Blaming Users, Bayes, Transparent Statistics

Matthew Kay

Lots of user-blaming rhetoric out there

You're doing science wrong! Just take another 3 years of stats!

And yet:

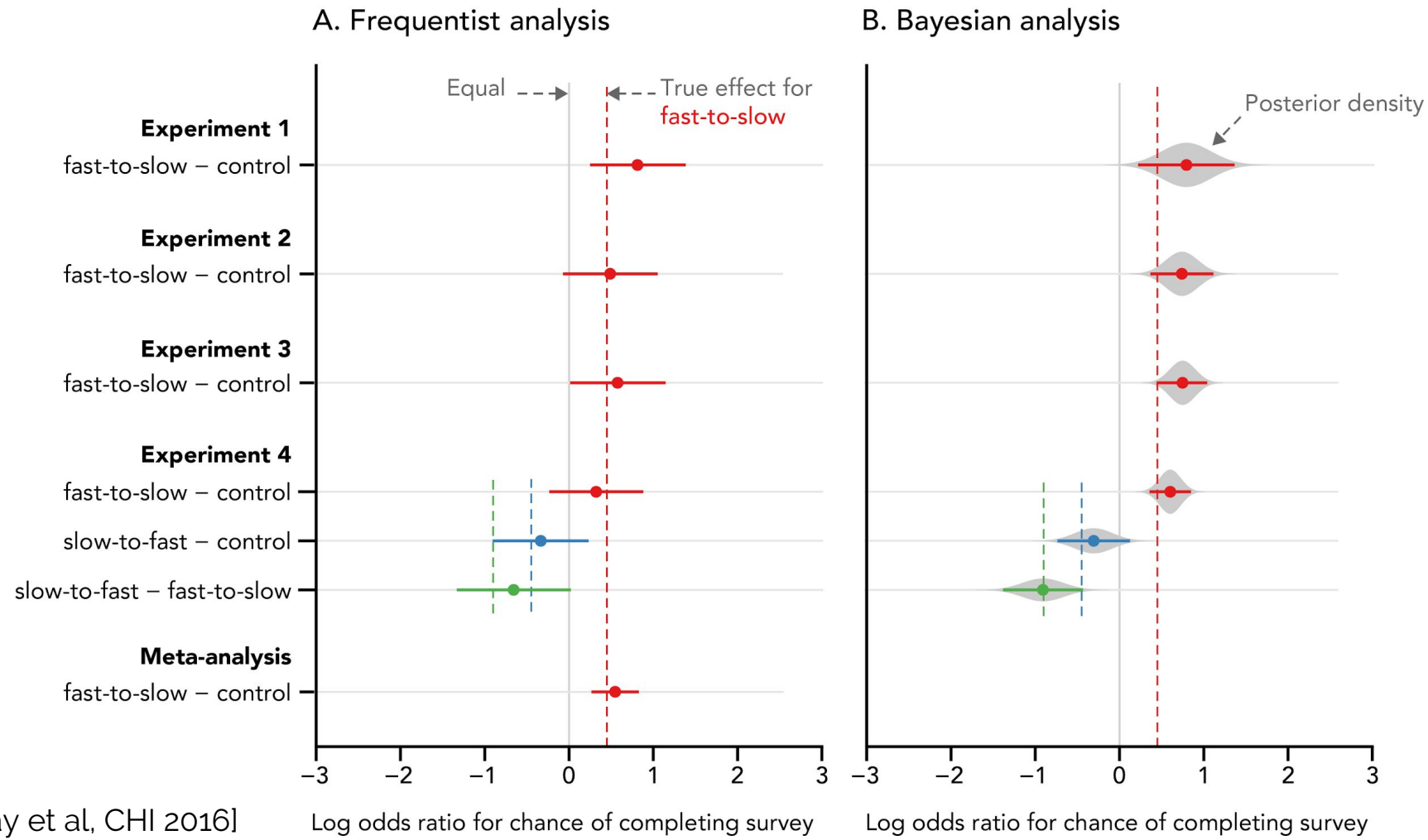
- Tools spit out results without guidance
- Tools encourage bad behavior
- Tools spit out difficult-to-understand representations of results
- P values, confidence intervals have counter-intuitive interpretations
- ...

Perhaps

1. We should use statistical methods that are not so counter-intuitive (Bayes?)

2. We should put on our Vis and HCI hats and build better tools!

Bayesian inference (demo)



[Kay et al, CHI 2016]

Bayes can help with:

Interpretability:

Pr(effect | model, evidence, previous knowledge)

Small samples: regularization / shrinkage

Building knowledge within single-novel-paper publishing incentives

Bayes can help with:

Reasoning about practical significance

Impact outside the field: posterior + cost function = effective decision-making

Easily accessible to business, practitioners

What does Vis have to say?

Statisticians have the solution: shame everyone!

To get out of blaming users, apply our Vis skills to this problem

What are the most effective representations?

Vis and HCI together can help help us build better tools for stats

But what can we do as authors?

As an author, how do I get my paper with new methods accepted?

- Practice defensive citation
- Practice VIS!

Practice defensive citation

In this section we describe a Bayesian variant of the censored log-linear model. In Bayesian modelling, we specify our *prior beliefs* about a model as probability distributions, and then *update* our beliefs based on observed evidence (the data collected in an experiment) [11]. These updated beliefs are called *posterior distributions*.

This approach yields a richer estimation of the parameters of interest – complete posterior probability distributions of all parameters – instead of point estimates and confidence intervals. Such posteriors offer an easy way for others to build on our work by using our posterior estimates to inform prior distributions in future work. As we will see, Bayesian estimation also provides a straightforward way to derive the expected performance of a visualization (with uncertainty) on any hypothetical dataset of correlations that can be expressed as a probability distribution over r. We largely adopt Kruschke's [12] approach to Bayesian experimental statistics by using 95% credibility intervals[5] of posterior distributions to estimate differences between parameters.

Practice VIS!

A mixed-design ANOVA with sex of face (male, female) as a within-subjects factor and self-rated attractiveness (low, average, high) and oral contraceptive use (true, false) as between-subjects factors revealed a main effect of sex of face, $F(1, 1276) = 1372$, $p < .001$, $\eta_p^2 = .52$. This was qualified by interactions between sex of face and SRA, $F(2, 1276) = 6.90$, $p = .001$, $\eta_p^2 = .011$ and between sex of face and oral contraceptive use, $F(1, 1276) = 5.02$, $p = .025$, $\eta_p^2 = .004$. The predicted interaction among sex of face, SRA and oral contraceptive use was not significant, $F(2, 1276) = 0.06$, $p = .94$, $\eta_p^2 < .001$. All other main effects and interactions were non-significant and irrelevant to our hypotheses, all $F \leq 0.94$, $p \geq .39$, $\eta_p^2 \leq .001$.

(APA Style)

Practice VIS!

Table 7
Stevens et al. 2006, table 2: Determinants of authoritarian aggression

Variable	Coefficient (Standard Error)
Constant	.41 (.93)
Countries	
Argentina	1.31 (.33)**[B,M]
Chile	.93 (.32)**[B,M]
Colombia	1.46 (.32)**[B,M]
Mexico	.07 (.32)[A,CH,CO,V]
Venezuela	.96 (.37)**[B,M]
Threat	
Retrospective egocentric economic perceptions	.20 (.13)
Prospective egocentric economic perceptions	.22 (.12)[#]
Retrospective sociotropic economic perceptions	−.21 (.12)[#]
Prospective sociotropic economic perceptions	−.32 (.12)*
Ideological distance from president	−.27 (.07)**
Ideology	
Ideology	.23 (.07)**
Individual Differences	
Age	.00 (.01)
Female	−.03 (.21)
Education	.13 (.14)
Academic Sector	.15 (.29)
Business Sector	.31 (.25)
Government Sector	−.10 (.27)
R^2	.15
Adjusted R^2	.12
N	500

**p < .01, *p < .05, #p < .10 (twotailed)

[A]Coefficient is significantly different from Argentina's at p < .05;

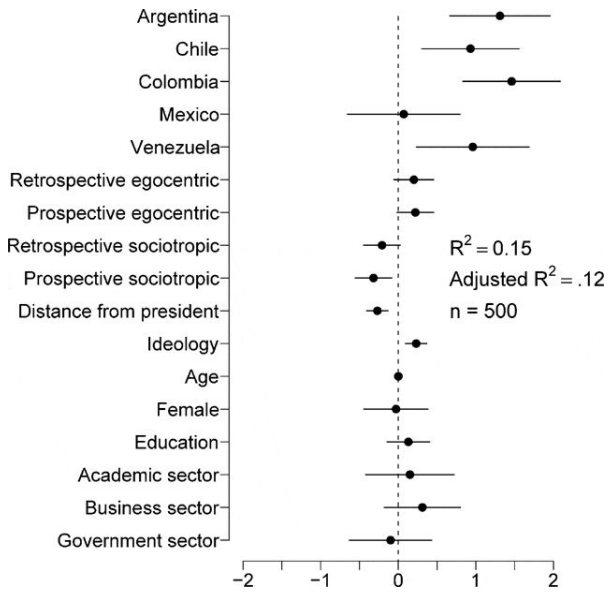
[B]Coefficient is significantly different from Brazil's at p < .05;

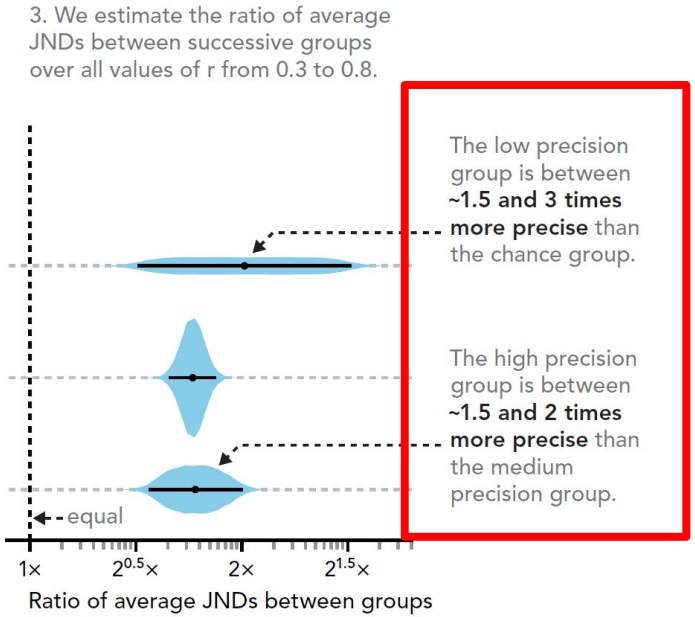
[CH]Coefficient is significantly different from Chile's at p < .05;

[CO]Coefficient is significantly different from Colombia's at p < .05;

[M]Coefficient is significantly different from Mexico's at p < .05;

[V]Coefficient is significantly different from Venezuela's at p < .05.





Jonathan P Kastellec and Eduardo L Leoni. 2007. Using Graphs Instead of Tables in Political Science. Perspectives on politics 5, 4: 755–771

Matthew Kay and Jeffrey Heer. 2016. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. InfoVIS 2015

But what can we do as reviewers?

If the field isn't there yet, can I reject a paper for using old approaches?

My current approach is to offer constructive feedback. Is that enough?

Some ideas from Transparent Stats in HCI

- Reviewing guidelines
- Exemplary papers for authors
- Reviewing MOOCs
- Badges for good practice (see also: OSF)
- Pre-registration (voluntary?)
- Flags in PCS
- …

What would work for Vis?

In summary

- Don't just blame researchers!
- Are some approaches more researcher-friendly?
- Can Vis help make empirical work better?
- Practice what we preach!
- What are the next steps…
  - …for each of us?
  - …for review processes?

Thanks! mjskay@umich.edu http://mjskay.com