Workshop: Bayesian Statistics in Numerical Cognition

Thomas J. Faulkenberry

Tarleton State University

MCLS 2019

Plan for today:

1. what is Bayesian inference?
2. Using JASP with examples:
   - t-tests (based on Verguts & De Moor, 2005)
   - linear regression (based on Holloway & Ansari, 2006)
   - factorial ANOVA (based on Campbell & Fugelsang, 2001)
3. estimating Bayes factors from summary statistics
4. Advanced topics (if there's time)
   - why do priors matter for Bayes factors?
   - the BIC approximation
   - optional stopping

All materials can be found at:

- `http://github.com/tomfaulkenberry/bayesMCLS`

The problem of inference

For any type of statistical inference, we fix a generative model

The problem of inference

For any type of statistical inference, we fix a generative model

(think sampling distributions)

The problem of inference

Given observed data, we then try to invert this model.

The problem of inference

Given observed data, we then try to invert this model.

The frequentist accepts or rejects $\mathcal{M}$ based on the likelihood of observing some data under a null hypothesis (i.e., the p-value)

The problem of inference

Given observed data, we then try to invert this model.

The frequentist accepts or rejects $\mathcal{M}$ based on the likelihood of observing some data under a null hypothesis (i.e., the p-value)

- bases decision criterion on controlling long-run error rates (i.e., α)

The problem of inference

Given observed data, we then try to invert this model.

The problem of inference

Given observed data, we then try to invert this model.

The Bayesian just directly asks: "What is the probability of this model $\mathcal{M}$, given that we've observed these data?"

The problem of inference

Given observed data, we then try to invert this model.

The Bayesian just directly asks: "What is the probability of this model $\mathcal{M}$, given that we've observed these data?"

- "posterior belief in model $\mathcal{M}$"

The problem of inference

Given observed data, we then try to invert this model.

The Bayesian just directly asks: "What is the probability of this model $\mathcal{M}$, given that we've observed these data?"

- "posterior belief in model $\mathcal{M}$"
- notation: $p(\mathcal{M} \mid \text{data})$

The problem of inference

Given observed data, we then try to invert this model.

The Bayesian just directly asks: "What is the probability of this model $\mathcal{M}$, given that we've observed these data?"

- "posterior belief in model $\mathcal{M}$"
- notation: $p(\mathcal{M} \mid \text{data})$
- no accept/reject decision

Bayes' Rule

$$\underbrace{p(\mathcal{M} \mid \text{data})}_{\substack{\text{Posterior beliefs} \\ \text{about model}}}$$

Bayes' Rule

$$\underbrace{p(\mathcal{M} \mid \text{data})}_{\substack{\text{Posterior beliefs} \\ \text{about model}}} = \underbrace{p(\mathcal{M})}_{\substack{\text{Prior beliefs} \\ \text{about model}}}$$

Bayes' Rule

$$\underbrace{p(\mathcal{M} \mid \text{data})}_{\substack{\text{Posterior beliefs}\\ \text{about model}}} = \underbrace{p(\mathcal{M})}_{\substack{\text{Prior beliefs}\\ \text{about model}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{M})}{p(\text{data})}}_{\text{predictive updating factor}}$$

Bayes' Rule

Natural action in science is to *compare* two models $\mathcal{M}_1$ and $\mathcal{M}_2$.

- Bayes' rule gives us a mathematical way to do this:

$$\frac{p(\mathcal{M}_1 \mid \text{data})}{p(\mathcal{M}_2 \mid \text{data})} =$$

Bayes' Rule

Natural action in science is to *compare* two models $\mathcal{M}_1$ and $\mathcal{M}_2$.

- Bayes' rule gives us a mathematical way to do this:

$$\frac{p(\mathcal{M}_1 \mid \text{data})}{p(\mathcal{M}_2 \mid \text{data})} = \frac{p(\mathcal{M}_1) \cdot \frac{p(\text{data}|\mathcal{M}_1)}{p(\text{data})}}{p(\mathcal{M}_2) \cdot \frac{p(\text{data}|\mathcal{M}_2)}{p(\text{data})}}$$

Bayes' Rule

Natural action in science is to *compare* two models $\mathcal{M}_1$ and $\mathcal{M}_2$.

- Bayes' rule gives us a mathematical way to do this:

$$\frac{p(\mathcal{M}_1 \mid \text{data})}{p(\mathcal{M}_2 \mid \text{data})} = \frac{p(\mathcal{M}_1) \cdot \frac{p(\text{data}|\mathcal{M}_1)}{p(\text{data})}}{p(\mathcal{M}_2) \cdot \frac{p(\text{data}|\mathcal{M}_2)}{p(\text{data})}}$$

$$= \frac{p(\mathcal{M}_1) \cdot p(\text{data} \mid \mathcal{M}_1)}{p(\mathcal{M}_2) \cdot p(\text{data} \mid \mathcal{M}_2)}$$

Bayes' Rule

Natural action in science is to *compare* two models $\mathcal{M}_1$ and $\mathcal{M}_2$.

- Bayes' rule gives us a mathematical way to do this:

$$\underbrace{\frac{p(\mathcal{M}_1 \mid \text{data})}{p(\mathcal{M}_2 \mid \text{data})}}_{\substack{\text{posterior beliefs}\\ \text{about models}}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\substack{\text{prior beliefs}\\ \text{about models}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{M}_1)}{p(\text{data} \mid \mathcal{M}_2)}}_{\text{predictive updating factor}}$$

Bayes factor

The predictive updating factor

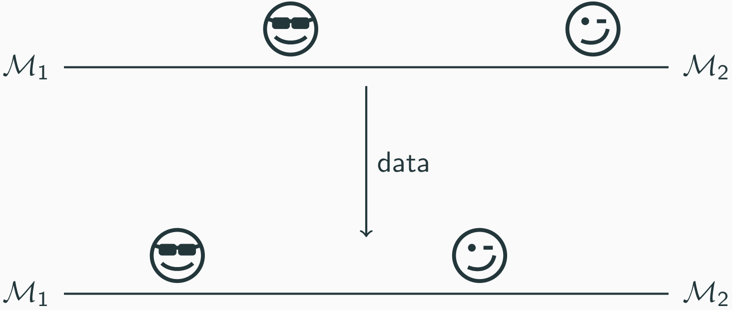
$$B_{12} = \frac{p(\text{data} \mid \mathcal{M}_1)}{p(\text{data} \mid \mathcal{M}_2)}$$

tells us how much better $\mathcal{M}_1$ predicts our observed data than $\mathcal{M}_2$.

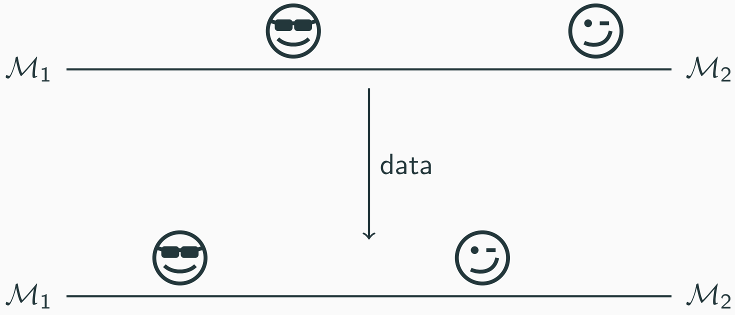
This ratio is called the **Bayes factor**

Bayes factors

Bayes factors



Bayes factors



Although 😎 and 😉 have different prior beliefs, they both shift their belief an equal amount toward $\mathcal{M}_1$.

Interpreting Bayes factors

Example 1: suppose $B_{12} = 10$.

Interpretation: the observed data are 10 times more likely under $\mathcal{M}_1$ than $\mathcal{M}_2$.

Interpreting Bayes factors

Example 1: suppose $B_{12} = 10$.

Interpretation: the observed data are 10 times more likely under $\mathcal{M}_1$ than $\mathcal{M}_2$.

Example 2: suppose $B_{12} = \frac{1}{10}$. Then $B_{21} = 10$.

Interpretation: the observed data are 10 times more likely under $\mathcal{M}_2$ than $\mathcal{M}_1$.

Interpreting Bayes factors

Example 1: suppose $B_{12} = 10$.

Interpretation: the observed data are 10 times more likely under $\mathcal{M}_1$ than $\mathcal{M}_2$.

Example 2: suppose $B_{12} = \frac{1}{10}$. Then $B_{21} = 10$.

Interpretation: the observed data are 10 times more likely under $\mathcal{M}_2$ than $\mathcal{M}_1$.

Example 3: suppose $B_{12} = 1$.

Interpretation: the observed data are equally likely under $\mathcal{M}_1$ and $\mathcal{M}_2$.

Bayes factors

Jeffreys (1961) proposed the following thresholds for evidence:

Bayes factor	Evidence
1-3	anecdotal
3-10	moderate
10-30	strong
30-100	very strong
¿100	extreme

Models $\leftrightarrow$ hypotheses

Full Bayesian inference requires specification of generative models for data. This is often difficult.

Also, we are typically trained to evaluate hypotheses about effects.

To reconcile the two, several teams (e.g., Rouder, Morey, Wagenmakers, et al.) have developed *default* Bayesian hypothesis tests. The key idea is that we define models on effect size.

Models $\leftrightarrow$ hypotheses

Specifying models on effect size

Models $\leftrightarrow$ hypotheses

Specifying models on effect size

- let $\delta = \dfrac{\mu}{\sigma}$ (think Cohen's *d*, but at the population level)

Models $\leftrightarrow$ hypotheses

Specifying models on effect size

- let $\delta = \dfrac{\mu}{\sigma}$ (think Cohen's *d*, but at the population level)
- define competing models on δ:

Models $\leftrightarrow$ hypotheses

Specifying models on effect size

- let $\delta = \dfrac{\mu}{\sigma}$ (think Cohen's *d*, but at the population level)
- define competing models on δ:
  - $\mathcal{H}_0 : \mu = 0$ (the effect size is 0)

Models $\leftrightarrow$ hypotheses

Specifying models on effect size

- let $\delta = \dfrac{\mu}{\sigma}$ (think Cohen's *d*, but at the population level)
- define competing models on δ:
  - $\mathcal{H}_0 : \mu = 0$ (the effect size is 0)
  - $\mathcal{H}_1 : \mu \neq 0$ (the effect size is not 0)

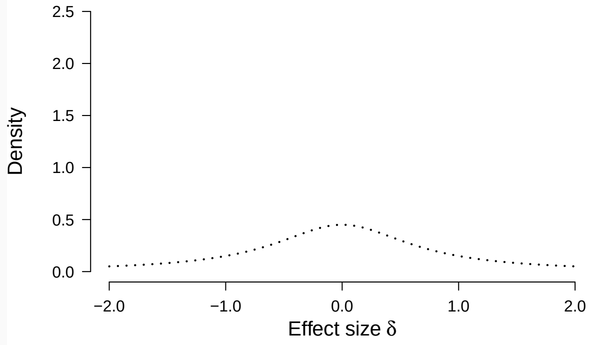
Models $\leftrightarrow$ hypotheses

Specifying models on effect size

- let $\delta = \frac{\mu}{\sigma}$ (think Cohen's d, but at the population level)
- define competing models on δ:
  - $\mathcal{H}_0 : \mu = 0$ (the effect size is 0)
  - $\mathcal{H}_1 : \mu \neq 0$ (the effect size is not 0)
- use Bayes' rule to compute

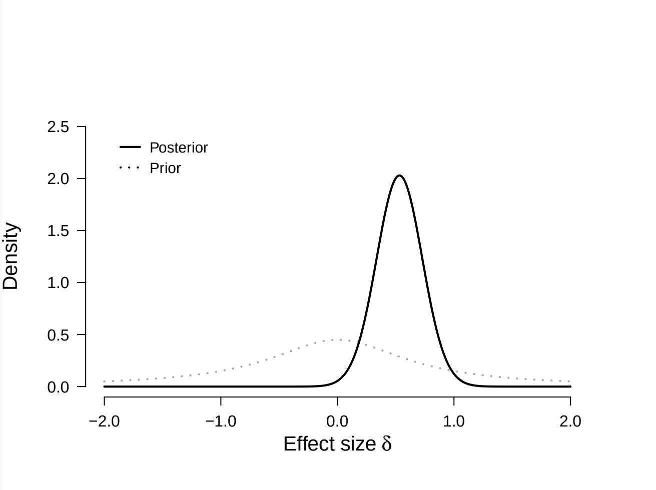
$$p(\mathcal{H}_1 \mid \text{data}) = p(\mathcal{H}_1) \times \frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data})}$$

Generic default Bayesian test



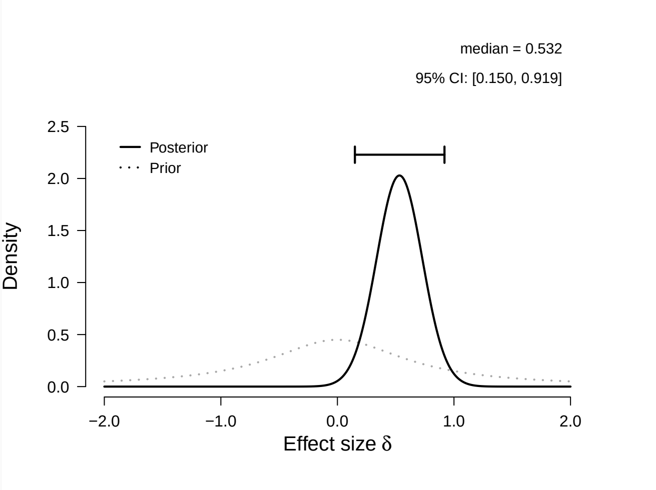
Start with prior belief about expected effect sizes δ.

Generic default Bayesian test



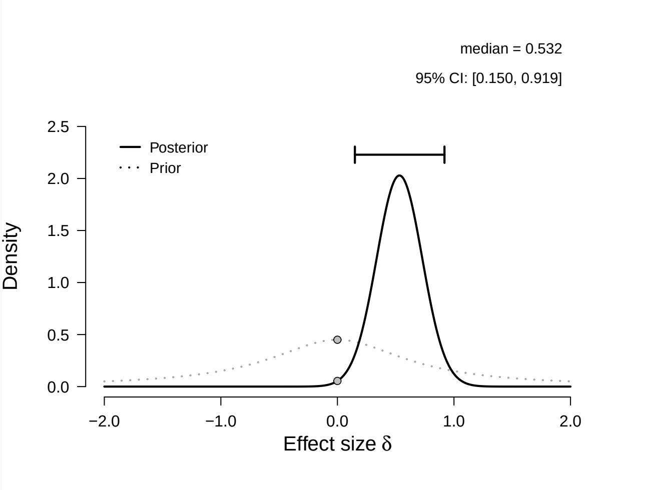
Observing data updates our prior to a posterior.

Generic default Bayesian test



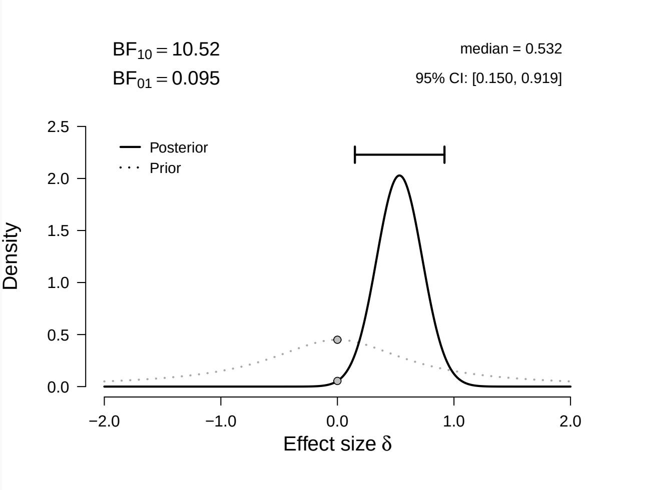
We can extract posterior estimates of δ

Generic default Bayesian test



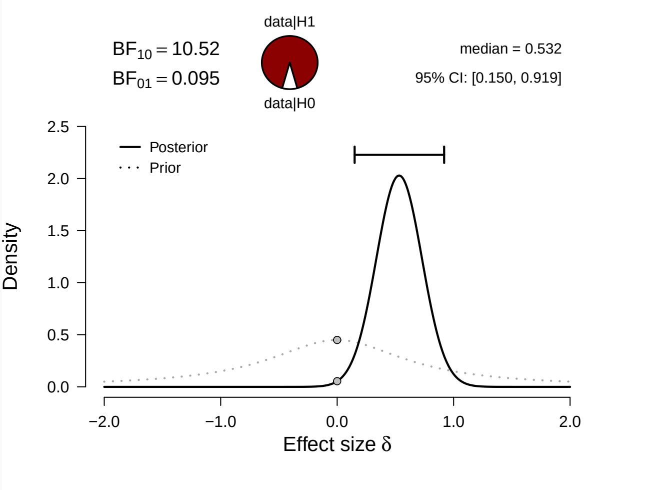
The Bayes factor is the ratio of the densities of $\delta = 0$ in the posterior and prior.

Generic default Bayesian test

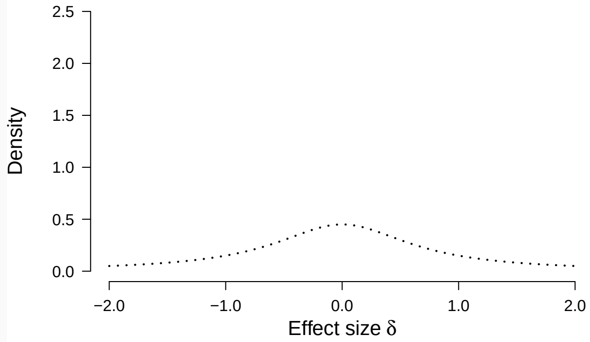


Observing data reduced our belief that $\delta = 0$ by a factor of 10.52

Generic default Bayesian test

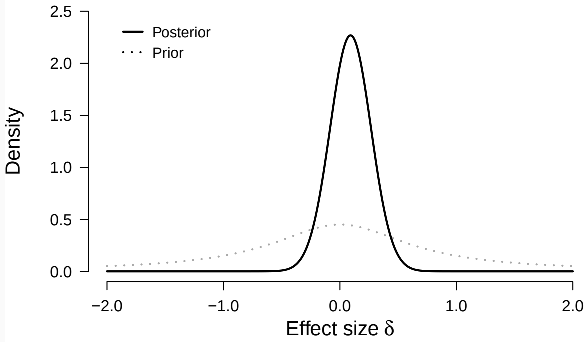


Generic default Bayesian test



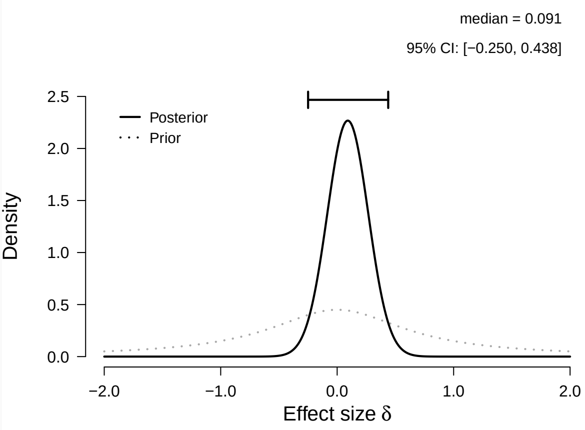
What happens if the null is supported instead?

Generic default Bayesian test



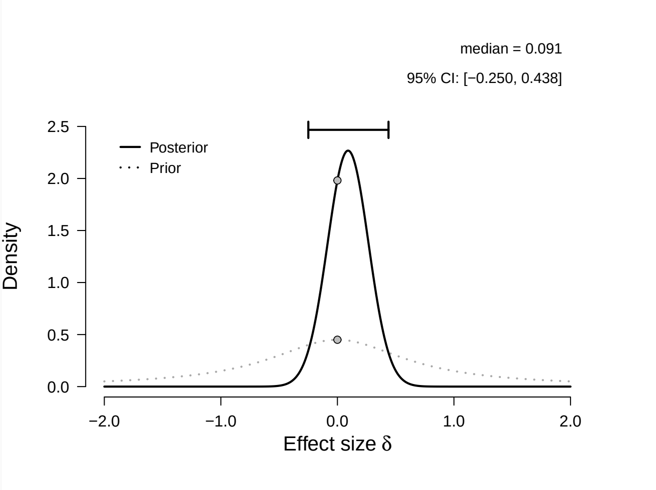
Observing data updates our prior to a posterior.

Generic default Bayesian test



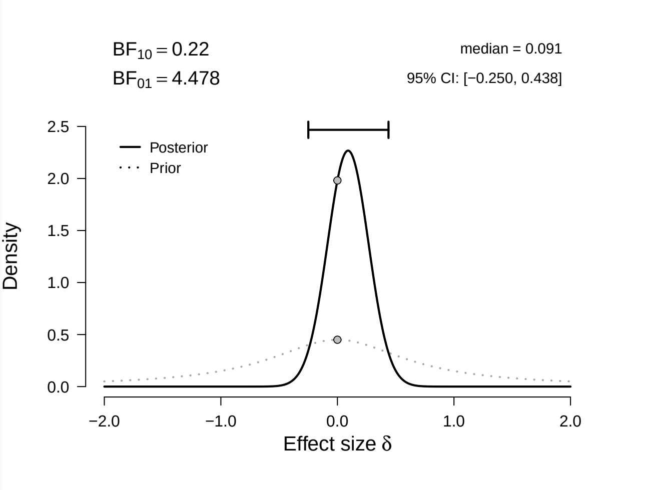
We can extract posterior estimates of δ

Generic default Bayesian test



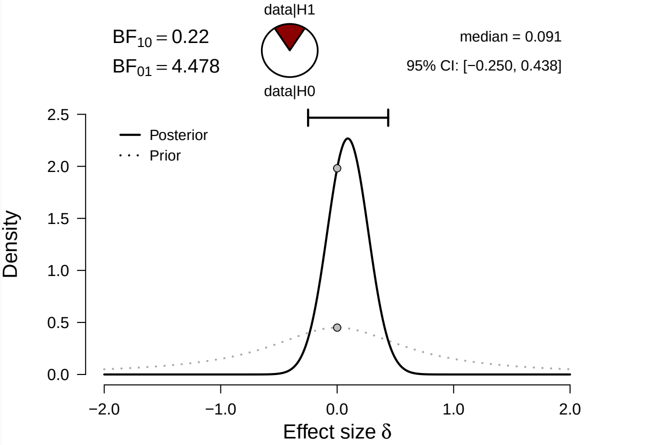
The Bayes factor is the ratio of the densities of $\delta = 0$ in the posterior and prior.

Generic default Bayesian test



Observing data increased our belief that $\delta = 0$ by a factor of 4.478

Generic default Bayesian test



Questions?

Now let's work some examples together.

All datasets can be downloaded at
http://github.com/tomfaulkenberry/bayesMCLS

http://jasp-stats.org

Advanced topics if we have time...

- why do priors matter for Bayes factors?
- the BIC approximation
- optional stopping

The role of the prior in Bayes factors

The role of the prior in Bayes factors

Recall that the Bayes factor is defined as a ratio of likelihoods:

$$B_{12} = \frac{p(\text{data} \mid \mathcal{M}_1)}{p(\text{data} \mid \mathcal{M}_2)}$$

This tells us how much better $\mathcal{M}_1$ predicts our observed data than $\mathcal{M}_2$.

The role of the prior in Bayes factors

But these likelihoods are only part of Bayes rule:

$$\underbrace{p(\mathcal{M} \mid \text{data})}_{\substack{\text{Posterior beliefs}\\\text{about model}}} = \underbrace{p(\mathcal{M})}_{\substack{\text{Prior beliefs}\\\text{about model}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{M})}{p(\text{data})}}_{\text{predictive updating factor}}$$

and they do not seem to involve the prior.

The role of the prior in Bayes factors

The Bayes factor is more accurately defined as a ratio of marginal likelihoods, where:

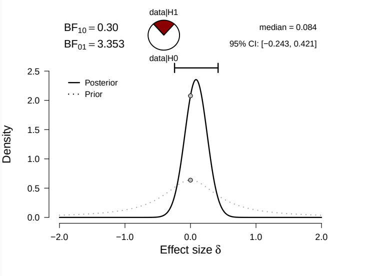
$$p(\text{data} \mid \mathcal{M}) = \int p(\text{data} \mid \theta, \mathcal{M}) p(\theta \mid \mathcal{M}) d\theta$$

Each marginal likelihood can be thought of as the average of an infinite family of data likelihoods, where each likelihood is computed for a specific value of some model parameter θ. This average is weighted by the prior probability of each θ

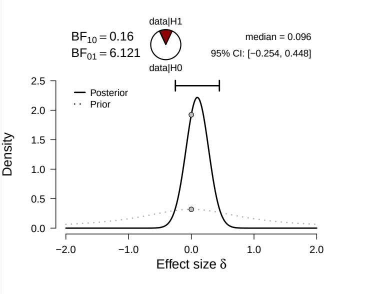
The role of the prior in Bayes factors

We can also see this by looking at the Savage-Dickey density ratio. Consider the one-sample t test from earlier using two different priors:

$\delta \sim \text{Cauchy}(0, \frac{1}{2})$



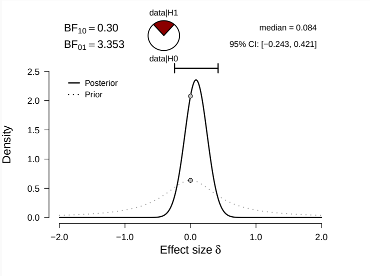
$\delta \sim \text{Cauchy}(0, 1)$



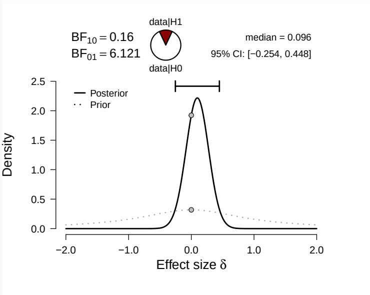
The role of the prior in Bayes factors

We can also see this by looking at the Savage-Dickey density ratio. Consider the one-sample t test from earlier using two different priors:

$\delta \sim \text{Cauchy}(0, \frac{1}{2})$



$\delta \sim \text{Cauchy}(0, 1)$



Moral: always report your priors and show that your results are consistent across a range of priors.

The BIC approximation

Classic ANOVA

Consider the test scores from students in three different treatment conditions:

- Treatment 1 - read and reread
- Treatment 2 - read, then answer prepared questions
- Treatment 3 - read, then create and answer questions

Classic ANOVA

Treatment 1	Treatment 2	Treatment 3
2	5	8
3	9	6
8	10	12
6	13	11
5	8	11
6	9	12
$M = 5$	$M = 9$	$M = 10$

Typical question – are there differences among these condition means?

Classic ANOVA

Standard approach:

- model $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$
- assume "null hypothesis" $\mathcal{H}_0 : \alpha_j = 0$
- compute probability of observing data Y_{ij} under $\mathcal{H}_0$
- if data is *rare* under $\mathcal{H}_0$, reject $\mathcal{H}_0$

Classic ANOVA

variance source	SS	df	MS	F
between treatments				
residual				
total				

Classic ANOVA

variance source	SS	df	MS	F
between treatments				
residual				
total	172			

$$
\begin{aligned}
SS_{\text{total}} &= \sum Y^2 - \frac{(\sum Y)^2}{N} \\
&= 1324 - \frac{144^2}{18} \\
&= 172
\end{aligned}
$$

Classic ANOVA

variance source	SS	df	MS	F
between treatments	84			
residual				
total	172			

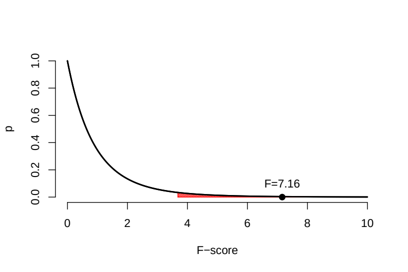
$$
\begin{aligned}
SS_{\text{bet tmts}} &= n\sum_{j=1}^{3}(\overline{Y}_j - \overline{Y})^2 \\
&= 6\Big[(5-8)^2 + (9-8)^2 + (10-8)^2\Big] \\
&= 84
\end{aligned}
$$

Classic ANOVA

variance source	SS	df	MS	F
between treatments	84	2	42	7.16
residual	88	15	5.87	
total	172	17		

Classic ANOVA

source	SS	df	MS	F
between treatments	84	2	42	7.16
within treatments	88	15	5.87	
total	172	17		



Since our data Y_{ij} is rare under $\mathcal{H}_0$ ($p = 0.007$), we reject $\mathcal{H}_0$ as an implausible model restriction.

What is the evidence for $\mathcal{H}_1$?

BIC approximation

With some assumptions, we can compute Bayes factors for ANOVA designs using a method due originally to Kass and Raftery (1995) (but also see Masson, 2011).

Basic idea:

BIC approximation

With some assumptions, we can compute Bayes factors for ANOVA designs using a method due originally to Kass and Raftery (1995) (but also see Masson, 2011).

Basic idea:

1. set up two models: $\mathcal{H}_0$ and $\mathcal{H}_1$

BIC approximation

With some assumptions, we can compute Bayes factors for ANOVA designs using a method due originally to Kass and Raftery (1995) (but also see Masson, 2011).

Basic idea:

1. set up two models: $\mathcal{H}_0$ and $\mathcal{H}_1$
2. compute BIC (Bayesian information criterion) for each model:

$$BIC = N \ln(SS_{\text{residual}}/N) + k \ln(N)$$

where

- N=total number of independent observations
- k=number of parameters in the model
- SS_{residual} = variance NOT explained by the model

BIC approximation

With some assumptions, we can compute Bayes factors for ANOVA designs using a method due originally to Kass and Raftery (1995) (but also see Masson, 2011).

Basic idea:

1. set up two models: $\mathcal{H}_0$ and $\mathcal{H}_1$
2. compute BIC (Bayesian information criterion) for each model:

$$BIC = N \ln(SS_{\text{residual}}/N) + k \ln(N)$$

   where

   - N=total number of independent observations
   - k=number of parameters in the model
   - SS_{residual} = variance NOT explained by the model

3. compute Bayes factor as $e^{\frac{\Delta BIC}{2}}$

BIC approximation

source	*SS*	*df*	*MS*	*F*
bet tmts	84	2	42	7.16
residual	88	15	5.87	
total	172	17		

We'll set up our two models:

Null model: $\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3$

BIC approximation

source	SS	df	MS	F
bet tmts	84	2	42	7.16
residual	88	15	5.87	
total	172	17		

We'll set up our two models:

Null model: $\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3$

- this model has $k = 1$ parameter (the data is explained by a SINGLE mean)

BIC approximation

source	*SS*	*df*	*MS*	*F*
bet tmts	84	2	42	7.16
residual	88	15	5.87	
total	172	17		

We'll set up our two models:

Null model: $\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3$

- this model has $k = 1$ parameter (the data is explained by a SINGLE mean)
- $SS_{\text{residual}} = 172$ (the model has only one mean, so **all** variance is left unexplained)

BIC approximation

source	*SS*	*df*	*MS*	*F*
bet tmts	84	2	42	7.16
residual	88	15	5.87	
total	172	17		

Null model: $\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3$

$$\begin{aligned} BIC_0 &= N \ln(SS_{\text{residual}}/N) + k \ln(N) \\ &= 18 \ln(172/18) + 1 \cdot \ln(18) \\ &= 43.52 \end{aligned}$$

BIC approximation

source	SS	df	MS	F
bet tmts	84	2	42	7.16
residual	88	15	5.87	
total	172	17		

Alternative model: $\mathcal{H}_1 : \mu_1 \neq \mu_2 \neq \mu_3$

BIC approximation

source	SS	df	MS	F
bet tmts	84	2	42	7.16
residual	88	15	5.87	
total	172	17		

Alternative model: $\mathcal{H}_1 : \mu_1 \neq \mu_2 \neq \mu_3$

- this model has $k = 3$ parameters (the data is explained by THREE means)

BIC approximation

source	SS	df	MS	F
bet tmts	84	2	42	7.16
residual	88	15	5.87	
total	172	17		

Alternative model: $\mathcal{H}_1 : \mu_1 \neq \mu_2 \neq \mu_3$

- this model has $k = 3$ parameters (the data is explained by THREE means)
- $SS_{\text{residual}} = 88$ (the model accounts for variance between treatments with the three means, so SS_{residual} is left unexplained)

BIC approximation

source	SS	df	MS	F
bet tmts	84	2	42	7.16
residual	88	15	5.87	
total	172	17		

Alternative model: $\mathcal{H}_1 : \mu_1 \neq \mu_2 \neq \mu_3$

$$\begin{aligned} BIC_1 &= N \ln(SS_{\text{residual}}/N) + k \ln(N) \\ &= 18 \ln(88/18) + 3 \cdot \ln(88) \\ &= 37.23 \end{aligned}$$

BIC approximation

Thus,

$$
\begin{aligned}
B_{10} &= e^{\frac{\Delta BIC}{2}} \\
&= e^{\frac{43.52-37.23}{2}} \\
&= 23.22
\end{aligned}
$$

BIC approximation

Thus,

$$\begin{aligned} B_{10} &= e^{\frac{\Delta BIC}{2}} \\ &= e^{\frac{43.52-37.23}{2}} \\ &= 23.22 \end{aligned}$$

This means that the data are approximately 23 times more likely under $\mathcal{H}_1$ than $\mathcal{H}_0$

BIC approximation

What about $p(\mathcal{H}_1 \mid \text{data})$?

It is easy to show

$$p(\mathcal{H}_1 \mid \text{data}) = \frac{B_{10}}{1 + B_{10}}$$

Thus, we have

$$\begin{aligned} p(\mathcal{H}_1 \mid \text{data}) &= \frac{22.87}{1 + 22.87} \\ &= 0.958 \end{aligned}$$

Optional stopping

Optional stopping

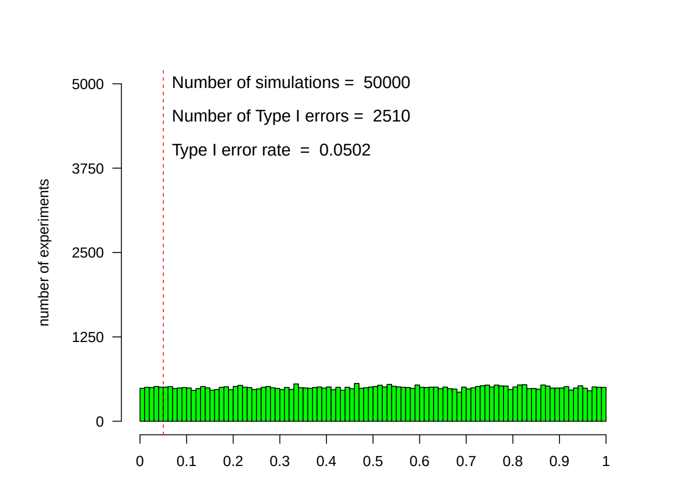
Optional stopping (the practice of stopping data collection when some desired threshold is obtained) is well known to be problematic in frequentist statistics.

To see why, consider the following simulation:

- consider a random sample from $\mathcal{N}(0, 1)$
- perform a single-sample t test against $\mu = 0$
- record the p-value
- do this many times
- count how many p-values are less than $\alpha = 0.05$ (Type I errors)

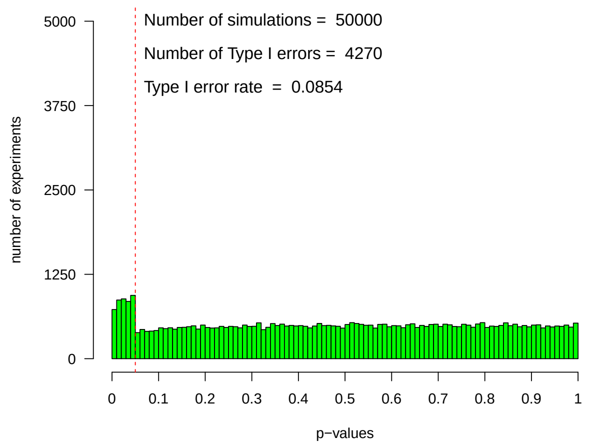
Optional stopping

If we only "look" at the data at the end (i.e., the full sample was collected), we see that the distribution of p-values is uniform, and Type I error rate is 5%



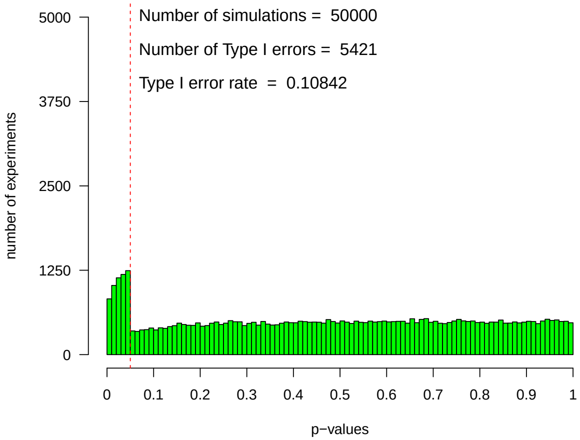
Optional stopping

Suppose we look halfway through data collection and stop if $p < 0.05$. Then we see that Type I error rate increases



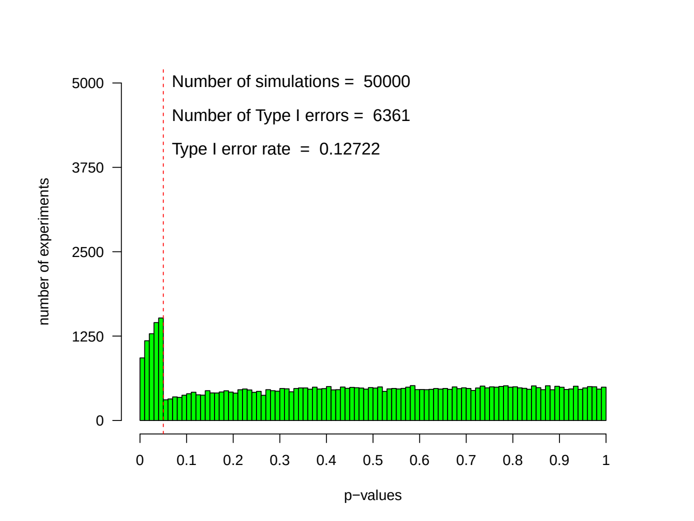
Optional stopping

A similar pattern continues with 3 looks...



Optional stopping

..and 4 looks...



Optional stopping

Some have argued (through similar simulations) that the same thing holds for Bayesians too..

Rouder (2014) counter-argues:

> *The critical error . . . is studying Bayesian updating conditional on some hypothetical truth rather than conditional on data. This error is easy to make because it is what we have been taught and grown familiar with in our frequentist training. (p. 308)*

Optional stopping

In other words, the Bayesian reasons about parameters, given observed data.

The correct question should be "Given that I've observed data Y, what is the relative probability that these data have come from Model 1 versus Model 2?"

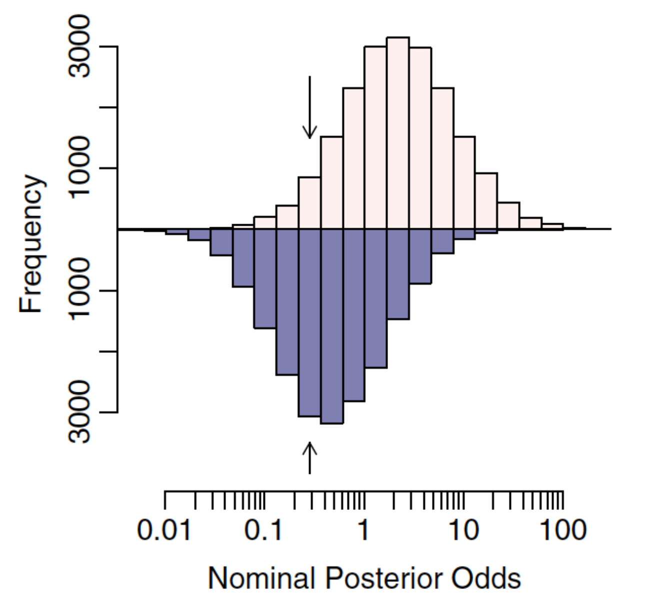
Rouder (2014) argues that optional stopping does not affect the answer to this question.

Optional stopping

To illustrate this, Rouder (2014) performed a simulation:

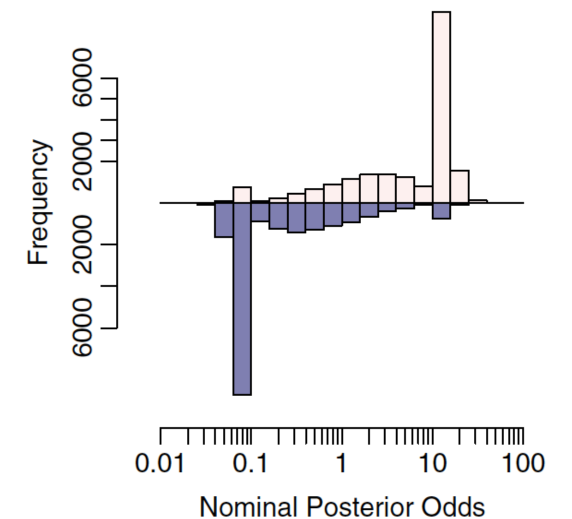
- start with two models, $\mathcal{H}_0$ and $\mathcal{H}_1$, a priori equally likely
- randomly pick one model and generate some data from it
- compute Bayes factor (which then equals posterior odds)

Optional stopping



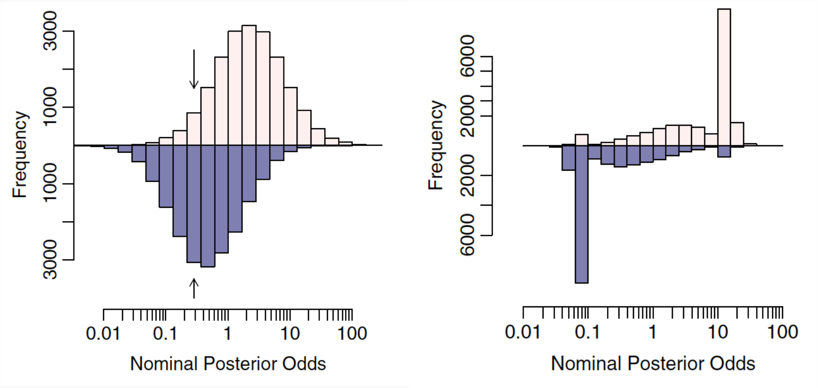
Optional stopping

Suppose next we stop sampling whenever we obtain a BF $\geq$ 10 in favor of either $\mathcal{H}_0$ or $\mathcal{H}_1$.



Optional stopping

Even though the distribution of Bayes factors is changed, the interpretation is the same. Conditional on observed data, the Bayes factor directly indexes the relative likelihood that the data came from either model.



Thank you!

- Thomas J. Faulkenberry
- Department of Psychological Sciences
- Tarleton State University
- `faulkenberry@tarleton.edu`
- Twitter: `@tomfaulkenberry`