Bayesian parameter estimation

2020 June 24
Why are you here?

(In any case, thanks for coming :)
Frequentist tests are confusing

“When you use a frequentist test, how confident are you usually that you chose the correct test for your statistical question?”

38 responses
Choosing a test

- Repeated measures
- One independent categorical variable with several levels
- One dependent binary outcome variable

36 responses

- ANOVA (with sphericity correction): 38.9%
- Logistic regression (MLE): 25%
- Bayesian ANOVA (Bayes' factor): 27.8%
- Bayesian logistic regression
- Hierarchical logistic regression
- Ugh
- Don't know
Frequentist tests are confusing

“When you use a frequentist test, how well do you usually feel you understand the test itself?”

37 responses
Hypothesis testing is your world

“For you, the *most important* component of a statistical analysis's results is…”

36 responses

- The significance of a statistical test for a hypothesis (38.9%)
- The magnitude of an estimated effect size (8.3%)
- The direction of an estimated effect size (8.3%)
- The confidence interval (8.3%)
- The quality of the fit of a statistical model (38.9%)
- I think stat significance is as important…
- Depends on the question/analysis
Statistics as posterior inference
Posterior inference?

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{p(D)}$$
Posterior inference?

“How surprising were our observations, given some parameters?”

“How credulous were we of some parameters?”

“How surprising were our observations under any parameters?”

“How credulous should we be of our parameters, given our observations?”

\[
P(\theta|D) = \frac{P(D|\theta)P(\theta)}{p(D)}
\]
Presidential heights, revisited

Recall from part 1:

Is there any evidence for a

What is the correlation between the heights and popular vote share of U.S. presidents?

\[ P(\rho | D) = P(D | \rho) \frac{P(\rho)}{P(D)} \]
Presidential heights, revisited

\[ \mu_{\text{height}} \sim N(\mu = 1, \sigma = 0.1) \]
\[ \mu_{\text{vote}} \sim \text{Beta}(\mu = 0.5, \sigma = 0.25) \]
\[ \sigma_{\text{height}} \sim \text{HalfCauchy}(\sigma = 2.5) \]
\[ \sigma_{\text{vote}} \sim \text{HalfCauchy}(\sigma = 2.5) \]
\[ \rho \sim U(l, b) \]

\[ \sigma = \begin{bmatrix} \sigma_{\text{height}} & \sigma_{\text{vote}} \end{bmatrix} \]
\[ \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \odot (\sigma \times \sigma^T) \]
\[ \mu = \begin{bmatrix} \mu_{\text{height}} & \mu_{\text{vote}} \end{bmatrix} \]

\[ \text{height, vote} \sim N(\mu, \Sigma) \]
Presidential heights, revisited

\[ \begin{align*}
\mu_{\text{height}} & \sim N(\mu = 1, \sigma = .1) \\
\mu_{\text{vote}} & \sim \text{Beta}(\mu = .5, \sigma = .25) \\
\sigma_{\text{height}} & \sim \text{HalfCauchy}(\sigma = 2.5) \\
\sigma_{\text{vote}} & \sim \text{HalfCauchy}(\sigma = 2.5) \\
\rho & \sim U(-1, b)
\end{align*} \]

\[ \sigma = \begin{bmatrix} \sigma_{\text{height}} & \sigma_{\text{vote}} \end{bmatrix} \]

\[ \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \otimes (\sigma \times \sigma^T) \]

\[ \mu = \begin{bmatrix} \mu_{\text{height}} \\ \mu_{\text{vote}} \end{bmatrix} \]

\[ \text{height, vote} \sim N(\mu, \Sigma) \]

Bivariate normal likelihood
Presidential heights, revisited

$$\begin{align*}
\mu_{\text{height}} &\sim N(\mu = 1, \sigma = .1) \\
\mu_{\text{vote}} &\sim \text{Beta}(\mu = .5, \sigma = .25) \\
\sigma_{\text{height}} &\sim \text{HalfCauchy}(\sigma = 2.5) \\
\sigma_{\text{vote}} &\sim \text{HalfCauchy}(\sigma = 2.5) \\
\rho &\sim U(l, b)
\end{align*}$$

$$\begin{align*}
\sigma &= \begin{bmatrix} \sigma_{\text{height}} & \sigma_{\text{vote}} \end{bmatrix} \\
\Sigma &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \odot (\sigma \times \sigma^T) \\
\mu &= \begin{bmatrix} \mu_{\text{height}} \\ \mu_{\text{vote}} \end{bmatrix} \\
\text{height, vote} &\sim N(\mu, \Sigma)
\end{align*}$$

Prior on correlation coefficient
Presidential heights, revisited

Priors on mean height and vote margin

\[
\begin{align*}
\mu_{\text{height}} &\sim N(\mu = 1, \sigma = .1) \\
\mu_{\text{vote}} &\sim \text{Beta}(\mu = .5, \sigma = .25) \\
\sigma_{\text{height}} &\sim \text{Half Cauchy}(\sigma = 2.5) \\
\sigma_{\text{vote}} &\sim \text{Half Cauchy}(\sigma = 2.5) \\
\rho &\sim U(l, b)
\end{align*}
\]

\[
\sigma = \begin{bmatrix}
\sigma_{\text{height}} & \sigma_{\text{vote}}
\end{bmatrix}
\]

\[
\Sigma = \begin{bmatrix}
1 & \rho \\
\rho & 1
\end{bmatrix} \odot (\sigma \times \sigma^T)
\]

\[
\mu = \begin{bmatrix}
\mu_{\text{height}} \\
\mu_{\text{vote}}
\end{bmatrix}
\]

\[
\text{height, vote} \sim N(\mu, \Sigma)
\]
Computing posteriors

Monte Carlo methods sample from the posterior.

Modern samplers don’t require much fussing.
Presidential Heights in PyMC3

Let’s walk through some code.
Great, but what are we actually doing?

Not hypothesis testing.
“What hypothesis?”

The posterior distribution is a full report for the parameter of interest.
Do scientists always need to test hypotheses?

Hypothesis testing is a decision procedure.

It’s used to produce binary decisions from uncertain beliefs.

Do scientists need to produce decisions?
Subjective Bayes and science

Probability theory extends classical logic to ideal reasoning under uncertainty.¹

\[
P(A|\neg B) = P(\neg B|A) \frac{P(A)}{P(\neg B)}
\]

¹ See Jaynes 2003 for more rigor
Weakly informative priors as null “hypotheses”

(Sort of.)

A weakly informative prior centered around a “null” value: the evidence must be sufficiently stronger than the prior.

Example: Ridge regression → Gaussian prior
Example: Ridge regression

MAP estimate with Gaussian prior

\[
\beta \sim N(0, \sigma_\beta) \\
y_i \sim N(\beta x_i, \sigma)
\]

\[
p(\beta|\tilde{y}, \tilde{x}) \propto p(\beta) \prod_{i} p(y_i | \beta, x_i)
\]

\[
\arg\max_{\beta} \sum_{i} \log p(y_i | \beta, x_i) + \log p(\beta)
\]

\[
\arg\min_{\beta} \frac{1}{\sigma} \sum_{i} (y_i - \beta x_i)^2 + \frac{1}{\sigma_\beta} \beta^2
\]
Presidential heights: informative priors

Narrow prior model
- median = 0.0017
- 42.8% < θ < 57.2%
- 95% HPD: [-0.1, 0.1]
- [-0.003, 0.002]

LKJ prior model
- median = 0.366
- 0.4% < θ < 99.6%
- 95% HPD: [0.115, 0.611]
- [-0.1, 0.1]

Full uniform prior model
- median = 0.383
- 0.5% < θ < 99.5%
- 95% HPD: [0.126, 0.618]
- [-0.1, 0.1]
Still want to reject a null?

We can do it with posterior estimates.
Rejecting a null “hypothesis”

Does the **ROPE** around a null value lie entirely outside the **HPDI**?
Rejecting a null “hypothesis”

Does the **ROPE** around a null value lie entirely outside the **HPDI**?

**Region Of Practical Equivalence**
Rejecting a null “hypothesis”

Does the ROPE around a null value lie entirely outside the HPDI?
Rejecting a null “hypothesis”

Does the **ROPE** around a null value lie entirely outside the **HPDI**?

No overlap? **Reject the null “hypothesis”**!
What does non-rejection look like?

Does the **ROPE** around a null value lie entirely outside the **HPDI**?

**Strongly misinformed prior**

**Prior: \( U(-1, 0) \)**

**median\(= -0.0363 \)**

**100.0\% <\( 0 <\ 0.0\%\)**

**95\% HPD**

**-0.1**

**0.1**

**-0.14139e-05**

**Partial overlap?**

**Cannot reject null.**
Savage-Dickey method

\[ BF_{01} = \frac{P(D|\theta \neq \theta_{null})}{P(D|\theta = \theta_{null})} = \frac{P(\theta \neq \theta_{null}|D)}{P(\theta \neq \theta_{null})} \]

Savage-Dickey method illustration

BF against ROPE: 5.450
Differences between BF and PE

<table>
<thead>
<tr>
<th>Bayes’ factor</th>
<th>Parameter estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Hypothesis testing as <strong>model comparison</strong></td>
<td>- “What hypothesis?”</td>
</tr>
<tr>
<td>- $H_0$ either a parameter value or a model</td>
<td>- Report full credibilities of magnitudes</td>
</tr>
<tr>
<td>- Imitate frequentist tests and assumptions</td>
<td>- Model is explicit</td>
</tr>
</tbody>
</table>
Bayes’ factors: YMMV

HDPI: (.27, .73)
BF: 3.14 for $\theta = .5$

Not a ton of data

Kruschke 2015
ANOVA

By popular demand :)
ANOVA tops the list

What frequentist tests would you consider replacing?

- ANOVA: 67%
- t-Tests: 55%
- Linear regressions: 45%
- Chi-Squared tests: 33%
- Other GLM regressions: 30%
- Logistic regressions: 30%
- Correlations: 21%
- Other between-groups differences tests: 21%

# Responses (34 total)
What is ANOVA for, anyway?

Technically, decomposing sources of variance

Loosely, often used to answer: “Did X influence Y?”

ANOVA is hypothesis testing and **only makes sense in that framework**
Standard ANOVA relies on a linear model

- “Explained” variance ratio
  - → $F$-statistic
- Group means == coefficients
  - → homogeneity of variance
  - → normality of outcomes
- “How likely is my $F$ given $\beta = 0$?”
  - → $p$-value

\[
F = \frac{\sum_{i=1}^{n}((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2/1}{\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2/(n - 2)}.
\]

CV Answer 1; CV Answer 2
Standard ANOVA relies on a linear model

- “Explained” variance ratio
  - $\rightarrow F$-statistic

- Group means $\equiv$ coefficients
  - $\rightarrow$ homogeneity of variance
  - $\rightarrow$ normality of outcomes

- “How likely is my $F$ given $\beta = 0$?”
  - $\rightarrow p$-value

$$F = \frac{\sum_{i=1}^{n}((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2/1}{\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2/(n-2)}.$$
Categorical variable coding example

<table>
<thead>
<tr>
<th>Level</th>
<th>Dummy Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sum Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>Sum Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>Sum Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>Sum Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-1</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>
Standard ANOVA relies on

- “Explained” variance ratio
  - → $F$-statistic

- Group means == coefficients
  - → homogeneity of variance
  - → normality of outcomes

- “How likely is my $F$ given $\beta = 0$?”
  - → $p$-value

What you **actually** care about: Model coefficients.

$$
\sum_{i=1}^{n} ((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2 + \sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta}x_i))^2.
$$

- explained
- unexplained

$$
F = \frac{\sum_{i=1}^{n} ((\hat{\alpha} + \hat{\beta}x_i) - \bar{y})^2/1}{\sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta}x_i))^2/(n-2)}.
$$

CV Answer 1; CV Answer 2
Who gives a Fischer?

- Doesn’t tell us about direction
- Ditto relationship *strength*
- Assumptions often don’t obtain
- Complicated designs are “fun”
Who gives a Fischer?

- Doesn’t tell us about direction
- Ditto relationship *strength*
- Assumptions often don’t obtain
- Complicated designs are “fun”

You need the model!
Everything you usually want to know is in the parameters of an appropriate regression model

So let’s do that.
Regression with categorical predictors

Figures lifted shamelessly from Kruschke 2015
Regression: ANOVA assumptions

Homogeneity of variance

Normality of outcomes
Regression: Pooled priors and shrinkage

Hyperprior on coefficient scales

Pooled prior on coefficients
Extension: Multiple categorical variables

\[
\begin{align*}
M_0 & \sim S_0 \\
\beta_0 & + \sum_j \beta_{1[j]} x_{1[j]}(i) + \sum_k \beta_{2[k]} x_{2[k]}(i) + \sum_{j,k} \beta_{1\times2[j,k]} x_{1\times2[j,k]}(i) \\
\mu_i & \sim \text{normal} \\
\sigma_y & \\
\gamma_i & \\
\end{align*}
\]

Between-group interaction coefficients
Extension: “Mixed effects” / ANCOVA

\[ y_i = \mu_i + \sum_j \beta_{ij} x_{ij}(i) + \beta_{cov} x_{cov}(i) + \epsilon_i \]

- \( M_0, S_0 \) normal
- \( 0, \sigma_\beta \) normal
- \( M_c, S_c \) normal
- \( \epsilon_i \) uniform

Continuous covariate
Extension: “Robust” errors model

Prior over degree of tail-heaviness

No more homogeneity of variance!

Heavy-tailed likelihood function
Extension: Bayesian logistic regression

\[
P(y_i) = \frac{1}{1 + e^{\beta x_i}}
\]

Set \( M_j = 0 \) and (maybe) put a prior on \( S_j \) for ANOVA-like model

Bernoulli likelihood
Change detection task
Recall task
Extension: Poisson regression

\[ \exp(\beta_0 + \sum_j \beta_{1[j]} x_{1[j]}(i) + \sum_k \beta_{2[k]} x_{2[k]}(i) + \sum_{j,k} \beta_{1\times2[j,k]} x_{1\times2[j,k]}(i)) \]

Note the exponent!
ETC

etc
What about t-tests?

Bayesian Estimation Supersedes the T-test
BEST model

Observations $i$

Groups $j$

Difference between groups 0 and 1: $\mu_1 - \mu_0$

Effect size: $\frac{(\mu_1 - \mu_0)}{\sqrt{\frac{(\sigma_1^2 + \sigma_0^2)}{2}}}$
Final thoughts
Prior choices?

Much ink already spilled on this topic.

Rule of thumb: weakly informative prior for a skeptical audience.
Other uses!

- Cognitive process models
  - ex: Kruschke & Vanpaemel 2015)

- Dynamical models
  - ex: SEIR, Lotka-Volterra, whatever

- Time series models

- Latent variable models
  - ex: Bayesian matrix factorization, LDA

- Exotic models
  - ex: Gaussian processes, Bayesian neural networks
Further reading

Kruschke, John K. *Doing Bayesian Data Analysis*. 2e, 2015.


If you get just one book, I recommend this one!
Appendix
Inference methods: What about SVI?

**HMC: Hamiltonian Monte Carlo**
- Most packages implement variants of the **No U-Turn Sampler (NUTS)**
- Works well without hand-tuning for a variety of common models
- Slow when there is lots of data (>1-10k observations)

**SVI: Stochastic Variational Inference**
- Approximate posterior by minimizing ELBO objective
- Many “automatic” posterior guides available
- Scales well and performs better on large datasets (>1-10k observations)
## Python packages

<table>
<thead>
<tr>
<th>2020 June</th>
<th>PyMC3</th>
<th>PyMC4</th>
<th>Pyro</th>
<th>NumPyro</th>
<th>(py)STAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stability</td>
<td>Mature</td>
<td>Pre-release</td>
<td>Mature</td>
<td>Development</td>
<td>Mature</td>
</tr>
<tr>
<td>Features</td>
<td>👍👍👍</td>
<td>👍👍</td>
<td>👍👍👍</td>
<td>👍👍</td>
<td>👍👍👍</td>
</tr>
<tr>
<td>Future development</td>
<td>❌</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Backend</td>
<td>Theano</td>
<td>Tensorflow</td>
<td>PyTorch</td>
<td>JAX</td>
<td>C++</td>
</tr>
<tr>
<td>MCMC Speed</td>
<td>🐫骆驼</td>
<td>🐫骆驼❓</td>
<td>💨</td>
<td>🐫骆驼驼驼驼</td>
<td>🐫骆驼驼驼驼</td>
</tr>
<tr>
<td>SVI Speed</td>
<td>🐫骆驼</td>
<td>🐫骆驼❓</td>
<td>🐫骆驼驼驼驼</td>
<td>🐫骆驼❓</td>
<td>🐫骆驼❓</td>
</tr>
<tr>
<td>GPU support</td>
<td>Model only</td>
<td>Yes?</td>
<td>Model only</td>
<td>Yes</td>
<td>Nope</td>
</tr>
</tbody>
</table>
Slide Graveyard