Bayesian parameter estimation

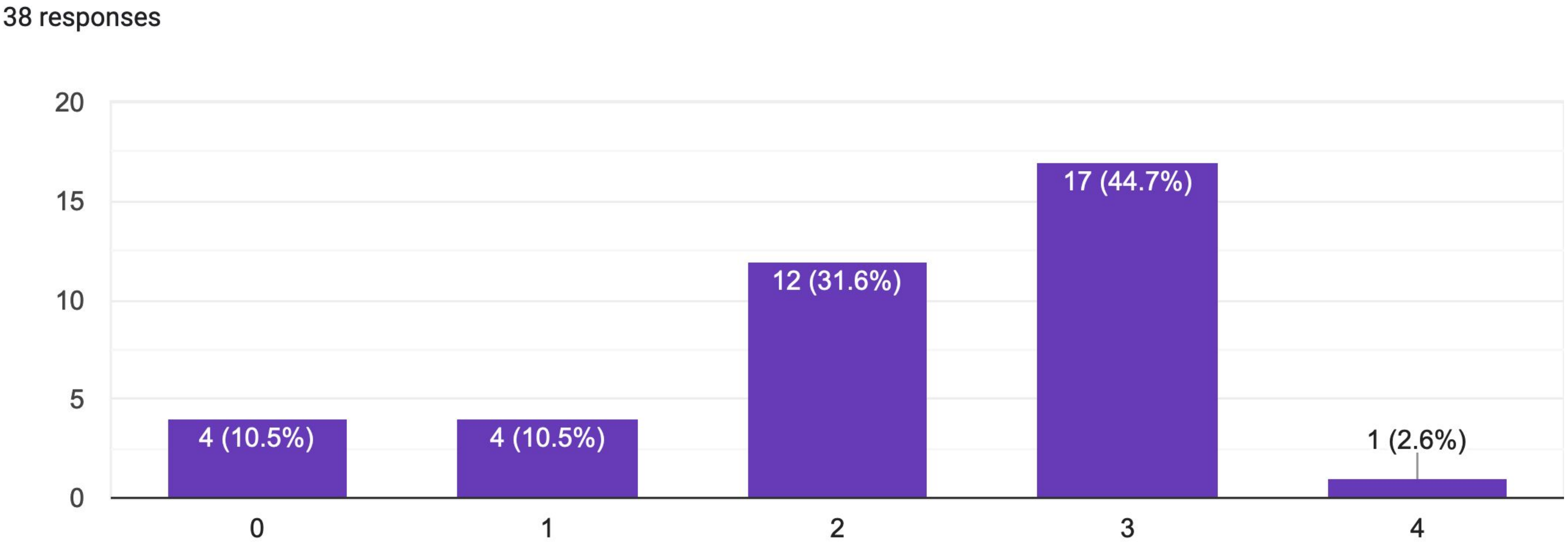
2020 June 24

Why are you here?

(In any case, thanks for coming :)

Frequentist tests are confusing

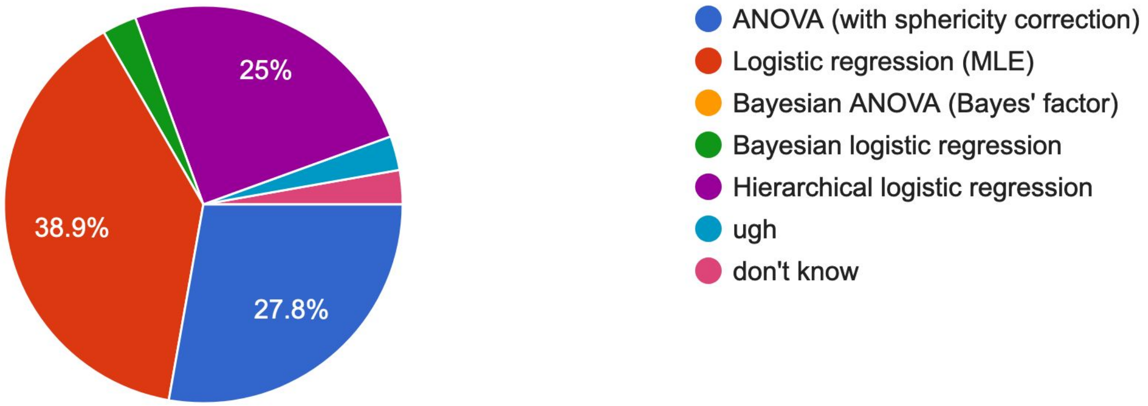
“When you use a frequentist test, how confident are you usually that you chose the correct test for your statistical question?”



Choosing a test

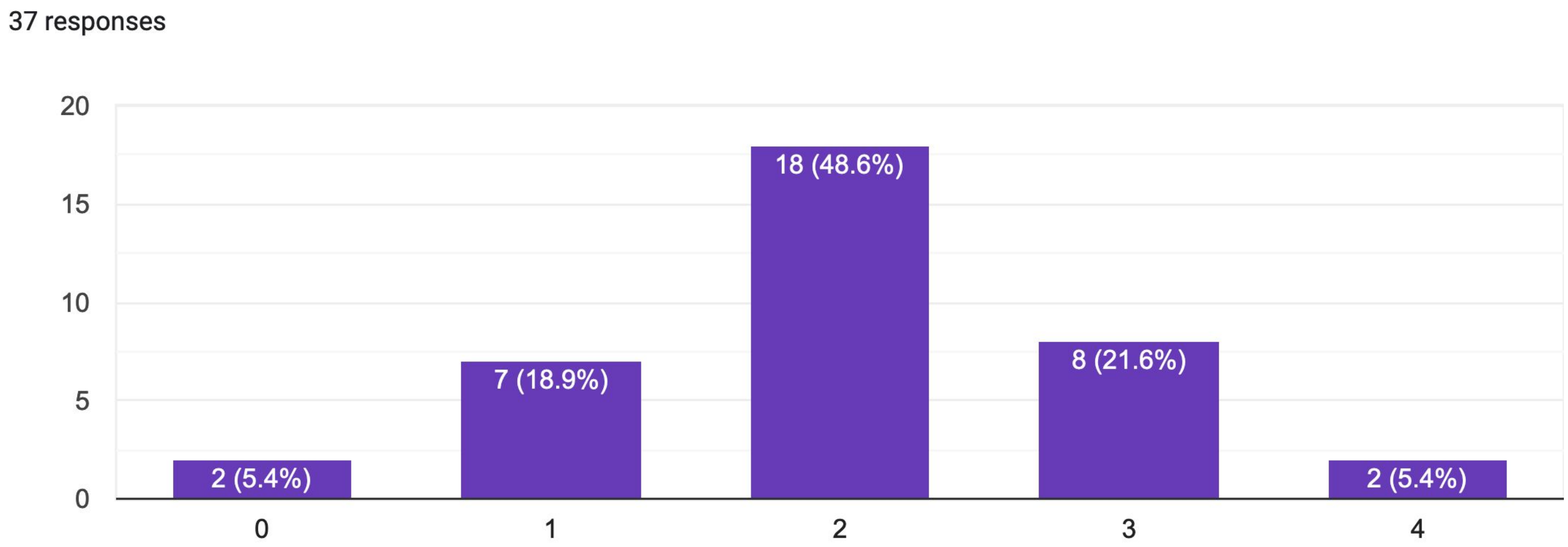
- Repeated measures
- One independent categorical variable with several levels
- One dependent **binary** outcome variable

36 responses



Frequentist tests are confusing

"When you use a frequentist test, how well do you usually feel you understand the test itself?"



Hypothesis testing is your world

“For you, the *most important* component of a **statistical** analysis's results is…”

36 responses

Statistics as posterior inference

Posterior inference?

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{p(D)}$$

Posterior inference?

“How credulous *should* we be of our parameters, given our observations?”

“How surprising were our observations, given some parameters?”

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{p(D)}$$

“How surprising were our observations under *any* parameters?”

“How credulous were we of some parameters?”

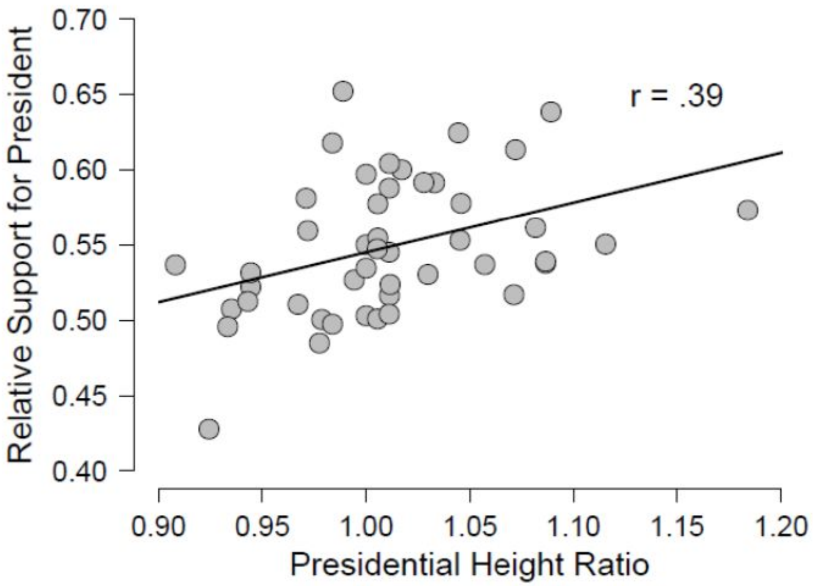
Presidential heights, revisited

Recall from part 1:

~~Is there any evidence for a~~

What is the correlation between the heights and popular vote share of U.S. presidents?

$$P(\rho|D) = P(D|\rho)\frac{P(\rho)}{P(D)}$$



Presidential heights, revisited

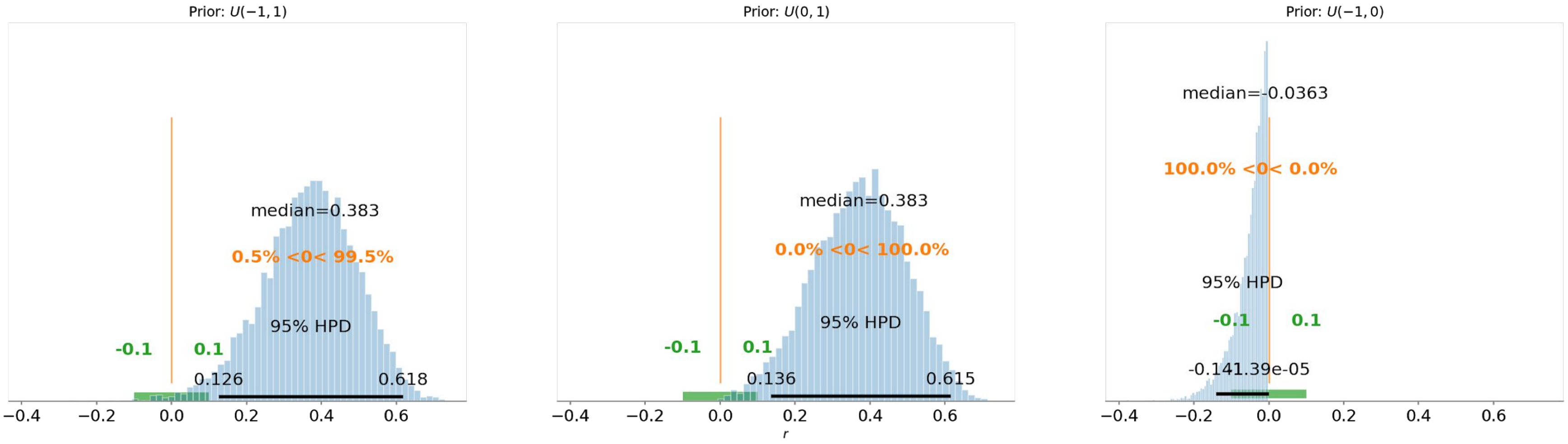
$$
\begin{aligned}
\mu_{height} &\sim N(\mu = 1, \sigma = .1) \\
\mu_{vote} &\sim Beta(\mu = .5, \sigma = .25) \\
\sigma_{height} &\sim HalfCauchy(\sigma = 2.5) \\
\sigma_{vote} &\sim HalfCauchy(\sigma = 2.5) \\
\rho &\sim U(l, b)
\end{aligned}
$$

$$\sigma = \begin{bmatrix} \sigma_{height} & \sigma_{vote} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \odot (\sigma \times \sigma^T)$$

$$\mu = \begin{bmatrix} \mu_{height} & \mu_{vote} \end{bmatrix}$$

$$height, vote \sim N(\mu, \Sigma)$$

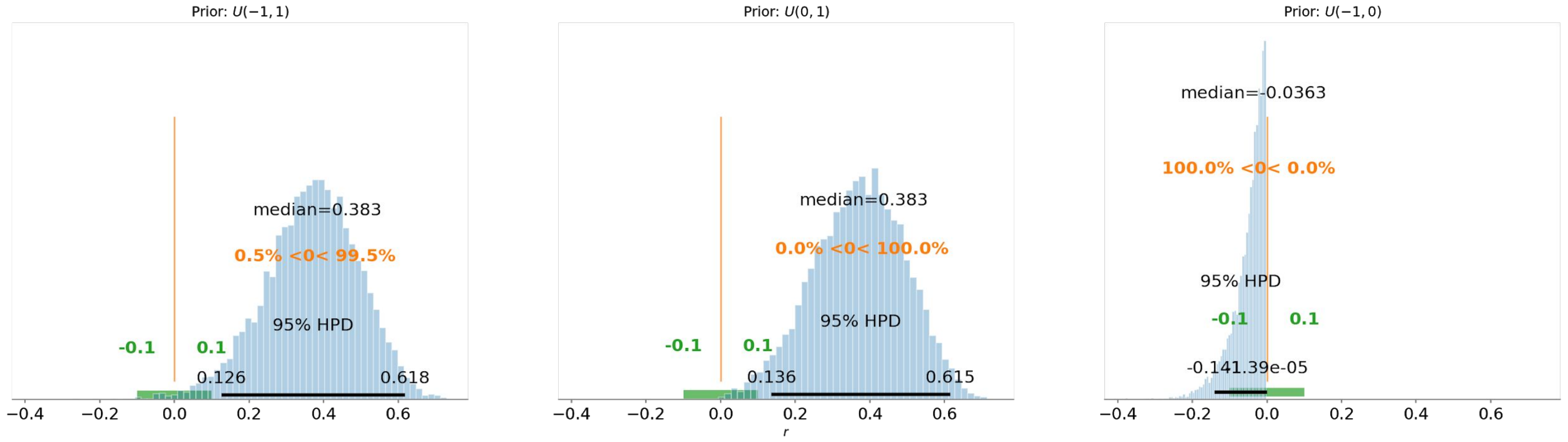


Presidential heights, revisited

$$\mu_{height} \sim N(\mu = 1, \sigma = .1)$$
$$\mu_{vote} \sim Beta(\mu = .5, \sigma = .25)$$
$$\sigma_{height} \sim HalfCauchy(\sigma = 2.5)$$
$$\sigma_{vote} \sim HalfCauchy(\sigma = 2.5)$$
$$\rho \sim U(l, b)$$

$$\sigma = \begin{bmatrix} \sigma_{height} & \sigma_{vote} \end{bmatrix}$$
$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \odot (\sigma \times \sigma^T)$$
$$\mu = \begin{bmatrix} \mu_{height} & \mu_{vote} \end{bmatrix}$$
$$height, vote \sim N(\mu, \Sigma)$$

Bivariate normal likelihood

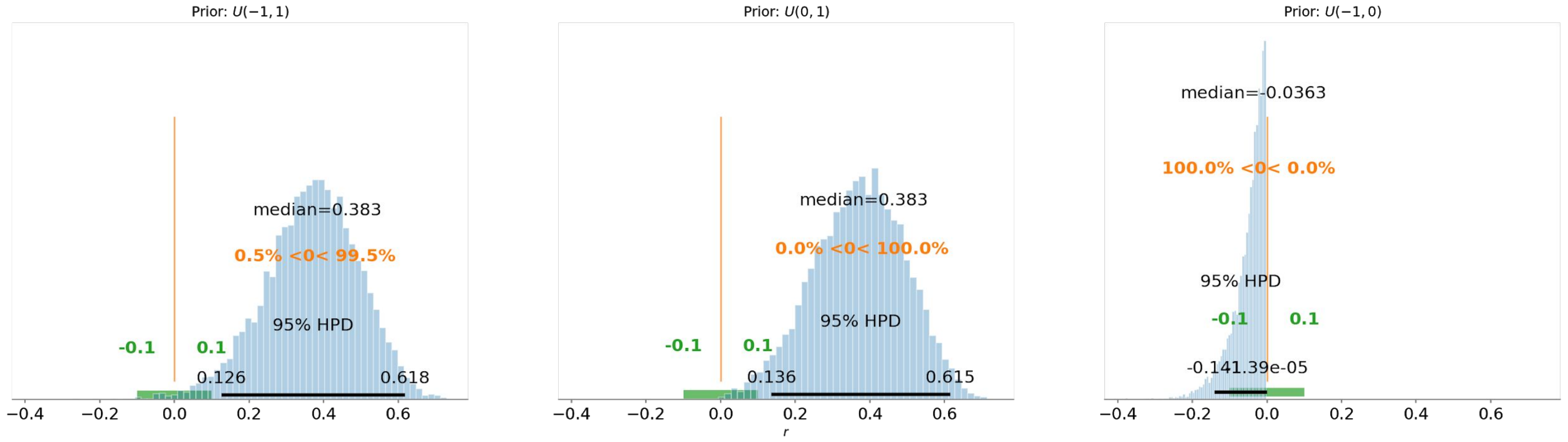


Presidential heights, revisited

Prior on correlation coefficient

$$\mu_{height} \sim N(\mu = 1, \sigma = .1)$$
$$\mu_{vote} \sim Beta(\mu = .5, \sigma = .25)$$
$$\sigma_{height} \sim HalfCauchy(\sigma = 2.5)$$
$$\sigma_{vote} \sim HalfCauchy(\sigma = 2.5)$$
$$\rho \sim U(l, b)$$

$$\sigma = \begin{bmatrix} \sigma_{height} & \sigma_{vote} \end{bmatrix}$$
$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \odot (\sigma \times \sigma^T)$$
$$\mu = \begin{bmatrix} \mu_{height} & \mu_{vote} \end{bmatrix}$$
$$height, vote \sim N(\mu, \Sigma)$$



Presidential heights, revisited

Priors on mean height and vote margin

$$\left\{\begin{aligned} \mu_{height} &\sim N(\mu = 1, \sigma = .1) \\ \mu_{vote} &\sim Beta(\mu = .5, \sigma = .25) \\ \sigma_{height} &\sim HalfCauchy(\sigma = 2.5) \\ \sigma_{vote} &\sim HalfCauchy(\sigma = 2.5) \end{aligned}\right.$$

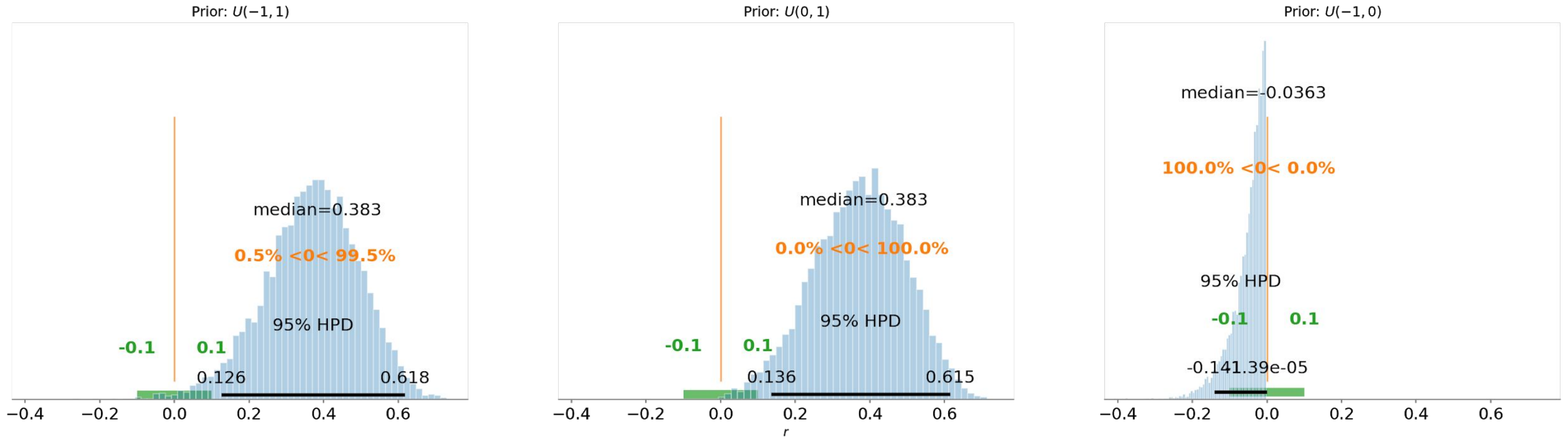
$$\rho \sim U(l, b)$$

$$\sigma = \begin{bmatrix} \sigma_{height} & \sigma_{vote} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \odot (\sigma \times \sigma^T)$$

$$\mu = \begin{bmatrix} \mu_{height} & \mu_{vote} \end{bmatrix}$$

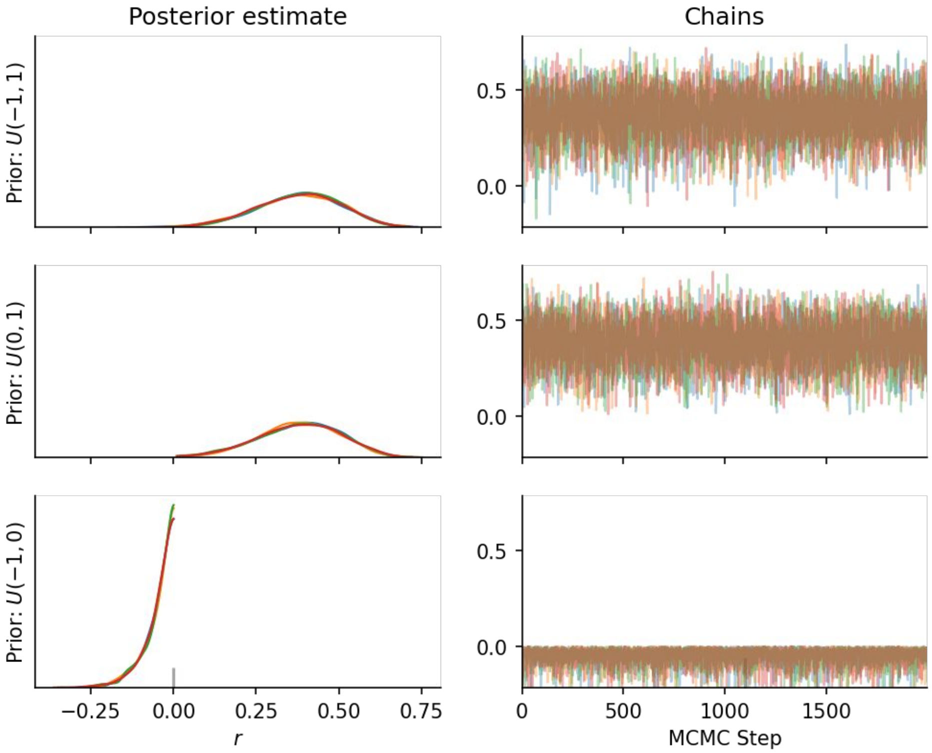
$$height, vote \sim N(\mu, \Sigma)$$



Computing posteriors

Monte Carlo methods sample from the posterior.

Modern samplers don't require much fussing.



Presidential Heights in PyMC3

Let’s walk through some code.

Great, but what are we actually doing?

Not hypothesis testing.

“What hypothesis?”

The posterior distribution is a full report for the parameter of interest.

Do scientists always need to test hypotheses?

Hypothesis testing is a **decision procedure**.

It’s used to produce binary decisions from **uncertain beliefs**.

Do scientists need to produce decisions?

Subjective Bayes and science

Probability theory **extends classical logic** to ideal reasoning under uncertainty.[1]

$$\frac{\begin{array}{r} A \rightarrow B \\ \neg B \end{array}}{\neg A}$$

$$\Big\downarrow$$

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

[1] See Jaynes 2003 for more rigor

Weakly informative priors as null “hypotheses”

(Sort of.)

A **weakly informative prior** centered around a “null” value:
the evidence must be sufficiently stronger than the prior.

Example: Ridge regression → Gaussian prior

Example: Ridge regression

MAP estimate with Gaussian prior

==

Likelihood function + L2 regularization

$$\beta \sim N(0, \sigma_\beta)$$

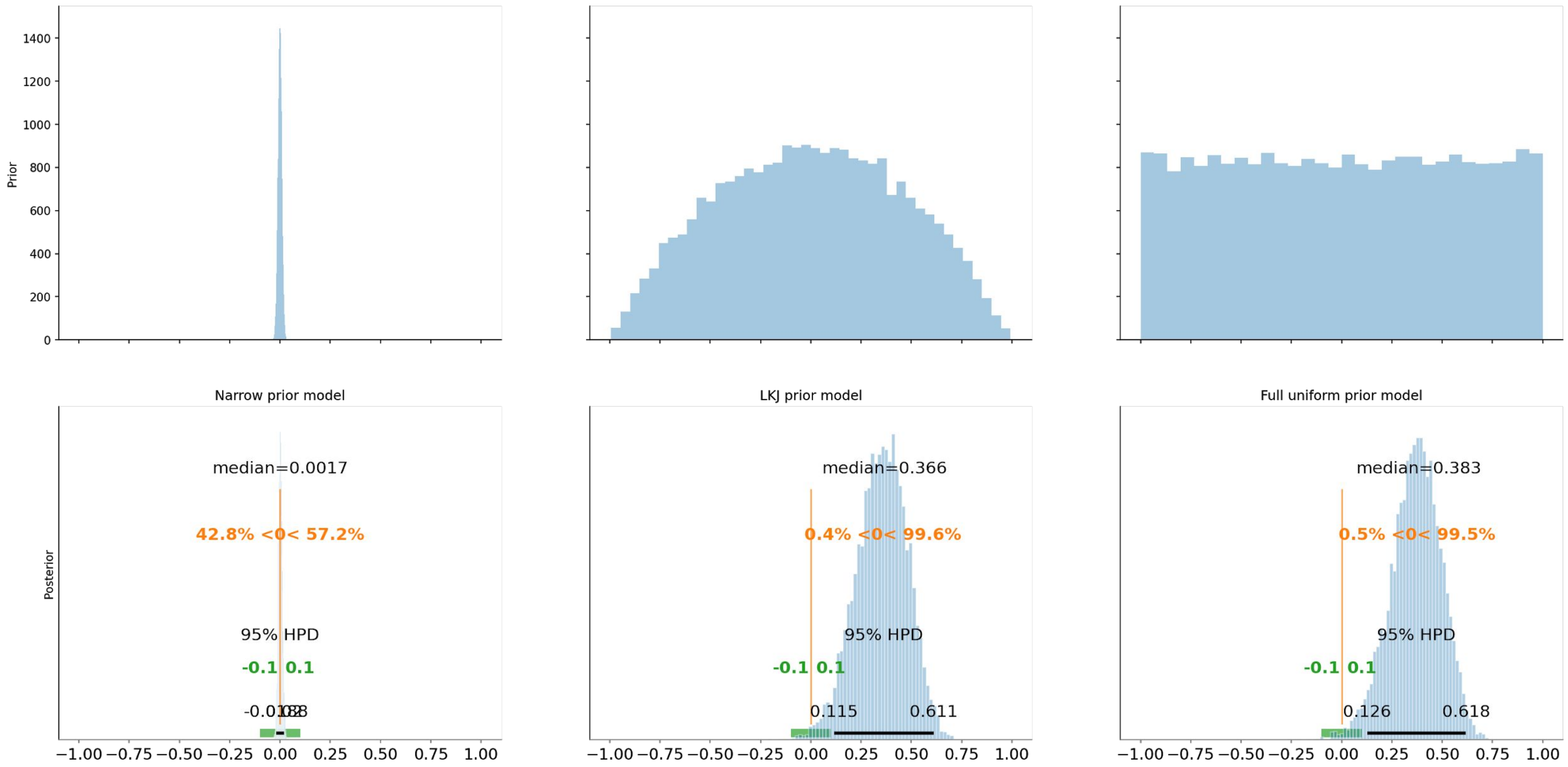
$$y_i \sim N(\beta x_i, \sigma)$$

$$p(\beta|\vec{y}, \vec{x}) \propto p(\beta) \prod_i^N p(y_i|\beta, x_i)$$

$$\operatorname*{argmax}_\beta \sum_i^N \log p(y_i|\beta, x_i) + \log p(\beta)$$

$$\operatorname*{argmin}_\beta \frac{1}{\sigma} \sum_i^N (y_i - \beta x_i)^2 + \frac{1}{\sigma_\beta} \beta^2$$

Presidential heights: informative priors

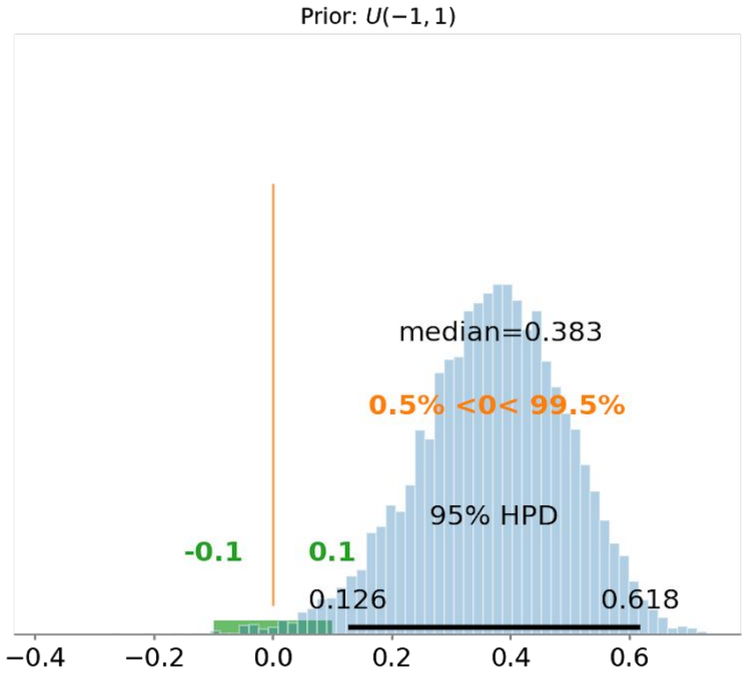


Still want to reject a null?

We can do it with posterior estimates.

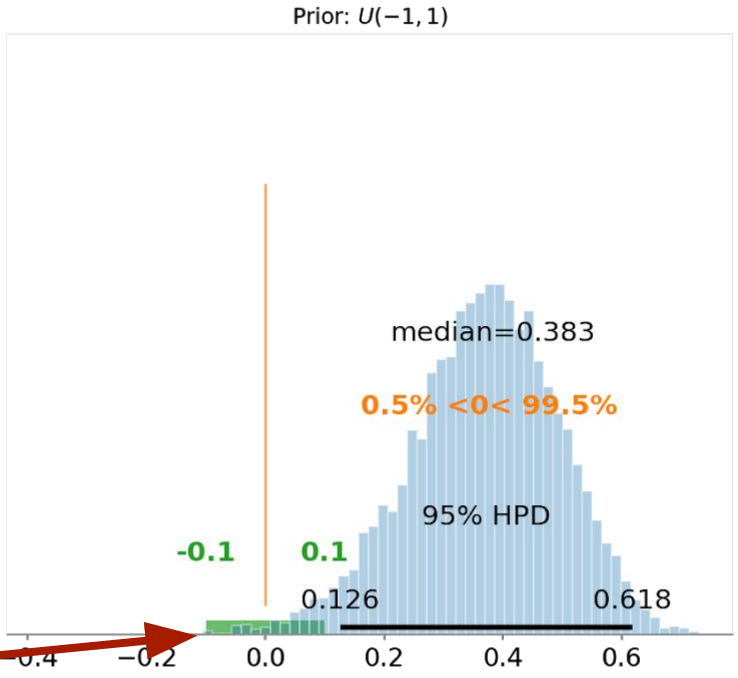
Rejecting a null "hypothesis"

Does the **ROPE** around a null value lie entirely outside the **HPDI**?



Rejecting a null “hypothesis”

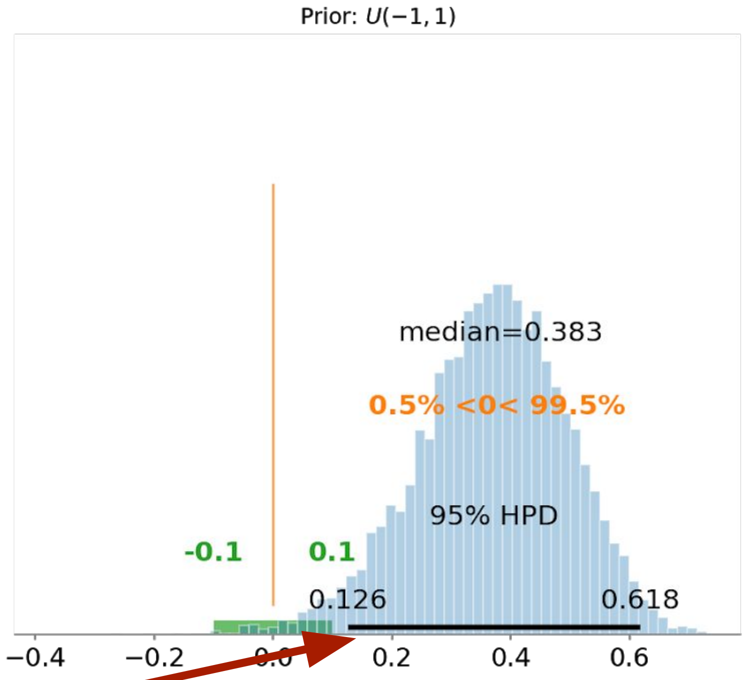
Does the **ROPE** around a null value lie entirely outside the **HPDI**?



Region Of Practical Equivalence

Rejecting a null “hypothesis”

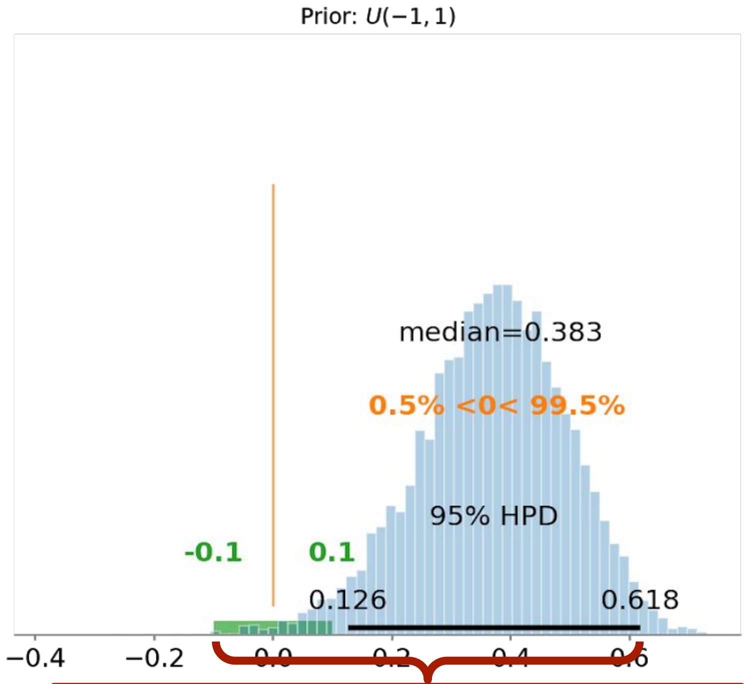
Does the **ROPE** around a null value lie entirely outside the **HPDI**?



Highest **P**robability **D**ensity **I**nterval

Rejecting a null “hypothesis”

Does the **ROPE** around a null value lie entirely outside the **HPDI**?

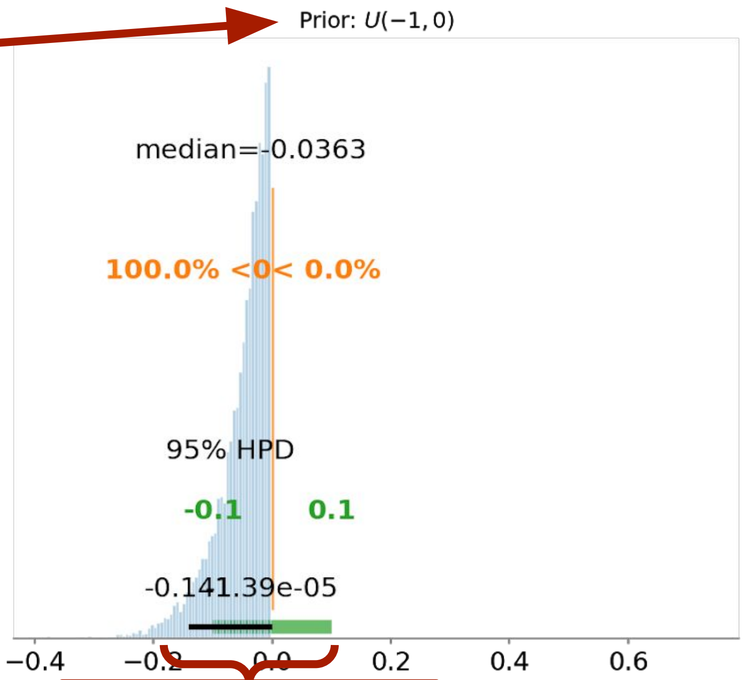


No overlap?
Reject the null “hypothesis”!

What does non-rejection look like?

Strongly misinformed prior

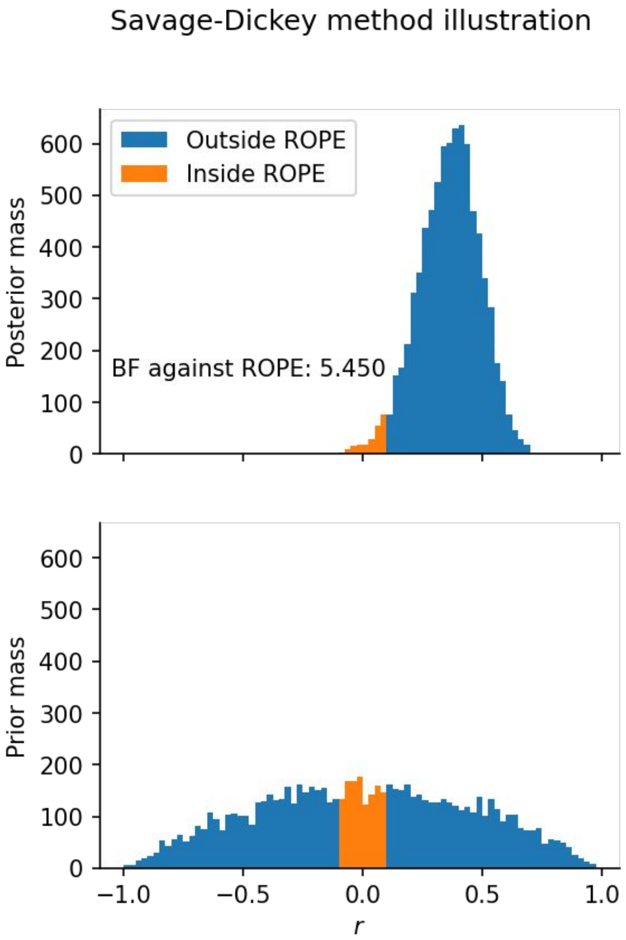
Does the **ROPE** around a null value lie entirely outside the **HPDI**?



Partial overlap?
Cannot reject null.

Savage-Dickey method

$$BF_{01} = \frac{P(D|\theta \neq \theta_{null})}{P(D|\theta = \theta_{null})} = \frac{P(\theta \neq \theta_{null}|D)}{P(\theta \neq \theta_{null})}$$



Differences between BF and PE

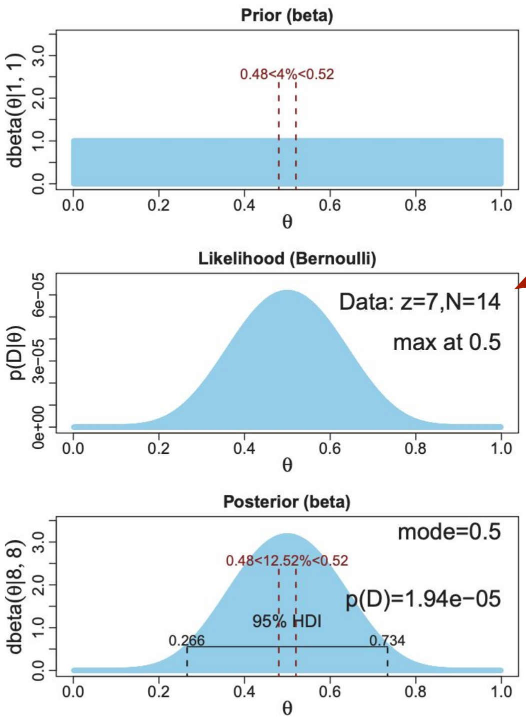
Bayes' factor

- Hypothesis testing as **model comparison**
- H_0 either a parameter value or a model
- Imitate frequentist tests and assumptions

Parameter estimation

- "What hypothesis?"
- Report full credibilities of magnitudes
- Model is explicit

Bayes' factors: YMMV



Not a ton of data

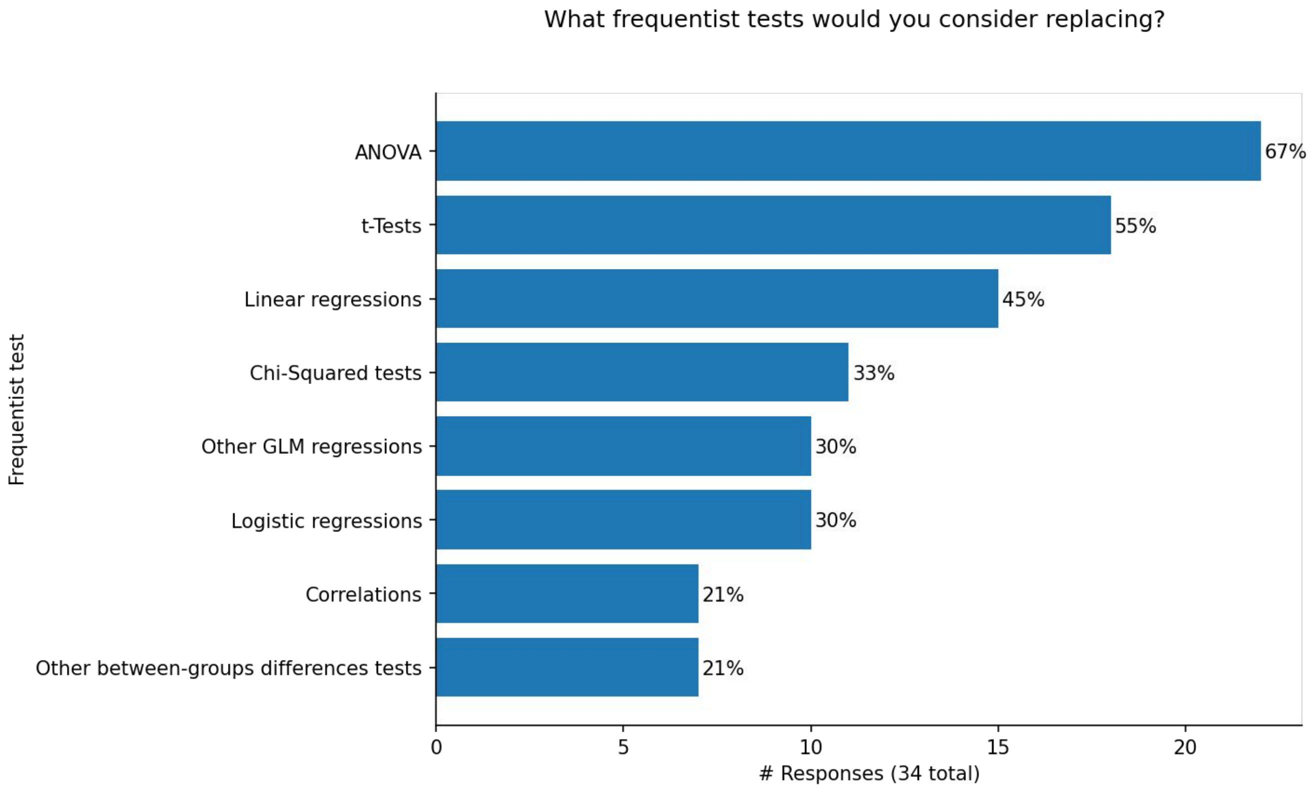
HDPI: (.27, .73)
BF: 3.14 *for* θ=.5

Kruschke 2015

ANOVA

By popular demand :)

ANOVA tops the list



What is ANOVA for, anyway?

Technically, decomposing sources of variance

Loosely, often used to answer: “Did X influence Y?”

ANOVA is hypothesis testing and **only makes sense in that framework**

Standard ANOVA relies on a linear model

- "Explained" variance ratio
    - → *F*-statistic
- Group means == coefficients
    - → homogeneity of variance
    - → normality of outcomes
- "How likely is my *F* given $\beta = 0$?"
    - → *p*-value

$$\underbrace{\sum_{i=1}^{n}((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2}_{\text{explained}} + \underbrace{\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2}_{\text{unexplained}}.$$

$$F = \frac{\sum_{i=1}^{n}((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2/1}{\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2/(n-2)}.$$

CV Answer 1; CV Answer 2

Standard ANOVA relies on a $\boxed{\text{linear model}}$

- "Explained" variance ratio
  - → *F*-statistic
- Group means == coefficients
  - → homogeneity of variance
  - → normality of outcomes
- "How likely is my *F* given $\boldsymbol{\beta} = \mathbf{0}$?"
  - → *p*-value

$$\underbrace{\sum_{i=1}^{n}((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2}_{\text{explained}} + \underbrace{\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2}_{\text{unexplained}}.$$

$$F = \frac{\sum_{i=1}^{n}((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2/1}{\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2/(n-2)}.$$

CV Answer 1; CV Answer 2

Categorical variable coding example

Level	Dummy Codes	
1	1	0
2	0	1
3	0	0

Level	Sum Codes	
1	1	0
2	0	1
3	-1	-1

Standard ANOVA relies on

What you **actually** care about:
Model coefficients.

$$\underbrace{\sum_{i=1}^{n}((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2}_{\text{explained}} + \underbrace{\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2}_{\text{unexplained}}.$$

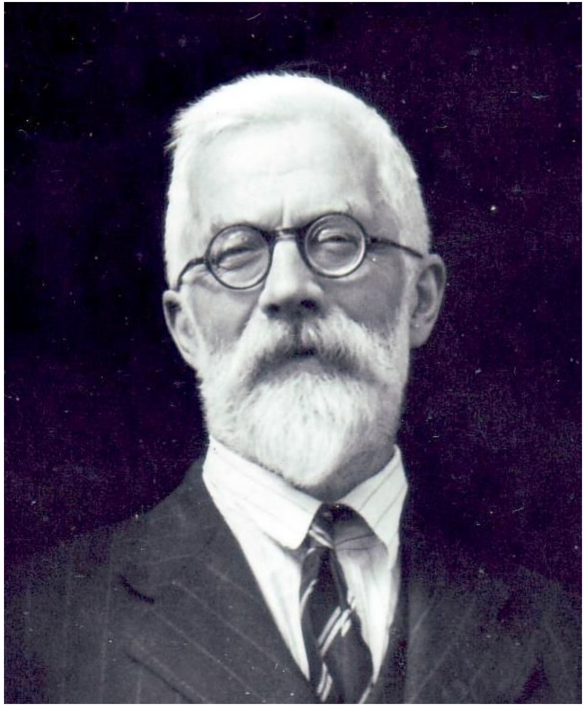
$$F = \frac{\sum_{i=1}^{n}((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2/1}{\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2/(n-2)}.$$

- “Explained” variance ratio
  - → *F*-statistic
- Group means == coefficients
  - → homogeneity of variance
  - → normality of outcomes
- “How likely is my *F* given $\boldsymbol{\beta} = \mathbf{0}$?”
  - → *p*-value

CV Answer 1; CV Answer 2

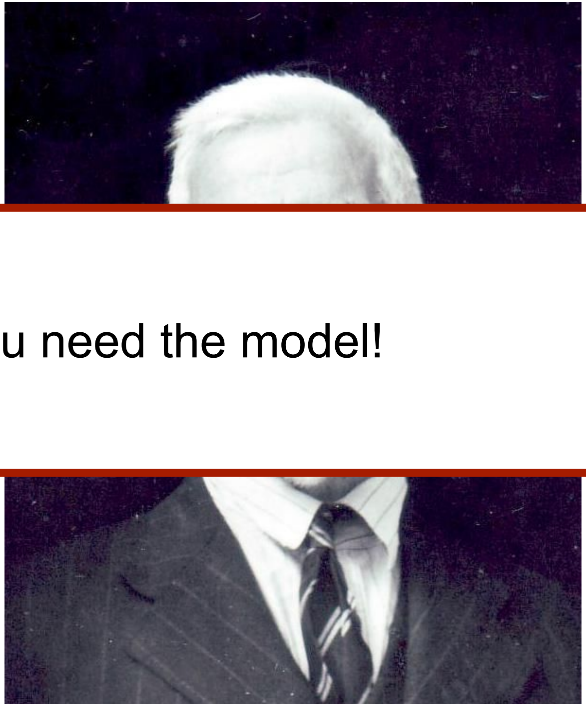
Who gives a Fischer?

- Doesn't tell us about direction
- Ditto relationship *strength*
- Assumptions often don't obtain
- Complicated designs are "fun"



Who gives a Fischer?

- Doesn't tell us about direction
- Ditto relationship *strength*
- Assumptions often don't obtain
- Complicated designs are "fun"

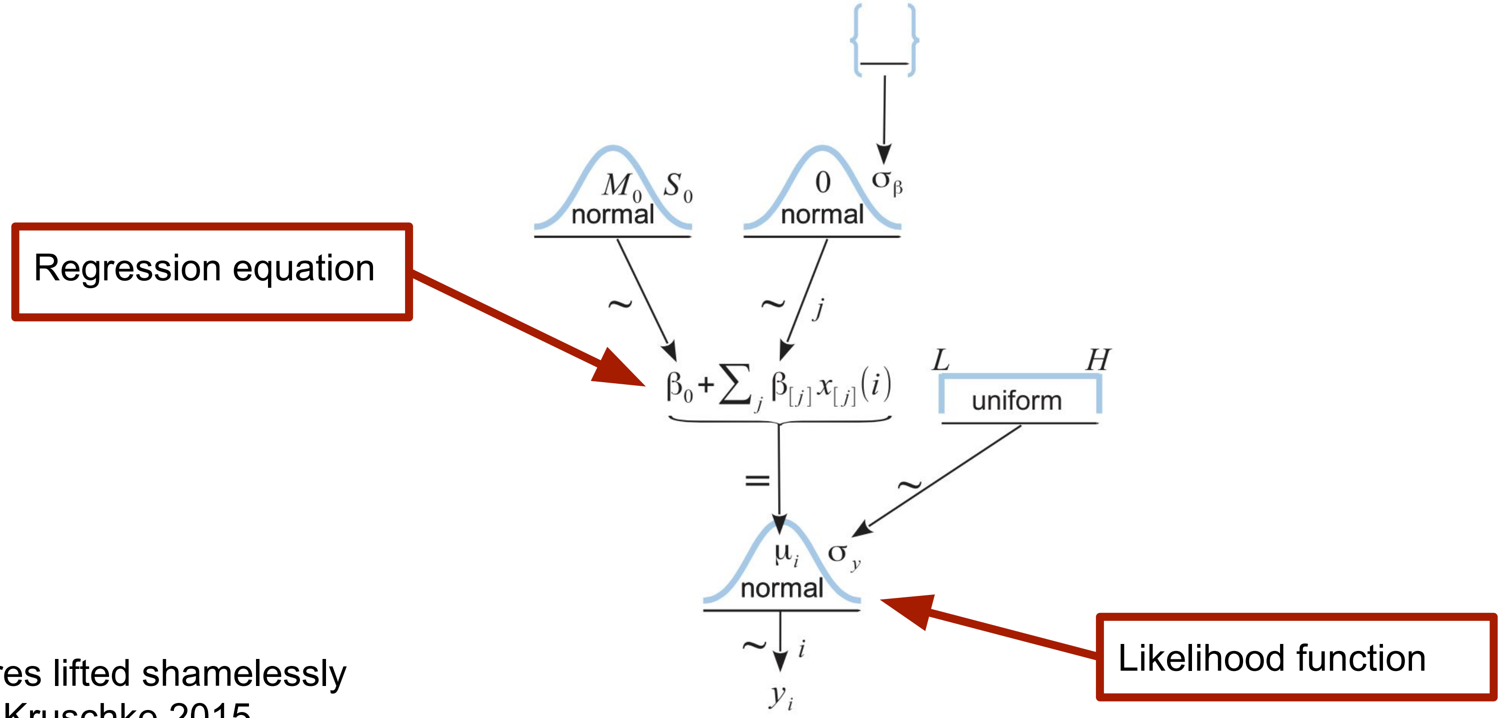


You need the model!

Everything you usually want to know is in the parameters of an appropriate regression model

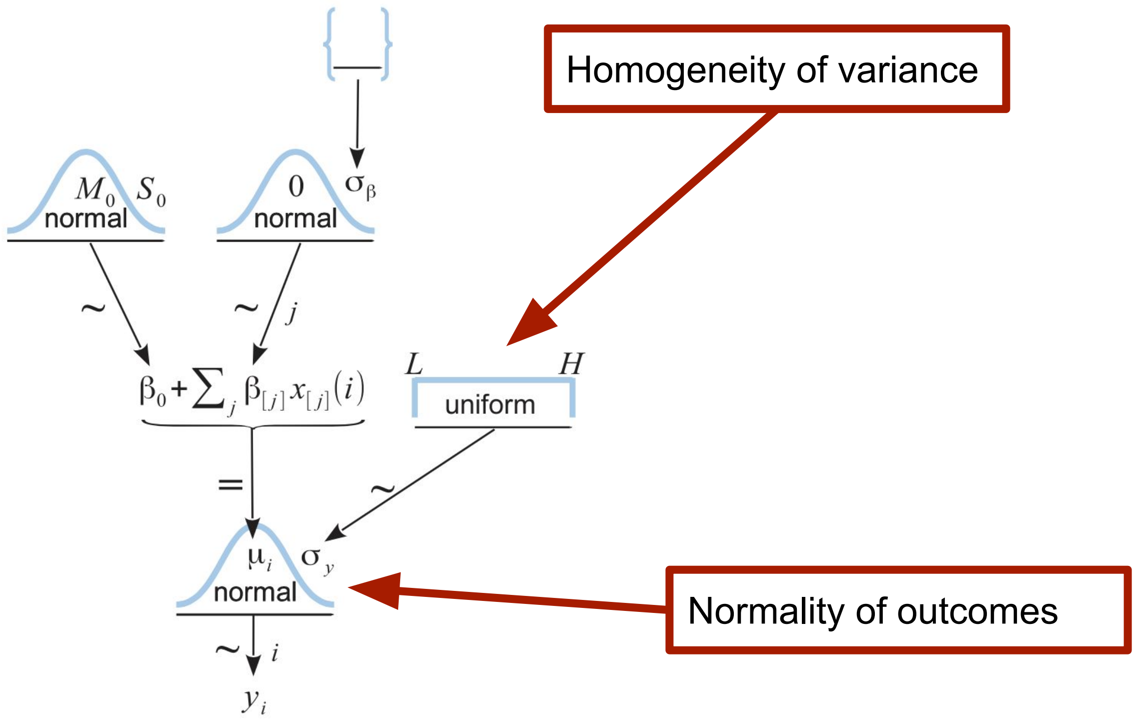
So let’s do that.

Regression with categorical predictors



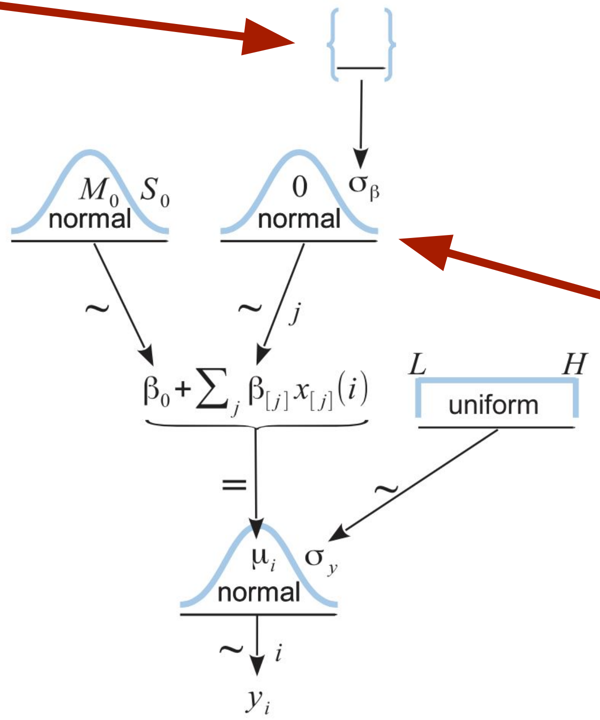
Figures lifted shamelessly
from Kruschke 2015

Regression: ANOVA assumptions



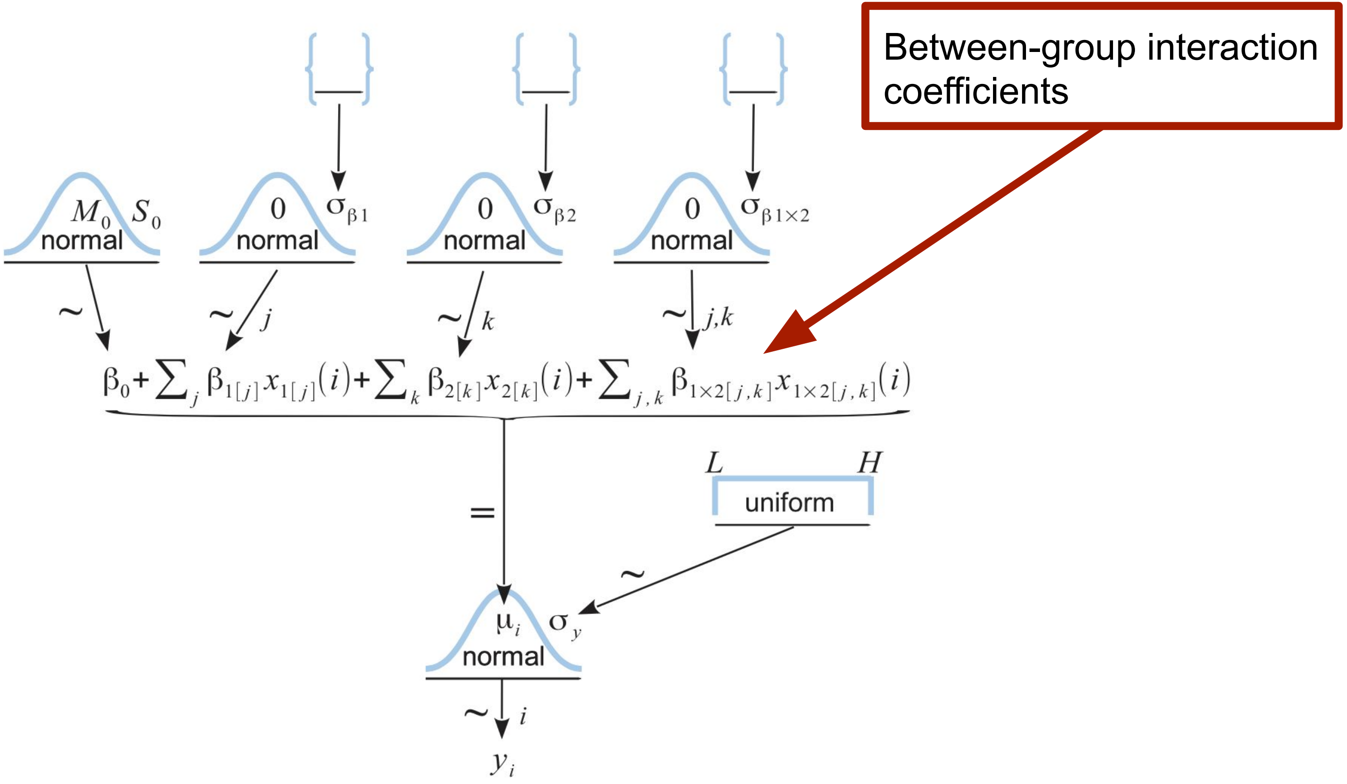
Regression: Pooled priors and shrinkage

Hyperprior on coefficient scales

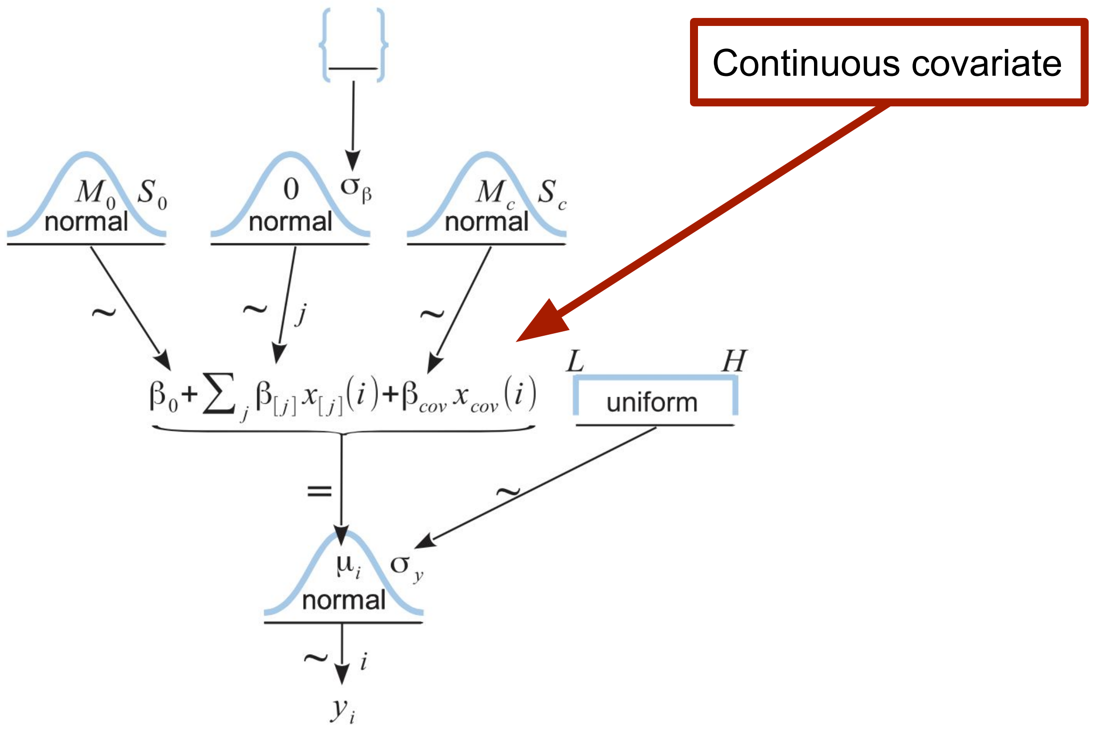


Pooled prior on coefficients

Extension: Multiple categorical variables

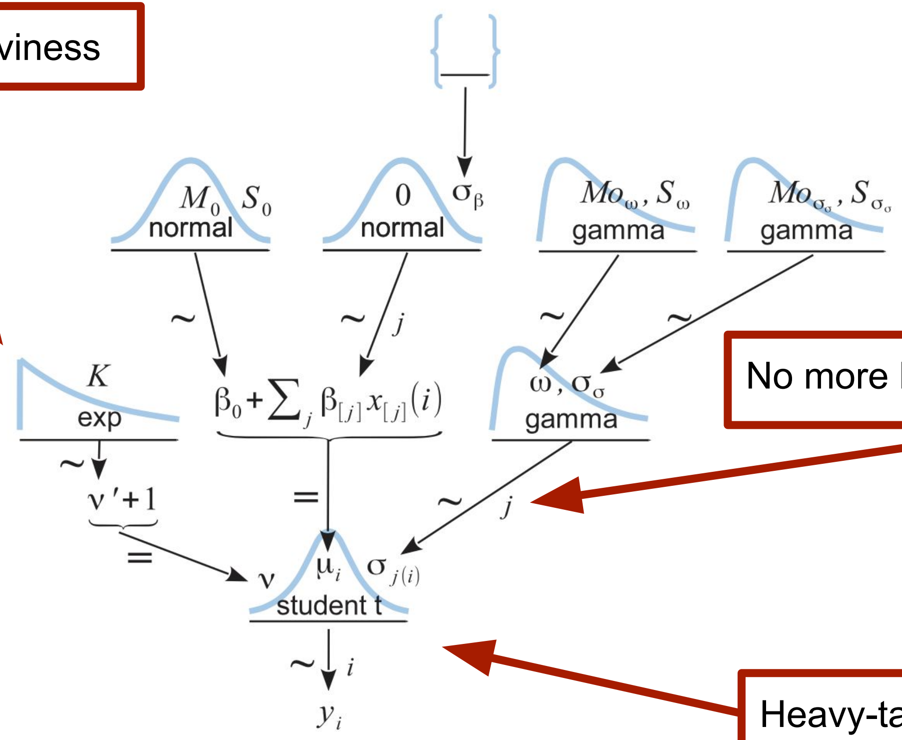


Extension: "Mixed effects" / ANCOVA



Extension: "Robust" errors model

Prior over degree of tail-heaviness



No more homogeneity of variance!

Heavy-tailed likelihood function

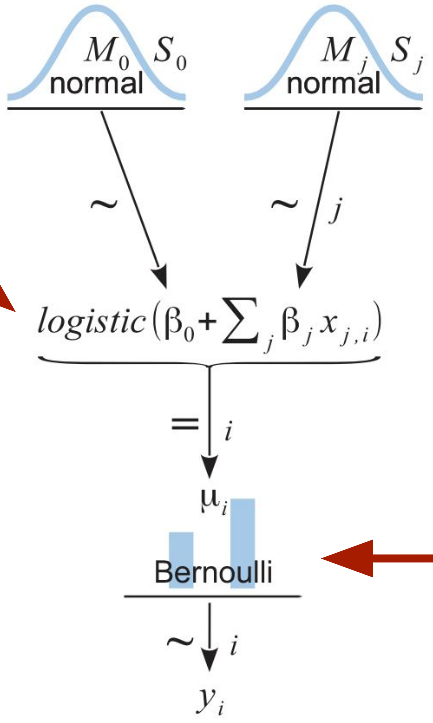
Extension: Bayesian logistic regression

Logistic regression

$$P(y_i) = logistic(\beta x_i) = \frac{1}{1 + e^{\beta x_i}}$$

==

$$\frac{\log P(y_i)}{\log P(1 - y_i)} = \beta x_i$$



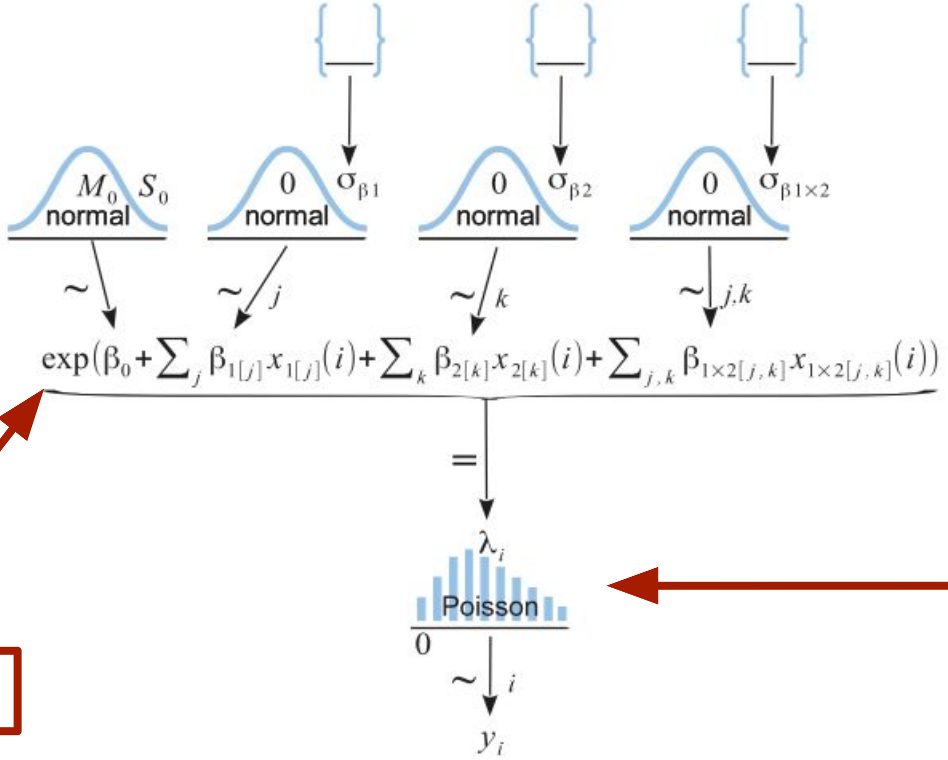
Set M_j = 0 and (maybe) put a prior on S_j for ANOVA-like model

Bernoulli likelihood

Change detection task

Recall task

Extension: Poisson regression



Note the exponent!

Poisson likelihood

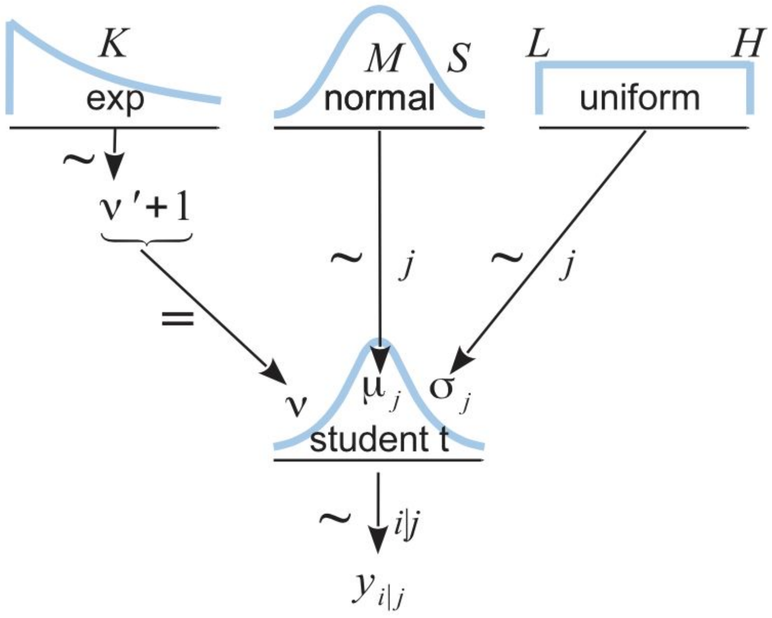
ETC

etc

What about t-tests?

Bayesian **E**stimation **S**upersedes the **T**-test

BEST model



Observations *i*

Groups *j*

Difference between groups 0 and 1: $\mu_1 - \mu_0$

Effect size: $(\mu_1 - \mu_0) / \sqrt{((\sigma_1^2 + \sigma_0^2) / 2)}$

Final thoughts

Prior choices?

Much ink already spilled on this topic.

Rule of thumb: **weakly informative prior for a skeptical audience.**

Other uses!

- Cognitive process models
    - ex: Kruschke & Vanpaemel 2015)
- Dynamical models
    - ex: SEIR, Lotka-Volterra, whatever
- Time series models
- Latent variable models
    - ex: Bayesian matrix factorization, LDA
- Exotic models
    - ex: Gaussian processes, Bayesian neural networks

Further reading

If you get just one book, I recommend this one!

Kruschke, John K. *Doing Bayesian Data Analysis.* 2e, 2015.

Jaynes, E.T. *Probability Theory: The logic of science*. 2003.

Gelman, Andrew. “Analysis of Variance: Why it is more important than ever”. *The Annals of Statistics*. 2005.

Wagenmakers, Eric-Jan; et al. “Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications”. *Psychological Bulletin Review*. 2018.

Appendix

Inference methods: What about SVI?

HMC: Hamiltonian **M**onte **C**arlo

- Most packages implement variants of the **N**o **U**-**T**urn **S**ampler (**NUTS)**
- Works well without hand-tuning for a variety of common models
- Slow when there is lots of data (>1-10k observations)

SVI: Stochastic **V**ariational **Inference**

- **Approximate** posterior by minimizing ELBO objective
- Many "automatic" posterior guides available
- Scales well and performs better on large datasets (>1-10k observations)

Python packages

2020 June	**PyMC3**	**PyMC4**	**Pyro**	**NumPyro**	**(py)STAN**
Stability	Mature	Pre-release	Mature	Development	Mature
Features	👍👍👍	👍👍	👍👍👍	👍👍	👍👍👍
Future development	❌	✔️	✔️	✔️	✔️
Backend	Theano	Tensorflow	PyTorch	JAX	C++
MCMC Speed	🐪🐪	🐪🐪🐪?	🐢	🐪🐪🐪🐪	🐪🐪🐪
SVI Speed	🐎🐎	🐎🐎🐎?	🐎🐎🐎🐎	🐎🐎🐎?	🐎🐎🐎?
GPU support	Model only	Yes?	Model only	Yes	Nope

Slide Graveyard