9.S918: Statistical Inference for Brain and Cognitive Sciences, Pset 3
due 7 May 2024

30 April 2024

1 Analyzing pilot data, power analysis, and preregistering an experiment

It's long been known that the wording of public opinion surveys can affect how people respond to those surveys. A particularly interesting example was first documented by Rugg (1941). Compare the two closely related questions in (1) below:

(1) a. Should the US government allow public speeches against democracy?
b. Should the US government forbid public speeches against democracy?

Perhaps surprisingly, the proportion of people who respond "yes" to the *allow* version of the question is lower than the proportion of people who respond "no" to the *forbid* version of the question.

One might also speculate that participants' responses are modulated by characteristics of their personality, such as the Big Five "openness to experience" dimension, which has been reported to be associated with tendencies toward right-wing authoritarianism (Butler, 2000).

The file

https://rlevy.github.io/statistical-inference-spring-2024/assets/assignments/pset_3/forbid_allow_openness_pilot_data.csv

contains hypothetical pilot data from 100 participants (50 in each condition) for a study attempting to reproduce Rugg's original results and also to look at the relationship between openness to experience (operationalized as a standardized continuous numeric variable where higher values indicate more openness to experience).

Tasks:

1. Consider the following scientific questions:

(a) Does Rugg's original generalization—that the choice of *allow* vs. *forbid* wording affects participant response patterns—still hold today?

(b) Do response patterns vary with openness to experience?

(c) Does the effect of wording choice (if it exists) vary with openness to experience?

Fit a logistic regression model to the provided pilot data such that answers to these three questions are provided by estimates and CIs on model parameters. What can we conclude regarding these questions from the pilot data?

2. Use your fitted model to conduct a power analysis for each of the three questions. Assuming that participants are randomly assigned to wording condition (in 50/50 proportions) and that the distribution of openness to new experience in the pilot data is reflective of the overall distribution in the population from which you will draw participants for your full experiment, plot statistical power curves for each of the three questions as a function of the total number N of new participants you will recruit in your full experiment.

3. Should you really divide your participant pool 50/50 between the two wording conditions, or would an uneven split improve statistical power? Answer this question using your intuitions. Then, choose a value of N and generate a power curve as a function of the proportion of participants assigned to the *allow* condition.

2 Parameter uncertainty and power analysis

In Problem 1, you probably used maximum likelihood-based POINT ESTIMATES of model parameters θ fitted to your pilot data D_{pilot} in your Monte Carlo-based power analysis. This approach, while common, does not take into account uncertainty about θ given the pilot data. One alternative that does take into account this uncertainty is to fit your model on the pilot data using Bayesian methods, and then **integrate out** the uncertainty about θ in estimating statistical power. Let us denote by γ the "design" of your hypothetical full experiment—the sample size, how those will be allocated to experimental conditions, clusters (participants/stimuli/other), and so forth. By definition we have:

$$\text{Power}(\gamma) = P(H_0 \text{ will be rejected} | \gamma, H_0 \text{ is false}).$$

Our belief state regarding θ based on our pilot data D_{pilot} is $P(\theta|D_{\text{pilot}})$. Integrating this out gives us:

$$\text{Power}(\gamma) = \int_\theta P(H_0 \text{ will be rejected} | \gamma, \theta, H_0 \text{ is false}) P(\theta | D_{\text{pilot}}, H_0 \text{ is false}) \, \mathrm{d}\theta$$

Usually we will not be able to compute this integral exactly: for example, as we have already seen in the class we often will not be able to compute $P(\theta|D_{\text{pilot}}, H_0 \text{ is false})$ exactly

but rather must approximate it using Markov Chain Monte Carlo samples. When we have samples $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(n)}$ from this posterior, we can use Monte Carlo to estimate statistical power taking into account uncertainty about θ, as follows:

$$\text{Power}(\gamma) \approx \frac{1}{n} \sum_{i=1}^{n} P(H_0 \text{ will be rejected} | \gamma, \theta = \hat{\theta}^{(i)}, H_0 \text{ is false})$$

This is a powerful and general technique.

Questions/Tasks:

1. In general, how do you think power estimates using this method will compare to power estimates to using maximum likelihood- or maximum a posteriori (MAP)-based point estimates of θ? Will one method generally be more optimistic regarding statistical power than the other? Or will it depend on the posterior and the point estimate? If you think it will depend, how will it depend?

2. Let's work through an example. Consider a hypothetical novel vocabulary intervention for children in first grade, where trained instructors use an app to interactively teach at-risk children vocabulary items in multimedia context for 30 minutes per session, 3 sessions per week, for 8 weeks. This is a resource-intensive intervention so one wants to have a clear-eyed picture of the ability to detect whatever effects the intervention might have. Imagine a hypothetical pilot dataset with $N = 20$ first graders, half of whom received the full intervention and half of whom received no intervention, and 4 weeks after the intervention ended a follow-up vocabulary assessment was performed using a standardized test that yields a score standardized to range between 0 and 10. The hypothetical data can be found here:

   https://rlevy.github.io/statistical-inference-spring-2024/assets/assignments/pset_3/hypothetical_intervention_data.csv

   Assume that residual noise is normally distributed. **Task:** compare the estimates of statistical power resulting from maximum likelihood-based point estimation versus Bayesian model fitting plus integrating out uncertainty about model parameters. In both cases, you can take advantage of the standard power calculation for normally distributed data, which is available in many software packages (e.g., `power.t.test()` in `R`); I won't go through the derivation here but you can look it up if you like. Plot statistical power estimates as a function of overall sample size N (assuming $N/2$ participants are assigned to each condition) for the two methods, and compare them. Do the results comport with the intuitions you expressed in the first part of the problem?

3. Suppose you didn't have a readily available lookup method for statistical power based on the estimated model parameters. How would you estimate statistical power using the Bayesian posterior method developed in this problem? (You don't have to actually do this; just explain how.)

3 Anti-conservativity of non-maximal random effects structure

I described in class on April 30 the criteria for "keeping it maximal" in mixed-effects models:

- For every theoretically critical fixed-effect term in your model whose average value varies *between* clusters (e.g., participants or items), include a random intercept for that clustering;
- For every theoretically critical fixed-effect term in your model that varies *within* clusters, include a random slope for that clustering.

"Design" your own study, define an underlying model generating the observations resulting from your study, choose hypothetical parameters for your model, and use Monte Carlo estimation to evaluate whether this criterion adequately avoids anti-conservativity in statistical analysis of the data that could result from your study. Note that the presence and degree of anti-conservativity may depend on your design choices, including the number of clusters, amount of data per cluster, and chosen model parameters, so try at least a couple of choices.

4 Small numbers of clusters: treat as random effects, fixed effects, or don't model at all?

Confusion often arises regarding when to treat clusters (that is, groupings of data reflecting repeated observations from a given source, such as multiple data points from the same participant or same stimulus) as "fixed effects" versus "random effects" in hierarchical/multilevel regression modeling, and what is at stake. Sometimes one might hear, for example, that one needs a relatively large number of clusters in order to treat them as a random effect, so that you can effectively estimate the distribution of that random effect. In this example you'll investigate the wisdom of that advice.

Consider a case where you are interested in inferring the relationship between a predictor x and response y using a dataset organized into four clusters of 20 observations each. For example, x might be the number of words in a word list presented to an experimental participant, ranging from 1 to 20, the clusters are four experimental participants, and y is the fMRI BOLD response in a particular brain region of interest. Assume that the number of words in the list varies from 1 to 20, that the design is perfectly balanced, so that each participant gets exactly one list of each length, and that the particular set of words in each list differs from participant to participant (so that there are no by-stimulus repeated measures). Assume a linear relationship between x and y, with by-participant random intercepts and random slopes drawn from $\mathcal{N}\left(0, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$, and trial-level residual error variance of 1. Simulate this dataset and then consider two possible regression analyses:

1. Don't include by-participant effects at all, since there are so few clusters.

2. Treat participants as a **fixed effect**, estimating the main effect of x in the presence of an interaction with participant (as well as a main effect of participant). Make sure you use the appropriate contrast coding for representing participant effects.

3. Treat participant as a **random effect**, using a multi-level model with maximal random-effects specification and fitting the model using maximum likelihood.

Are the point estimates of the effect of x the same or different among the three analyses? What about the standard errors of those estimates? Using Monte Carlo, estimate the Type I error rate of the three approaches when the null is true ($\beta_x = 0$). Interpret your results.

References

Butler, J. C. (2000). Personality and emotional correlates of right-wing authoritarianism. *Social Behavior and Personality: an international journal*, *28*(1), 1–14.

Rugg, D. (1941). Experiments in wording questions: Ii. *Public Opinion Quarterly*, *5*(1), 91.