The Advantages of Bayesian Statistics

in the Study of Second Language Acquisition

Guilherme D. Garcia

Ball State University

guilhermegarcia.github.io

AAAL 2018
Chicago

Overview

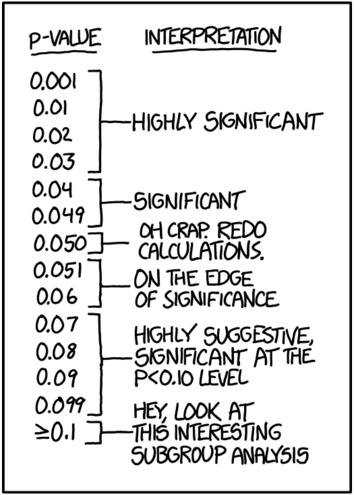
Evolution

Tests $\rightarrow$ Models $\rightarrow$ Hierarchical models

Our tools to analyze data are much better now, **but...**

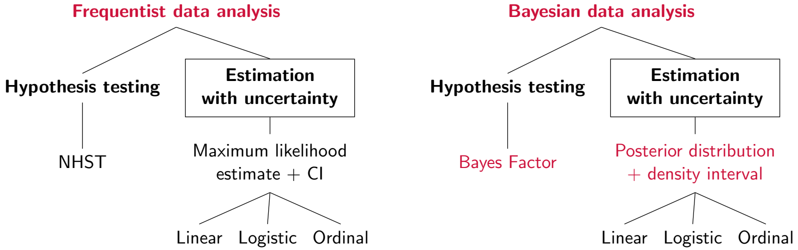
1. Collect and explore data
2. Run test/model
3. **Check p-value**
   - $p < 0.05 \rightarrow$ stop and publish
   - $p > 0.05 \rightarrow$ back to step 1

☞ **we still focus too much on step 3**


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Big picture

The typical tools we use



☞ **Why should we change from Frequentist to Bayesian?**

Some issues with Frequentist statistics

Old stats

- Results either significant or not significant
  - As stipulated by an arbitrary threshold (commonly $\alpha = 0.05$)
- Focus on p-values instead of what really matters: **effect sizes**
  - p-values are highly sensitive to sample sizes $\rightarrow$ p hacking

☞ **The "New Statistics" clearly helped**
  - **From:** Null Hypothesis Significance Testing (NHST)
  - **To:** Estimation based on effect sizes, CIs (Cumming 2014)

Some issues with Frequentist statistics

New stats

- Overall, Frequentist methods have important issues

  Let's check **three** of them:

  - Counter-intuitive interpretation
  - Lack of flexibility
  - Naïve assumptions

Non-intuitive interpretation

Frequentist approach:

A p-values: we get $p(D|\theta)$ under H_0

B Confidence intervals: counter-intuitive interpretation

C Effect size is a point estimate (single value)

Bayesian approach:

A No p-values: we get $p(\theta|D)$

B Credible intervals (e.g., HDI)[1] $\rightarrow$ easy interpretation

C Effect size is a (posterior) distribution of credible values

[1] Highest Density Interval

Lack of flexibility

Frequentist approach:

- We can't really change what a test/model assumes

E.g.: Outliers often removed from dataset to enforce normality

E.g.: Homogeneity of variance: unrealistic and unchangeable

Bayesian approach:

- Model adapted to our needs

E.g.: Keep outliers; choose non-normal distribution[2]

E.g.: Variance is also estimated

[2] Cf. frequentist robust regressions.

Naïve assumptions

Frequentist approach:

- Can't incorporate what is known about a phenomenon
- Every study (model) "starts from zero"

Bayesian approach:

- Can be informed by priors
- Studies can feed from previous findings

☞ **Intuition**

"Extraordinary claims require extraordinary evidence"[3]

[3] Laplace, but also Hume and Sagan

Going Bayesian

Frequentist approach:

- Probability of data given parameter (under H_0) $\rightarrow p(D|\theta)$

Bayesian approach:

- Probability of **parameter** given data $\rightarrow p(\theta|D)$

☞ + meaningful: we're interested in the parameter, not the data

- $p(\theta)$ calculated using Bayes' Theorem:*

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

Example

- Assume two groups of learners
  - **A** mean score $= 0.8$, $s = 0.5$, $n = 100$
  - **B** mean score $= 0.3$, $s = 0{,}5$ $n = 100$
- Parameter of interest = difference of means $= \mu_B - \mu_A$

☞ **Estimate** = $\boxed{\text{-0.43, 95\% HDI = [-0.56, -0.30]}}$ (no p-value)

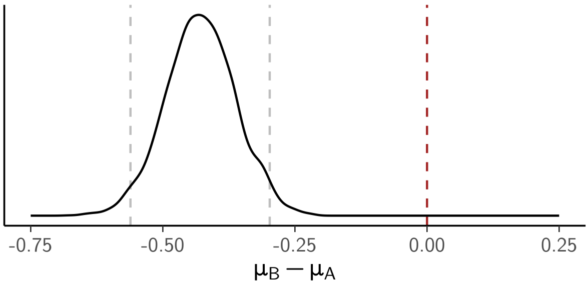
- The most probable parameter value is -0.43

☞ But we're given an entire **distribution** of credible values

- We can also easily visualize this distribution with a plot

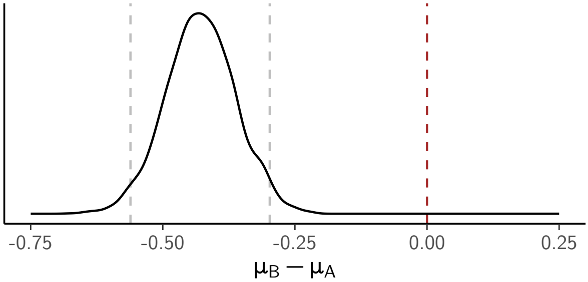
Informative output

Posterior distribution + 95% HDI [-0.56, -0.30]



Interpretation

- Values closer to the peak are more credible given the data



We can use the 95% HDI as a decision tool: (Kruschke 2015)

☞ 95% HDI doesn't include zero $\rightarrow \neq$ is statistically credible

- *Note that 95% is an arbitrary number*

Flexibility

- Prior expectations incorporated in the model
  - Realistic (we rarely start from absolute zero knowledge)
  - Effective (helps the model focus on plausible parameter values)
- Normality is **not** necessary
  - A set of distributions to choose from
- Variance is also estimated (more later)
  - When do experimental groups have equal variance?

L1-L2 transfer

- L1 as initial state (Schwartz and Sprouse 1996, White 2000)

☞ Expect certain L2 deviations based on L1 grammar

E.g.: Spanish speakers learning English: **penult stress bias**

E.g.: Italian speakers learning French: **pro-drop bias**

Example I: L1-L2 transfer

☞ **We can add these biases to the model!**

- We can even compare our model to a naïve model
  And check which one best fits the data

E.g.: Spanish $\rightarrow$ English: $p(\textit{penult}) > 0.5$

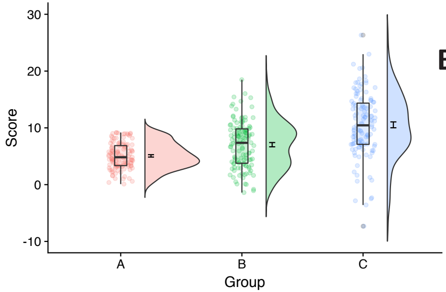
E.g.: Italian $\rightarrow$ French: $p(\textit{drop}) > 0.5$

☞ This also applies to universal biases: *we rarely start from zero*

Variance matters

- We know that different groups often have different variance
- ☞ A Bayesian model also estimates $p(\sigma)$

  In the form of a complete posterior distribution



E.g.: **Three groups** of students

120 obs (some test score)

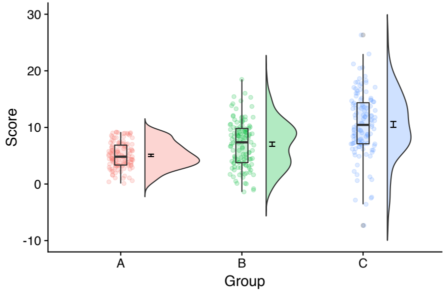
Different $\bar{x}$: 5, 7, 9

☞ Different s: 2, 4, 6

Variance matters

- We know that different groups often have different variance
- ☞ A Bayesian model also estimates $p(\sigma)$

  In the form of a complete posterior distribution



Frequentist model

- A $\neq$ B: $p < 0.05$;
- CI $=$ [0.58, 3.36]

Bayesian model

- $\neq$ less credible
- HDI $=$ [-0.07, 3.95]

▸ More

Final remarks

5 advantages of a Bayesian approach

1. Priors incorporate theoretical assumptions (L1-L2 transfer)
2. Meaningful and intuitive interpretation
   - $p(\theta|D)$ instead of $p(D|\theta)$ (under H_0)
   - Directly compatible with various theories of learning
3. Comprehensive output: posterior distribution
4. More flexibility with assumptions (outliers, U-shaped learning)
5. No p-values (avoids simplistic interpretations; NHST errors)

Disadvantages?

1. Computationally demanding: here, 0.02s *vs*. 42s
2. Not widespread in our field(s) yet (journals, pee-review)
3. More flexibility and power require more technical knowledge
    - But: getting more and more accessible

Where to start?

- **R, Python, Stata, Matlab**

Kruschke's↑ *Doing Bayesian Data Analysis* (+ intro papers)
McElreath's↑ *Statistical Rethinking* (+ lecture series)
Gelman et al.'s↑ *Bayesian Data Analysis* (+ blog etc.)

Bayes + Applied Linguistics: Plonsky's bibliography↑

Thank you!

References I

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, 25(1):7–29.

Kruschke, J. (2015). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan*. London: Academic Press, 3rd edition.

Schwartz, B. D. and Sprouse, R. (1996). L2 cognitive states and the full transfer/full access model. *Second Language Research*, 12(1):40–72.

White, L. (2000). Second language acquisition: from initial to final state. In Archibald, J., editor, *Second language acquisition and linguistic theory*, pages 130–155. Oxford: Wiley-Blackwell.

Appendix i

Tools

R	`rstan, rstanarm, brms, rjags`
Python	`PyStan`
Stata	
Matlab	`MatlabStan`

Appendix ii

Going Bayesian

- Calculating $p(\theta)$ not always computationally possible

☞ **Solution:** sample from posterior using a **sampler**

- Currently, `Stan`↑ (but see also JAGS and BUGS)

  Stan is a language for statistical modeling
- Fortunately, we don't actually need to learn it*

Appendix iii

Code

- ☞ **Models run:** `Score ~ Group + (1 | Subject)`
  - Data simulation:

```
set.seed(2)

df = data.frame(Group = as.factor(rep(c("A", "B","C"),
                                  each = 120)),
                Subject = rep(paste("subject",
                                    seq(1, 9),
                                    sep = "_"),
                              each = 40),
                Score = c(rnorm(120, 5, 2),
                          rnorm(120, 7, 4),
                          rnorm(120, 9, 6)))
```

Appendix iv

Variance: Why the Bayesian model is superior

- More closely approximates empirical sampling distributions:

☞ coefficients + residual standard error

- We still see the trend generated

☞ But our certainty shifts (i.e., more conservative)

- In part because our Bayesian model is not conditional on H_0: it's averaging across **all** possible values of σ^2