Understanding Statistical Power & Avoiding *P*-Hacking

Ethan Brown, Research Methodology Consulting Center, brow3821@umn.edu

Key terms

Suppose you are doing an experiment to test whether there is a difference between an experimental group mean and a control group mean in a study. **Statistical power** is the probability of correctly detecting difference of two groups statistically, if there is an effect. It is a function of four things:

1. **Effect size.** How big do you think the actual difference between means is?
2. **Within-group variability.** How much variation in means is there in each group?
3. **Sample size.** How many participants will you collect in your experimental and control groups?
4. **Significance level (α)**. What is the significance level? In other words, how low does your *p*-value need to be to reject the null hypothesis? (Often arbitrarily set at 0.05.)

The bigger the effect size, the smaller the within-group variability, and the larger the sample size, the more power. Cohen's *d* is a **scale-free effect size**, which already incorporates within-group variability.

Power Analysis

Usually, the purpose of a **power analysis** is to determine the sample size that you would need to detect a specific kind of effect. Often, a power of 0.80 is (arbitrarily) considered adequate, depending on the field and the hypothesis. However, you need to determine what effect size you would like your study to be able to detect—this may come from pilot data or previous literature. Another kind of power analysis is, given a sample size, determining the **minimum detectible effect (MDE)**. Given your study's parameters, how large an effect can you determine with sufficient power?

Power analysis can be done in software like **G*Power, R** (with the *pwr* package), **Optimal Design**, and many others. They can also be done in many software by **simulation**: repeatedly simulating performing the study with random error, and calculating how many of those simulated studies lead to correctly rejecting the null hypothesis. Simulation can be used in situations where mathematical solutions are not realistic.

Lack of Power and *P*-Hacking

Many studies in social science are **underpowered**, meaning they have too small a sample size to reliably detect the effects that they are trying to find. (Most effect sizes in social science are small.) This means that their findings may be spurious and simply due to random variation. Conducting a large number of underpowered analyses, and only reporting on the ones where $p < .05$, is known as ***p*-hacking**, and contributes to false findings entering the research literature.

One of the primary practices to avoid *p*-hacking is to do power analysis in advance, and to *preregister* the study: specify the main study analyses ahead of time and clarify what is **exploratory** and what is **confirmatory**. This can be done on sites like aspredicted.org and osf.io.

About RMCC

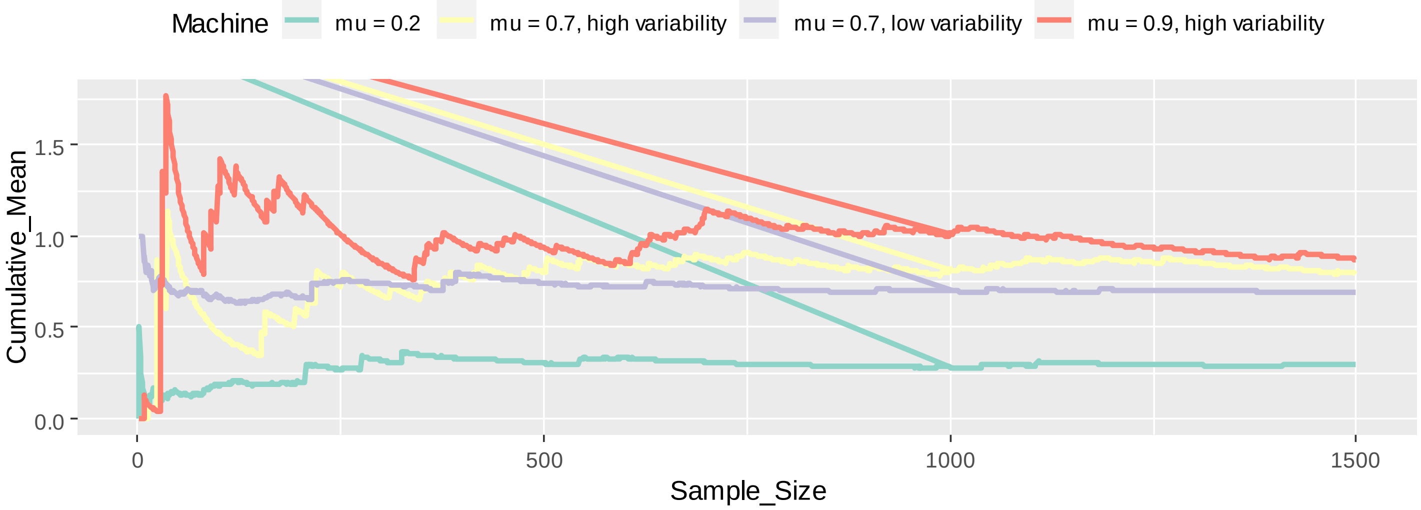
The **Research Methodology Consulting Center (RMCC)** is a new CEHD service unit with specialized expertise in quantitative research methods. With a combination of services unique to education and human development research, we offer consulting on everything from developing a hypothesis and selecting the appropriate statistical analysis to the development of data collection protocols and reporting templates. Visit cehd.umn.edu/research/consulting for more information and to request a consultation. CEHD faculty/staff have 6 free consultations per year for unfunded research, and CEHD graduate students have 4 free consultations per year for Masters theses and PhD dissertations.

Lunch & Learn: "All Power Analyses are Wrong. How to Be More Right", Chris Desjardins, RMCC
March 6, 12-1PM, Burton Hall 227 (Lunch Provided)

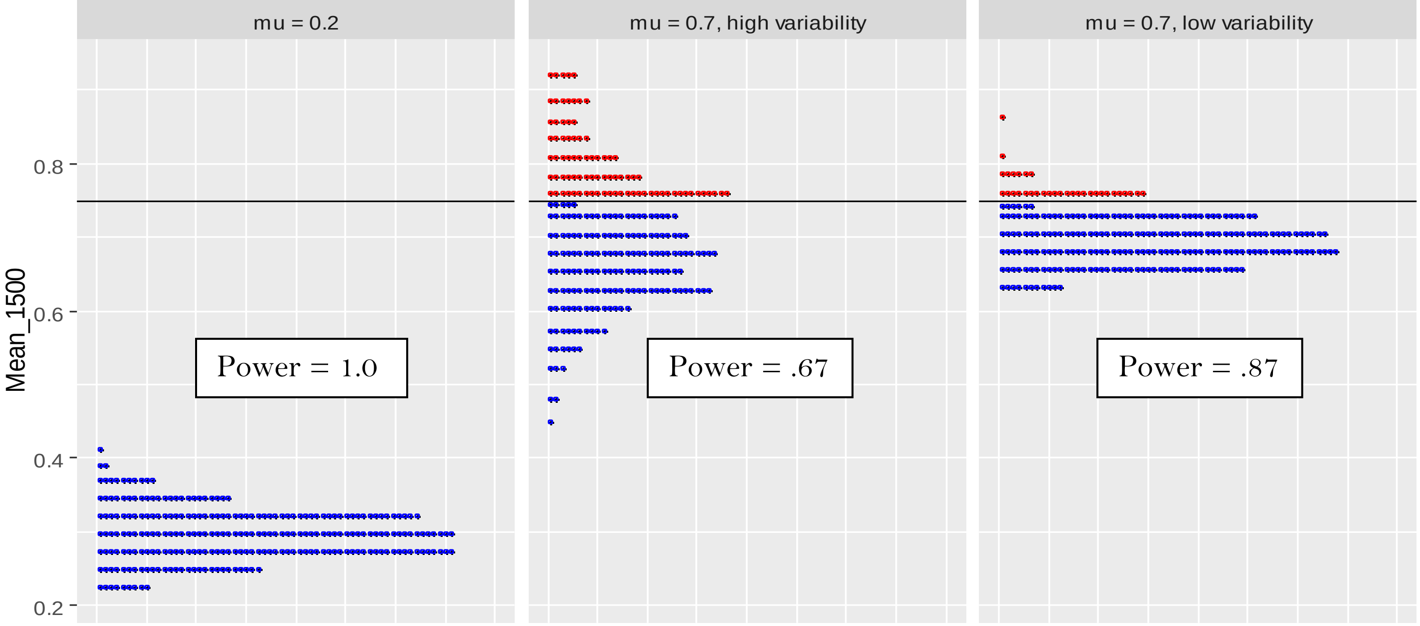
Example

You work for the state casino regulation committee. Your job is to ensure that casinos are accurately reporting to customers the average winnings from slot machines. Suppose one slot machine pays out $0, $1, or $20 on each game, and the machine claims that the average payout is $0.90. You can play the slot machine as many times as you want, but it costs money each time. **What is the minimum sample size needed to detect whether the machine's claim is accurate?**

Graph of the means of 4 different machines as they grow from *n* = 1 to *n* = 1500



Reject H₀ for *n* = 1500 at α = .05 when M < 0.75


mu = 0.2
mu = 0.7, high variability
mu = 0.7, low variability
Mean_1500
Power = 1.0
Power = .67
Power = .87