BayesTesting.jl
Bayesian Hypothesis Testing without Tears

Jeff Mills
Lindner College of Business
University of Cincinnati

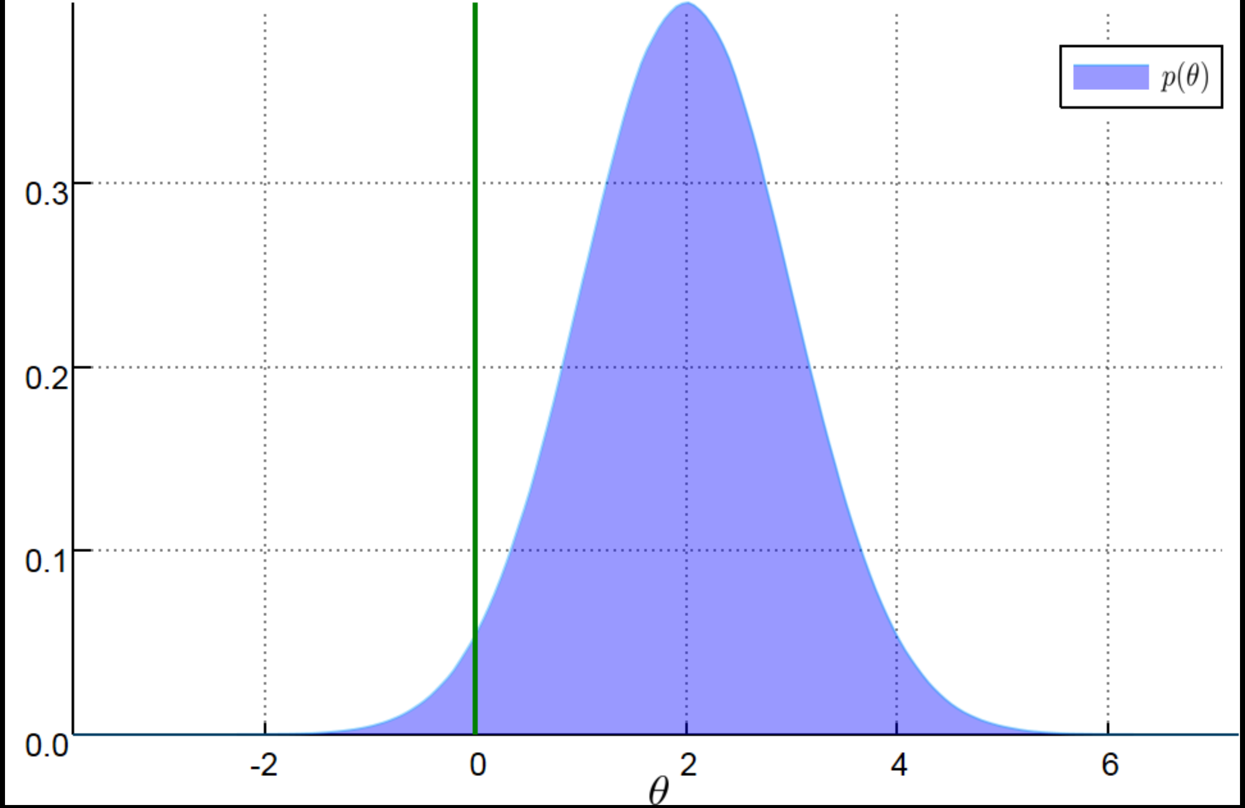
JuliaCon, 2018

A fresh approach to hypothesis testing using a fresh approach to numerical computing.

We wish to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ (usually with $\theta_0 = 0$).

A picture paints a thousand statistics.

Where's zero?

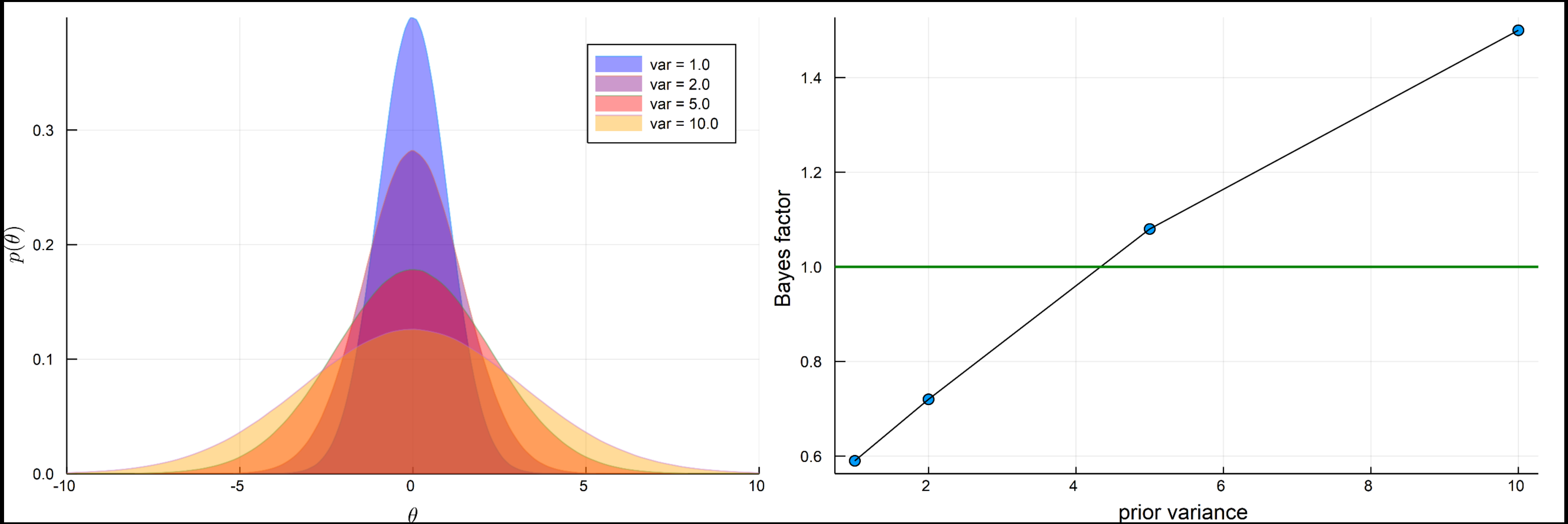


- **Test** $\boldsymbol{H_0: \theta = \theta_0}$ **vs.** $\boldsymbol{H_1: \theta \neq \theta_0}$**.** **Why the tears?**

- **Frequentist NHST and *p*-values** - wide agreement that unsatisfactory [Gelman and Carlin, 2017, *JASA*]
  We created a monster. And we keep feeding it, hoping that it will stop doing bad things. It is a forlorn hope. - Don Berry (2017) A *p*-Value to Die For. *JASA*

- **Bayes factors** are arguably worse - suffer from several theoretical problems that have serious practical implications. (Bernardo, 1999, Cousins, 2017).
  - assigning nonzero prior probability to a quantity of zero measure
  - inability to use the same priors as used for inference
  - prior is conditional on the null hypothesis
  - the Jeffreys-Lindley-Bartlett (JLB) paradox

The Jeffreys-Lindley-Bartlett paradox



- As the prior variance increases, the Bayes factor (odds in favor of the null
- hypothesis) increases *regardless of the data.*

- **Bayesian testing literature**: find `priors’ that mitigate the practical impact of the Jeffreys-Lindley-Bartlett paradox - very limited success [Cousins, 2017, *Synthese*].

- Bayesian *inference* with a wide variety of conventional priors works well.

- Attempting to solve the testing problem by modifying the prior has not worked.

- **The problem is not the prior**.

What is the correct alternative hypothesis: one, or all?

Is Serena Williams the best female tennis player in the world?

To test this hypothesis, we do not play **Serena vs. the world** all at once!

We have everyone play a series of **one-on-one** matches.

The problem is the poorly defined alternative hypothesis, $H_1: \theta \neq \theta_0$.

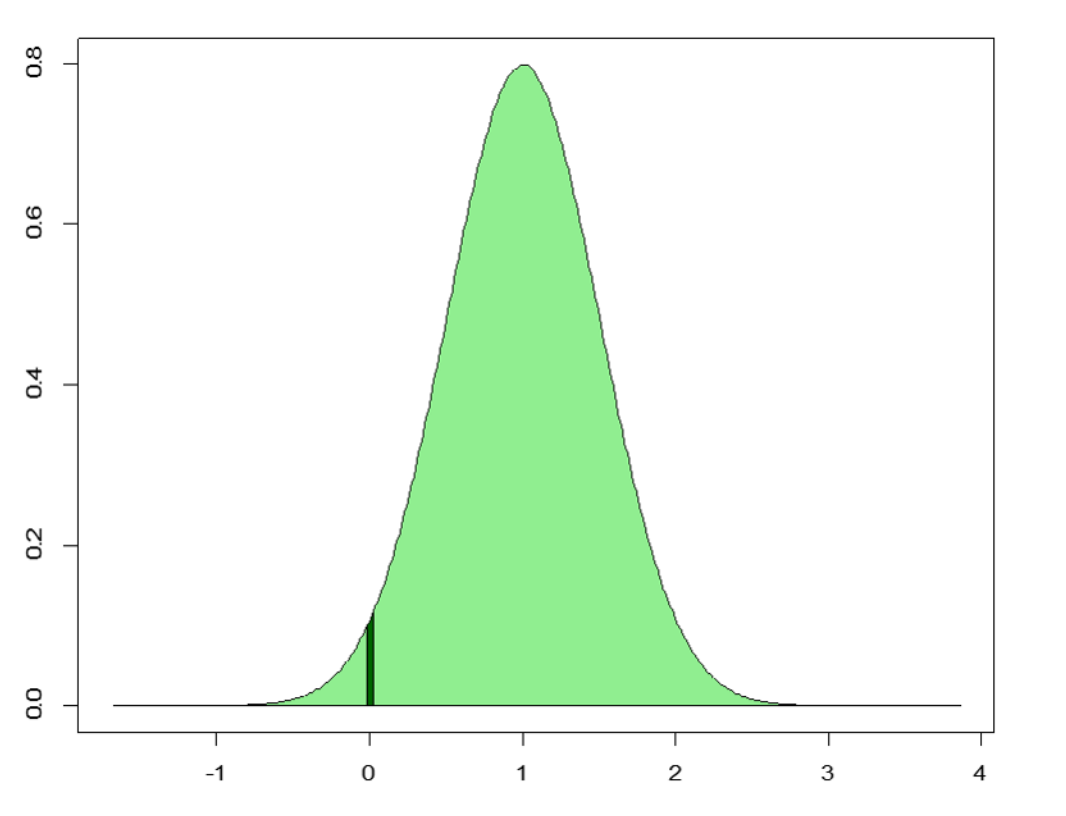
- **Prevailing Bayesian Hypothesis testing procedure**

$H_0\!: \theta = \theta_0$ (or $H_0\!: |\theta - \theta_0| < \varepsilon$)
($\boldsymbol{prob.} = \mathbf{0}$)

$H_1\!: \theta \neq \theta_0$ (or $H_1\!: |\theta - \theta_0| > \varepsilon$)
($\boldsymbol{prob.} = \mathbf{1}$) *(one vs. many)*

Leads to the JLB paradox

Bayes factor, $\frac{p(H_1|D)}{p(H_0|D)} \to \infty$ as $\varepsilon \to 0.$



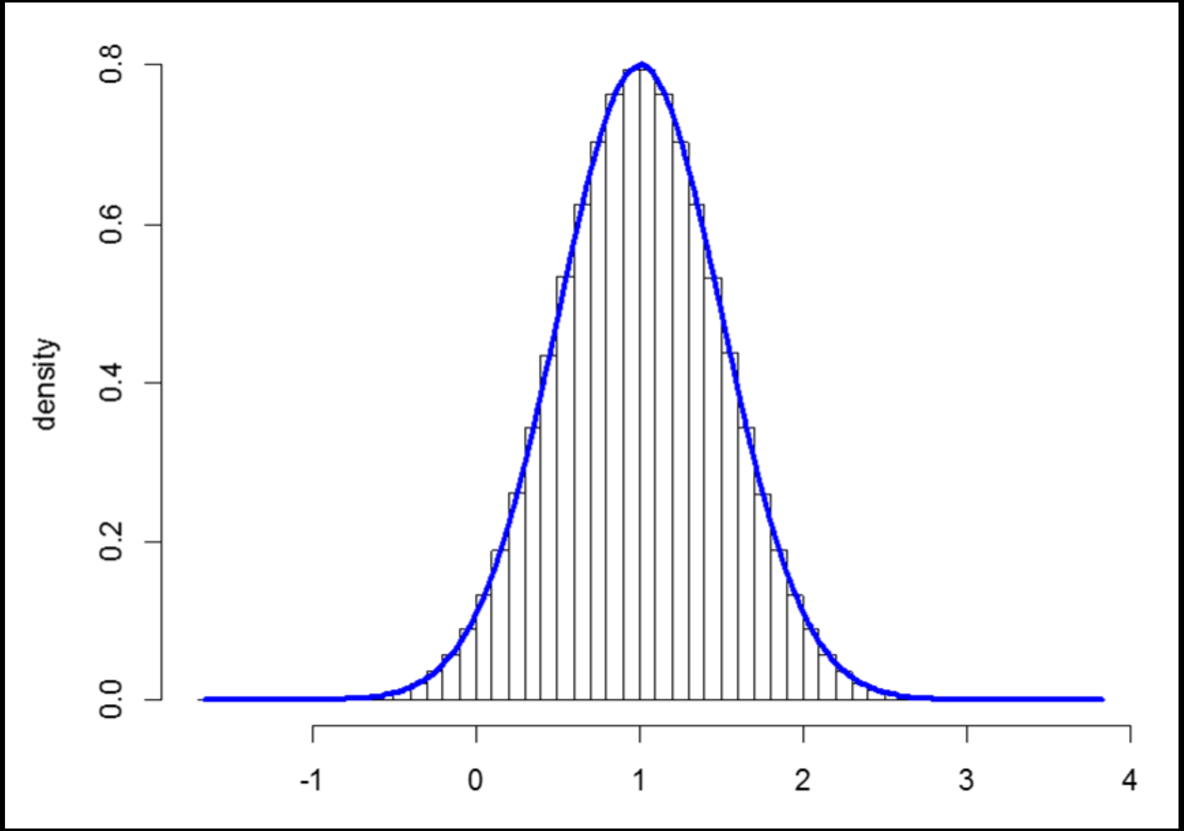
Alternative Bayesian Hypothesis testing procedure

For any ε,

$H_0: |\theta - \theta_0| < \varepsilon$
$(\boldsymbol{prob.} > \mathbf{0})$

$H_z: |\theta - \theta_z| < \varepsilon$ (*one vs. one*)
$(\mathbf{0} \leq \boldsymbol{prob.} \leq \mathbf{1})$

$\theta_z \in \{\Theta: \theta_z = \theta_{z-1} + 2\varepsilon, z \in \mathbb{Z}\}$,
$z \in \mathbb{Z}$, the integers.



H_z partitions the hypothesis space.

Minimize expected loss from decision to choose H_0 vs. H_z, for all z.

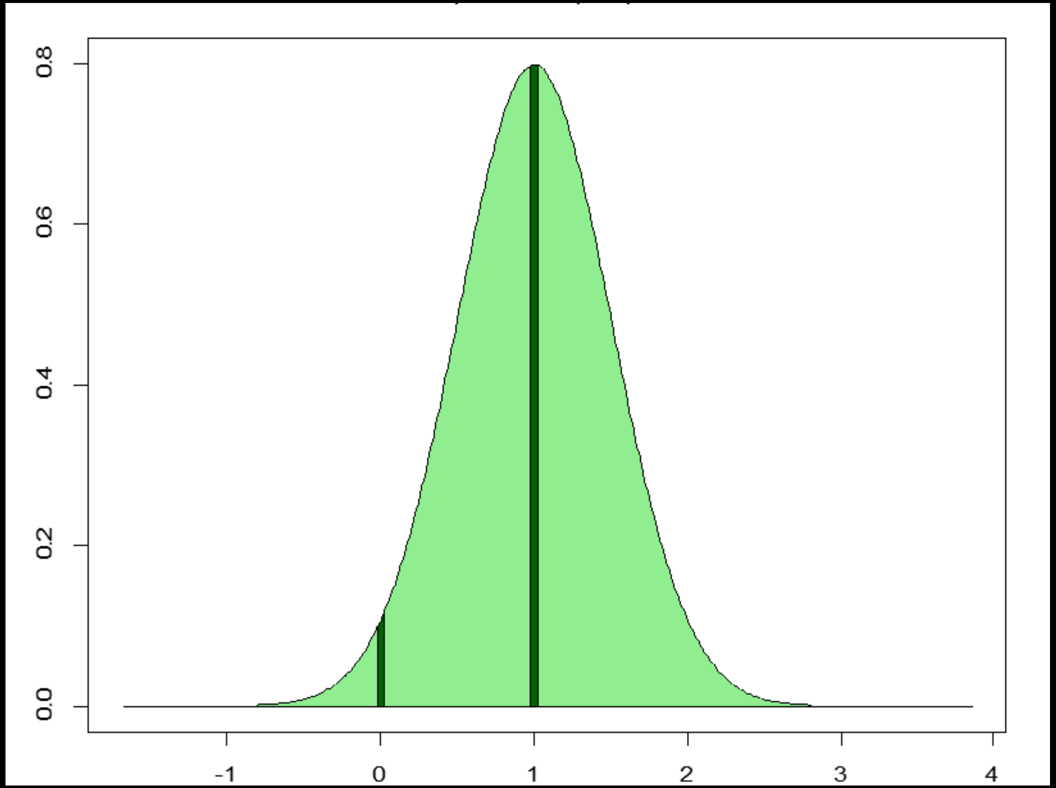
Decision rule: reject H_0 if

$$\frac{p(H_z|D)}{p(H_0|D)} \geq \frac{c_0}{c_1}$$

c_0 = cost of type I error,
c_1 = cost of type II error.

Equivalent to: reject H_0 if

$$\sup_z \left(\frac{p(H_z|D)}{p(H_0|D)} \right) \geq \frac{c_0}{c_1}.$$



$$\text{As } \varepsilon \to 0, \qquad \sup_z \left(\frac{p(H_z|D)}{p(H_0|D)}\right) \to \frac{\sup\limits_z p(\theta = \theta_z\,|D)}{p(\theta = \theta_0\,|D)}.$$

In practice, evaluate the posterior density at θ_0 and at the MAP estimate,

$$O = \frac{p(H_z|D)}{p(H_0|D)} = \frac{p(\theta = \bar{\theta}\,\big|D)}{p(\theta = \theta_0|D)},$$

$$\bar{\theta} = \text{argmax}_\theta\, p(\theta|D) \quad \text{(MAP estimate)}$$

Odds against the null hypothesis compared to the most likely alternative

- A new (Bayesian) testing procedure that does not suffer from the flaws of the current standard approaches

- The JLB paradox does not occur
- Can use same priors for inference and testing
- Priors are not dependent on the null hypothesis
- Axioms of probability are not violated
- Odds are easier than *p*-values to interpret correctly
- A UMP, robust (in the Bayesian sense) test
- Easy to combine with posterior simulation allowing exact inference and testing
- It works!   Be skeptical, but look at the evidence

  try it - a Julia package: BayesTesting.jl

- **A package for ease of use: BayesTesting.jl**
- https://github.com/tszanalytics/BayesTesting.jl
- DOI: 10.13140/RG.2.2.18951.70564

- Generic function for computing posterior odds given an MC (or MCMC) sample from the posterior

```
using BayesTesting
mcodds(mc_sample)         # default null hypothesis of zero
mcodds(mc_sample,h0=2.0)  # set the value in the null
```

- Additional functions for comparison of means, proportions, difference in differences, regression coefficients, etc.

A few canonical examples …

Testing a mean: $H_0 : \mu = \mu_0$

With unknown variance and uninformative prior so that $p(\bar{\mu})/p(\mu_0) = 1$.

For a sample from *any* distribution (CLT and Maxent), the posterior density of the mean with unknown variance is a Student-t.

$$O = \left(1 + \frac{t^2}{v}\right)^{\frac{v+1}{2}}, \qquad t = \sqrt{(\bar{\mu} - \mu_0)^2 / s^2}$$, the usual *t*-statistic.

For evidence against *H₀*, instead of *p*-value, use:

Odds of **4:1** $\approx$ **10**% *p*-value, **7:1** $\approx$ **5**% *p*-value, **30:1** $\approx$ **1**% *p*-value

Linear regression coefficient

Data on presidential candidates’ height difference and difference in vote popularity.

$$pop_i = \alpha + \beta ht_i + u_i, \qquad u_i \sim N(0, \sigma^2)$$

$pop_i =$ the proportion of votes received by winner.

$ht_i =$ height ratio of the winner to other candidate

On some great and glorious day the plain folks of the land will reach their heart's desire at last, and the White House will be adorned by a downright moron.

- H. L. Mencken, 26 July, 1920

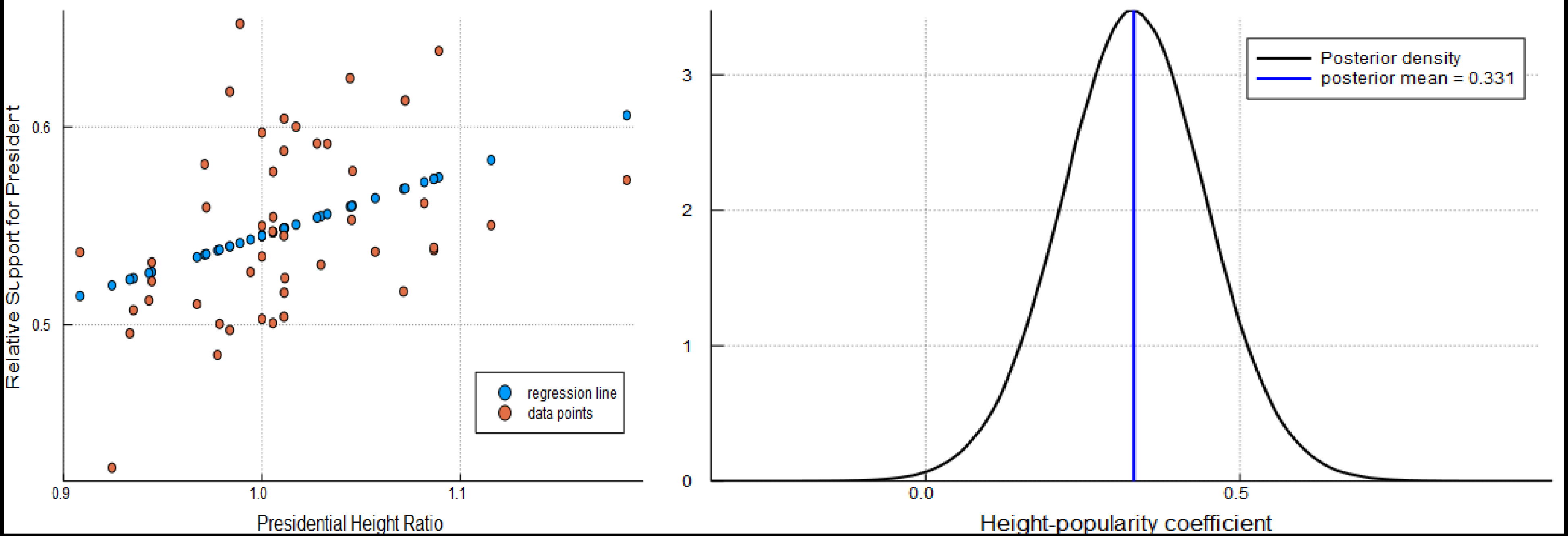
Example code

```
using CSV, BayesTesting
dat = CSV.read(“Presidents.csv")
ht = Array(dat[2])
pop = Array(dat[3])

# using blinreg.jl function:
bhat, seb, odds, pval, s, R2 = blinreg(pop,ht)
```

RESULTS:

		intercept	slope
coeffs	=	0.214	0.331
s.e's	=	0.116	0.114
odds	=	5.384	51.119
p-value	=	0.071	0.0058

s^2 (eqn. variance) = 0.0017, R-squared = 0.155



Test of $H_0: \beta = 0.0,$ posterior odds = 51.12 (*p*-value = 0.006)

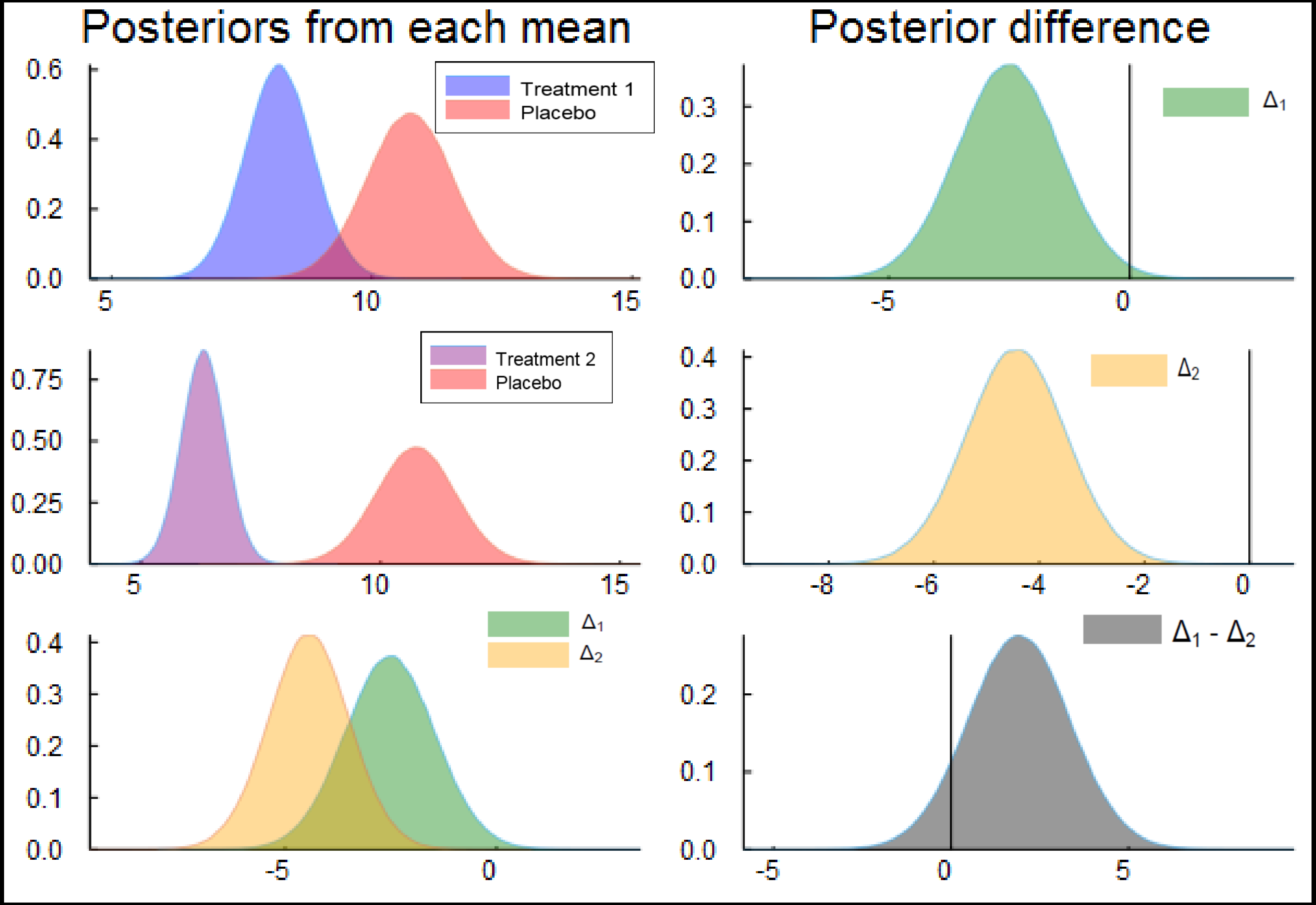
Prevailing Bayes factor approach: 6.33:1 against the null hypothesis

- very sensitive to the prior

```
using Distributions, BayesTesting

# function   x, y are two samples to compare means

plt, diff_mean, qnts, tst = compare_means(x,y)

# OR

compare_means(m1, m2, s1, s2, n1, n2) # m1, m2, are sample means
                                      # s1, s2, are sample SDs
                                      # n1, n2 are sample obs
# difference differences in means
d_means = diff_mean1 – diff_mean2     # analytically intractable
mcodds(d_means)

mean(d_means);  std(d_means)   # also compute quantiles, plot, etc.
```

Difference in differences (analytical exact distribution unknown)

	Odds	*p*-value
Δ_1	15.73	0.0194
Δ_2	24400.8	<0.0001
Δ_3	2.45	0.1832



- ***Theory:*** *Mills (2018) Objective Bayesian Precise Hypothesis Testing.*

- ***Package:*** *\github\tszanalytics\BayesTesting.jl*

- ***Applications:***
  - *RCTs methodology: Mills et al. (2018)*
  - *Comparison of means and proportions: Strawn et al. (2017)*
  - *Hypothesis testing with only summary statistics: Strawn et al. (2017)*
  - *Meta-analysis: Strawn et al. (2017)*
  - *ANOVA testing: Mills and Namavari (2017)*
  - *Unit root testing: Mills (2016)*
  - *Cointegration: Mills and Namavari (2018)*
  - *Predictive model selection: Cornwall and Mills (2018)*

  - ***Thank you!***

Supplementary material

Use posterior simulation - No need to rely on large sample assumptions.

Apply the same asymptotic theory to numerical approximation of the exact (small sample) posterior density.

LLN: for a sample of M draws from the posterior for θ, as $M \to \infty$,

$$\frac{1}{M}\sum_{r=1}^{M} f\left(\theta^{(r)}\right) \to E\left(f(\theta)\right),$$

where $\theta^{(r)}$ is the rth pseudo-sample value.

Joint testing and Rao-Blackwellization: the posterior odds ratio for testing a set of linear restrictions, $H_0\colon R\theta = r$,

$$O = \frac{\int p(\bar{\theta}_R, \phi \big| R\theta - r = 0, D)\, d\phi}{\int p(\bar{\theta}, \phi \big| D)\, d\phi}$$

This is computed as,

$$O = \frac{\sum_{i=1}^{M} p(\bar{\theta}_R, |\phi^{(i)}, R\theta - r = 0, D)}{\sum_{i=1}^{M} p(\bar{\theta}|\phi^{(i)}, D)}$$

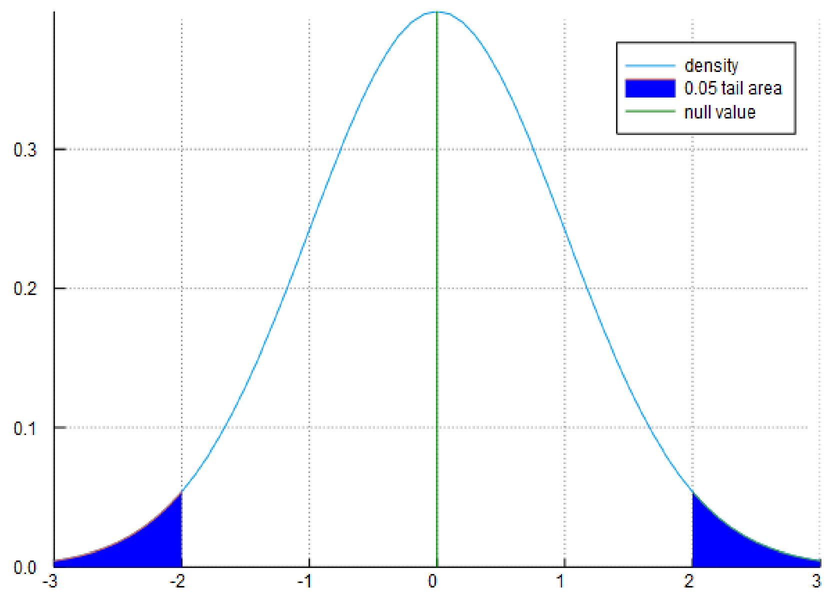
ϕ = vector of nuisance parameters,

$\bar{\theta}_R$ is a vector of MAP estimates from the restricted joint posterior

$\bar{\theta}$ is a vector of MAP estimates from the unrestricted joint posterior

- “The null hypothesis significance testing procedure (NHSTP)" is problematic – see all the hand wringing in the stats and psych. literature, *JASA* 2017 (and previously), etc.

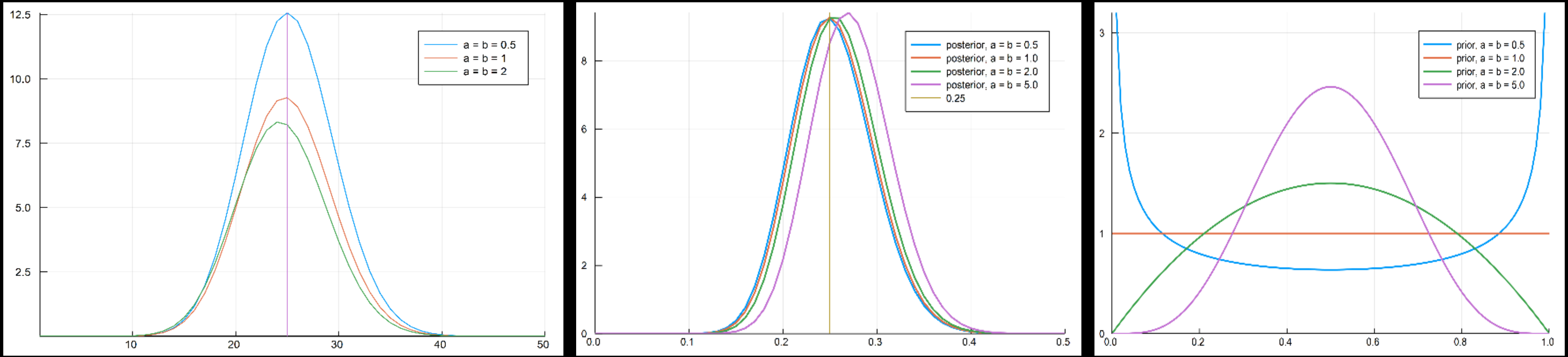
- The null hypothesis is assumed true to derive a test procedure, then the null can be rejected …, logically that means the test procedure is no longer valid because it is conditional on a hypothesis that has been rejected. (Jaynes, 2003, p.524)



- Areas in the tail were not observed - only one of the inner boundaries.

- Positive test statistic can be rejected because we include area in negative tail.

Sensitivity to variations in the prior

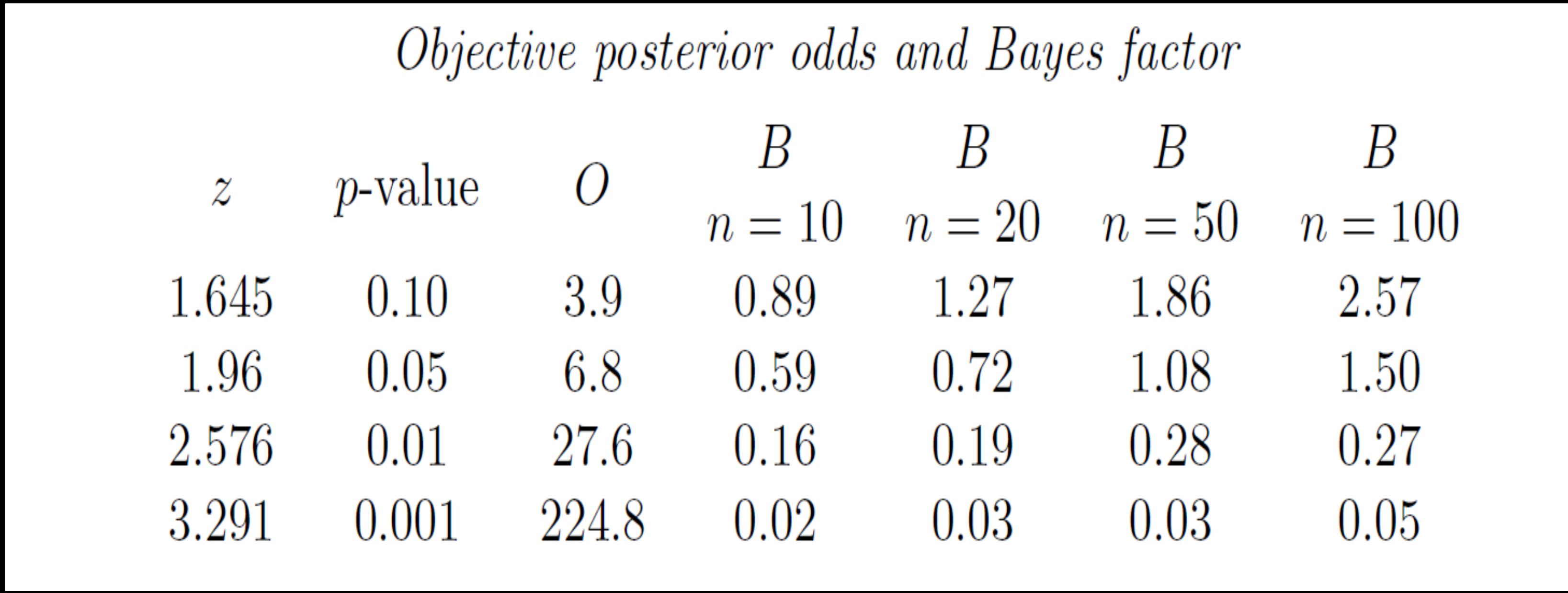


Even with an informative prior that is incorrect (a = b = 5) (right panel), the posterior (middle panel) with 30 observations remains close to the uninformative posterior.

The Standard Bayes factor is very sensitive to the prior parameter values (left panel). Objective odds (ratio of height of posterior) are robust .

Testing a mean: $H_0 \colon \mu = \mu_0$ with known variance, using an uninformative prior so that $p(\mu_0) = p(\bar{\mu})$, the posterior odds ratio is,

$$O = \exp(z^2/2\sigma^2)$$



Objective posterior odds and Bayes factor

z	p-value	O	B $n = 10$	B $n = 20$	B $n = 50$	B $n = 100$
1.645	0.10	3.9	0.89	1.27	1.86	2.57
1.96	0.05	6.8	0.59	0.72	1.08	1.50
2.576	0.01	27.6	0.16	0.19	0.28	0.27
3.291	0.001	224.8	0.02	0.03	0.03	0.05

For a z-statistic of 1.96:

p-value = 0.05

posterior odds = 6.8

Bayes factor varies with N: giving odds of

1.7:1 ***against*** H_0 with 10 obs.

1.5:1 ***in favor*** of H_0 with 100 obs.