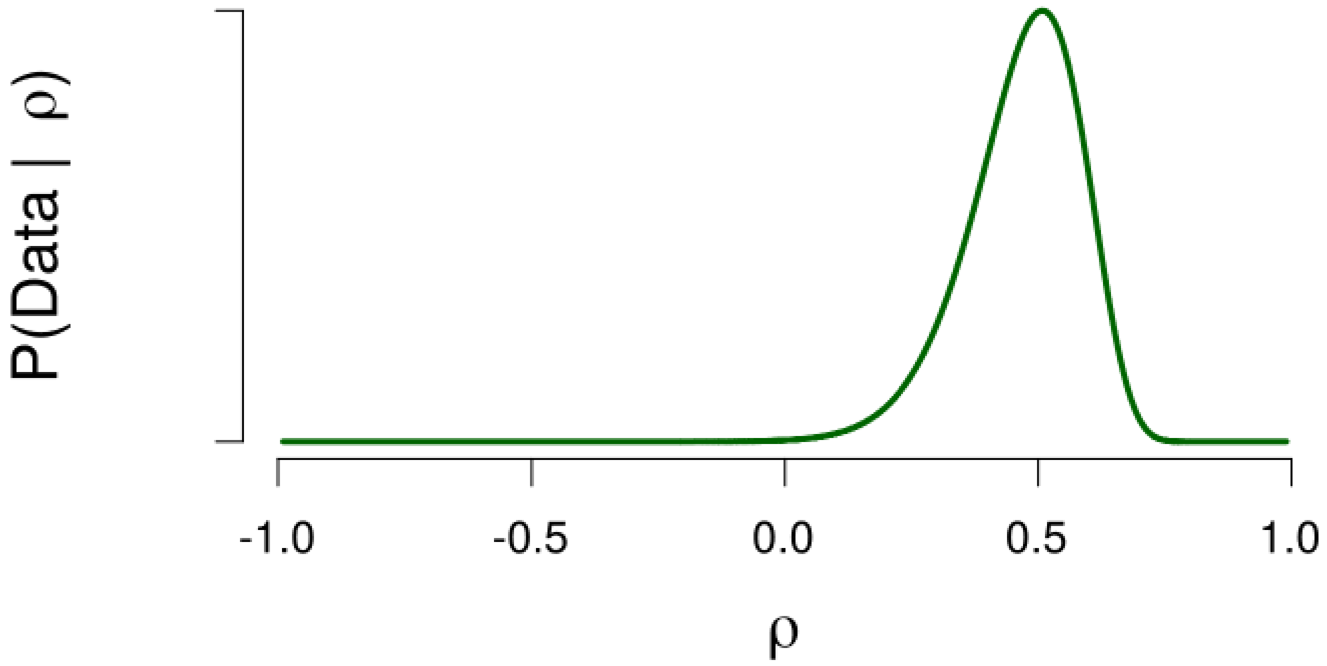
Bayesian Correlation: Data



Bayesian Correlation: Likelihood

$$P(\text{data} \mid \rho)$$

Likelihood of the observed data, for each value of ρ



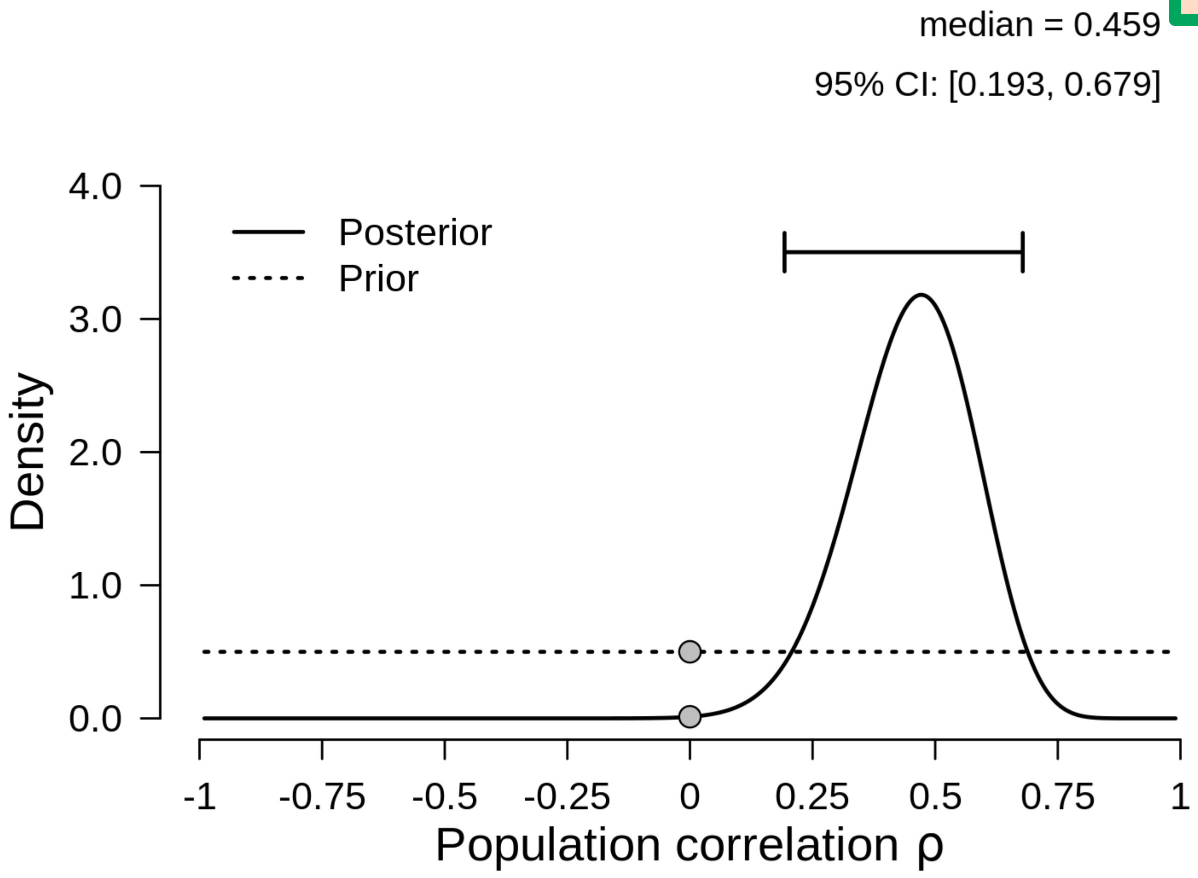
The likelihood grew more narrow, compared to when we had a lower sample size!

Bayesian Correlation: Posterior Distribution

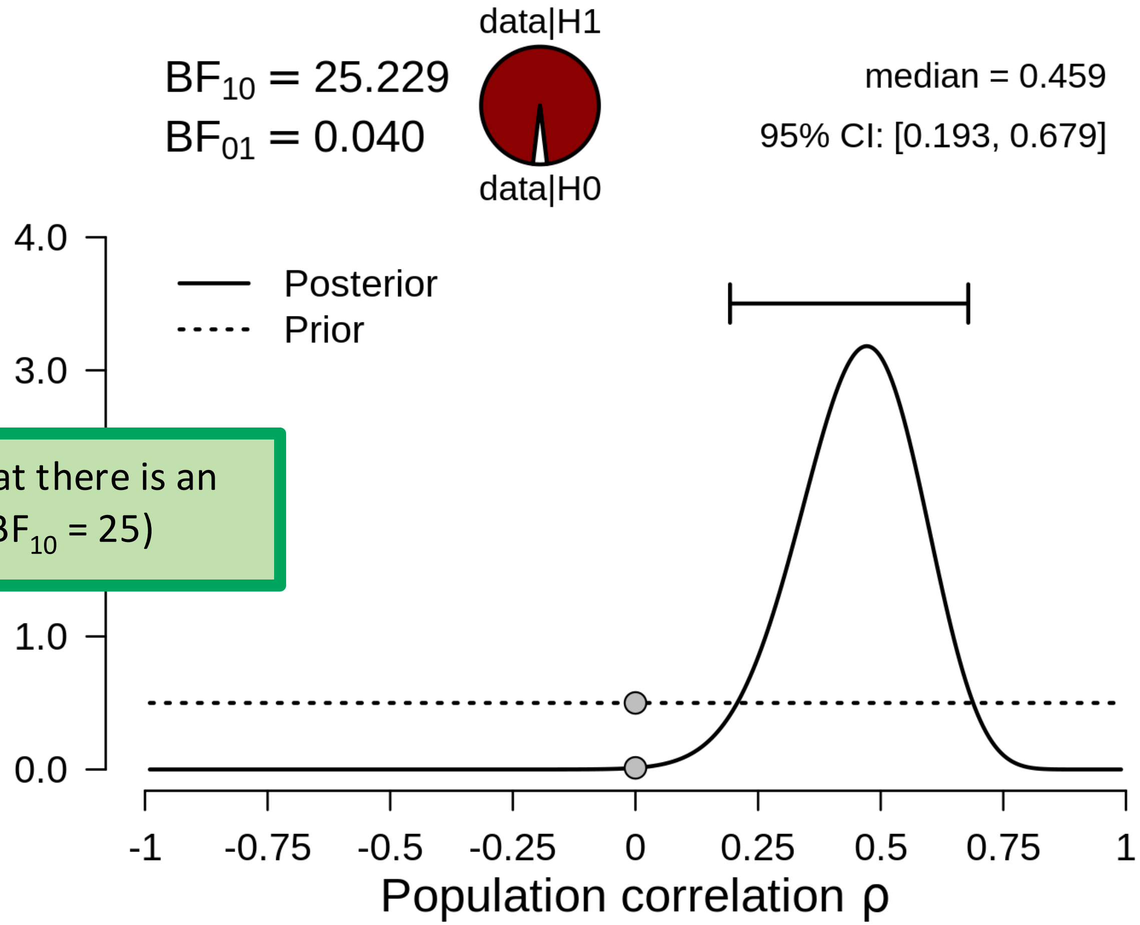
The posterior is also more narrow!
This also leads to a more narrow credible interval (still 95%, but the two numbers are closer together (e.g., 0 is no longer in the interval)

We can make a more specific prediction, with the same certainty (95%)

95% Credible interval:
"95% probability that ρ is between 0.193 and 0.679"

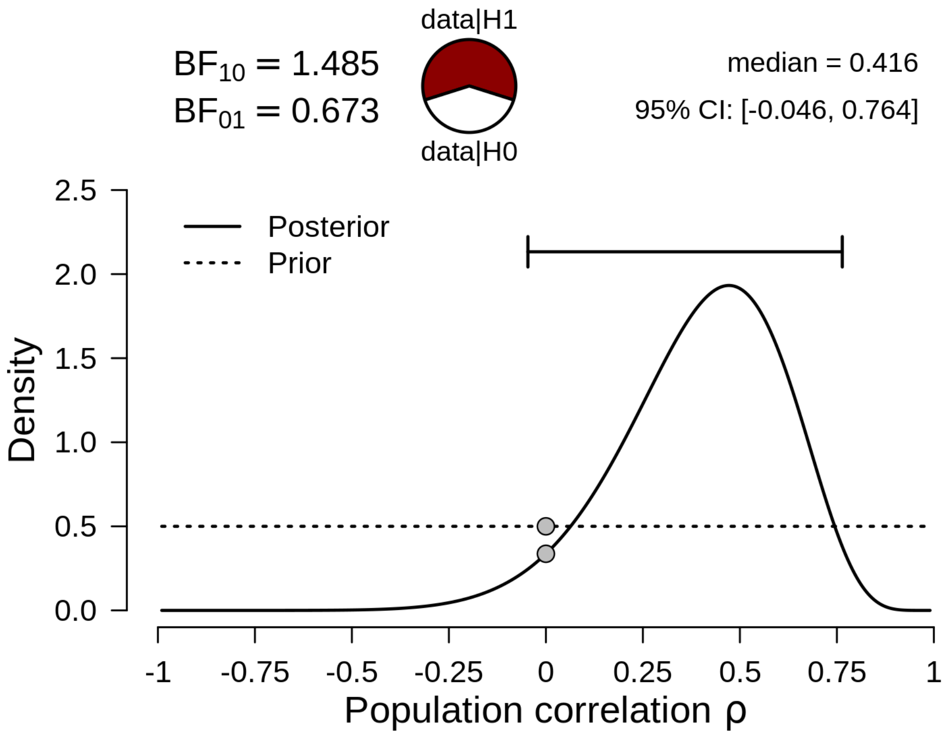


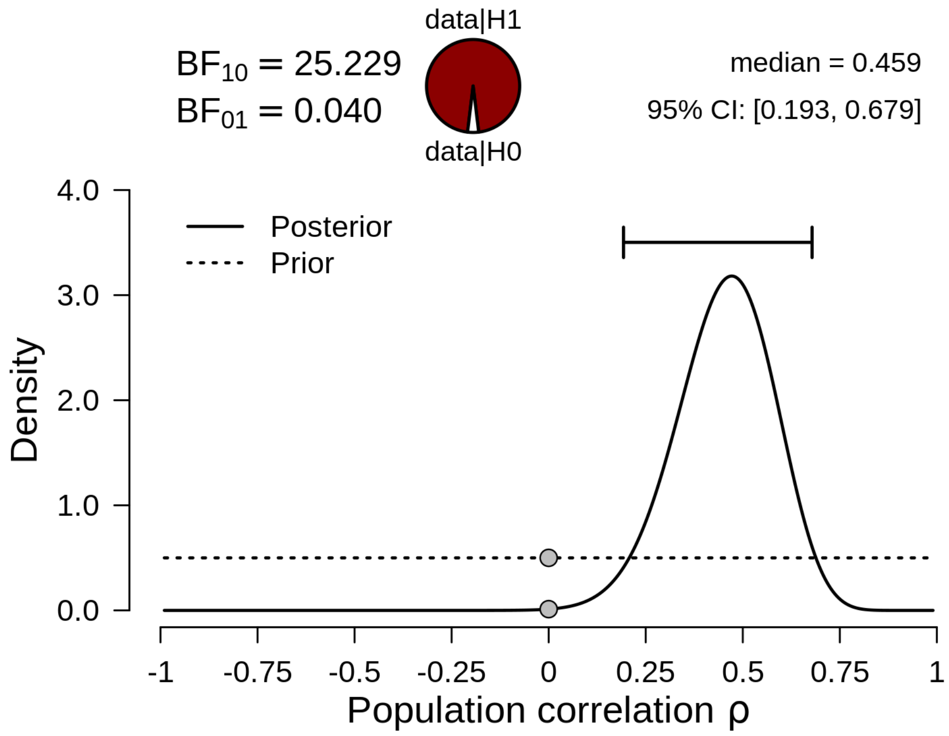
Bayesian Correlation: Bayes Factor



Now we have pretty strong evidence that there is an association between the variables (BF_{10} = 25)

Bayesian Correlation: Bayes Factor





N = 15, observed correlation = 0.485

N = 40, observed correlation = 0.485

Today

- Recap of last week
- Bayesian Inference
  - Correlation
  - **T-test**
- Bayes & Frequentism
- Statistics in the Wild
- Recap
  - Example exam question

Bayesian T-Test

- What about differences between means?
- In psychology, differences between groups are characterized by **δ** (“delta”, also known as Cohen’s *d*): a *standardized* difference between groups

Effect size	***d***	**Reference**
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988

As with the Bayes factor, this is a table to provide some intuition for the magnitude of an effect – there are no hard cut-off values here!
(don’t need to study this table for exam)

For further reading and formulas (not exam material):
https://en.wikipedia.org/wiki/Effect_size#Difference_family:_Effect_sizes_based_on_differences_between_means

Bayesian T-Test

$$P(\delta \mid \text{data}) = P(\delta) \ \frac{P(\text{data} \mid \delta)}{P(\text{data})}$$

Bayesian T-Test: Hypothesis

$$\mathcal{H}_0 : \delta = 0$$

$$\mathcal{H}_1 : \delta \neq 0$$

$$\mathcal{H}_+ : \delta > 0$$

$$\mathcal{H}_- : \delta < 0$$

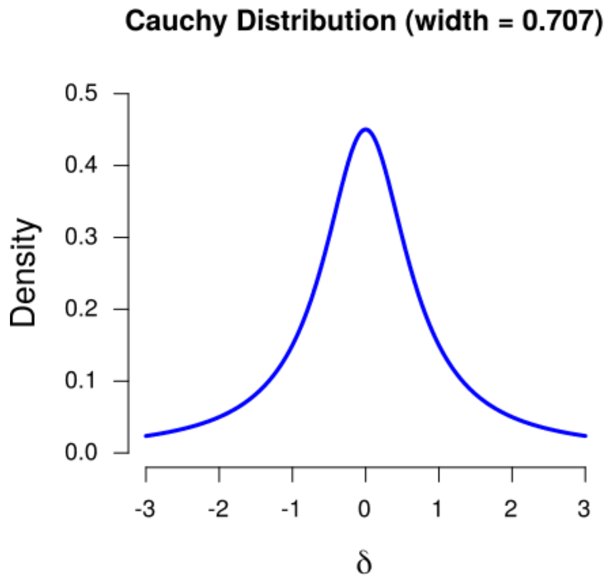
"Alternative Hypothesis"

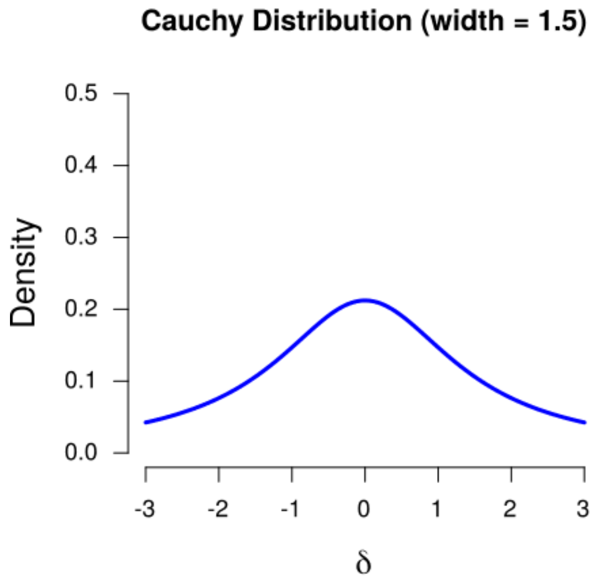
Bayesian T-Test: Prior Distribution

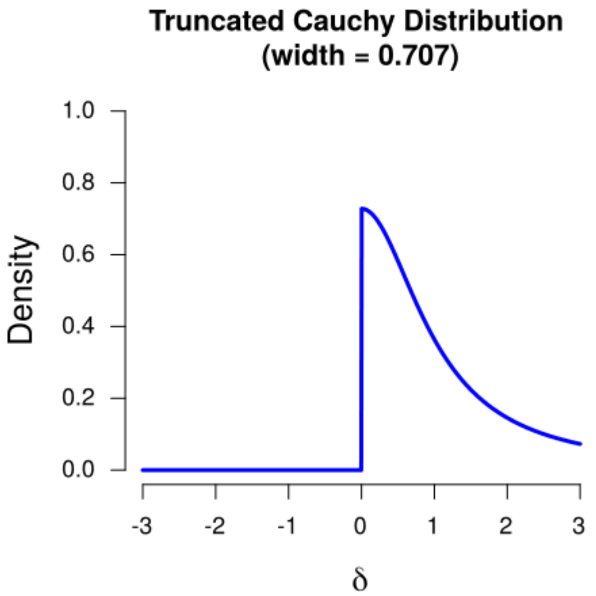
The prior distribution is on the same domain as the parameter of interest: so we need a distribution that is between [-∞, ∞].
A normal distribution could work, but the convention is to use a t-distribution with df = 1. This is known as the **Cauchy** distribution

A uniform distribution would not work here, because delta has an infinite domain

The Cauchy distribution is governed by a single shape parameter that determines how wide it is







Bayesian T-Test: Prior Distribution

A prior distribution that reflects the belief that 50% of the values of δ are located between -0.707 and 0.707, **a priori**

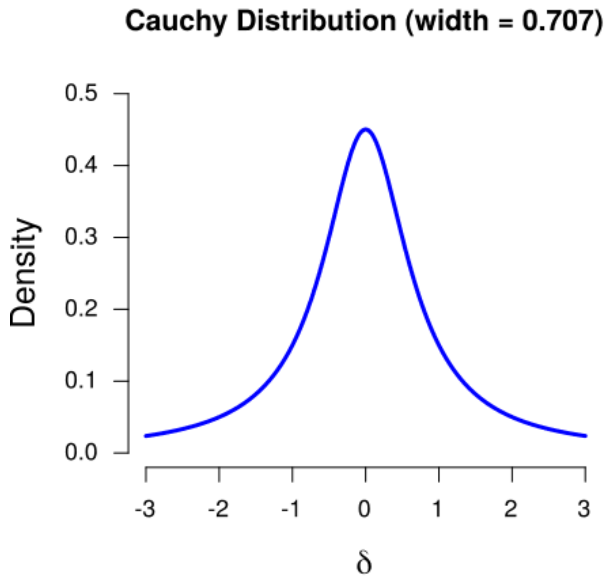
This is often selected as the "uninformative" option. Why? Unfortunately that is a long and technical story, not suited for this course :(

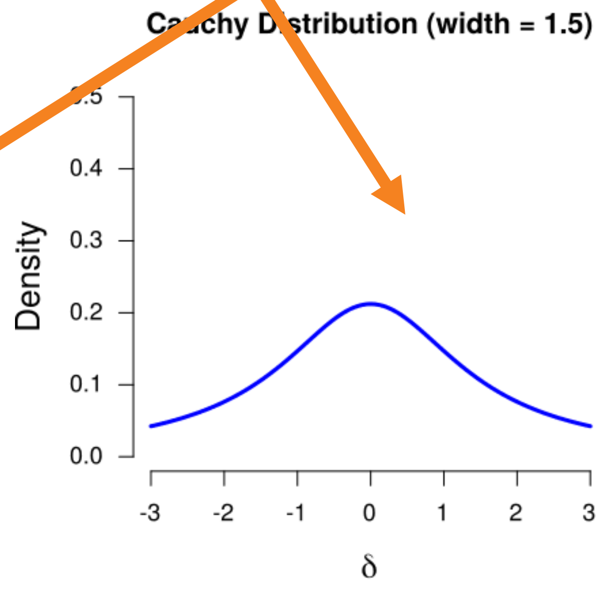
A prior distribution that reflects the belief that 50% of the values of δ are located between -1.5 and 1.5 , **a priori**

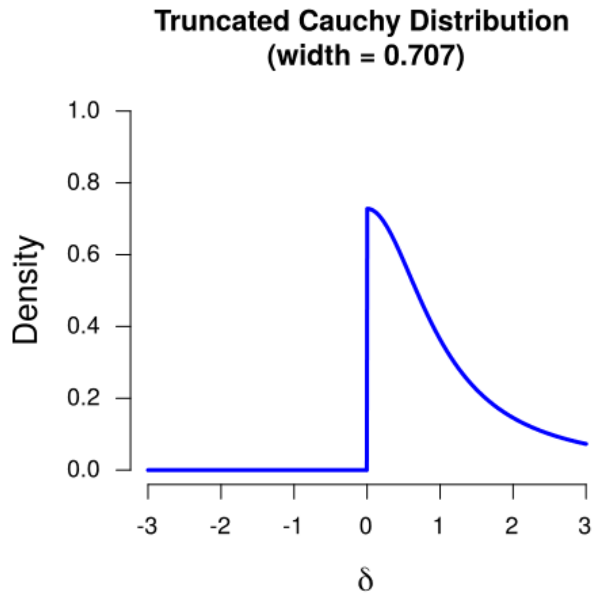
A prior distribution that reflects the belief that only positive values of δ are possible and that 50% of the values of δ are between 0 and 0.707, **a priori**

$$\mathcal{H}_1 : \delta \neq 0$$

$$\mathcal{H}_+ : \delta > 0$$





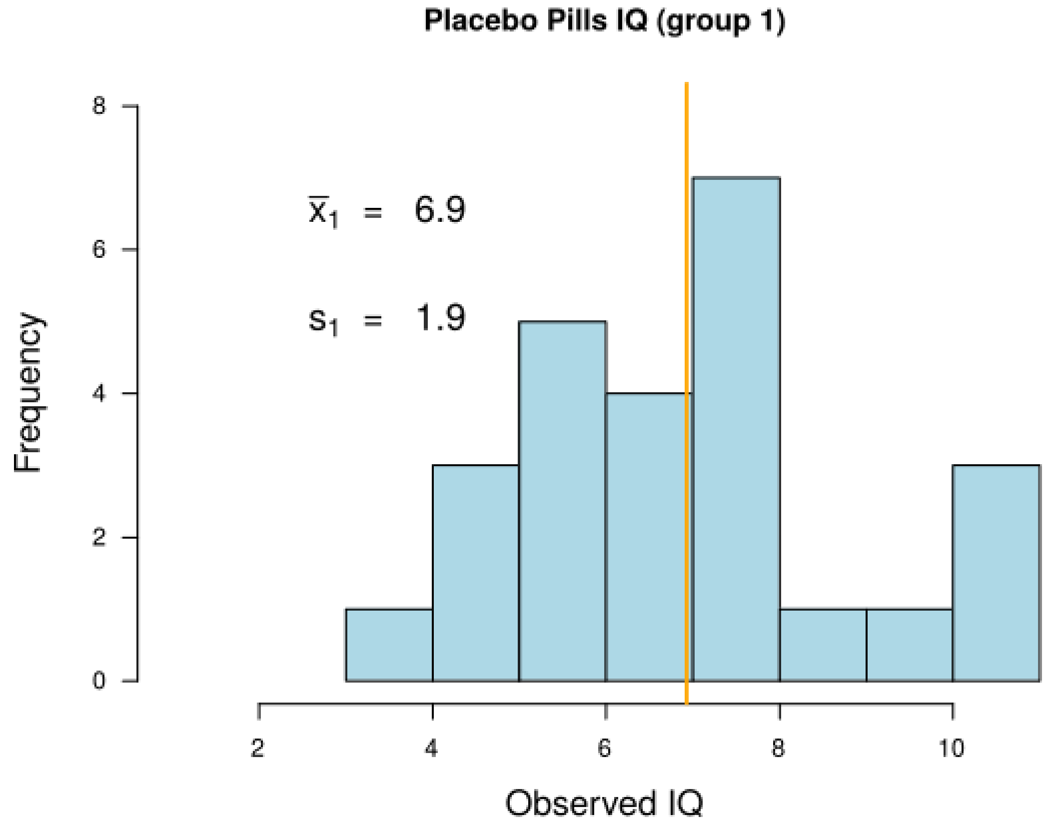


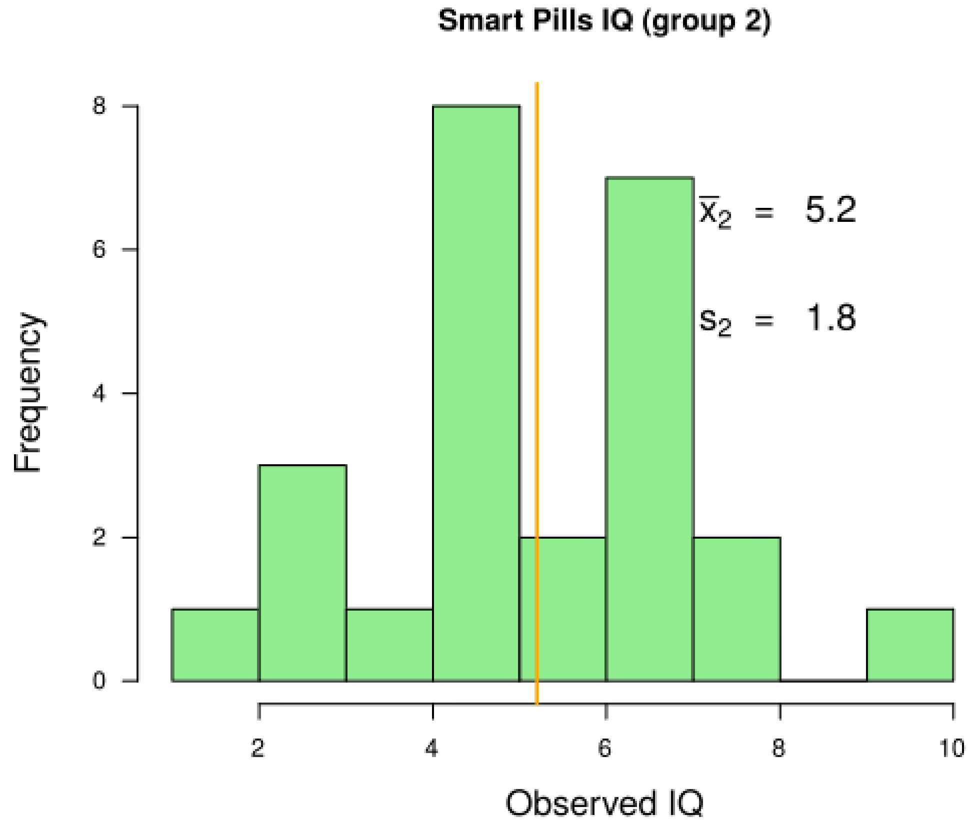
Bayesian T-Test: Data

Observed δ = 0.95

(t = 3.27)

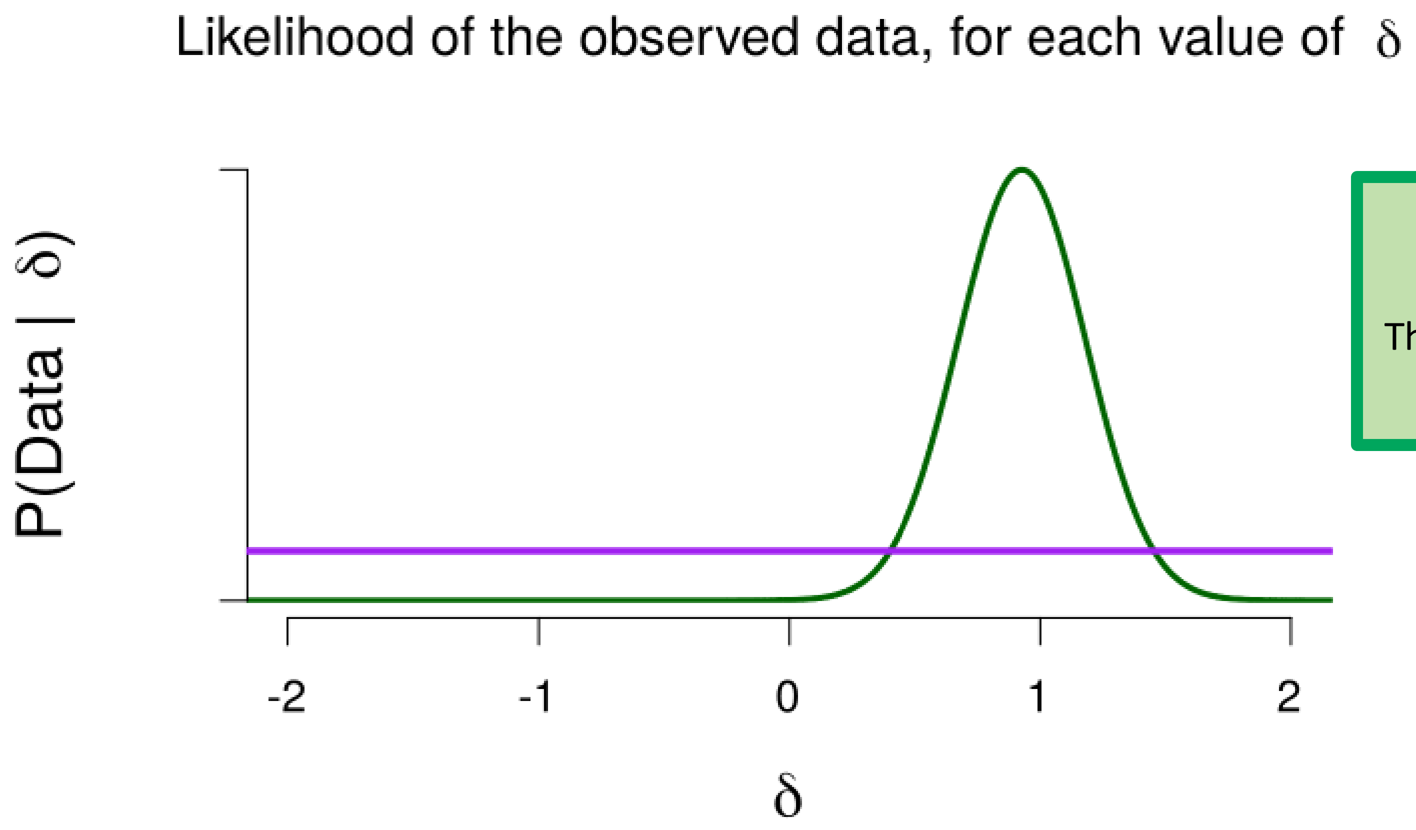
n = 25





Bayesian T-Test: Likelihood

$$P(\text{data} \mid \delta)$$

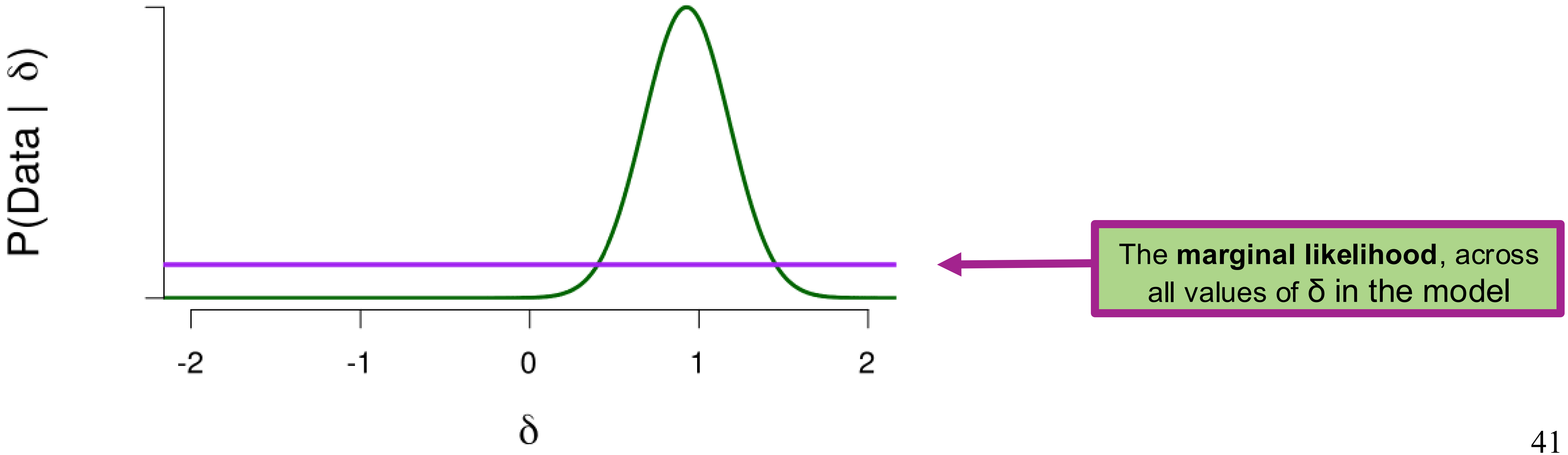


We see that the data are likely for values of δ close to 1.
This makes sense, because the observed delta (i.e., the data) is equal to 0.95!

Bayesian T-Test: Likelihood

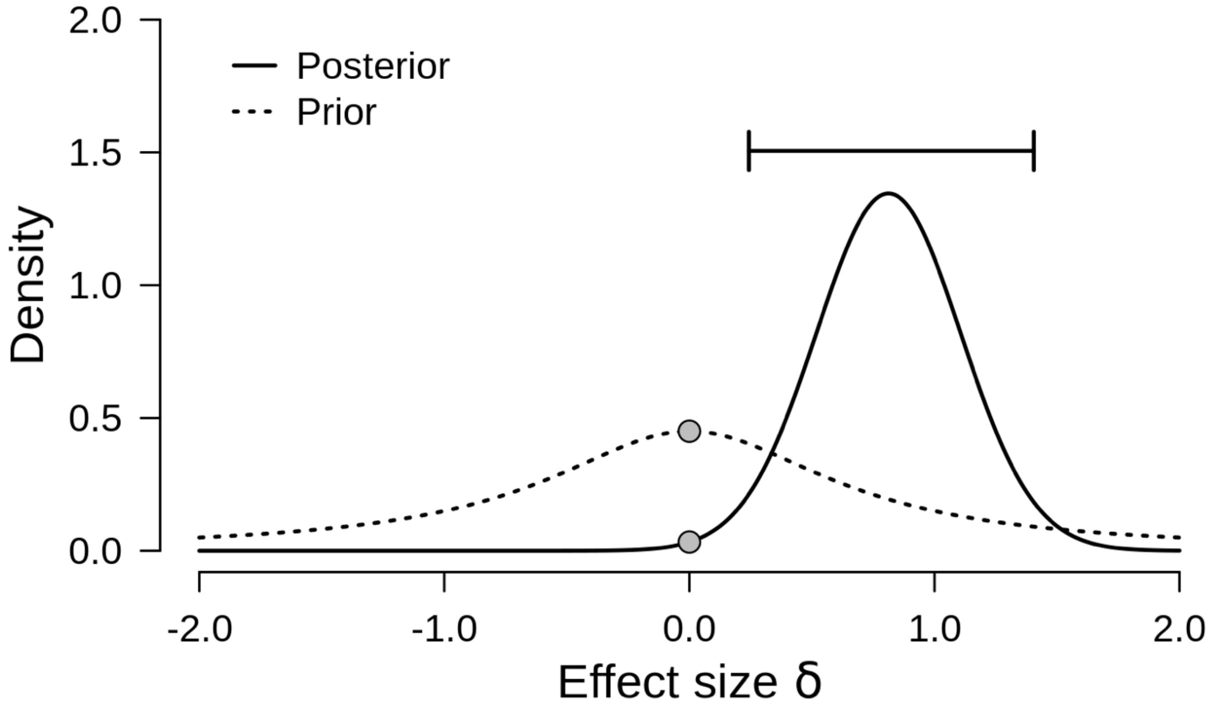
$$P(\text{data} \mid \delta)$$

Likelihood of the observed data, for each value of δ

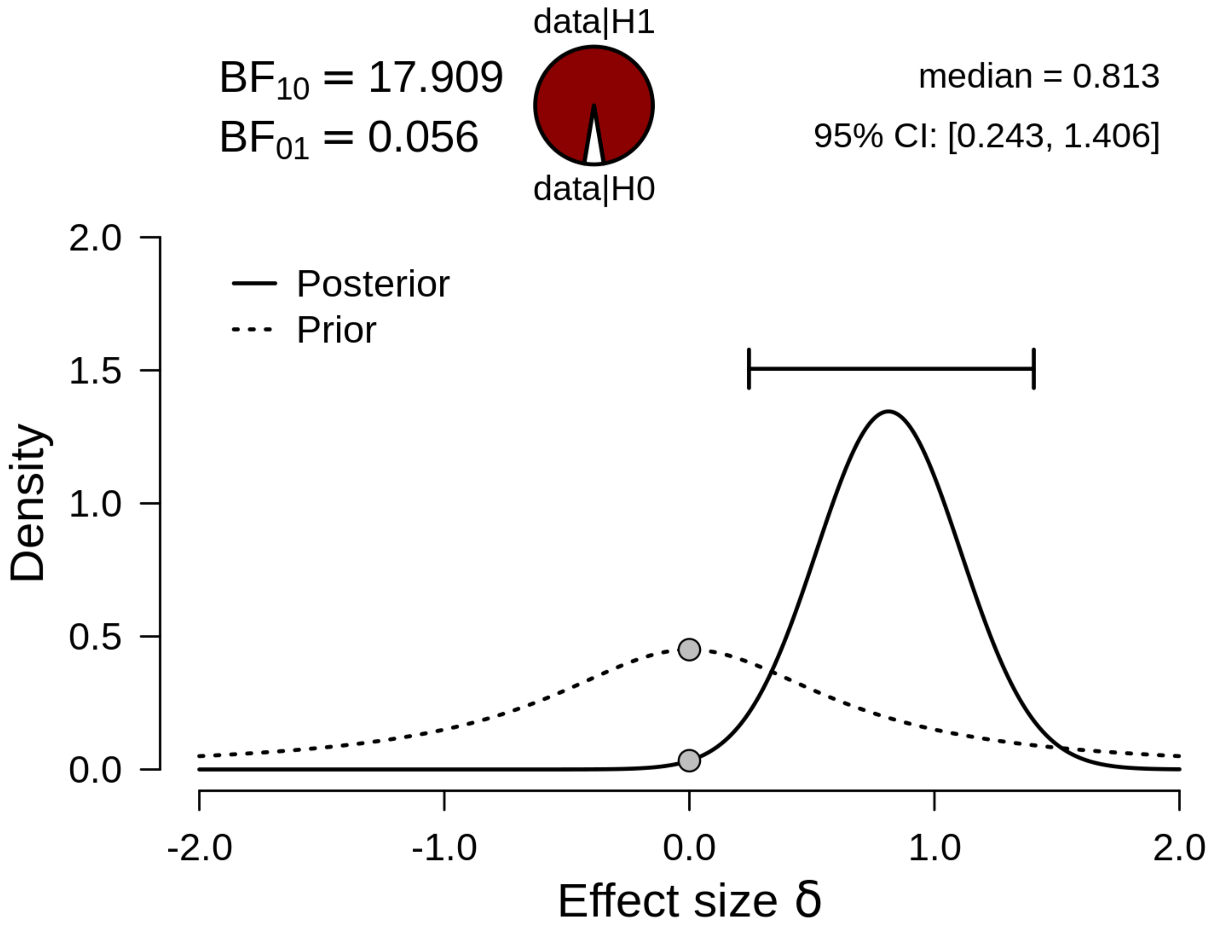


Bayesian T-Test: Posterior Distribution

$$P(\delta \mid \mathrm{data})$$



Bayesian T-Test: Bayes Factor

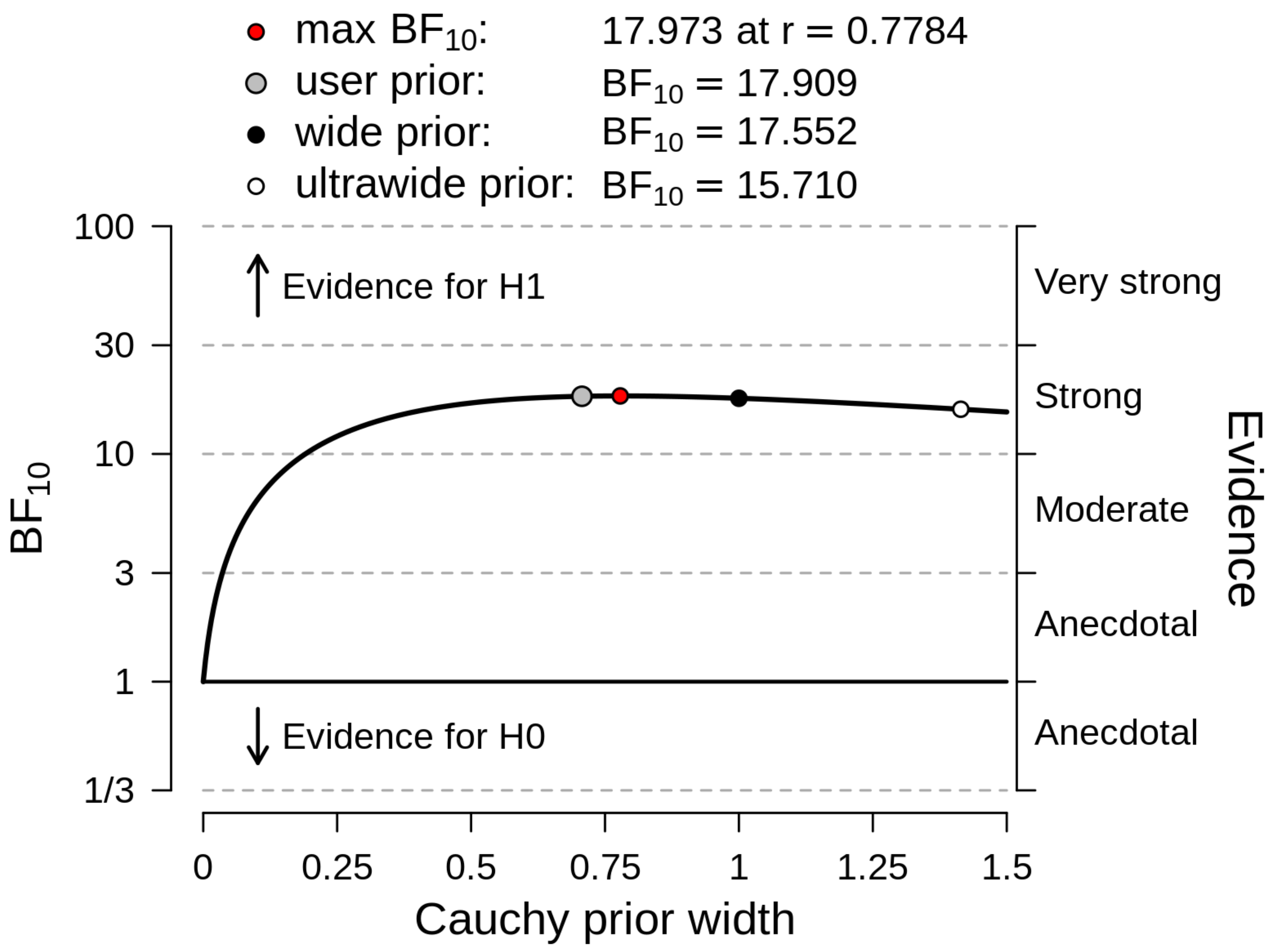


Bayesian T-Test: Robustness Check

Because choosing the prior setting is fairly subjective, we can explore what would have happened if we had chosen a different value for the prior width. This is known as a **Bayes factor robustness check**, **or sensitivity analysis**

We see here that the Bayes factor is pretty stable (i.e., flat line) across many values for the prior width. Except for very strong prior settings (0.15 and lower) does the evidence for H_1 decrease

"Here, the Bayes factor does not qualitatively differ for a wide range of prior specifications (0.25 – 1.5)



Bayesian T-Test - Demonstration

www.jasp-stats.org

Today

- Recap of last week
- Bayesian Inference
  - Correlation
  - T-test
- **Bayes & Frequentism**
- Statistics in the Wild
- Recap
  - Example exam question

Bayes & Frequentism

Frequentist Hypothesis testing

1) **Specify alpha**
2) Assumptions
3) Hypothesis
4) **Test statistic + sampling distribution**
5) **P-value**
6) Conclusion

Bayesian Hypothesis testing

1) Assumptions
2) Hypothesis
3) **Set prior distribution**
4) **Compute likelihood**
5) **Bayes factor**
6) Conclusion

Bayes & Frequentism

Bayes	Frequentism
Compare the predictive quality of two hypotheses	Minimize type 1 error (incorrectly reject H_0)

Bayes & Frequentism

Bayes	Frequentism
Compare the predictive quality of two hypotheses	Minimize type 1 error (incorrectly reject H_0)
Can monitor the results as data accumulate	Can only look at the data once: looking twice doubles the risk of type 1 error

Bayes & Frequentism

Bayes	Frequentism
Compare the predictive quality of two hypotheses	Minimize type 1 error (incorrectly reject H_0)
Can monitor the results as data accumulate	Can only look at the data once: looking twice doubles the risk of type 1 error
Check assumptions	Check assumptions

Bayes & Frequentism

Bayes	Frequentism
Compare the predictive quality of two hypotheses	Minimize type 1 error (incorrectly reject H_0)
Can monitor the results as data accumulate	Can only look at the data once: looking twice doubles the risk of type 1 error
Check assumptions	Check assumptions
Still incomplete (but steadily expanding)	Has wide range of types of analyses

Bayes & Frequentism

Bayes	Frequentism
Compare the predictive quality of two hypotheses	Minimize type 1 error (incorrectly reject H_0)
Can monitor the results as data accumulate	Can only look at the data once: looking twice doubles the risk of type 1 error
Check assumptions	Check assumptions
Still incomplete (but steadily expanding)	Has wide range of types of analyses
Sampling plan: can monitor the results and stop when satisfied	Have to specify sampling plan before, based on hypothesized effect size

Dichotomous reasoning of p-values is arbitrary

[“It’s funny how significance make a difference” - MF DOOM]

- Compare:
  - Group A has a higher on verbal IQ than group B
    - **(p = 0.049)**
  - There is no difference between the groups in verbal IQ
    - **(p = 0.051)**

- Fisher (1935):
  - Founder of p-value
  - In a footnote: *“You may think of the idea of rejecting the null-hypothesis at some value, **for instance** at 0.05”*

Not intuitive

- Common mistakes when working with p-values
  - ‘evidence for H_A’
  - ‘probability that H_0 is true’

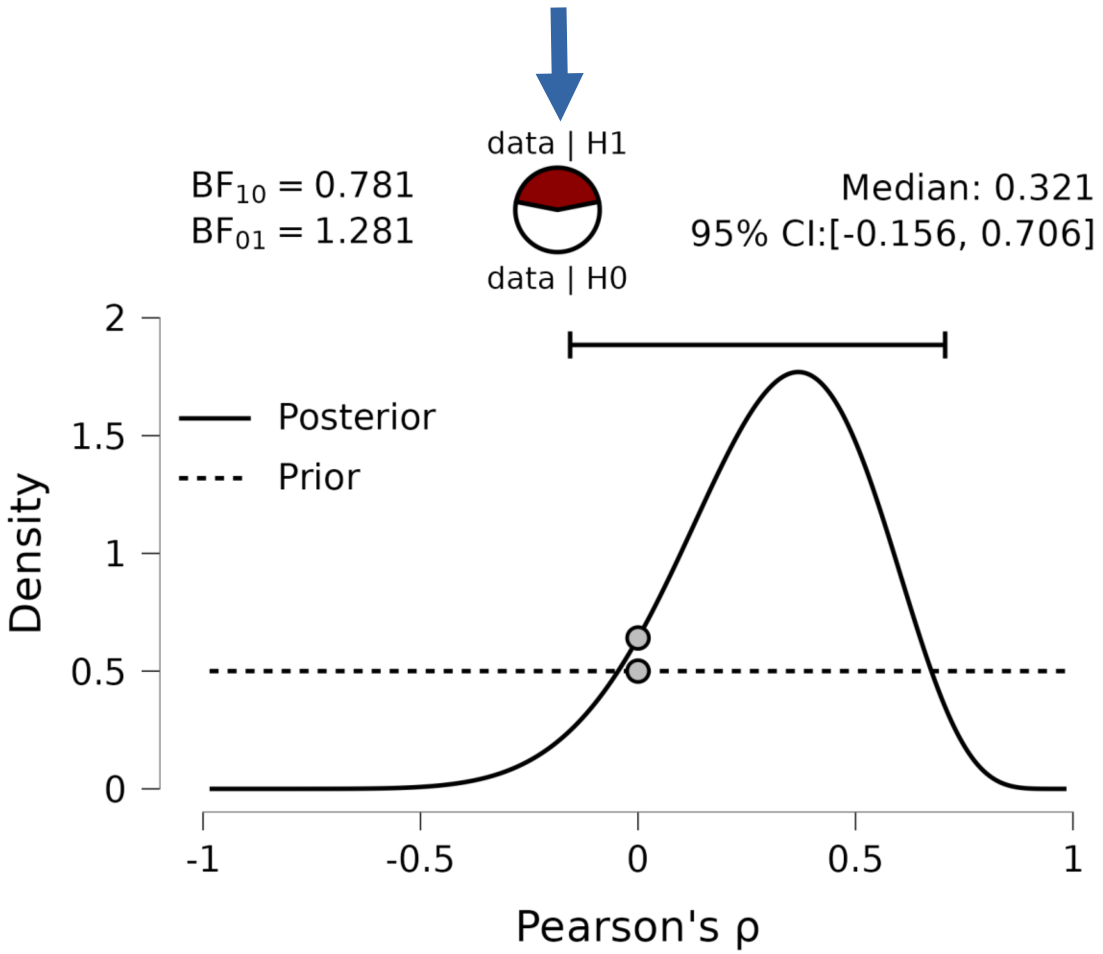
These are **incorrect** interpretations!

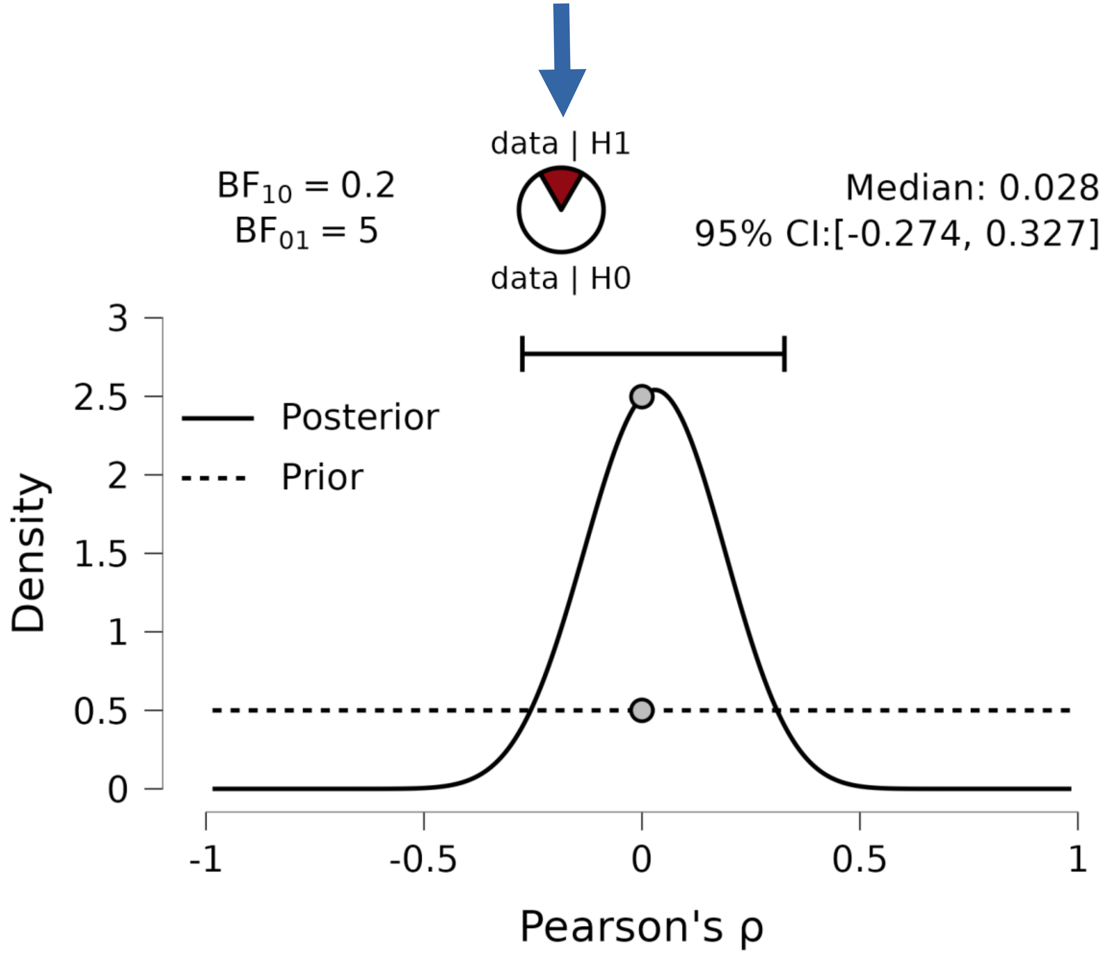
- Understandable because this is what we actually want to know

- What is the p-value again?

  ‘The probability of these *data* or more extreme if H_0 is true’

“Absence of evidence” vs “evidence of absence”





Both analyses would yield a non-significant p-value – but what does that actually mean?

Today

- Recap of last week
- Bayesian Inference
    - Correlation
    - T-test
- Bayes & Frequentism
- **Statistics in the Wild**
- Recap
    - Example exam question

Statistics in the Wild

Often, there is great pressure to publish articles, and to publish, “non-null effects” are needed...

So what if you spent a year of research collecting a data set, but your result is not significant? Or your Bayes factor is inconclusive (i.e., close to 1)?

Just test some extra subjects (and not do Bayes)

My data is not really normal, but who cares about assumptions?

I will just test another hypothesis...

If I only report this alternative prior distribution, my Bayes factor is way higher!

If I conduct a t-test instead of a Wilcoxon test, I can reject H_0!

Hey, my computer can simulate data!

Deleting this one “outlier” makes my effect significant!

Source:
Pixabay.com

Statistics in the Wild

- There has been an increase in statistical rigor since the Reproducibility crisis:
  - Center for Open Science tried to replicate 100 published psychological studies
  - Only 35 studies replicated

- https://science.sciencemag.org/content/349/6251/aac4716.full?ijkey=1xgFoCnpLswpk&keytype=ref&siteid=sci
- https://en.wikipedia.org/wiki/Reproducibility_Project)

Statistics in the Wild

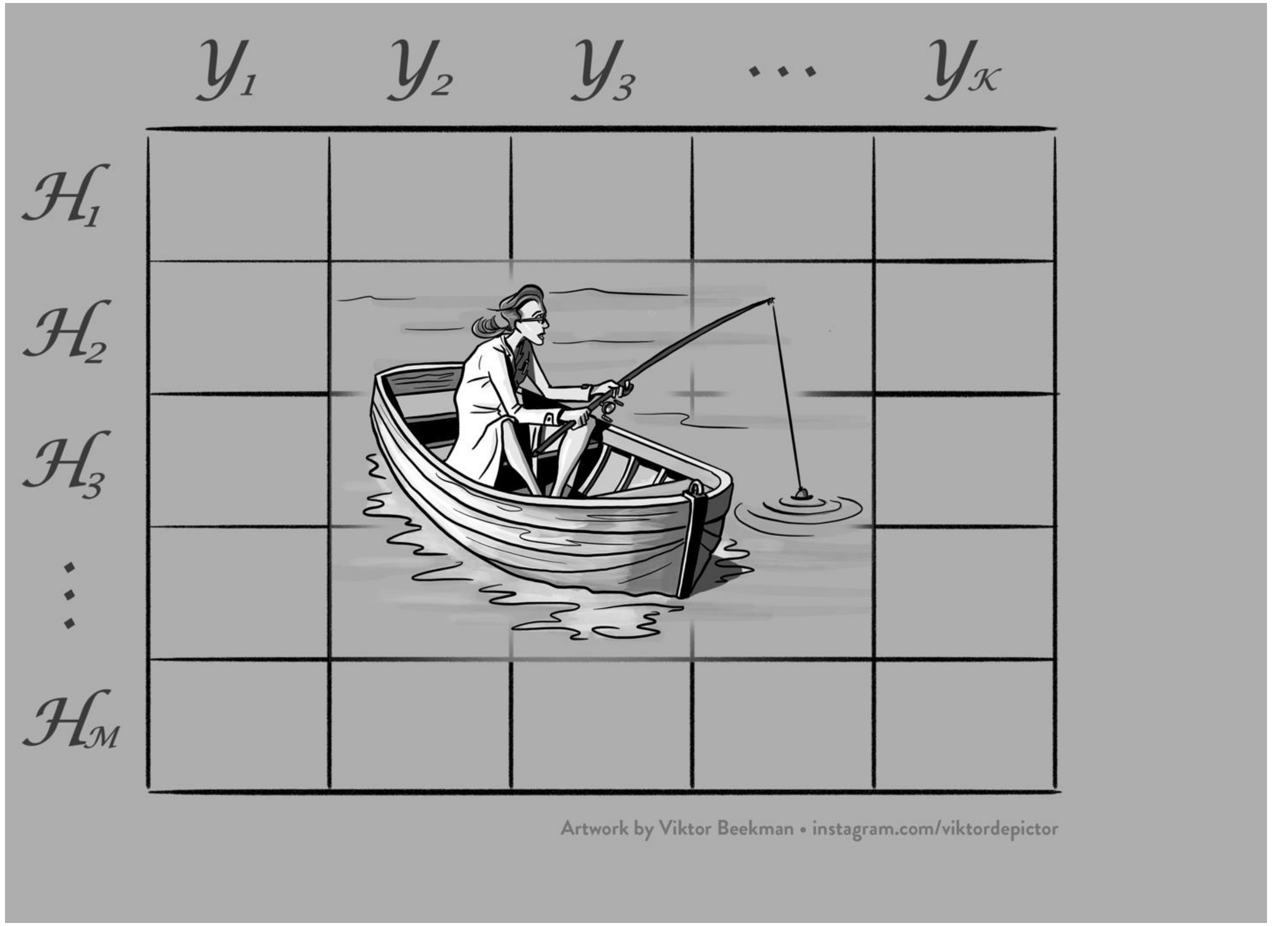
- A powerful initiative to reduce bias in reporting is preregistration
  - Before you conduct your study, you preregister several pieces of information:
  - Description of study and sampling plan
  - Hypotheses (+ one-sided/two-sided)
  - Analysis plan: what statistical tests, and what settings: prior and Bayes factor / alpha-level for accepting or rejecting hypothesis

- Munafò, M., Nosek, B., Bishop, D. et al. A manifesto for reproducible science. Nat Hum Behav 1, 0021 (2017) doi:10.1038/s41562-016-0021

Statistics in the Wild: Preregistration

Preregistration prevents **Data Dredging/Fishing/Hacking:** trying all sorts of different analyses/settings in order to get a significant result

“Oh, this two-sided t-test was not significant, let’s try a one-sided t-test instead! Hey, now it’s significant!”



Source:
https://www.bayesianspectacles.org
Pixabay.com

Statistics in the Wild: Preregistration

Preregistration prevents **presenting exploratory research as confirmatory research**: *exploratory* research is used to generate theories and hypotheses; *confirmatory* research is used to test those

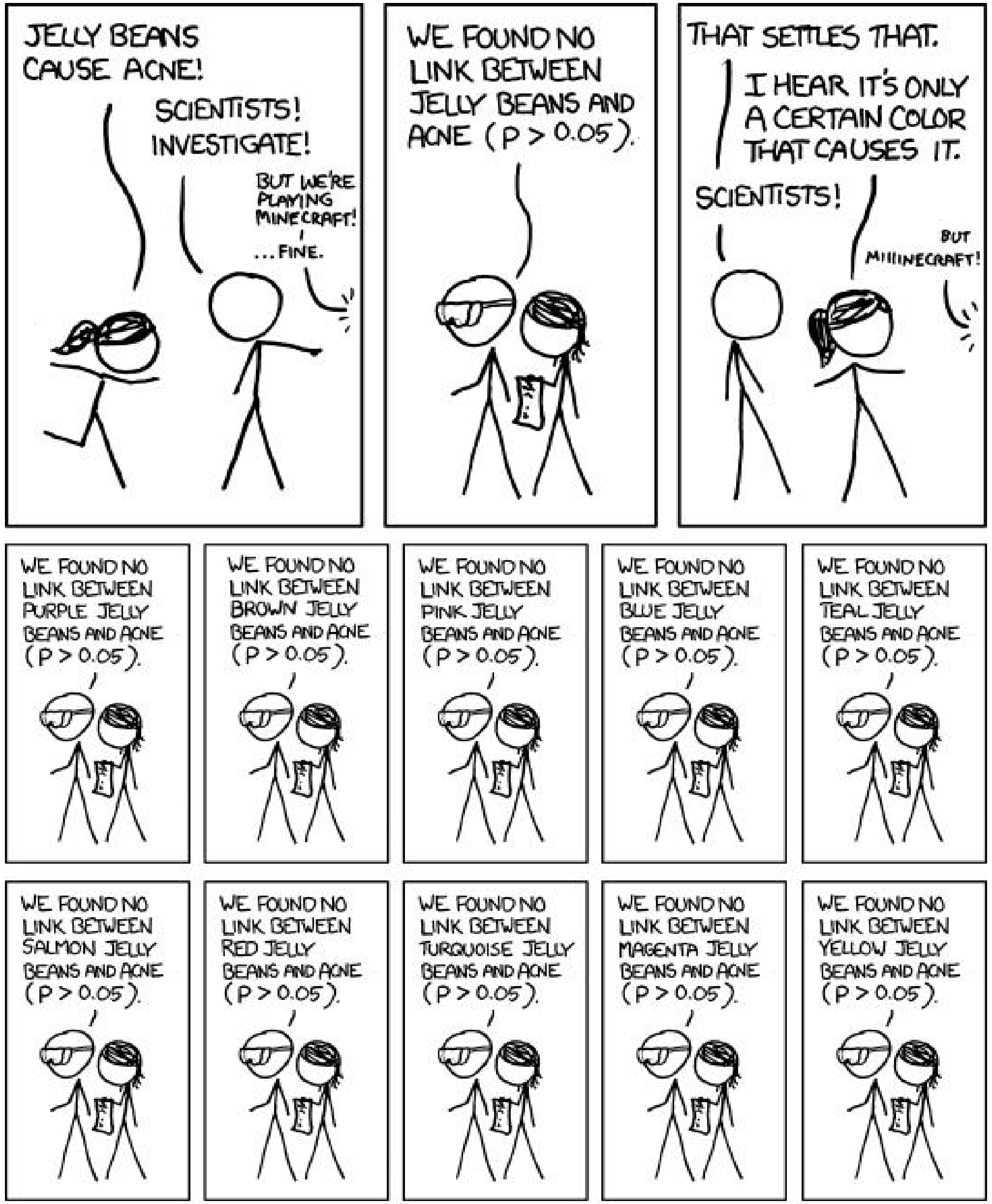
Exploratory: Let's measure all big 5 personality traits and look at all possible correlations to see which ones correlate

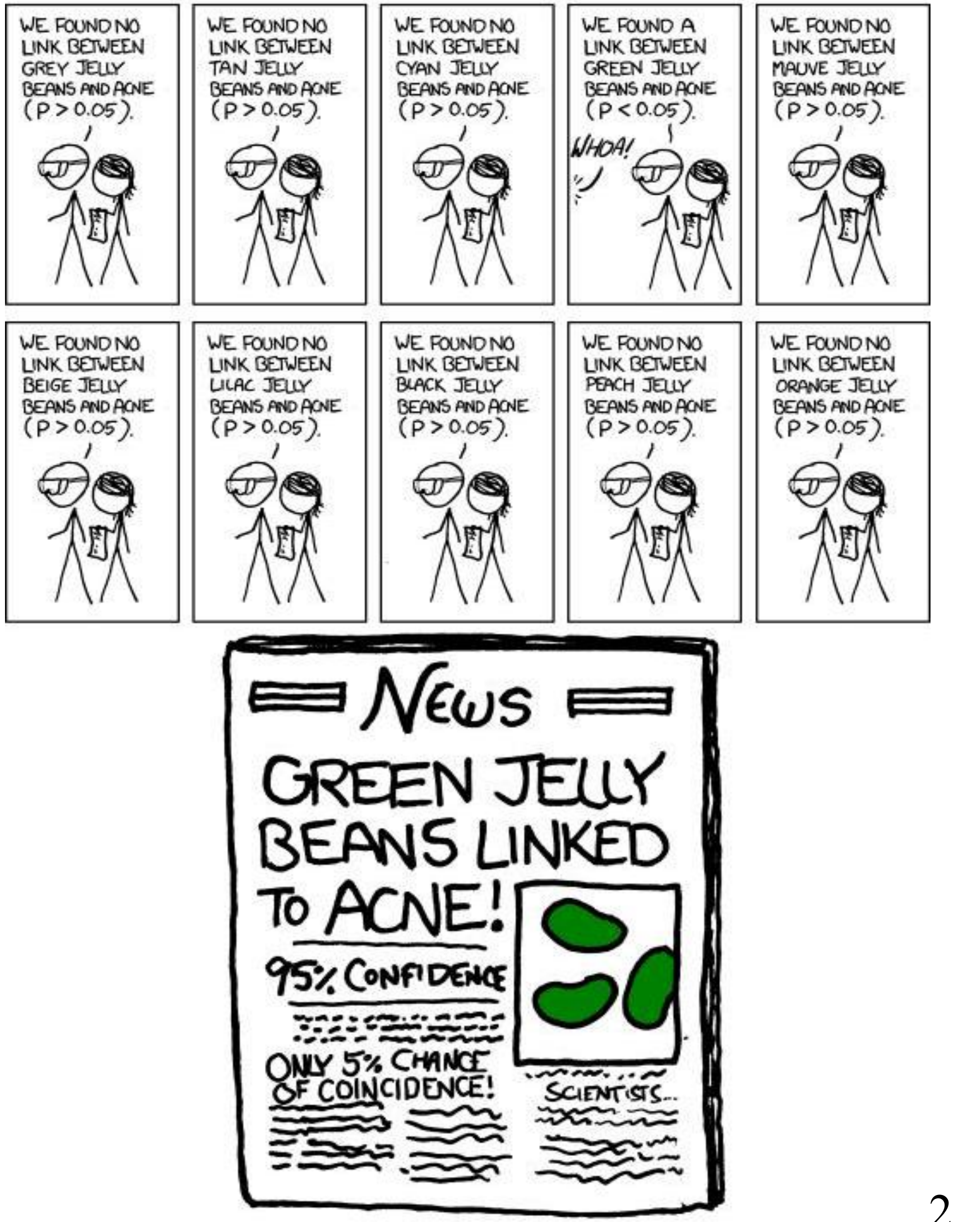
Confirmatory: Let's measure all big 5 personality traits, because my theory postulates that C & N, O & E, and E & A traits are positively

Presenting exploratory as confirmatory: "Oh I found a positive correlation between E & A – that was my theory all along so my theory is now confirmed!"

Source: Pixabay.com

Statistics in the Wild: Preregistration





https://xkcd.com/882/

Statistics in the Wild

- Luckily, journals are also cooperating
  - Registered reports:
    - Write most of your paper+preregistration, without conclusions, when it gets accepted by the journal you conduct the experiment fill in the results and conclusion section of your paper based on the results
  - Increased support for null findings
- Transparency is key: sharing analysis files/code, making data available

Today

- Recap of last week
- Bayesian Inference
  - Correlation
  - T-test
- Bayes & Frequentism
- Statistics in the Wild
- **Recap**
  - Example exam question

Recap

- The same flow of inference we saw for the proportion can also be extended to other tests: for instance correlation and t-test
- We can investigate the effect of the prior on the Bayes factor: robustness check or sensitivity analysis
- We have seen cases where $p < 0.05$, yet the Bayes factor is close to 1

Recap

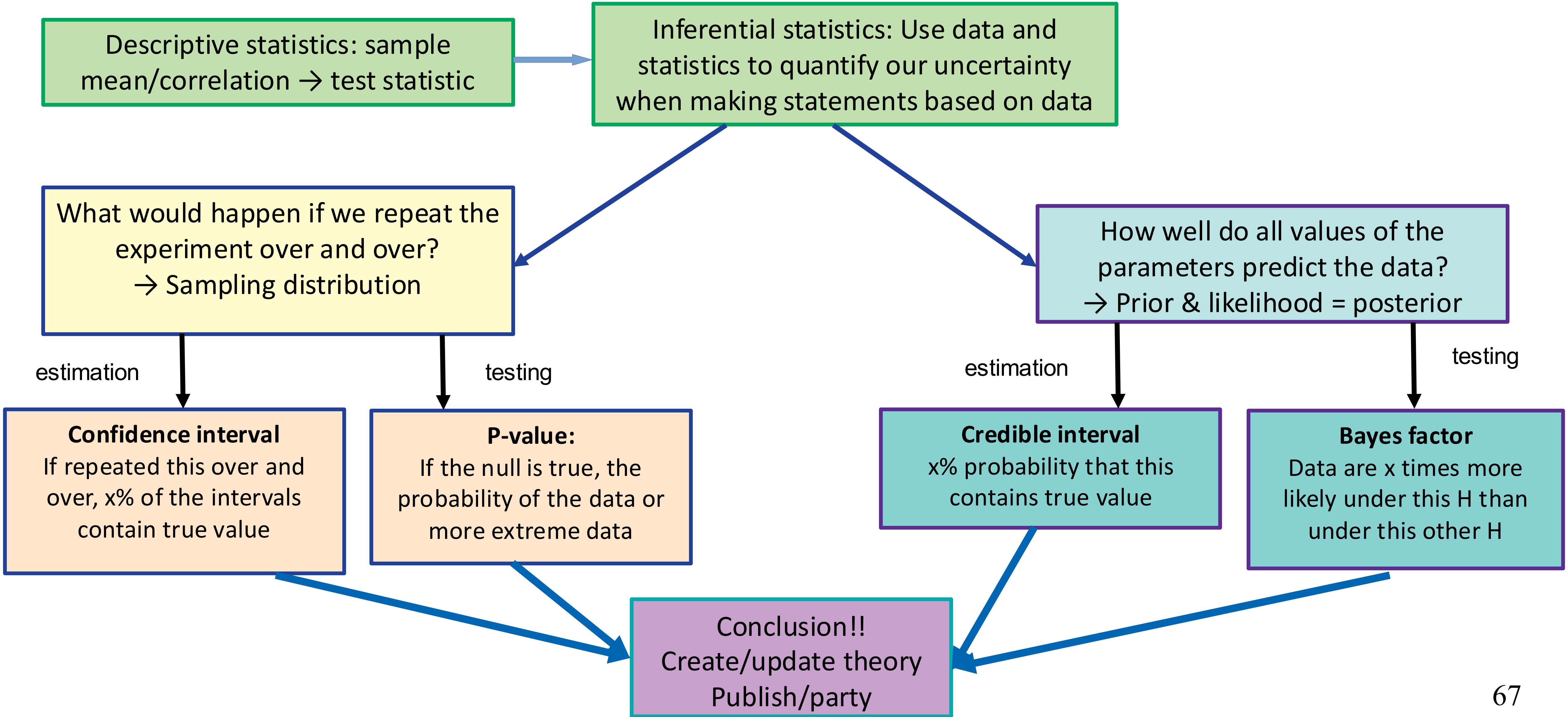
- The Bayes factor is a relative metric! Pay attention to what is being compared to what

- Bayes does not fix all your problems: equally vulnerable to assumption violations and statistical malpractice!

- Preregistration and Open Science can help with this

Big Recap

We have applied this to multiple methodological designs: 1/2/3+ groups, quantitative/categorical/ordinal variables



More Cool Stuff to Read

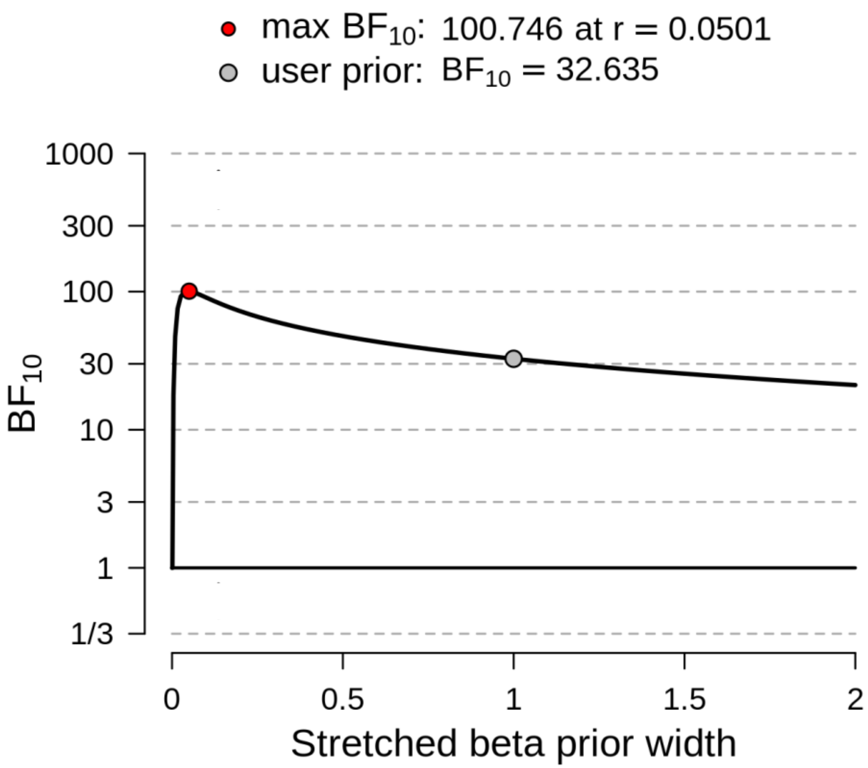
- One statistical analysis must not rule them all
  - Any single analysis hides an iceberg of uncertainty. Multi-team analysis can reveal it.
- The rise and fall of peer review
  - Why the greatest scientific experiment in history failed, and why that's a great thing
- Good Research Practices (PhD dissertation from our department)
  - Research practices and reform ideas aiming to combat the crisis of confidence in psychology
- Strong public claims may not reflect researchers' private convictions
  - Small questionnaire we sent around to empirical researchers to gauge how confident they are in their papers' claims

Today

- Recap of last week
- Bayesian Inference
  - Correlation
  - T-test
- Bayes & Frequentism
- Statistics in the Wild
- Recap
  - **Example exam question**

Example Exam Question

We measure two variables: Openness and Extraversion. After a Bayesian correlation analysis we look at the effect of the prior we chose. What statement is true?



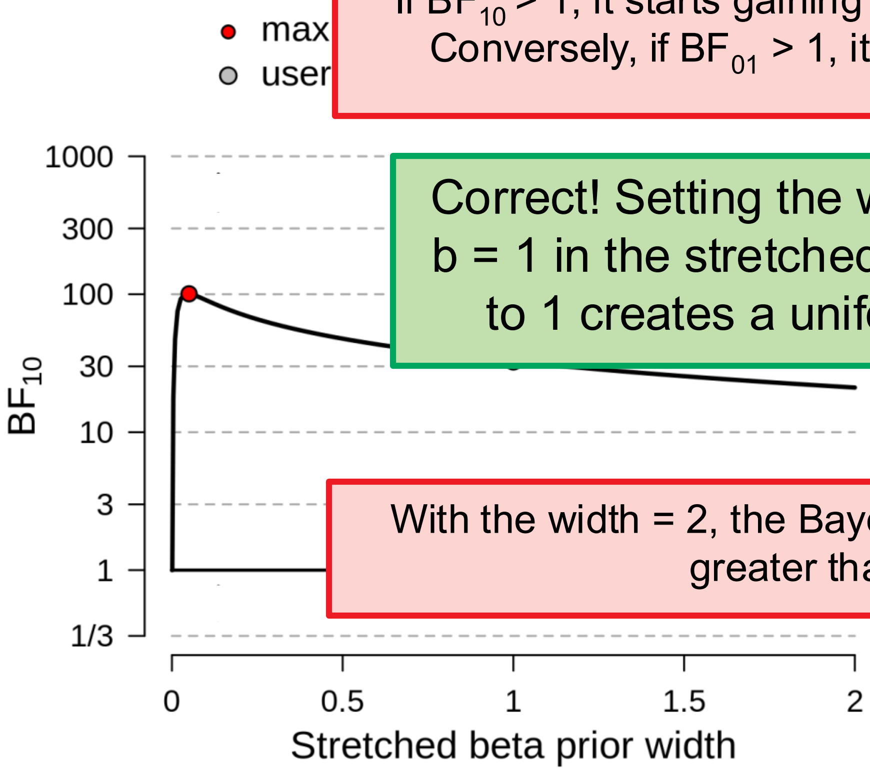
a) We gained evidence in favor of the null hypothesis

b) We used a uniform prior for the correlation

c) If we had specified the prior width equal to 2, the Bayes factor would be inconclusive (i.e., < 3)

Example Exam Question

We measure two variables: Openness and Extraversion. After a Bayesian correlation analysis we look at the effect of the prior we chose. What statement is true?



If $BF_{10} > 1$, it starts gaining towards supporting H_1
Conversely, if $BF_{01} > 1$, it starts supporting H_0

Correct! Setting the width to 1 (i.e., a = b = 1 in the stretched beta distribution) to 1 creates a uniform distribution

With the width = 2, the Bayes factor would still be greater than 10

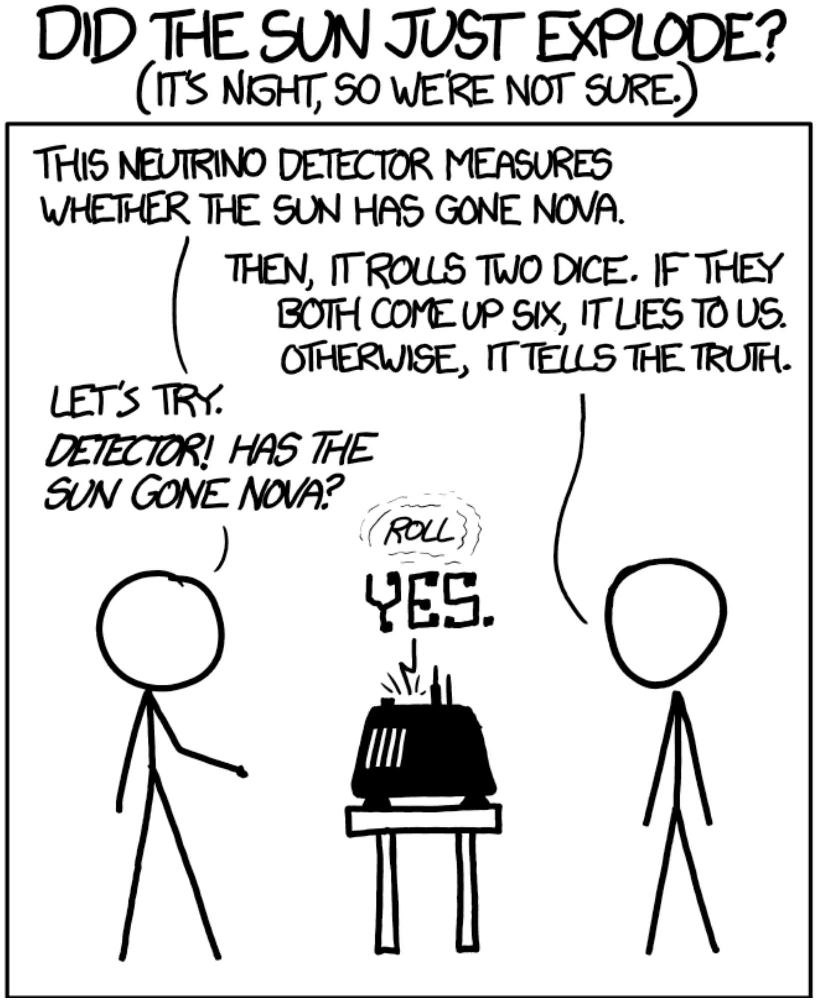
a) We gained evidence in favor of the null hypothesis

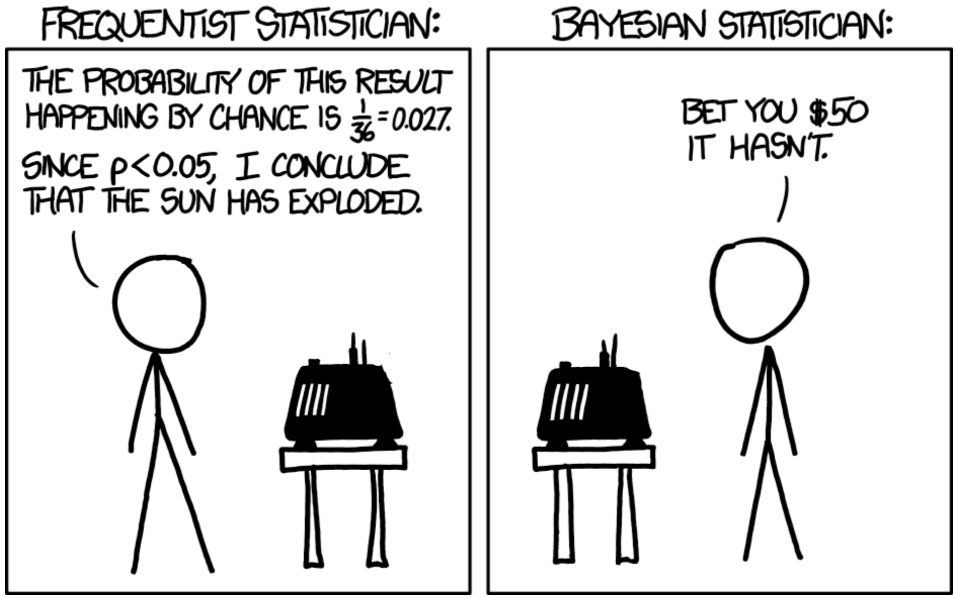
b) We used a uniform prior for the correlation

c) If we had specified the prior width equal to 2, the Bayes factor would be inconclusive (i.e., < 3)

Questions?

Thank you for your attention





https://xkcd.com/1132/

Bonus Article

There is a lot more cool stuff happening in Psychological Methods than just Bayesians criticizing Frequentists. Here at the UvA, Denny Borsboom and his lab have developed network models for dealing with complex psychological phenomena, such as psychiatric disorders. Based on correlations between many variables, they can visualize how such complex systems function.

Here is a cool interactive network plot of how psychiatric symptoms in the DSM V are connected:
Symptoms are represented as nodes and connected by an edge whenever they figure in the same disorder. The color of nodes represents the DSM-IV chapter in which they occur most often. The link above links to the article itself.

https://www.annualreviews.org/doi/suppl/10.1146/annurev-clinpsy-050212-185608/suppl_file/cp9_borsboom_fig3.svg