Investigating the Relationship Between the Bayes Factor and the Separation of Credible Intervals

Zhengxiao Wei [1]
Dr. Farouk S. Nathoo [1]
Dr. Michael E. J. Masson [2]

[1]Department of Mathematics and Statistics
[2]Department of Psychology
University of Victoria

October 12, 2022

Outline

Motivation

- ☑ Null-hypothesis significance testing and p-value
- ✎ Bayes factor and credible interval for within-subject designs

Problems with p-values

- p-values depend on **hypothetical data** that were never observed,
  e.g., uncensored data from a *censored* volt-meter.
- p-values depend on possibly unknown **subjective intentions**,
  e.g., optional stopping.
- p-values do **not quantify** statistical evidence,
  e.g., p-postulate.

(Wagenmakers, 2007;
Wasserstein and Lazar, 2016;
Greenland et al., 2016;
Goodman, 2008)

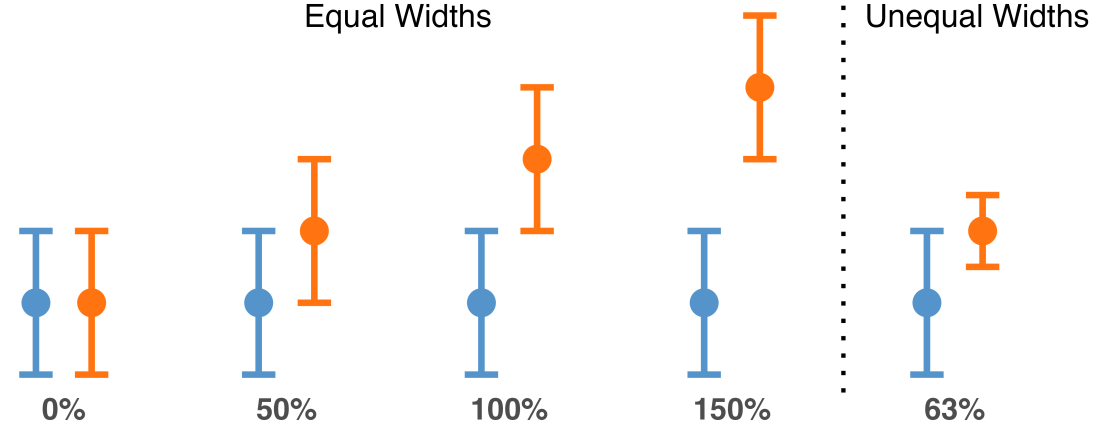
Method

- Bayes factor (**BF**) $\longrightarrow$ measure of evidence $\longrightarrow$ model selection

$$\underbrace{\frac{p(\mathcal{H}_0 \mid \text{Data})}{p(\mathcal{H}_1 \mid \text{Data})}}_{\text{posterior odds}} = \underbrace{\frac{p(\text{Data} \mid \mathcal{H}_0)}{p(\text{Data} \mid \mathcal{H}_1)}}_{\text{Bayes factor } BF_{01}} \cdot \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\text{prior odds}} \tag{1}$$

- Interval estimates of effect sizes (rather than the p-value dichotomy)
  - ▷ Frequentist: confidence interval (**CI**)
  - ▷ Bayesian: credible interval (CrI)
    e.g., highest-density interval (**HDI**)

- Graphical display
  - ▷ Two intervals' separation percentage



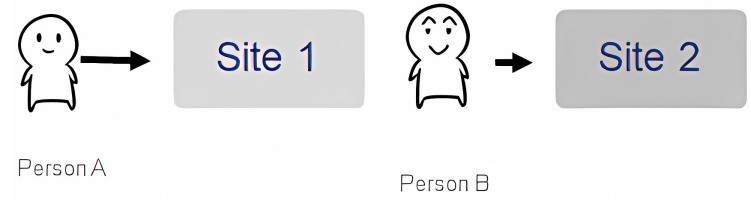
Experimental Designs

- A one-way **between-subjects design**

$$\mathcal{M}_1: \; Y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad (\textbf{let } \tau_i = \sigma_\epsilon t_i) \tag{2}$$

versus $\mathcal{M}_0: \; Y_{ij} = \mu + \epsilon_{ij},$

$$\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2) \text{ for } i = 1, \cdots, a; \; j = 1, \cdots, n.$$



- A "one"-way **within-subject** (repeated-measures) **design**

$$\mathcal{M}_1: \; Y_{ij} = \mu + \sigma_\epsilon(t_i + b_j) + \epsilon_{ij} \tag{3}$$

versus $\mathcal{M}_0: \; Y_{ij} = \mu + \sigma_\epsilon b_j + \epsilon_{ij}$

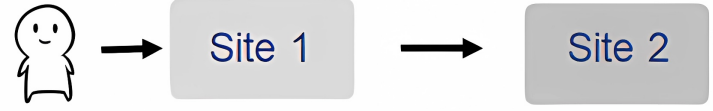


Image: Nielsen Norman Group

Variance Partitioning

R COMMANDS

Remove the between-subjects variability in within-subject designs

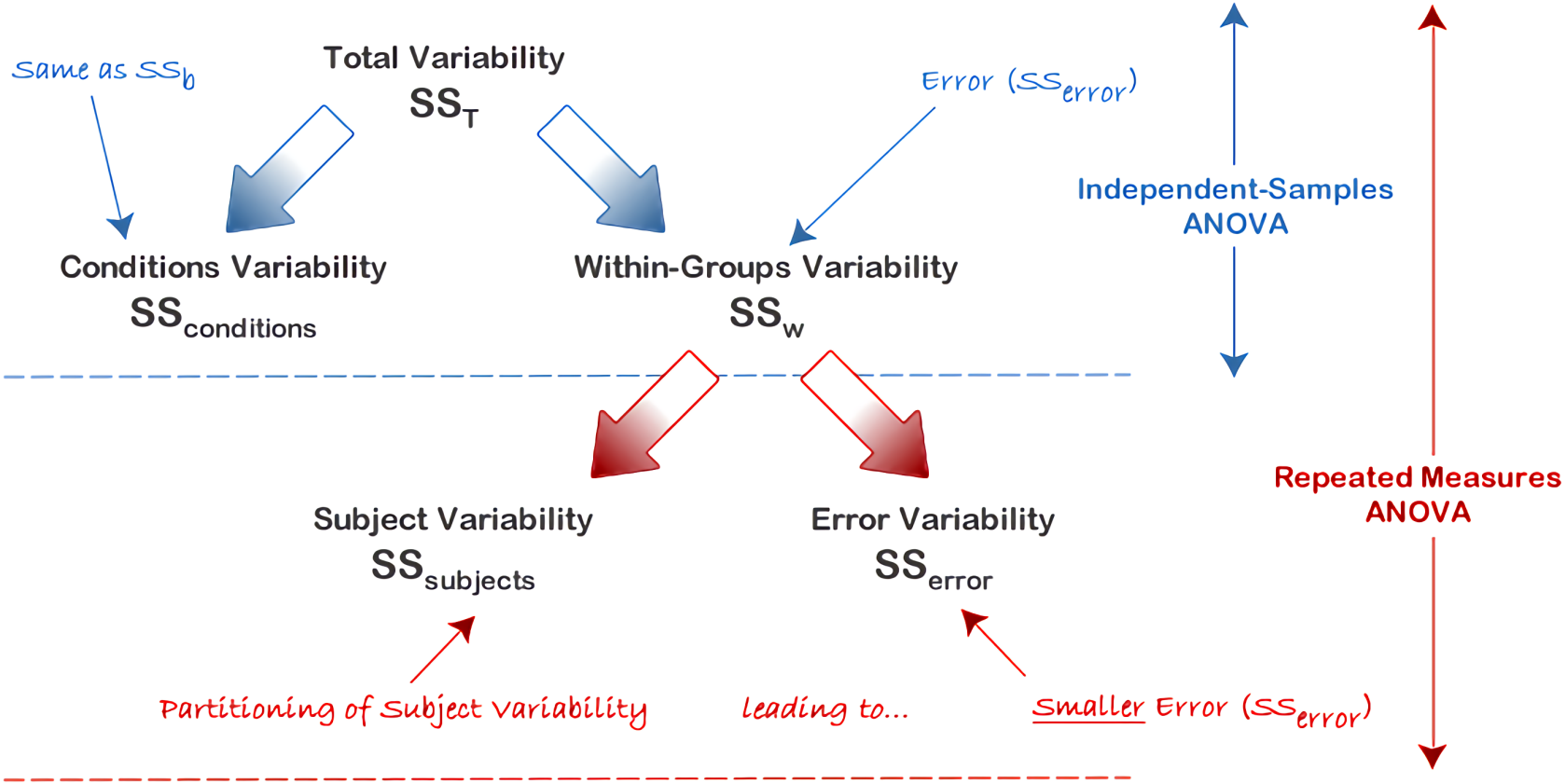


Image: Laerd Statistics

Within-Subject Interval Estimates

- Significant difference between two sample means: **both designs**
  - ☑ $|M_{p\cdot} - M_{q\cdot}| > \sqrt{2} \times$ CI width (Loftus & Masson, 1994) (4)
  - ☑ $\sqrt{2}/2 \approx 71\%$ CI separation (5)

- **Four within-subject interval estimates**

Label	Equation	Description
CI	$M_{i\cdot} \pm \sqrt{\frac{SS_{\mathrm{W}}}{n(n-1)a}} \cdot t^*_{1-\frac{\alpha}{2},\ a(n-1)}$	standard confidence interval
HDI	Markov chain Monte Carlo sampling of μ_i	standard highest-density interval
LM-CI	$M_{i\cdot} \pm \sqrt{\frac{SS_{\mathrm{S\times C}}}{n(n-1)(a-1)}} \cdot t^*_{1-\frac{\alpha}{2},\ (n-1)(a-1)}$	within-subject CI
NKM-HDI	$M_{i\cdot} \pm \sqrt{\frac{SS_{\mathrm{S\times C}}}{n(n-1)a}} \cdot t^*_{1-\frac{\alpha}{2},\ a(n-1)}$	conditional within-subject HDI
LH- or JZS-HDI	$\mathbb{E}\left[\mu_i \pm \frac{\sigma_\epsilon}{\sqrt{n}} \cdot t^*_{1-\frac{\alpha}{2},\ a(n-1)} \mid \text{Data}\right]$	modification of NKM-HDI

Hierarchical Specification

Priors (Rouder et al., 2012 & 2017)

- Jeffreys prior, $\pi(\mu, \sigma_\epsilon^2) \propto \frac{1}{\sigma_\epsilon^2}$; (6)
- $t_i \mid g \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g)$; $\quad b_j \mid g_b \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_b)$; (7)
  which treat the effect sizes as random effects and allow for shrinkage;
- $g \sim$ Scale-inv-$\chi^2(1, h^2)$; $\quad g_b \sim$ Scale-inv-$\chi^2(1, h_b^2)$; (8)
- By default, $h = 0.5$ for the fixed effects
  and $h = 1$, $h_b = 1$ for the random effects (9)

The marginal prior density of the column vector $(t_1, \cdots, t_a)^\top$ is a multivariate Cauchy distribution.

Bayes factor

$$BF_{10} = \frac{p(\text{Data} \mid \mathcal{M}_1)}{p(\text{Data} \mid \mathcal{M}_0)} \tag{10}$$

Hierarchical Specification

Priors (Rouder et al., 2012 & 2017)

- Jeffreys prior, $\pi(\mu, \sigma_\epsilon^2) \propto \frac{1}{\sigma_\epsilon^2}$; (6)
- $t_i \mid g \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g)$; $b_j \mid g_b \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_b)$; (7)
  which treat the effect sizes as random effects and allow for shrinkage;
- $g \sim \text{Scale-inv-}\chi^2(1, h^2)$; $g_b \sim \text{Scale-inv-}\chi^2(1, h_b^2)$; (8)
- By default, $h = 0.5$ for the fixed effects
  and $h = 1$, $h_b = 1$ for the random effects (9)

The marginal prior density of the column vector $(t_1, \cdots, t_a)^\top$ is a multivariate Cauchy distribution.

Bayes factor

$$BF_{10} = \frac{p(\text{Data} \mid \mathcal{M}_1)}{p(\text{Data} \mid \mathcal{M}_0)} \tag{10}$$

Modeling Fixed Effects

Projecting a set of a effects $(t_1, \cdots, t_a)$
into $a-1$ parameters $(t_1^\star, \cdots, t_{a-1}^\star)$
with the property that the marginal prior on all a effects is identical

- $t_i^\star \mid g \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g)$; (7′)
- $(t_1^\star, \cdots, t_{a-1}^\star) = (t_1, \cdots, t_a) \cdot \mathbf{Q}$; (11)
- $\mathbf{I}_a - a^{-1}\mathbf{J}_a = \mathbf{Q} \cdot \mathbf{Q}^\top$; (12)

e.g., $t_1^\star = \dfrac{\sqrt{2}}{2}(t_1 - t_2)$ when $a = 2$;

where $\mathbf{Q}$ is an $a \times (a-1)$ matrix of the $a-1$ eigenvectors of unit length corresponding to the nonzero eigenvalues of the left side term in (12).

Jeffreys-Zellner-Siow Bayes Factor

The resulting *JZS-BF*$_{10}$ from (6)-(9) avoids critical issues:

- Bartlett's paradox

  The Bayes factor approaches zero as the prior variance increases.

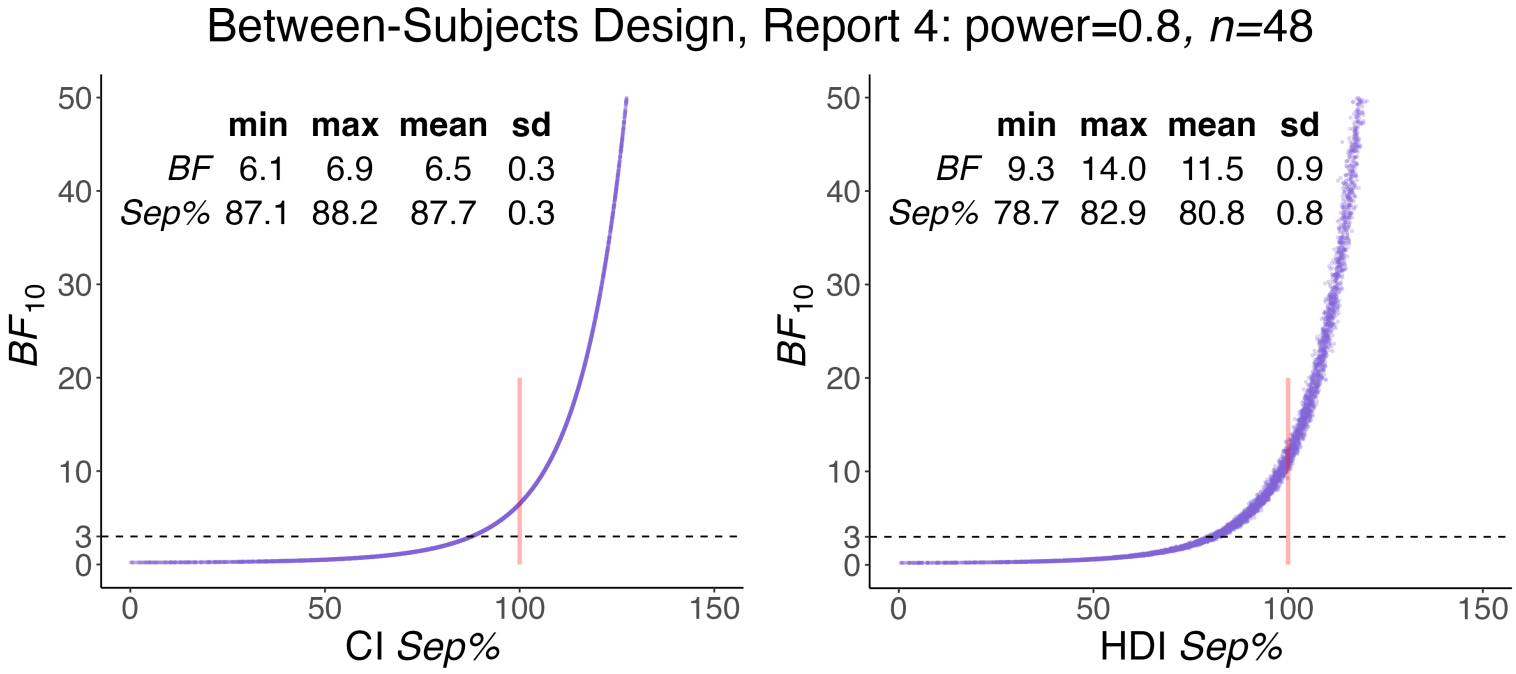
- Information paradox

  The Bayes factor tends to be bounded, given overwhelming information.

Another consideration is the Pearson Bayes factor (in an analytic form; Faulkenberry, 2021).

Results: Between-Subjects Designs

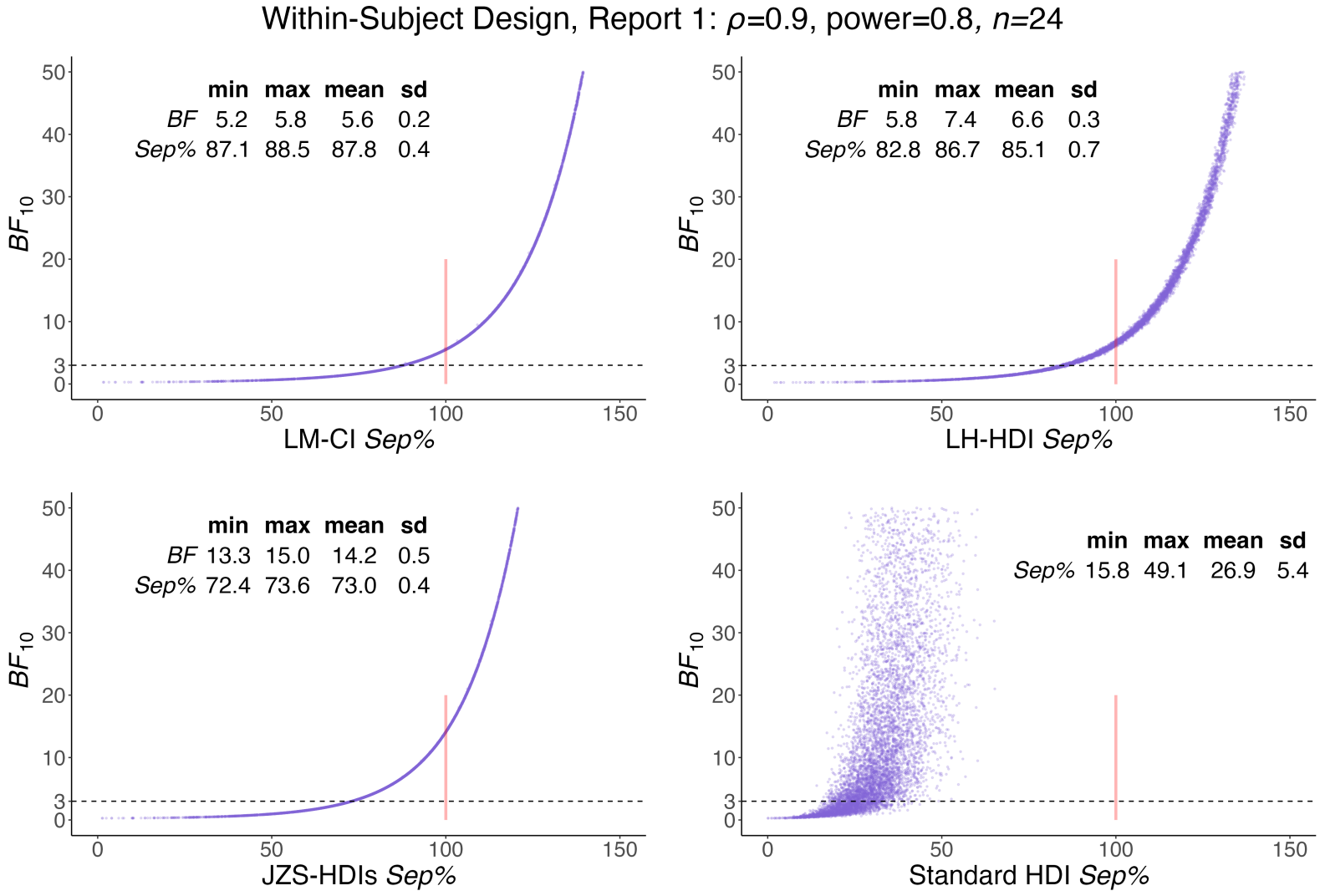
71% CI separation (based on $\alpha = .05$) (5)



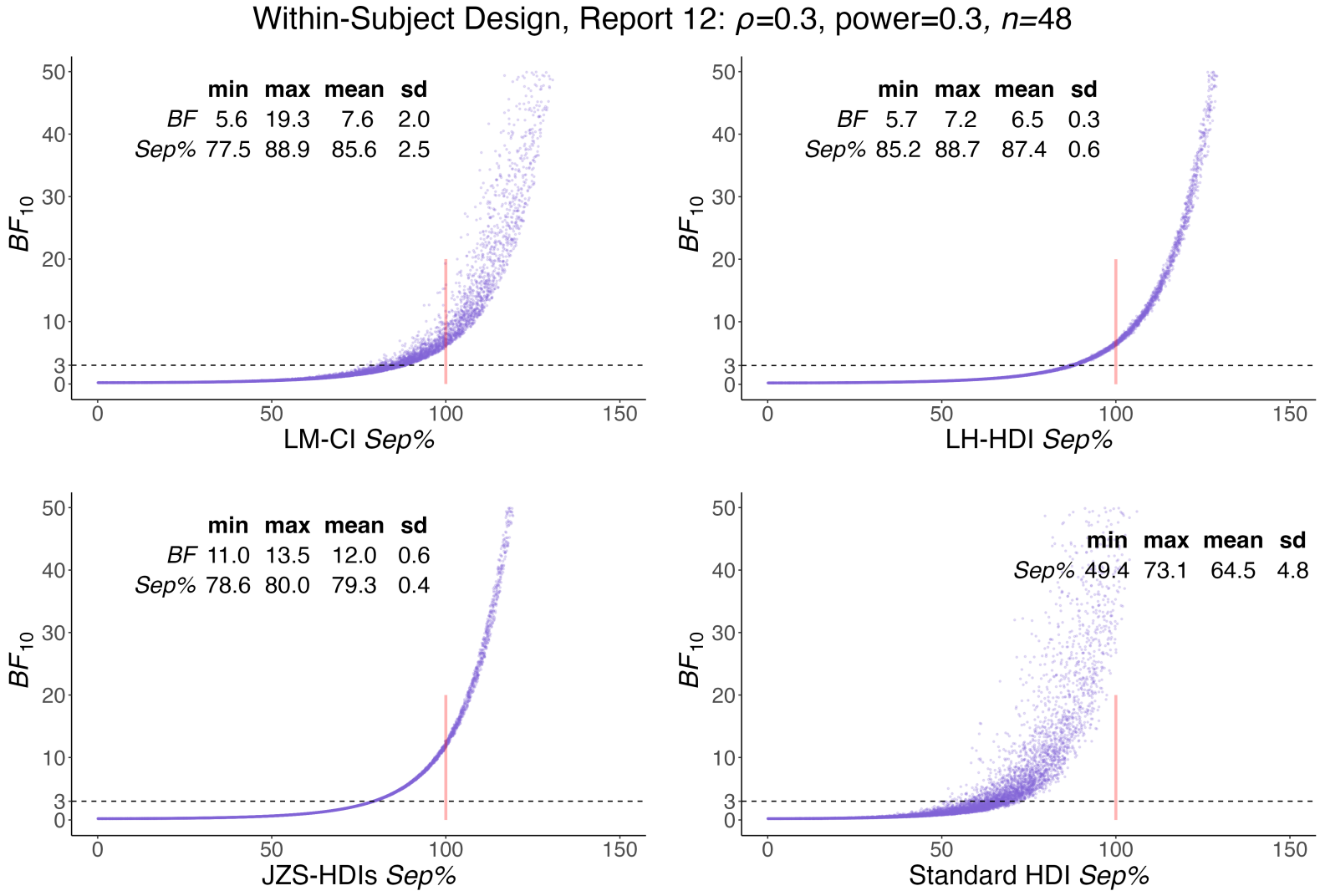
Why is Bayesian inference more conservative?

The p-value of **.01** may correspond to the Bayes factor value of **3** for some n.

Results: Within-Subject Designs

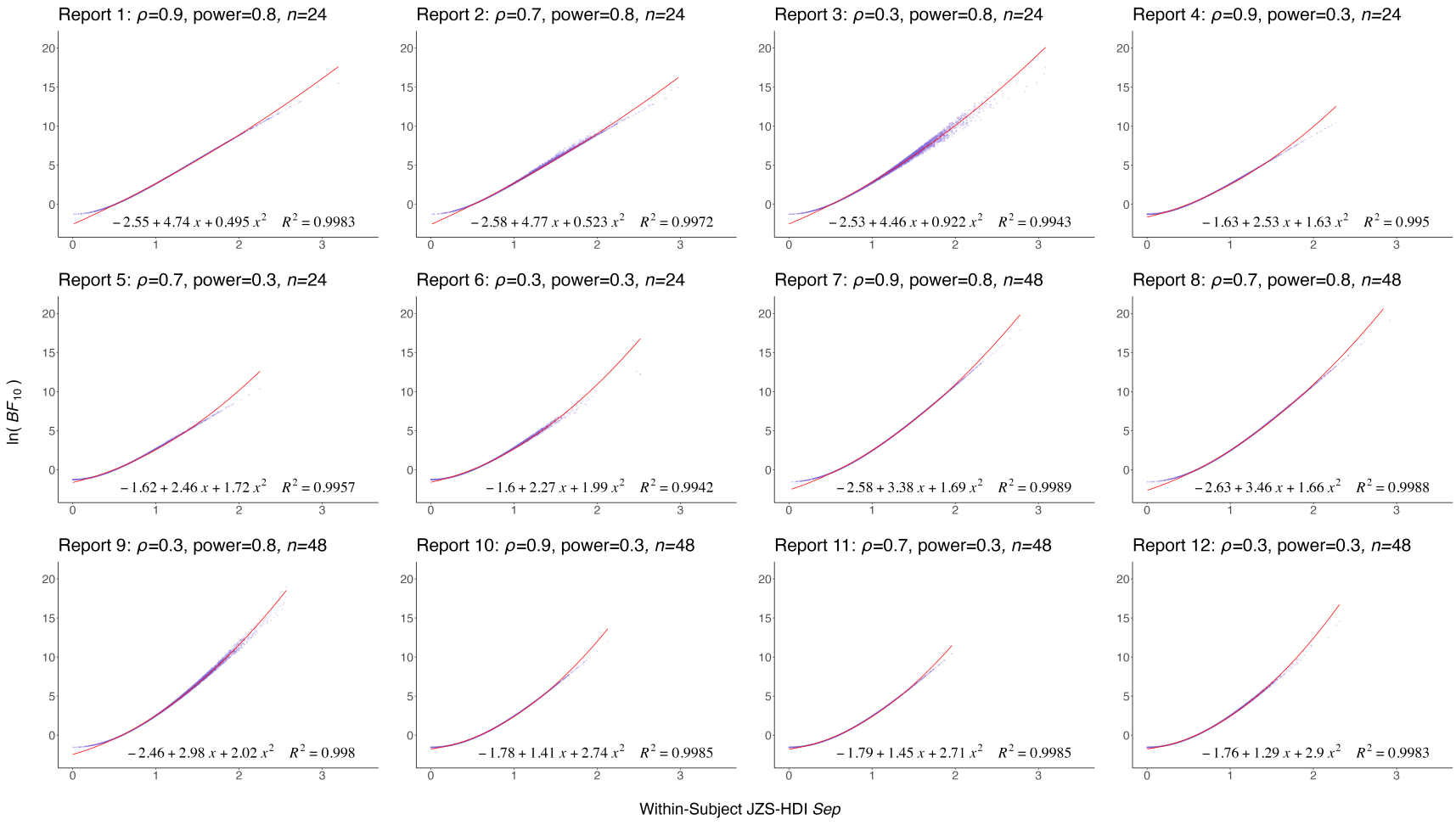


Results: Within-Subject Designs (continued)

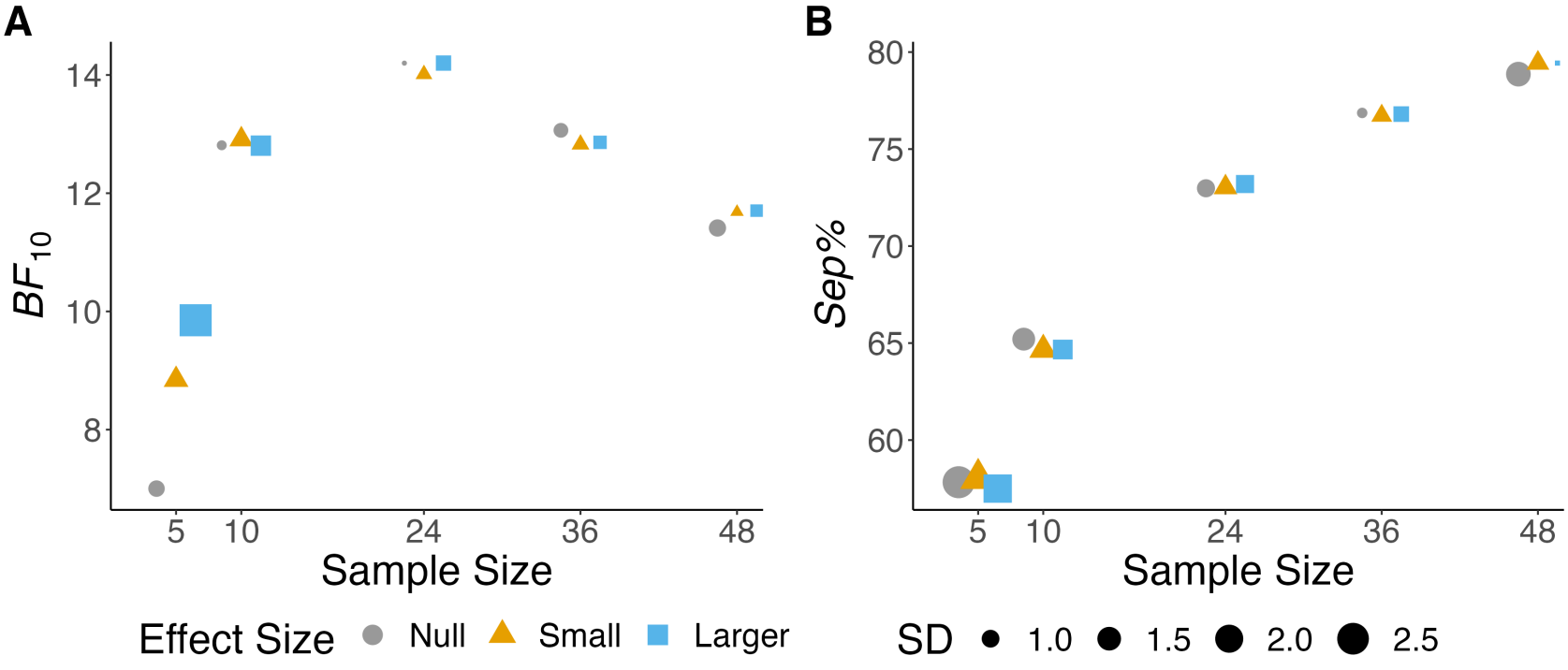


Curve Fitting for Case JZS-HDI

$$\ln BF \sim a + b \times Sep + c \times Sep^2$$



Varying Sample Sizes



Where We Were

Initiated Sept. 23, 2020

- Open Science Framework

  `https://osf.io/x2pvw/`

- *Statistics Jan. 19, 2022* and *CaBS Oct. 1, 2021* seminars
- *WNAR Jun. 14, 2021* and *SSC Jun. 7, 2021* conferences

- '*rmBayes*' `R` package CRAN GitHub pkgdown RPubs

  Submitted Sept. 14, 2021; Accepted Sept. 15, 2021

  $>$ install.packages("rmBayes", type $=$ "binary")

- *Psychonomic Bulletin & Review* journal

  Submitted May 10, 2022; Accepted Apr. 16, 2023

Concluding Remarks

- We discovered a **quadratic exponential** relationship,
  dependent on sample size,
  between the Bayes factor and the separation of credible intervals
  in the linear mixed-effects model for a within-subject design.

Contribution

- Practitioners can adopt our approach,
  the modified within-subject Bayesian interval using JZS-HDI
  (which has less Monte Carlo variability compared to the existing methods)
  and examine the separation percentage
  to assess the strength of evidence for effects.

Limitation

- Our current method assumes the equal variance.
  We further examine the subject-centering transformation, *i.e.*,
  subtracting from the original response a corresponding subject
  mean minus the overall mean, to tackle the heteroscedastic data.

Future Work

When it comes to multiway ANOVA

- Model specifications

  Two-factor factorial designs with blocking

  $$Y_{ijk} = \mu + s_k + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{13}$$

  Two-way repeated-measures designs

  $$Y_{ijk} = \mu + s_k + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (\alpha s)_{ik} + (\beta s)_{jk} + \epsilon_{ijk} \tag{14}$$

- Assumptions

  Compound symmetry $\implies$ Sphericity (Cirgularity)

- Bayes factor strategies

  Top-down analysis violates the principle of marginality

References

Faulkenberry, T. J. (2021). The Pearson Bayes factor: An analytic formula for computing evidential value from minimal summary statistics. *Biom. Lett., 58,* 1-26. https://doi.org/10.2478/bile-2021-0001

Heck, D. W. (2019). Accounting for estimation uncertainty and shrinkage in Bayesian within-subject intervals: A comment on Nathoo, Kilshaw, and Masson (2018). *J. Math. Psychol., 88,* 27–31. https://doi.org/10.31234/osf.io/whp8t

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychon. Bull. Rev., 1*, 476–490. https://doi.org/10.3758/BF03210951

Morey, R. D., & Rouder, J. N. (2023). *BayesFactor: Computation of Bayes factors for common designs*. R package version 0.9.12-4.6. https://cran.r-project.org/package=BayesFactor

Nathoo, F. S., Kilshaw, R. E., & Masson, M. E. J. (2018). A better (Bayesian) interval estimate for within-subject designs. *J. Math. Psychol., 86*, 1–9. https://doi.org/10.1016/j.jmp.2018.07.005

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *J. Math. Psychol., 56*, 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychol. Methods, 22*, 304-321. https://doi.org/10.1037/met0000057

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev., 14*, 779–804. https://doi.org/10.3758/BF03194105

Wei, Z., Nathoo, F. S., & Masson, M. E. J. (2022). *rmBayes: Performing Bayesian inference for repeated-measures designs*. R package version 0.1.15. https://cran.r-project.org/package=rmBayes

Thank you.

✉ zhengxiao@uvic.ca (Z. Wei) https://orcid.org/0000-0003-1866-2320

nathoo@uvic.ca (F.S. Nathoo) https://orcid.org/0000-0002-2569-3507

mmasson@uvic.ca (M.E.J. Masson) https://orcid.org/0000-0002-5430-6078

Acknowledgements

This work was supported by discovery grants to Farouk Nathoo and Michael Masson from the Natural Sciences and Engineering Research Council of Canada.

Farouk Nathoo holds a Tier II Canada Research Chair in Biostatistics for Spatial and High-Dimensional Data.

Backup

Two Approaches in Parameter Estimation

BACK

Frequentist

- Probability is "long-run frequency of repeated events"
- $\mathbb{P}(X \mid \theta)$
  is a sampling distribution
  - function of X with θ fixed
  - Assuming repetitive sampling from the population to find the single true value for a parameter governing the population
- Null-hypothesis significance testing (NHST) and p-value
- Confidence interval

Bayesian

- Probability is "degree of certainty about values"
- $\mathbb{P}(X \mid \theta)$
  is a likelihood
  - function of θ with X fixed
  - Assuming the probability distribution of a parameter value and its reliability is increased by increasing the sample size
- Prior
- Posterior $\propto$
  Prior $\times$ Likelihood
- Credible interval

Confidence Interval and Credible Interval

BACK

Confidence Interval (CI)

parameters are fixed but unknown and data are random;

the lower and upper bounds are random

A 95% confidence interval means that with a large number of repeated samples, 95% of such calculated confidence intervals would include the true value of the parameter.

Credible Interval (CrI)

parameters are random and data are fixed;

the lower and upper bounds are fixed

A 95% credible interval is an interval within which an unobserved parameter value falls with a 95% probability.

Choice of Words

BACK

Within-subject (6,180,000 results)

Bub et al. (2021); Cousineau (2019); Nathoo et al. (2018); Rouder et al. (2012); Loftus and Masson (1994).

Within-subjects (3,410,000 results)

Rouder et al. (2017); Franz and Loftus (2012); Wagenmakers (2010).

Both

Heck (2019); Masson and Loftus (2003).

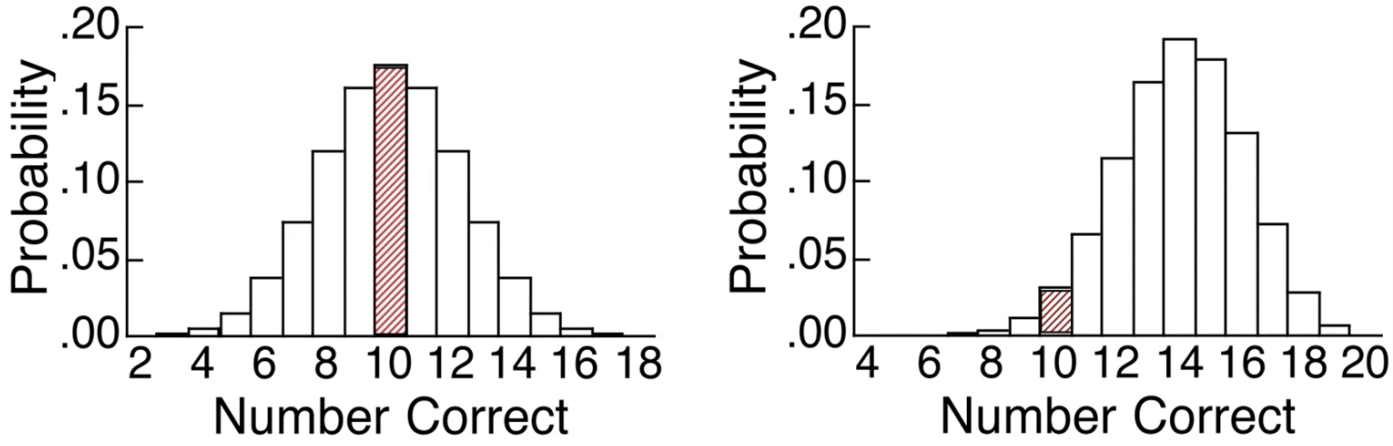
BACK

```
> summary(aov(Response~Level, recall.long)) #between-subjects
            Df Sum Sq Mean Sq F value Pr(>F)
Level        2   52.3   26.13    0.74  0.487
Residuals   27  953.6   35.32
> summary(aov(Response~Level+Error(Subject/Level), recall.long)) #within-subject

Error: Subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9  942.5   104.7

Error: Subject:Level
          Df Sum Sq Mean Sq F value   Pr(>F)
Level      2  52.27  26.133   42.51 1.52e-07 ***
Residuals 18  11.07   0.615
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> set.seed(277); anovaBF(Response~Level, recall.long, progress=F) #between-subjects
Bayes factor analysis
--------------
[1] Level : 0.3410478 ±0.01%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS

> set.seed(277); anovaBF(Response~Level+Subject, recall.long, whichRandom="Subject",
 progress=F) #within-subject
Bayes factor analysis
--------------
[1] Level + Subject : 35959.93 ±0.56%

Against denominator:
  Response ~ Subject
---
Bayes factor type: BFlinearModel, JZS
```

Example 1 of Computing the Bayes Factor

BACK

$\mathcal{H}_0$: the probability of correctly guessing the outcome $\theta = .5$ versus $\mathcal{H}_1 : \theta = .7$. Data: 10 outcomes out of 20 coin flips are predicted correctly.



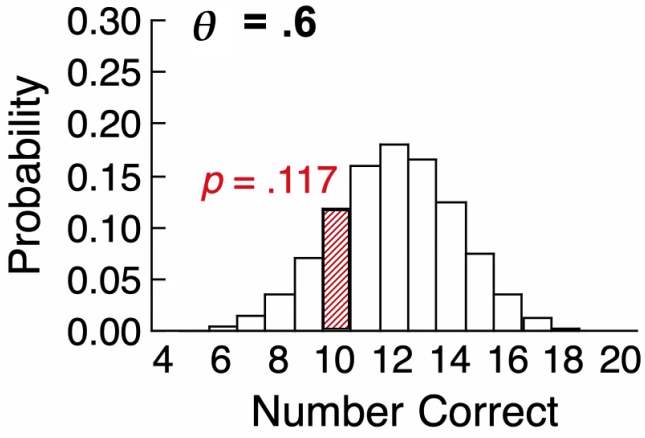
$$LR = BF_{01} = \frac{p(\text{Data} \mid \mathcal{H}_0)}{p(\text{Data} \mid \mathcal{H}_1)} = \frac{\binom{20}{10} \cdot 0.5^{20}}{\binom{20}{10} \cdot 0.7^{10} \cdot 0.3^{10}} = 5.718$$

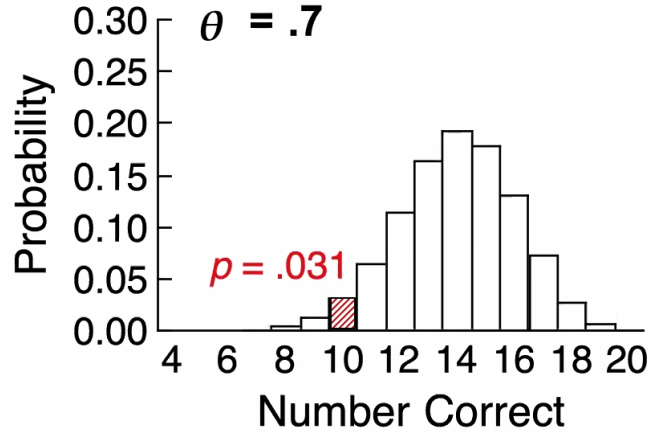
• **Evidence** (supporting $\mathcal{H}_0$) is anecdotal (BF 1-3), moderate (BF 3-10), strong (BF 10-30), or very strong (BF 30-100).

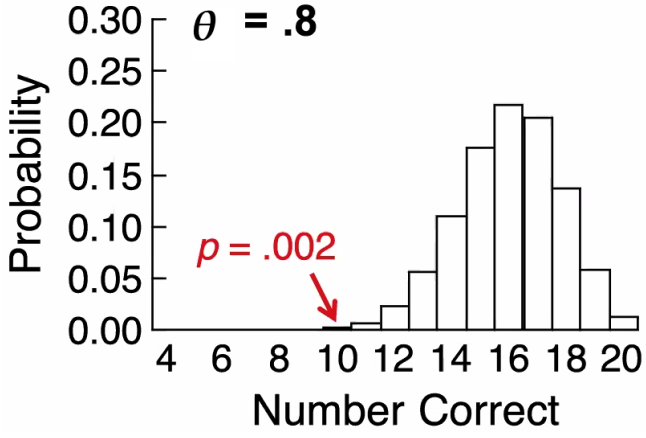
Example 2 of Computing the Bayes Factor

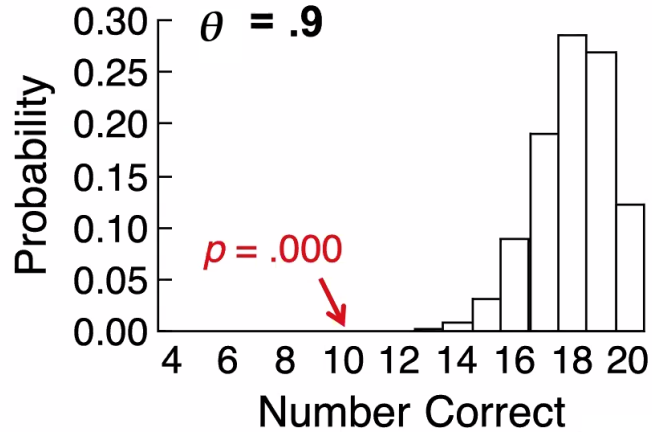
BACK

Now, $\mathcal{H}_1 : \theta \in \{.6,\ .7,\ .8,\ .9\}$ and they are equally plausible.









$$p(\text{Data} \mid \mathcal{H}_1) = (.117).25 + (.031).25 + (.002).25 + (.000).25 = .037$$

$$BF_{01} = \frac{p(\text{Data} \mid \mathcal{H}_0)}{p(\text{Data} \mid \mathcal{H}_1)} = 4.699$$

Complexity in Computing $p(\text{Data} \mid \mathcal{H}_1)$

Discrete Data

$$BF_{01} = \frac{\sum_i p(\text{Data} \mid \mathcal{H}_0, \theta_i) \cdot p(\theta_i \mid \mathcal{H}_0)}{\sum_j p(\text{Data} \mid \mathcal{H}_1, \theta_j) \cdot p(\theta_j \mid \mathcal{H}_1)} \tag{A1}$$

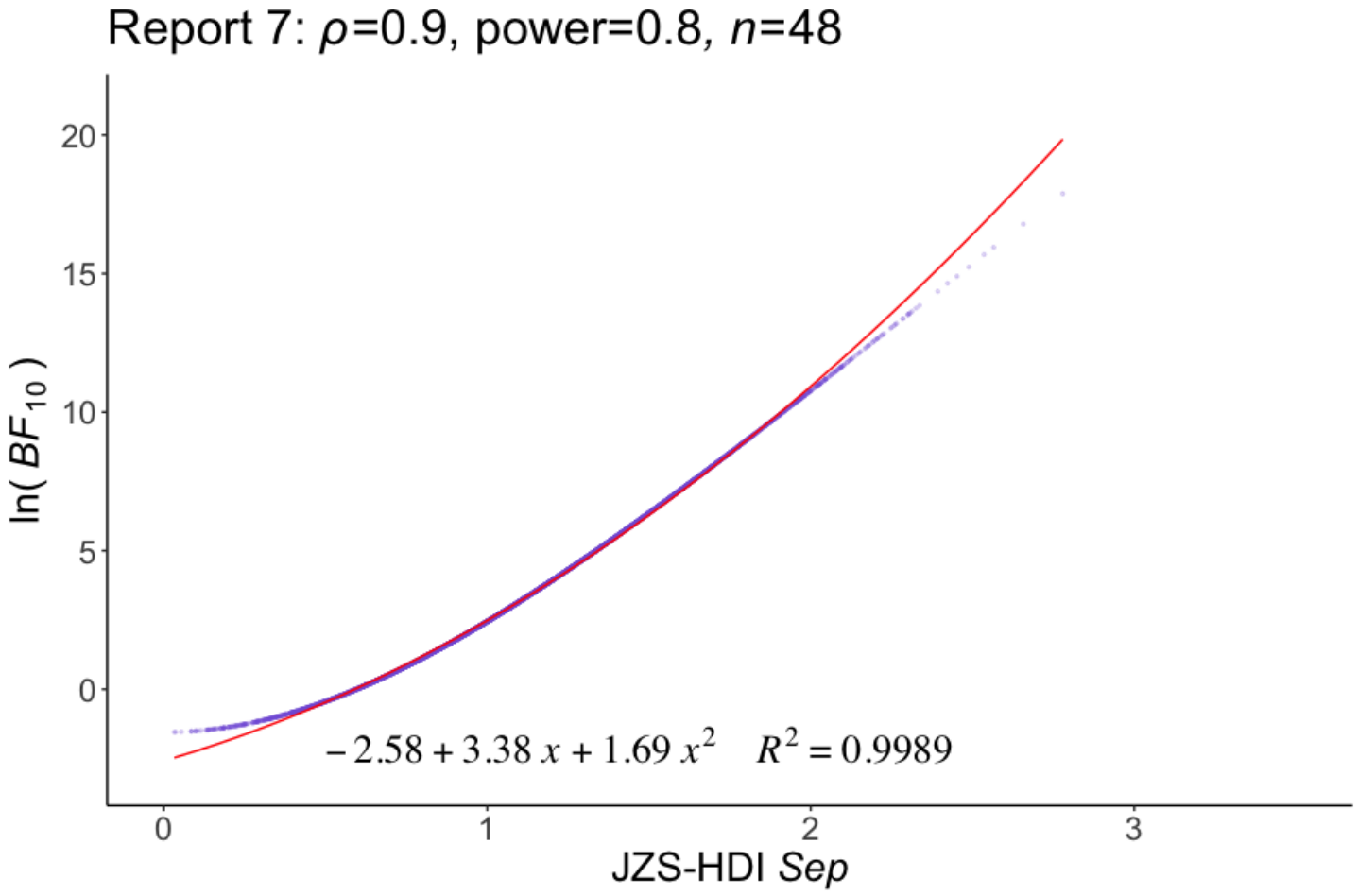
Continuous Data

$$BF_{01} = \frac{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta_0}} p(\text{Data} \mid \mathcal{H}_0, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} \mid \mathcal{H}_0) \mathrm{d}\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta_1}} p(\text{Data} \mid \mathcal{H}_1, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} \mid \mathcal{H}_1) \mathrm{d}\boldsymbol{\theta}} \tag{A2}$$

Log Bayes Factor Versus Separation

ZOOM OUT

$$\ln BF \sim a + b \times Sep + c \times Sep^2$$



Log Bayes Factor Versus Separation

ZOOM OUT

$$\ln BF \sim a + b \times Sep + c \times Sep^2$$