A gentle introduction to the case for Bayesian data analysis

Thom Baguley
Psychology
Nottingham Trent University
thomas.baguley@ntu.ac.uk

Abstract

After a long period of controversy, the use of Bayesian data analysis (BDA) in psychology has reached a tipping point. Psychology - as with many other sciences - is beginning to accept that Bayesian methods offer an important and useful set of tools for answering certain questions and dealing with complex, messy real-world data. One implication is that in future, all psychologists - even those expert in and happily using classical, frequentist methods - need to have a basic Bayesian grounding in order to keep abreast of a literature where BDA is commonplace. Furthermore, it may become increasingly difficult to publish in the top empirical journals without adding some Bayesian methods to one's toolbox.

Thanks to ...

Mark Andrews

Jens Roeser

ESRC (Advanced Training Initiative)

Overview

1. The turn to Bayes in Psychology
2. So ... how does Bayesian data analysis work?
3. A simple illustrative example: *predicting the US presidential election*
4. Another simple example: *finding evidence to support the null hypothesis*
5. Using Bayesian tools: *inference in multilevel models*
6. Complex Bayesian models: *cognitive processes in language production*
7. Some amateur futurology

Psychology is embracing Bayesian data analysis (again)

On taking up the post of editor of the *Journal of Experimental Psychology: General* - arguably the most prestigious outlet for experimental work in psychology - the editor announced that the journal

> *will seek to attract work in specialized areas critical to the development of our science, for example [...] the adoption of Bayesian methods in data analysis.*

[Gauthier, 2012]

What is wrong with classical (frequentist) methods?

Leaving aside philosophical problems (e.g., that inference is based on unobserved data) here are a few of the things that have made me reconsider frequentist methods ...

- p values (and related quantities) are of limited informational value

What is wrong with classical (frequentist) methods?

Leaving aside philosophical problems (e.g., that inference is based on unobserved data) here are a few of the things that have made me reconsider frequentist methods ...

- p values (and related quantities) are of limited informational value
- null hypothesis tests are ill-suited to certain problems

What is wrong with classical (frequentist) methods?

Leaving aside philosophical problems (e.g., that inference is based on unobserved data) here are a few of the things that have made me reconsider frequentist methods ...

- p values (and related quantities) are of limited informational value
- null hypothesis tests are ill-suited to certain problems

  e.g., quantifying support *for* a null hypothesis

What is wrong with classical (frequentist) methods?

Leaving aside philosophical problems (e.g., that inference is based on unobserved data) here are a few of the things that have made me reconsider frequentist methods ...

- p values (and related quantities) are of limited informational value
- null hypothesis tests are ill-suited to certain problems

  e.g., quantifying support *for* a null hypothesis
- relevant external information enters a frequentist analysis in an *ad hoc* way

So ... how does Bayesian data analysis work?

A Bayesian analysis involves three things:

1. A probability model for the data (the likelihood)

So ... how does Bayesian data analysis work?

A Bayesian analysis involves three things:

1. A probability model for the data (the likelihood)
2. A probability model of relevant prior information (the prior)

So ... how does Bayesian data analysis work?

A Bayesian analysis involves three things:

1. A probability model for the data (the likelihood)
2. A probability model of relevant prior information (the prior)
3. A probability model that *combines* the data and the prior information (the posterior)

Normal distribution example

Does a single exposure to an US flag influence voting intention?

Carter et al. (2011) report a priming study in which a single exposure to a US flag changes voting intentions (making them more Republican)

They report an 'effect size' of over a quarter of a standard deviation:

$$\hat{\mu}_E - \hat{\mu}_C = 0.142, \quad t_{181} = 2.02, \quad p = .04 \quad d = 0.298$$

Normal distribution example

Does a single exposure to an US flag influence voting intention?

Carter et al. (2011) report a priming study in which a single exposure to a US flag changes voting intentions (making them more Republican)

They report an 'effect size' of over a quarter of a standard deviation:

$$\hat{\mu}_E - \hat{\mu}_C = 0.142, \quad t_{181} = 2.02, \quad p = .04 \quad d = 0.298$$

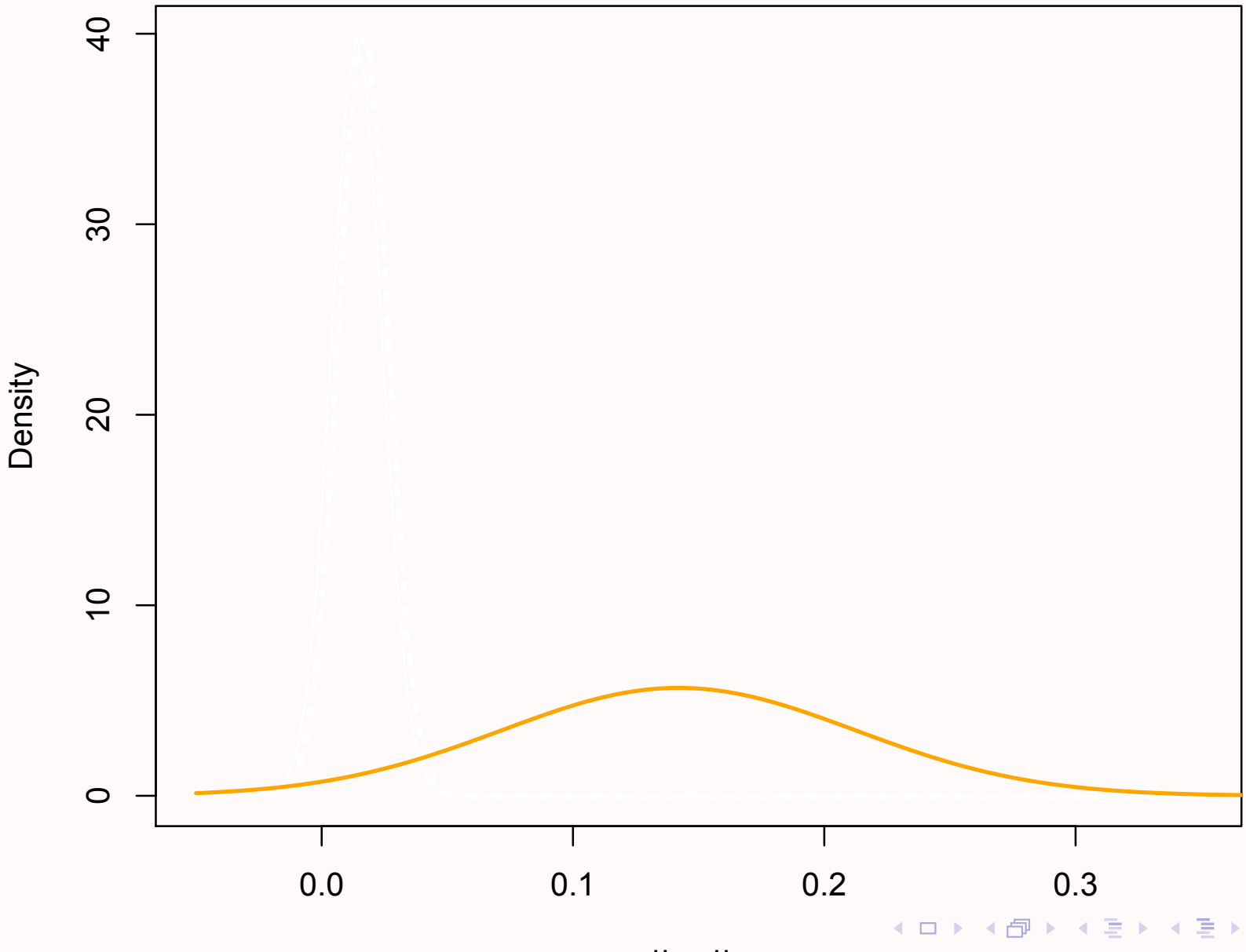
... this seems unrealistically large e.g., being associated with a 10.7% swing

The likelihood

The likelihood is a mathematical function that models the probability of the observed data as a function of a parameter (e.g., a population mean) or a set of parameters

The observed mean is 0.142 with a standard error of 0.07

Assuming a normal distribution for the effect, the likelihood would look like this ...

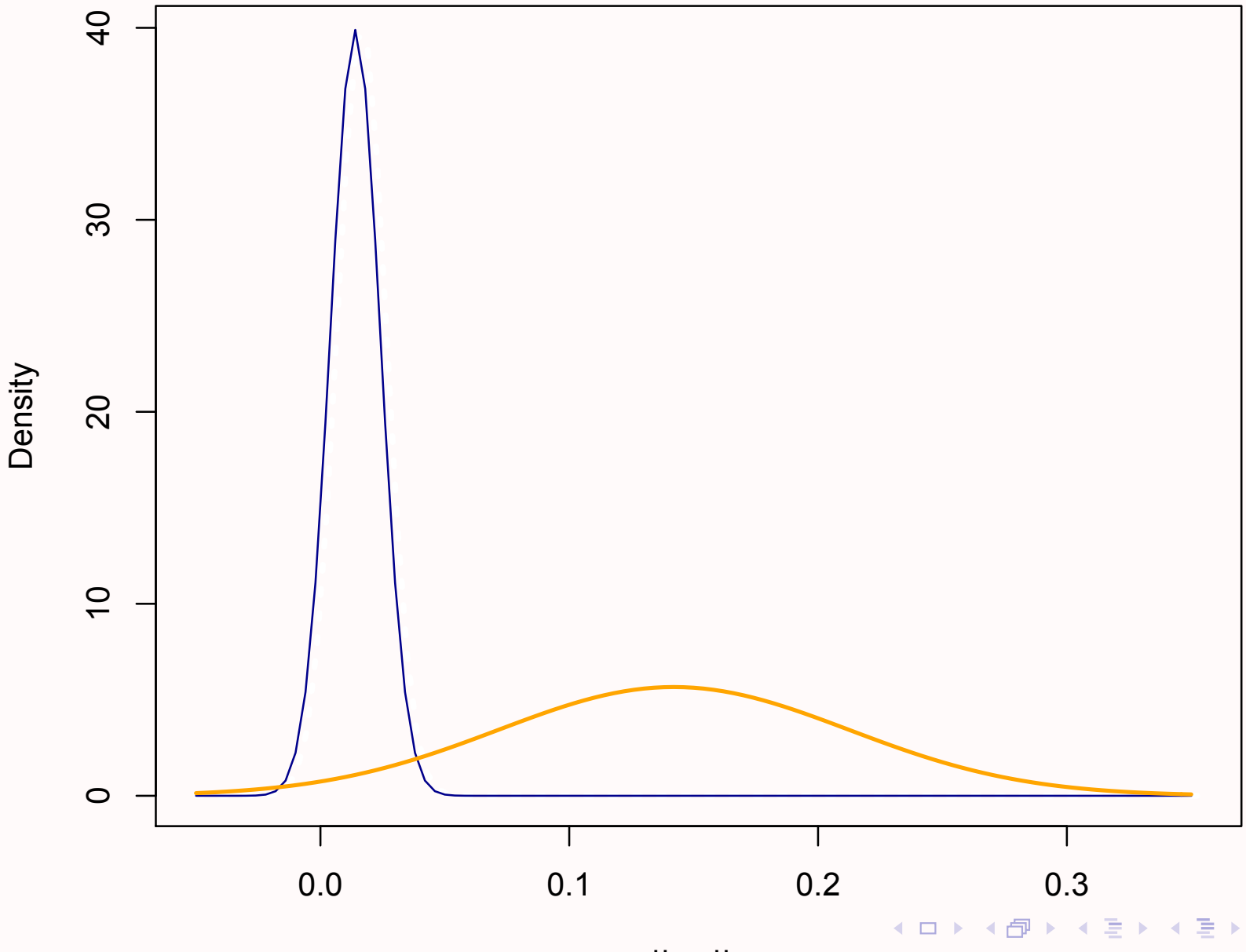

Density
0
10
20
30
40
0.0
0.1
0.2
0.3

The prior

The prior is a probability distribution that reflects relevant information about a parameter (e.g., a population mean)

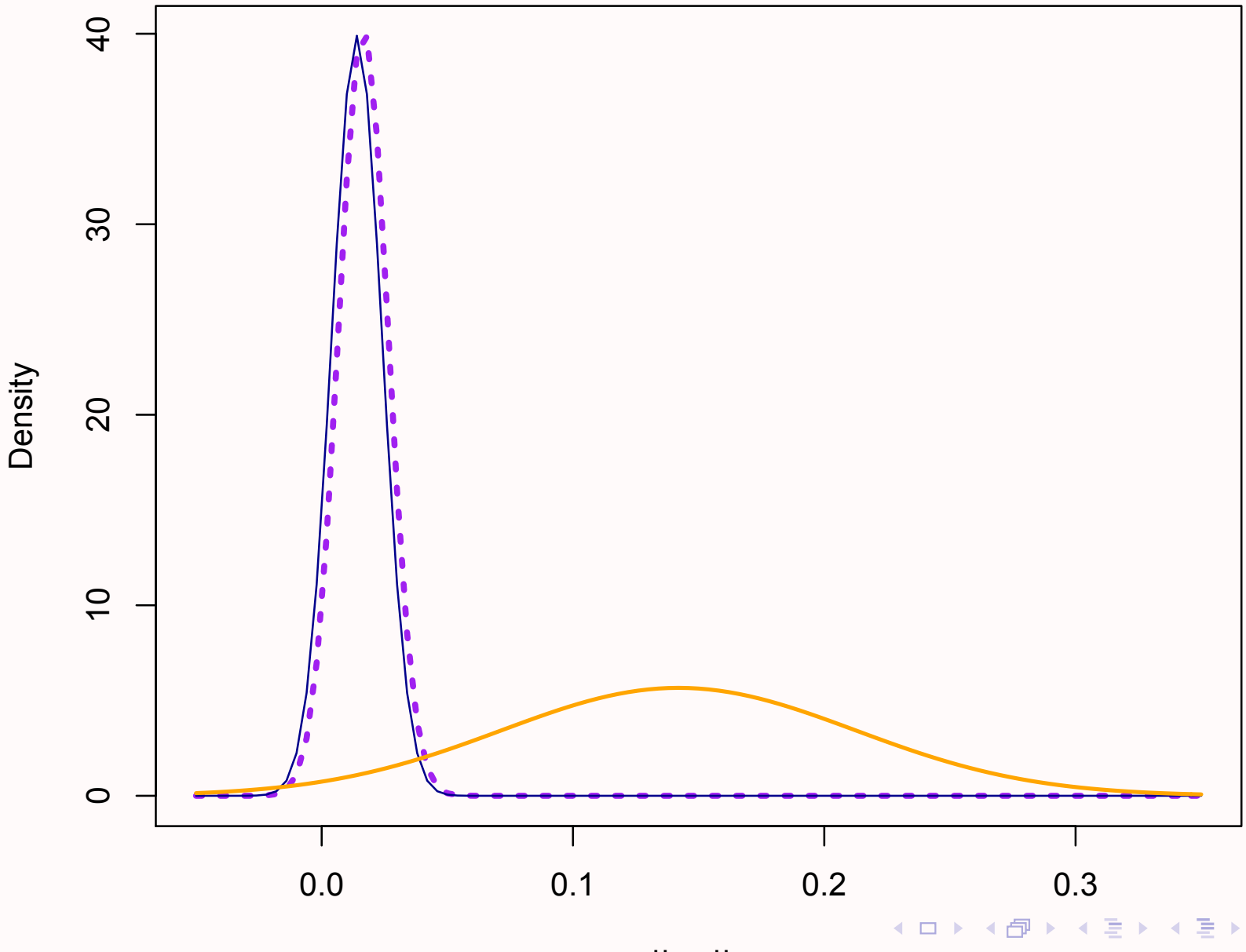
e.g.,
$M = 0.014$ with corresponds to roughly a 1% swing
$SD = 0.01$ implies 95% chance of a swing from -0.5% to 2.5%

Again, assuming a normal distribution for the prior ...


Density
0
10
20
30
40
0.0
0.1
0.2
0.3

Obtaining the posterior using Bayes' theorem ...

$$\overbrace{\mathrm{P}(\theta|\mathcal{D},\Omega)}^{\text{Posterior}} = \frac{\overbrace{\mathrm{P}(\mathcal{D}|\theta)}^{\text{Likelihood}}\overbrace{\mathrm{P}(\theta|\Omega)}^{\text{Prior}}}{\int \mathrm{P}(\mathcal{D}|\theta)\mathrm{P}(\theta|\Omega)d\theta}$$


Density
0
10
20
30
40
0.0
0.1
0.2
0.3

Frequentist confidence interval

$\hat{\mu}_1 - \hat{\mu}_2 = 0.142$, 95% CI [0.003, 0.281], $p = .04$

Frequentist confidence interval

$\hat{\mu}_1 - \hat{\mu}_2 = 0.142$, 95% CI [0.003, 0.281], $p = .04$

Bayesian posterior probability interval and Bayes factor

$\hat{\mu}_1 - \hat{\mu}_2 = 0.0165$, 95% probability interval [-0.003, 0.036]

Frequentist confidence interval

$\hat{\mu}_1 - \hat{\mu}_2 = 0.142$, 95% CI [0.003, 0.281], $p = .04$

Bayesian posterior probability interval and Bayes factor

$\hat{\mu}_1 - \hat{\mu}_2 = 0.0165$, 95% probability interval [-0.003, 0.036] ...
or using default prior on effect size $BF_{10} = 1.05$

Why Bayesian data analysis (BDA)?

- More efficient use of available information

Why Bayesian data analysis (BDA)?

- More efficient use of available information
- The right tool for certain problems

Why Bayesian data analysis (BDA)?

- More efficient use of available information
- The right tool for certain problems

(e.g., assessing evidence in favour of H_0)

Why Bayesian data analysis (BDA)?

- More efficient use of available information
- The right tool for certain problems

  (e.g., assessing evidence in favour of H_0)

- More flexibility in modelling complex, messy (real world) data sets

Why Bayesian data analysis (BDA)?

- More efficient use of available information
- The right tool for certain problems

  (e.g., assessing evidence in favour of H_0)

- More flexibility in modelling complex, messy (real world) data sets
- Greater emphasis on transparency in modelling

A simple illustrative example: *predicting the US presidential election*

- Guam 'straw poll' predicted result every result since 1984
- 2016: 71.3% Clinton vs. 24.3% Trump

A simple illustrative example: *predicting the US presidential election*

- Guam 'straw poll' predicted result every result since 1984
- 2016: 71.3% Clinton vs. 24.3% Trump
- So 8 out 8 elections predicted the result correctly

A simple illustrative example: *predicting the US presidential election*

- Guam 'straw poll' predicted result every result since 1984
- 2016: 71.3% Clinton vs. 24.3% Trump
- So 8 out 8 elections predicted the result correctly
- Naive estimate: that $P(\text{Cinton victory}) = 1$

A simple illustrative example: *predicting the US presidential election*

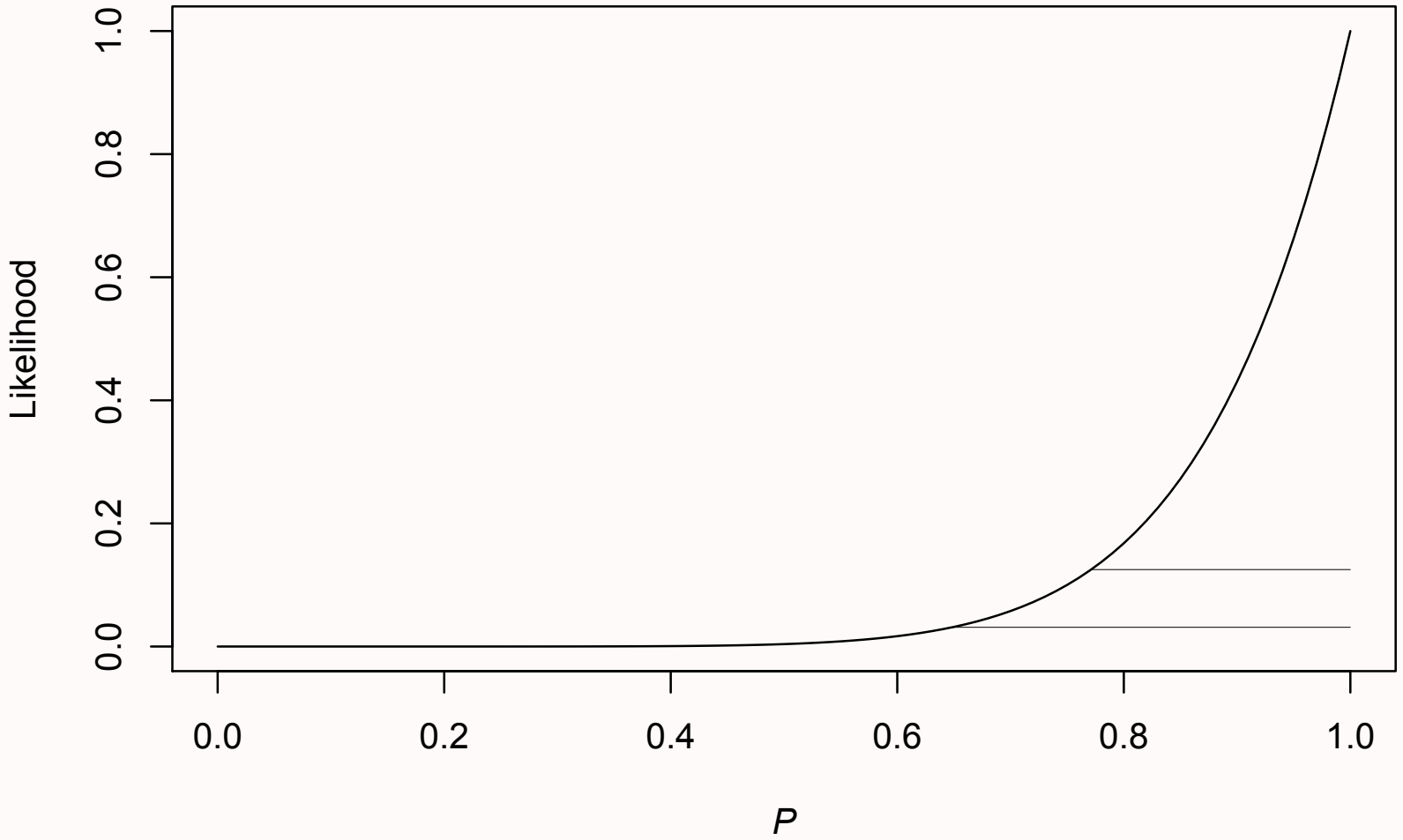
- Here the likelihood is from a binomial distribution

A simple illustrative example: *predicting the US presidential election*

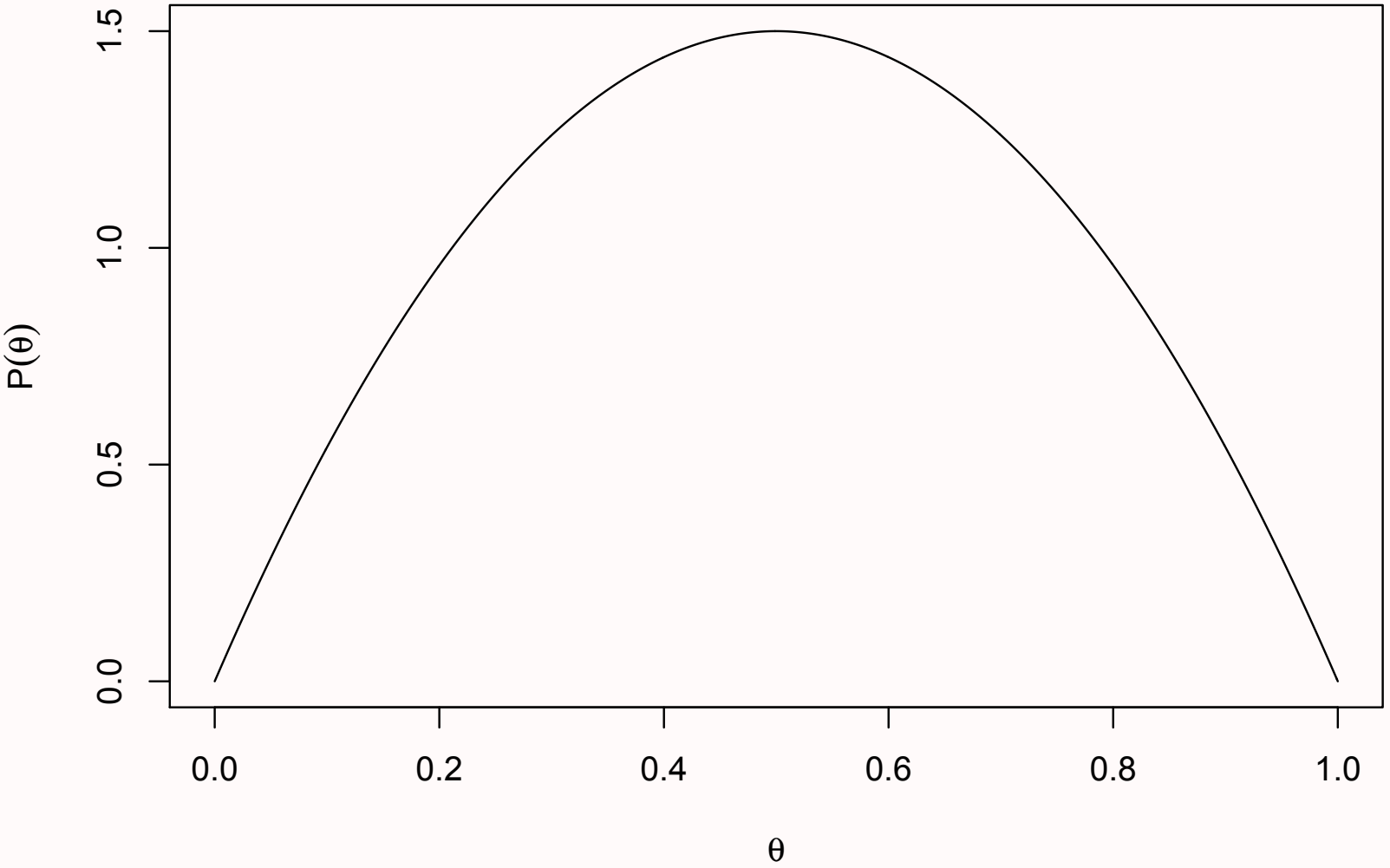
- Here the likelihood is from a binomial distribution
- For the prior we use a Beta distribution
- *Beta*(α, β)

A simple illustrative example: *predicting the US presidential election*

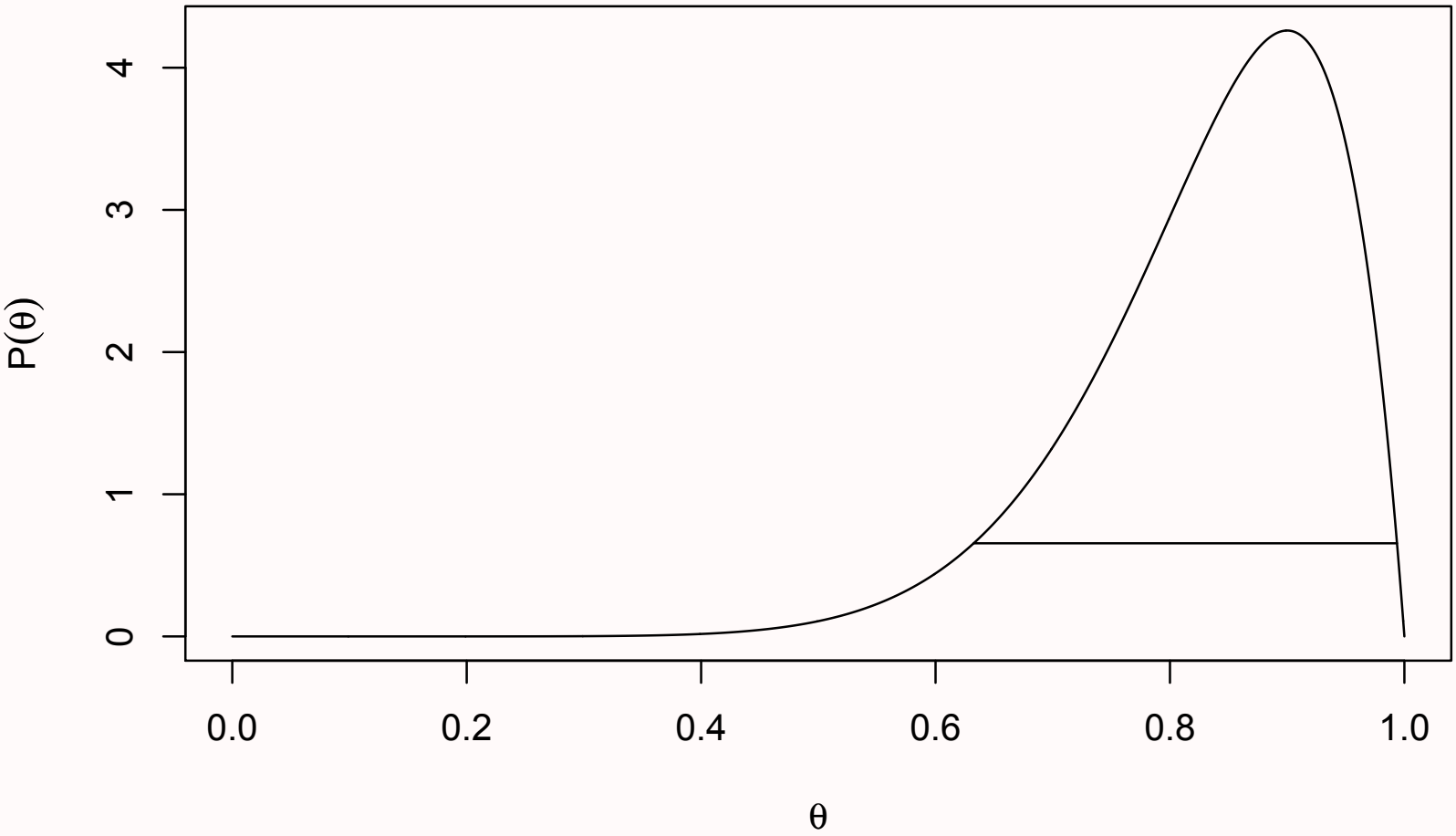
- Here the likelihood is from a binomial distribution
- For the prior we use a Beta distribution
- *Beta*(α, β)
- Combined with likelihood for a binomial proportion



Likelihood for 8 out of 8 wins



Prior with $Beta(2,2)$ distribution



Posterior distributed as $Beta(10, 2)$

Comparing the Bayesian and frequentist solutions:

```
> binom.bayes(8,8, prior.shape1= 2, prior.shape2 = 2)
  method x n shape1 shape2      mean     lower upper  sig
1  bayes 8 8     10      2 0.8333333 0.6356405     1 0.05

> binconf(8,8,method='asymptotic')
 PointEst Lower Upper
        1     1     1

> binconf(8,8,method='wilson')
 PointEst     Lower Upper
        1 0.6755924     1

> binconf(10,12,method='asymptotic')
  PointEst     Lower    Upper
 0.8333333 0.6224748 1.044192
```

The Bayesian interval estimate has some attractive properties

The Bayesian interval estimate has some attractive properties

- Interval aways stays within the 0 to 1 range

The Bayesian interval estimate has some attractive properties

- Interval aways stays within the 0 to 1 range
- Implicitly implements *shrinkage*

The Bayesian interval estimate has some attractive properties

- Interval aways stays within the 0 to 1 range
- Implicitly implements *shrinkage*
- Avoids *ad hoc* fixes

Another simple example: *finding evidence to support the null hypothesis*

- Bayesian interval estimate
- Krushke's Bayesian alternative to the t test (BEST package in R)

Another simple example: *finding evidence to support the null hypothesis*

- Bayesian interval estimate
- Krushke's Bayesian alternative to the t test (BEST package in R)
- Can test hypothesis using a ROPE

Another simple example: *finding evidence to support the null hypothesis*

- Bayesian interval estimate
- Krushke's Bayesian alternative to the t test (BEST package in R)
- Can test hypothesis using a ROPE

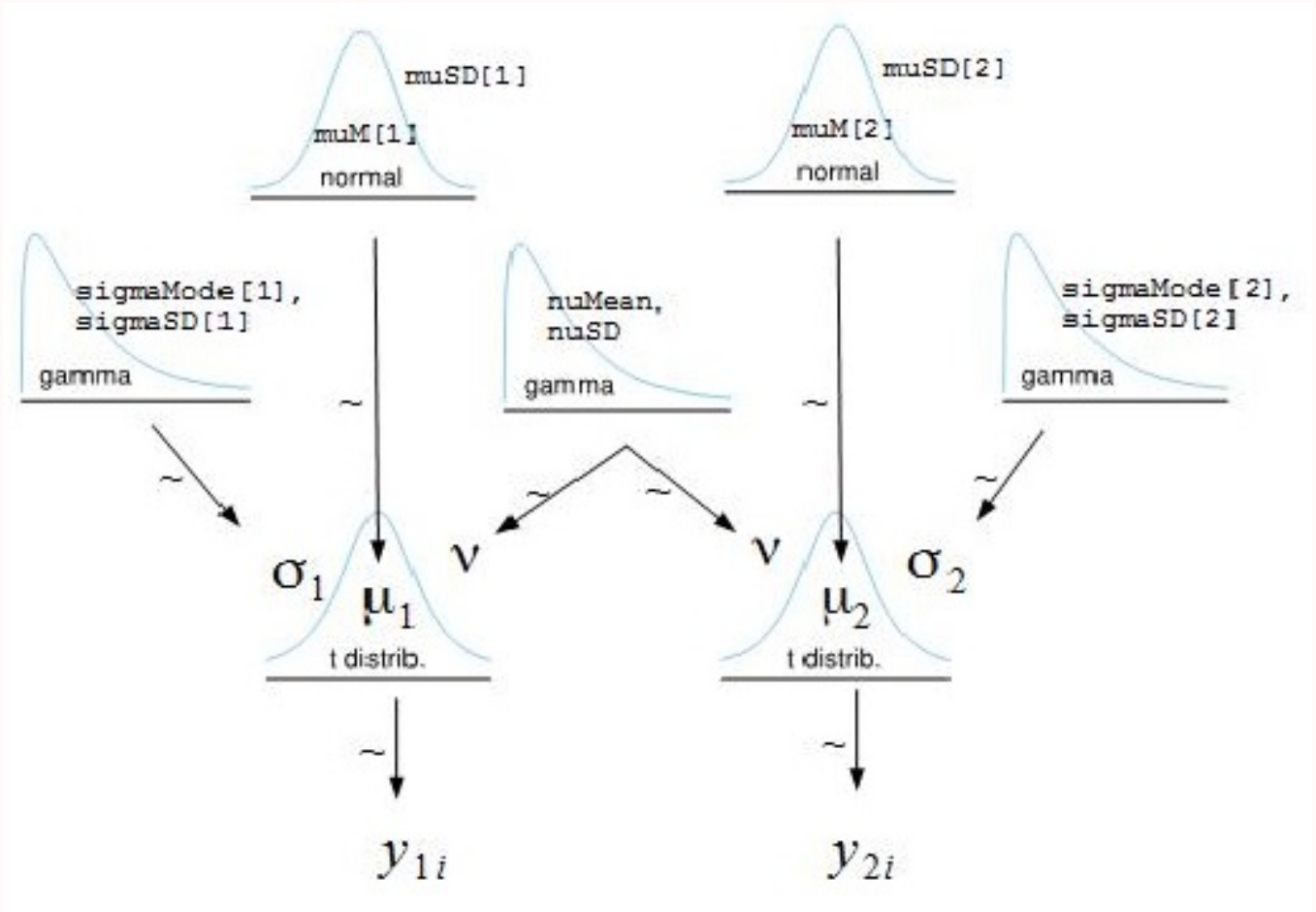
- Bayes factors default (JZS) priors
- Rouder and Morey's approach (BayesFactor package in R)

Another simple example: *finding evidence to support the null hypothesis*

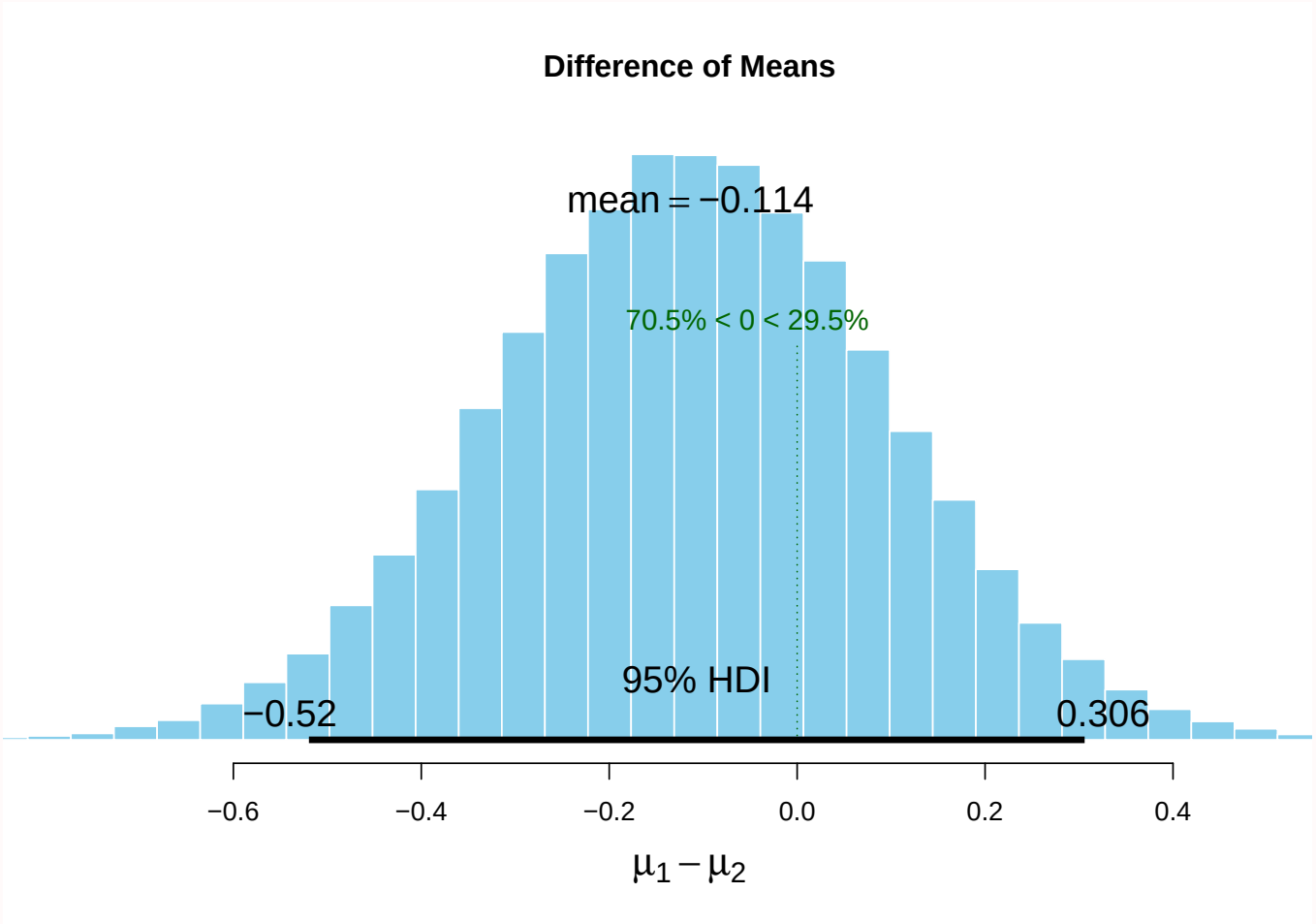
- Bayesian interval estimate
- Krushke's Bayesian alternative to the t test (BEST package in R)
- Can test hypothesis using a ROPE

- Bayes factors default (JZS) priors
- Rouder and Morey's approach (BayesFactor package in R)

Krushke's robust Bayesian alternative to the t test

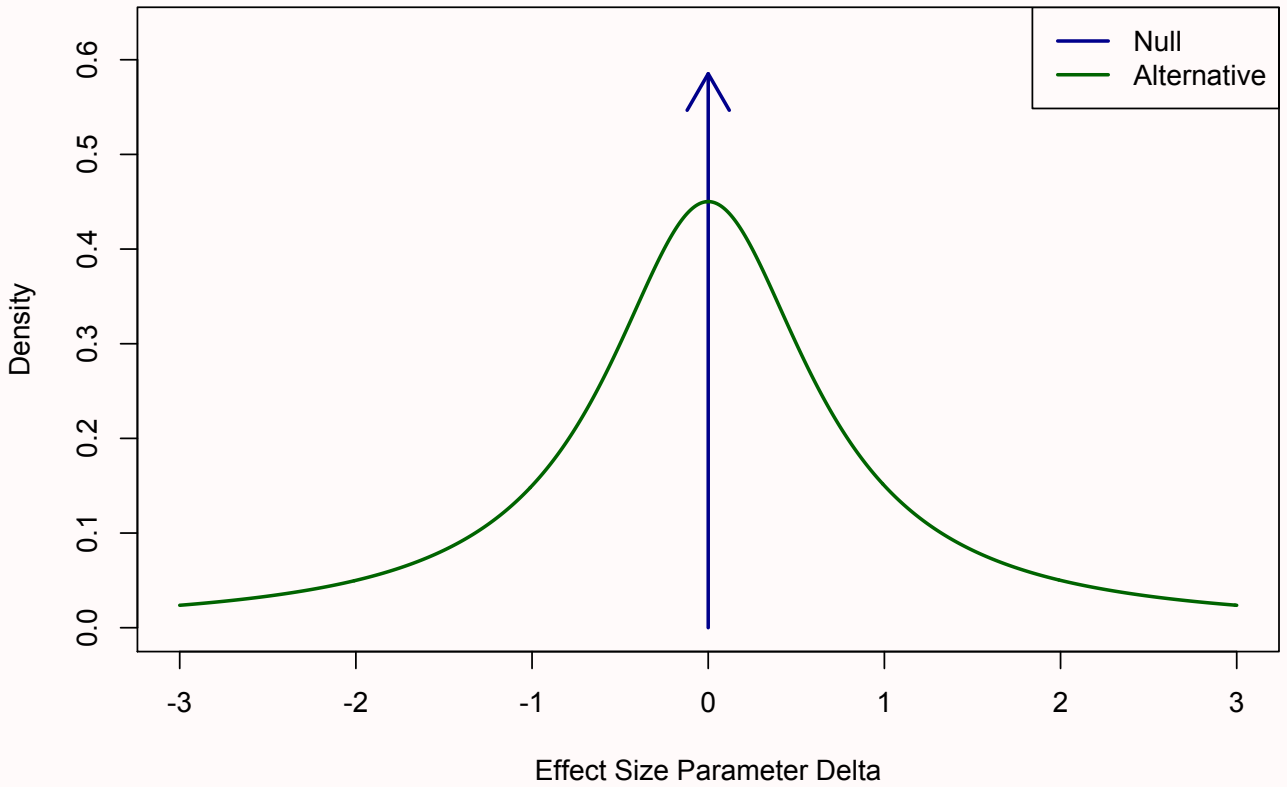


Flag priming replication from ManyLabs

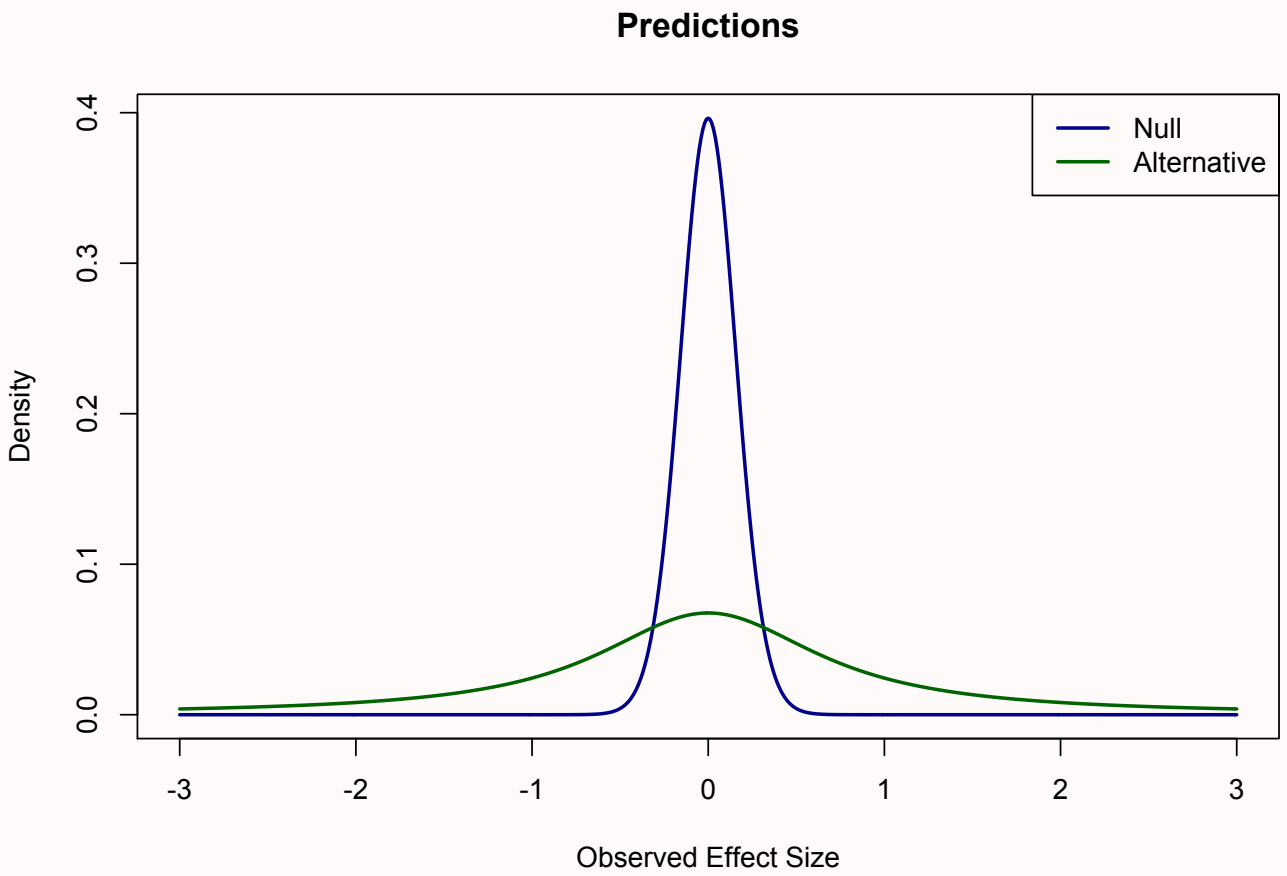


Highest posterior density interval for difference

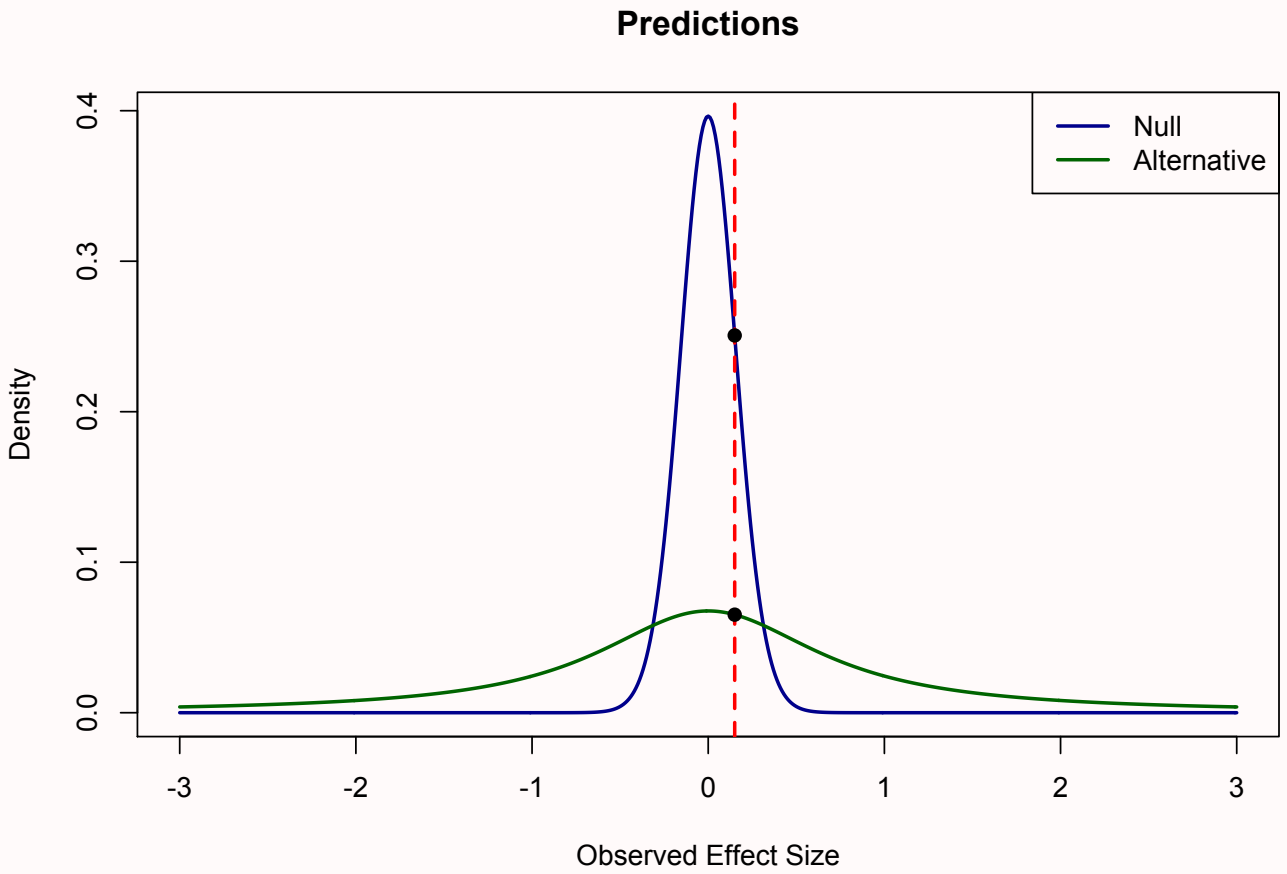
The models being compared



The predictions from each model



The Bayes factor for the observed effect size



Bayes factor with default JZS prior

```
> library(BayesFactor)
>
> ttestBF(formula=flagdv ~ flagGroup, data=flag.data)

Bayes factor analysis
--------------
[1] Alt., r=0.707 : 0.2825635 +/-0.02%

Against denominator:
  Null, mu1-mu2 = 0
---
Bayes factor type: BFindepSample, JZS
```

Bayes factor with default JZS prior

```
> library(BayesFactor)
>
> ttestBF(formula=flagdv ~ flagGroup, data=flag.data)

Bayes factor analysis
--------------
[1] Alt., r=0.707 : 0.2825635 +/-0.02%

Against denominator:
  Null, mu1-mu2 = 0
---
Bayes factor type: BFindepSample, JZS
```

The Bayes Factor is $1/0.283 = 3.54$ in favour of H_0

Why do the methods differ?

Why do the methods differ?

- Hypothesis testing vs. interval estimation

Why do the methods differ?

- Hypothesis testing vs. interval estimation

Why do the methods differ?

- Hypothesis testing vs. interval estimation
- Bayes factors useful tools in some situations

Why do the methods differ?

- Hypothesis testing vs. interval estimation
- Bayes factors useful tools in some situations
- Can be very sensitive to priors

Why do the methods differ?

- Hypothesis testing vs. interval estimation
- Bayes factors useful tools in some situations
- Can be very sensitive to priors

Why do the methods differ?

- Hypothesis testing vs. interval estimation
- Bayes factors useful tools in some situations
- Can be very sensitive to priors
- JZS priors reflect universe in which big effects are rare

Why do the methods differ?

- Hypothesis testing vs. interval estimation
- Bayes factors useful tools in some situations
- Can be very sensitive to priors
- JZS priors reflect universe in which big effects are rare

Using Bayesian tools: Inference in multilevel models

- for standard repeated measures ANOVA models it is possible to use t and F statistics

Using Bayesian tools: Inference in multilevel models

- for standard repeated measures ANOVA models it is possible to use t and F statistics
- if a complex covariance structure or unbalanced model this may be problematic owing to:

Using Bayesian tools: Inference in multilevel models

- for standard repeated measures ANOVA models it is possible to use t and F statistics
- if a complex covariance structure or unbalanced model this may be problematic owing to:
  a) difficulty estimating the error df

Using Bayesian tools: Inference in multilevel models

- for standard repeated measures ANOVA models it is possible to use t and F statistics
- if a complex covariance structure or unbalanced model this may be problematic owing to:
  a) difficulty estimating the error *df*
  b) boundary effects for variance estimates

Using Bayesian tools: Inference in multilevel models

- for standard repeated measures ANOVA models it is possible to use t and F statistics
- if a complex covariance structure or unbalanced model this may be problematic owing to:
  a) difficulty estimating the error *df*
  b) boundary effects for variance estimates

Possible solutions

- asymptotic approximations (in large samples)

Possible solutions

- asymptotic approximations (in large samples)
- corrections such as the Kenwood-Rogers approximation

Possible solutions

- asymptotic approximations (in large samples)
- corrections such as the Kenwood-Rogers approximation (e.g., using `pbkrtest`)

Possible solutions

- asymptotic approximations (in large samples)
- corrections such as the Kenwood-Rogers approximation (e.g., using `pbkrtest`)
- bootstrapping

Possible solutions

- asymptotic approximations (in large samples)
- corrections such as the Kenwood-Rogers approximation (e.g., using `pbkrtest`)
- bootstrapping
- MCMC estimation

Possible solutions

- asymptotic approximations (in large samples)
- corrections such as the Kenwood-Rogers approximation (e.g., using `pbkrtest`)
- bootstrapping
- MCMC estimation (e.g., using `MCMCglmm`)

... with MCMC methods being the preferred approach (being generally both safe and versatile)

Data: The effect of placement on academic performance

Do placements improve academic performance?

Data: The effect of placement on academic performance

Do placements improve academic performance?

An inherently multilevel problem because ...

Data: The effect of placement on academic performance

Do placements improve academic performance?

An inherently multilevel problem because ...

- Students ($n = 15,732$) nested in ($m = 199$) courses

Data: The effect of placement on academic performance

Do placements improve academic performance?

An inherently multilevel problem because ...

- Students ($n = 15,732$) nested in ($m = 199$) courses
- Students are sampled from a population

Data: The effect of placement on academic performance

Do placements improve academic performance?

An inherently multilevel problem because ...

- Students ($n = 15,732$) nested in ($m = 199$) courses
- Students are sampled from a population
- Courses are sampled from a population

Data: The effect of placement on academic performance

Do placements improve academic performance?

An inherently multilevel problem because ...

- Students ($n = 15,732$) nested in ($m = 199$) courses
- Students are sampled from a population
- Courses are sampled from a population

Is final degree mark impacted by the placement?

Data: The effect of placement on academic performance

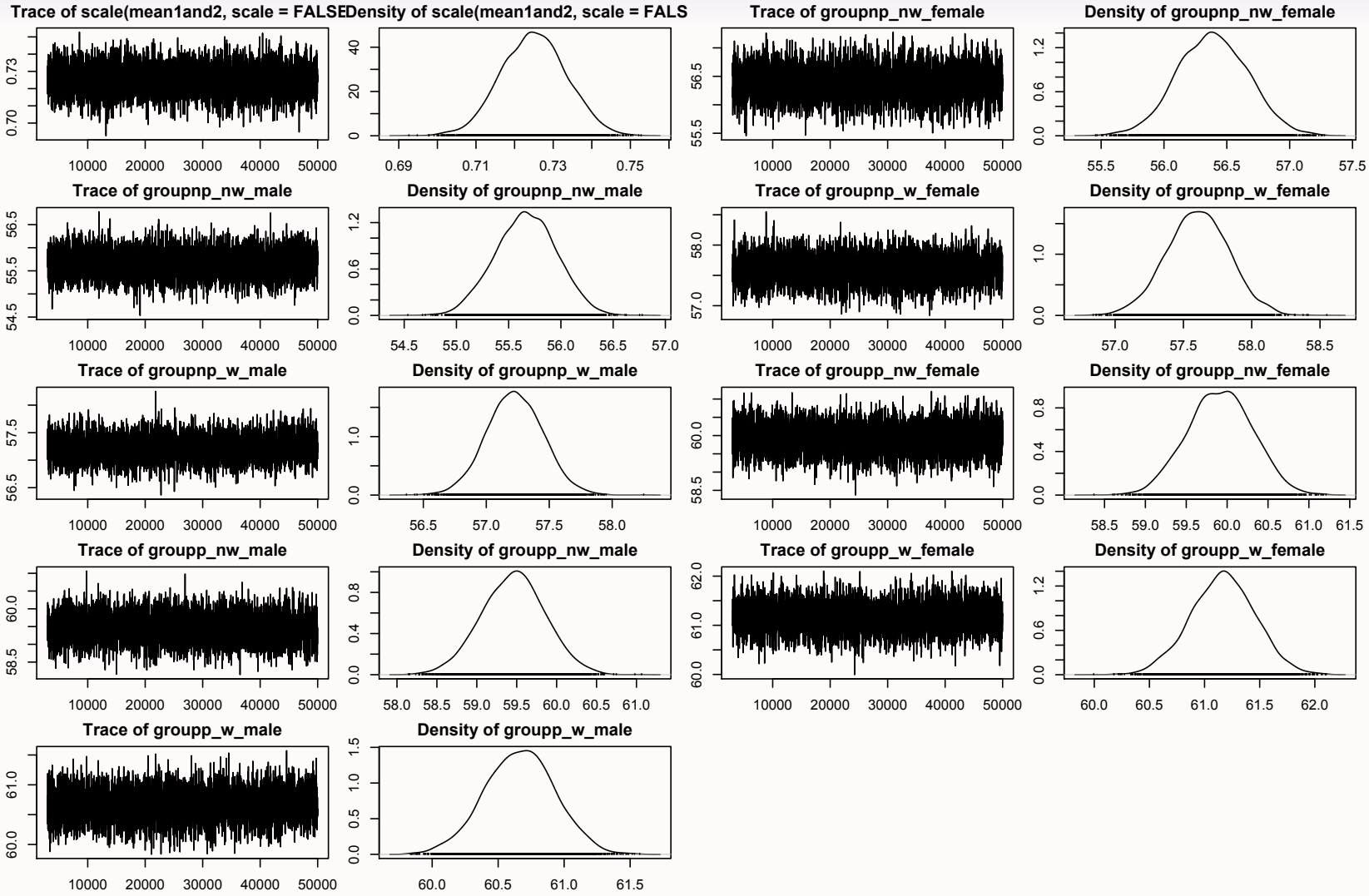
Do placements improve academic performance?

An inherently multilevel problem because ...

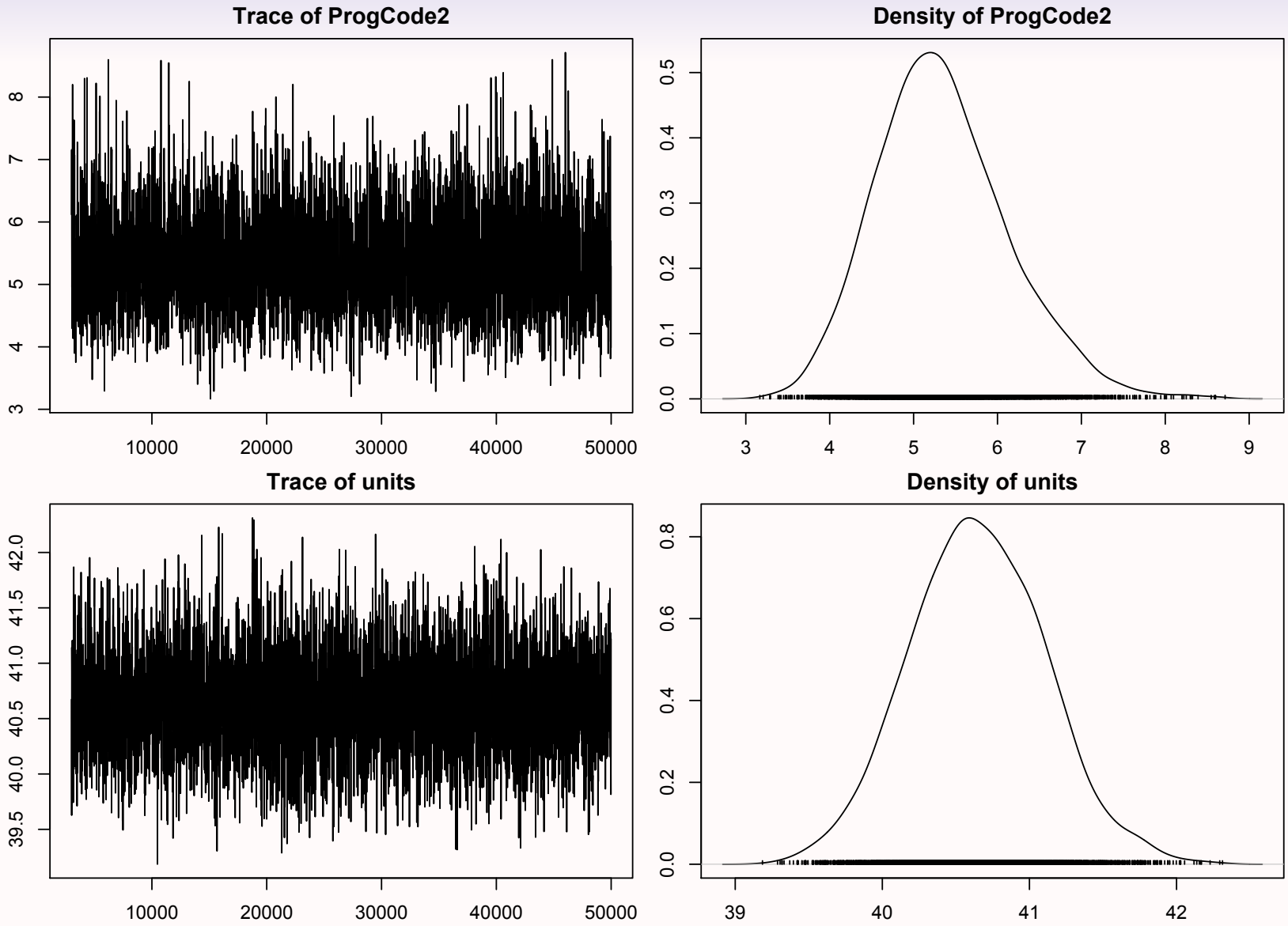
- Students ($n = 15,732$) nested in ($m = 199$) courses
- Students are sampled from a population
- Courses are sampled from a population

Is final degree mark impacted by the placement?

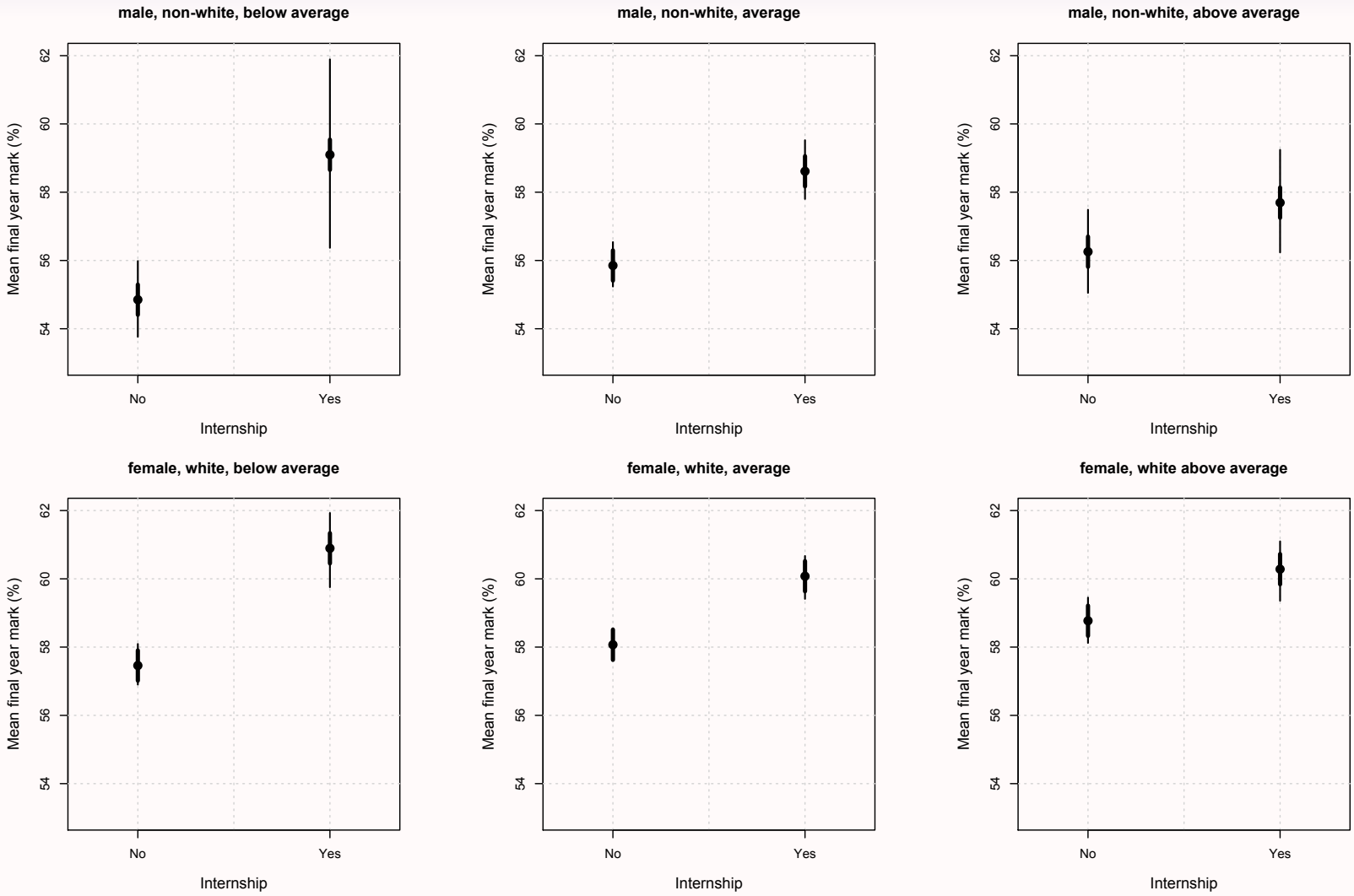
(... after controlling for year 1 and year 2 mean mark)



Trace of MCMC chains for fixed effects



Trace of MCMC chains for random effects



HPD intervals

Advantages of Bayesian estimation

HPD intervals in this example have good frequentist properties

Advantages of Bayesian estimation

HPD intervals in this example have good frequentist properties

Frequentist alternatives not employed because:

- inaccurate (e.g., Wald or profile CIs)

Advantages of Bayesian estimation

HPD intervals in this example have good frequentist properties

Frequentist alternatives not employed because:

- inaccurate (e.g., Wald or profile CIs)
- implementation issues (e.g., parametric bootstrap)

Advantages of Bayesian estimation

HPD intervals in this example have good frequentist properties

Frequentist alternatives not employed because:

- inaccurate (e.g., Wald or profile CIs)
- implementation issues (e.g., parametric bootstrap)
- performance unknown for generalized multilevel models

Advantages of Bayesian estimation

HPD intervals in this example have good frequentist properties

Frequentist alternatives not employed because:

- inaccurate (e.g., Wald or profile CIs)
- implementation issues (e.g., parametric bootstrap)
- performance unknown for generalized multilevel models

Complex Bayesian models: *cognitive processes in language production*

Evidence (in speech) that people plan one phrase at a time

Complex Bayesian models: *cognitive processes in language production*

Evidence (in speech) that people plan one phrase at a time

- Does the phrasal scope effect for speech replicate in writing?

Complex Bayesian models: *cognitive processes in language production*

Evidence (in speech) that people plan one phrase at a time

- Does the phrasal scope effect for speech replicate in writing?
- Is lexical planning required for the entire phrasal planning unit

Complex Bayesian models: *cognitive processes in language production*

Evidence (in speech) that people plan one phrase at a time

- Does the phrasal scope effect for speech replicate in writing?
- Is lexical planning required for the entire phrasal planning unit
- . . . in both spoken and written sentence production?

- Modality: spoken vs. written (i.e., typed)
- First noun phrase: simple vs. complex
- Codability: high vs. low codability of N2

  $DV =$ Onset latency

- Modality: spoken vs. written (i.e., typed)
- First noun phrase: simple vs. complex
- Codability: high vs. low codability of N2

  $DV =$ Onset latency

i. **Peter** moved up and the *N2* and Tania moved down.
ii. **Peter and the *N2*** moved up and Tania moved down.

Why Bayesian data analysis?

Why Bayesian data analysis?

Multiple random effects

Why Bayesian data analysis?

Multiple random effects

High proportion of lost trials (missing data)

Why Bayesian data analysis?

Multiple random effects

High proportion of lost trials (missing data)

Convergence issues

Why Bayesian data analysis?

Multiple random effects

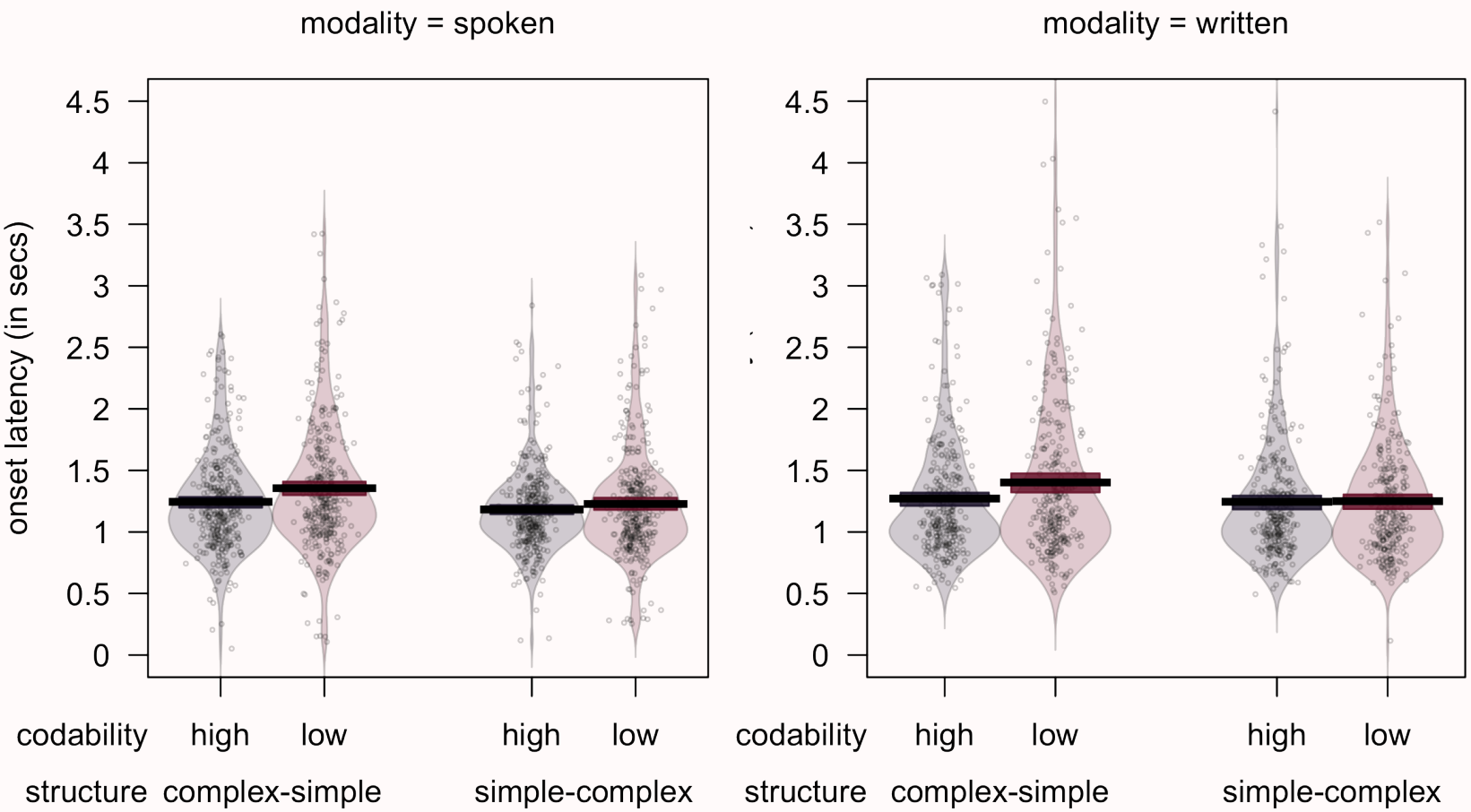
High proportion of lost trials (missing data)

Convergence issues

Wider range of alternative models (e.g., mixture models)

Setting up the model using `stan_lmer()` from `rstanarm`

```
stan_lmer(onset ~ struc + coda + moda + stru_coda + stru_mod  + coda_mod + coda_stru_mod
               + (struc + coda + moda + stru_coda + stru_mod + coda_mod + coda_stru_mod | sub
               + (struc + coda + moda + stru_coda + stru_mod + coda_mod + coda_stru_mod | iter
               prior_intercept = student_t(df = 7, 0),
               prior = student_t(df = 1, 0),
               prior_covariance = decov(regularization = 2),
               data = d,
               chains = 4,
               iter = 2000,
               cores = 4,
               adapt_delta = .97)
```

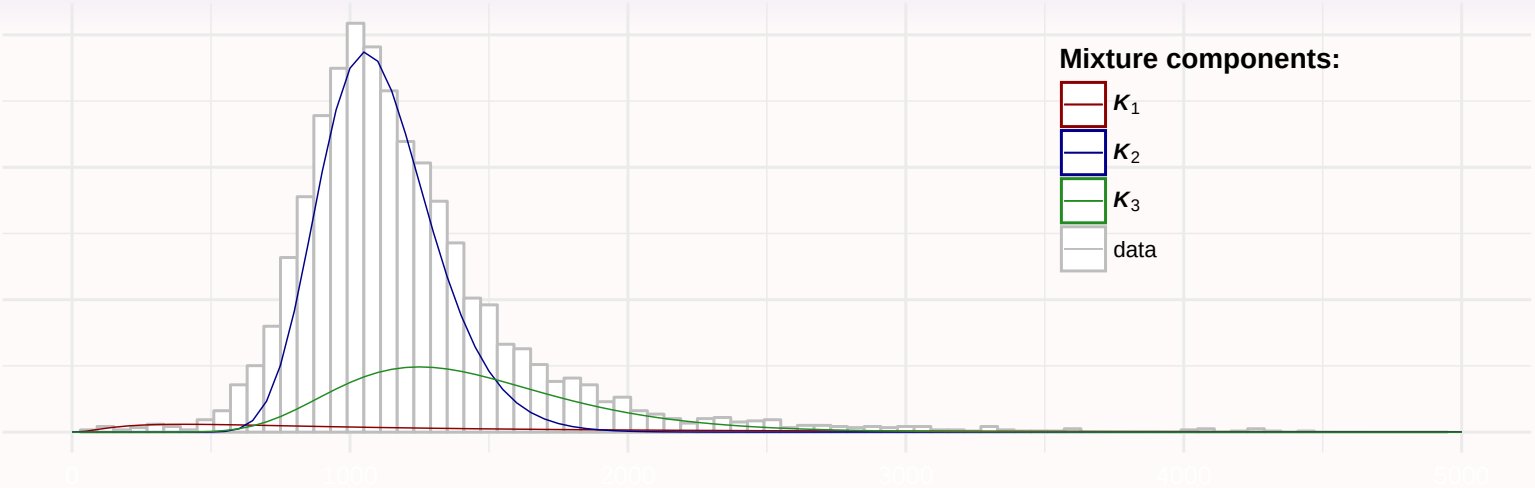


Onset data with HPD intervals

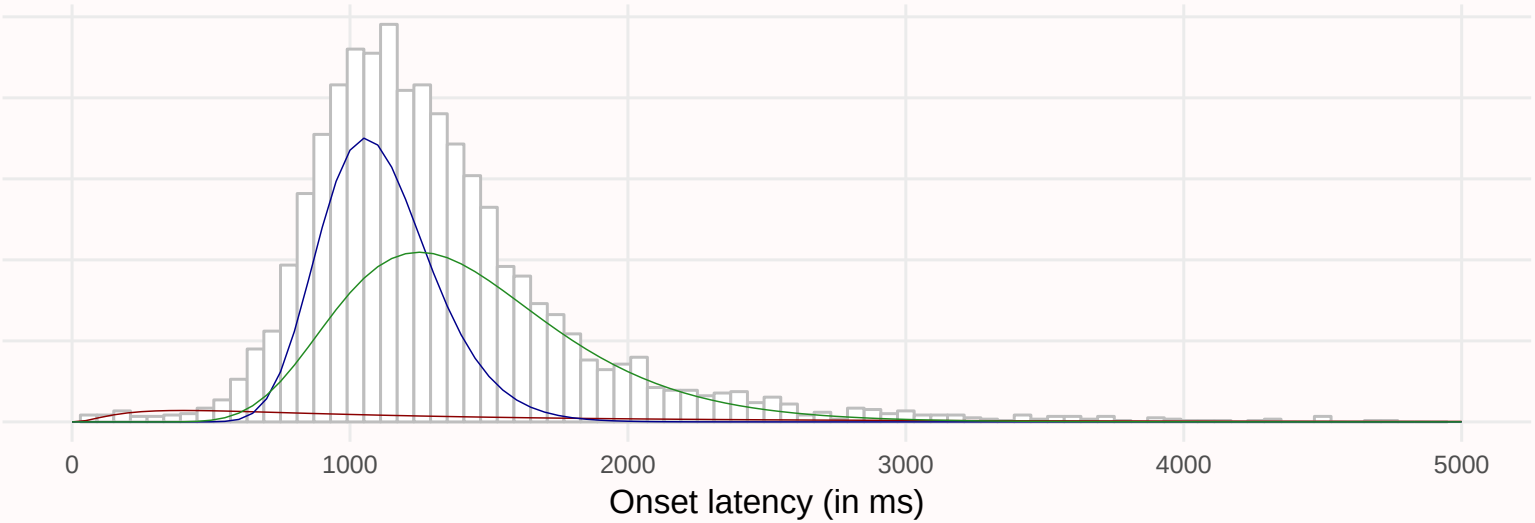
Parameter estimates with HDP intervals

	Median	2.5%	97.5%	P<0
(Intercept)	1273.89	1209.32	1337.22	0.00
struc	45.12	23.44	67.00	0.00
coda	36.11	10.71	62.35	0.00
moda	22.96	-32.17	78.18	0.20
stru_coda	23.55	3.44	43.72	0.01
stru_mod	0.59	-20.92	22.37	0.48
coda_mod	-3.80	-24.53	17.02	0.64
coda_stru_mod	7.80	-11.49	27.12	0.21
sigma	413.79	402.18	426.69	0.00

Simple NP



Conjoined NP



Mixture model exploring components of onset latency

Conclusions ...

- Bayesian data analysis provides useful tools for modelng data

Conclusions ...

- Bayesian data analysis provides useful tools for modelng data
- You don't need to be Bayesian to use these tools

Conclusions ...

- Bayesian data analysis provides useful tools for modelng data
- You don't need to be Bayesian to use these tools
- ... but it is worth being open to Bayesian ideas

Conclusions ...

- Bayesian data analysis provides useful tools for modelng data
- You don't need to be Bayesian to use these tools
- ... but it is worth being open to Bayesian ideas
- New software (almost) eliminates the practical obstacles

What the future holds ...

Increasing acceptance of alternatives to frequentist methods

What the future holds ...

Increasing acceptance of alternatives to frequentist methods

- Likelihood approaches (e.g., AIC, BIC)

What the future holds ...

Increasing acceptance of alternatives to frequentist methods

- Likelihood approaches (e.g., AIC, BIC)
- Empirical Bayes

What the future holds ...

Increasing acceptance of alternatives to frequentist methods

- Likelihood approaches (e.g., AIC, BIC)
- Empirical Bayes

Bayesian methods commonplace (and dominating) in some fields

What the future holds ...

Increasing acceptance of alternatives to frequentist methods

- Likelihood approaches (e.g., AIC, BIC)
- Empirical Bayes

Bayesian methods commonplace (and dominating) in some fields

Greater emphasis on data sharing and publication of code

Bayesian statistics is difficult in the sense that thinking is difficult.

Donald A. Berry