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

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
Why change?
Going Bayesian
Examples & implementation

Interpretation
Lack of flexibility
Naïve assumptions

---

### P-VALUE vs. INTERPRETATION

<table>
<thead>
<tr>
<th>P-VALUE</th>
<th>INTERPRETATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 0.001</td>
<td><strong>HIGHLY SIGNIFICANT</strong></td>
</tr>
<tr>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td><strong>SIGNIFICANT</strong></td>
</tr>
<tr>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>0.050</td>
<td><strong>OH CRAP. REDO CALCULATIONS.</strong></td>
</tr>
<tr>
<td>0.051</td>
<td><strong>ON THE EDGE OF SIGNIFICANCE</strong></td>
</tr>
<tr>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>0.07</td>
<td><strong>HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P&lt;0.10 LEVEL</strong></td>
</tr>
<tr>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>0.099</td>
<td><strong>HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS</strong></td>
</tr>
<tr>
<td>≥ 0.1</td>
<td></td>
</tr>
</tbody>
</table>
Why change?
Going Bayesian
Examples & implementation

Lack of flexibility
Naïve assumptions

Big picture
The typical tools we use

Frequentist data analysis
- Hypothesis testing
  - NHST
  - Estimation with uncertainty
    - Maximum likelihood estimate + CI
      - Linear
      - Logistic
      - Ordinal

Bayesian data analysis
- Hypothesis testing
  - Bayes Factor
- Estimation with uncertainty
  - Posterior distribution + density interval
  - Linear
  - Logistic
  - Ordinal

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)\(^1\) → 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
  
  **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 $H_0$) → $p(D|\theta)$

**Bayesian approach:**
- Probability of *parameter* given data → $p(\theta|D)$
  + 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** = -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:
- 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
  - 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*

(Schwartz and Sprouse 1996, White 2000)
Example I: L1-L2 transfer

We can add these biases to the model!
  o We can even compare our model to a naïve model
    And check which one best fits the data

E.g.: Spanish $\rightarrow$ English: $p(\text{penult}) > 0.5$
E.g.: Italian $\rightarrow$ French: $p(\text{drop}) > 0.5$

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]
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, peer-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\(^\uparrow\) *Doing Bayesian Data Analysis* (+ intro papers)
McElreath’s\(^\uparrow\) *Statistical Rethinking* (+ lecture series)
Gelman et al.’s\(^\uparrow\) *Bayesian Data Analysis* (+ blog etc.)

Bayes + Applied Linguistics: Plonsky’s bibliography\(^\uparrow\)
Thank you!
References I


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\(^\dagger\) (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 \sim Group + (1 \mid Subject)

- Data simulation:

```r
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)))
```

Garcia
The Advantages of Bayesian Statistics
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 $\sigma^2$