Lab 2

Power by Simulation

Sean Sylvia

2025-02-13

Table of contents

Overview and Learning Objectives 1
The Hospital Drama Begins 2
The Plot Thickens: Power Calculations Without Clustering 2
Understanding the Sources of Uncertainty 5
Power Calculations With Clustering 6
The Truth About *p*-Values 9
(Optional) Plotting the Minimum Detectable Effect (MDE) 9
Final Words from the Hospital Staff 10
Submission Instructions 10

Overview and Learning Objectives

In this lab, you will explore the concept of **statistical conclusion validity** by conducting **power calculations via simulation**. Specifically, you will:

1. Conduct a power simulation for a simple randomized experiment **without clustering** (i.e., ignoring the fact that participants may be grouped within clinics).
2. Extend that simulation to **account for clustering** (i.e., clinics as the unit of randomization).
3. Examine the impact of **sample size**, **effect size**, and **clustering** on statistical power, and visualize how minimum detectable effect sizes (MDE) change with sample size.

By the end of this lab, you will have a deeper understanding of: - How **sources of uncertainty** (sampling variation, variance in potential outcomes across participants, and measurement error) affect your study's outcomes. - Why we must beware of **Type I** (false positive) and **Type II** (false negative) errors—especially if Dr. P-Hacker is anywhere near our data. - How

p-values should (and should not!) be interpreted. - How to **simulate data** that include cluster-level effects and adjust for these effects in your analysis.

The Hospital Drama Begins

Welcome to **St. Null's Memorial Hospital**, where a brand-new quality improvement intervention is about to be tested. The hospital's network of clinics, collectively called **The Null Distribution**, is famous for its dedication to meticulous data collection—and also for some questionable statistical practices performed by a rather infamous staff member.

- **CEO Barnaby Beta** has championed a new, cost-intensive intervention aimed at improving patient satisfaction.
- **Nurse Random** insists on conducting a proper **randomized control trial (RCT)**, but quickly realizes **Dr. P-Hacker** might have already meddled with the initial power calculations.
- **Dr. P-Hacker** is known to declare victory ("Significant at the 5% level!") before even cleaning the data, and is rumored to own a golden "`p < 0.05`" sign that he waves in staff meetings.
- **Dr. Doub R. Obust** is the voice of reason, reminding everyone of fundamental statistical principles.

In this lab, you (the consultants) have been summoned to **salvage** the situation. Let's see how this plays out.

The Plot Thickens: Power Calculations Without Clustering

Scene 1: The Mysterious Spreadsheet

Nurse Random bursts into the conference room, clutching a color-coded spreadsheet.

> **Nurse Random:** "Dr. P-Hacker, your power calculations look suspiciously high. Did you account for the fact that we're randomizing by clinic, not by individual patient?"
>
> **Dr. P-Hacker:** "Of course not, Nurse Random. Look, a random patient is a random patient—no matter which clinic they come from! Besides, I love high power. It makes me feel like my study can conquer the world!"

> **Dr. Doub R. Obust (rolling eyes):** “We’re going to need to run a simulation that properly reflects how the intervention is assigned at the **clinic** level, not just among individual patients. Otherwise, your calculations will be as overconfident as your new P-value neon sign.”

Nonetheless, **Dr. P-Hacker** shares his “method” with you first. Let’s start by **replicating** his approach (ignoring clustering). This will be our baseline—**what not to do** if clinics are truly the unit of randomization.

Step 1: Flawed Power Simulation (Ignoring Clustering)

We’ll simulate a dataset **as if** individual patients are randomly assigned to treatment or control. We’ll compute the statistical power for detecting a **true treatment effect** of a specified size, ignoring any clinic-level differences.

Task 1: Flawed Simulation Setup

Fill in the code below to acheive a power of around 0.8 (may not be exact, but get as close as you can):

1. Set a **sample size** for the total number of patients.
2. Specify a **treatment effect** (`effect`).
3. Decide on the **proportion** of participants to receive treatment (`prop`).
4. Select a **significance level** (`t_alpha`).
5. Run multiple simulations (`sim.size`).

Then, run a linear regression on each simulated dataset, gather the p-values, and compute how often the null hypothesis is rejected (i.e., estimate power).
Note: The code below is not executable. You need to change the `eval: false` to `eval: true` to make it work and fill in the blanks.

```
library(data.table)
set.seed(123456)

# Define parameters
sample_size <-      # e.g. 500
```

```
effect <-           # e.g. 0.2
prop <-             # e.g. 0.5
t_alpha <-          # e.g. 0.05
sim.size <-         # e.g. 2000

# Initialize storage for results
reject_t <- numeric(sim.size)

# Simulation loop
for (i in 1:sim.size) {
  dt <- data.table(id = 1:sample_size)

  # Assign treatment individually (incorrect for cluster randomization!)
  dt[, treatment := rbinom(.N, 1, prop)]

  # Simulate outcome (10 is baseline, 'effect' is added if treatment=1)
  dt[, outcome := 10 + effect * treatment + rnorm(.N, mean = 0, sd = 1)]

  # Estimate the effect
  fit <- lm(outcome ~ treatment, data = dt)
  p_value <- summary(fit)$coefficients[2,4]

  # Check if we reject H0: (p-value < alpha)
  reject_t[i] <- ifelse(p_value < t_alpha, 1, 0)
}

# Compute estimated power
power_flawed <- mean(reject_t)
cat("Estimated Power (Ignoring Clustering):", power_flawed, "\n")
```

Discussion of Step 1

- **Nurse Random** sighs: "We got an estimated power of 0.8 (yours might be slightly different). But do we trust this number?"
- **Dr. Doub R. Obust** chimes in: "Nope. We're ignoring that patients within the same clinic might be more similar to each other than to patients in other clinics. Our standard errors are artificially small."

At this point, **CEO Barnaby Beta** perks up: "Artificially small standard errors? That means we're basically claiming more precision than we actually have, right?"

Yes, indeed. If we disregard the **clustering** of patients, we risk making a **Type I error** more often than we realize. Dr. P-Hacker, in typical fashion, responds:

> **Dr. P-Hacker:** "Type I error? Isn't that just when we see something interesting that's obviously there?!"
> **Dr. Doub R. Obust (groaning):** "No. A Type I error is a *false positive*—we conclude there *is* an effect even though, in truth, there isn't. Like thinking you've discovered a rare golden banana flavor at the cafeteria soda machine when really it's just seltzer water with a weird label!"

Understanding the Sources of Uncertainty

Before we fix our simulation, **Dr. Doub R. Obust** insists that you reflect on **why** ignoring clinic-level clustering is a problem. He ticks off sources of variability that a real experiment faces:

1. **Sampling Variation**: Even if you have a large population of patients, the sample you select is just one draw from a bigger population. Different samples might give different estimates.

2. **Variance in Potential Outcomes**: Not all patients respond to treatments in the same way. Some might be strongly affected, some not at all—leading to variation in the outcomes.

3. **Measurement Error**: If your measurement tool for patient well-being is noisy (e.g., patients often mis-report how they feel, or staff record data incorrectly), it introduces extra variability that can muddy your effect estimates.

4. **Clustering**: Patients in the *same clinic* share certain characteristics, environmental factors, or staff practices. This correlation must be accounted for in the design and analysis.

> **Nurse Random:** "So if we ignore that last point—clustering—our estimate of the variability (and thus the standard error) is off, and we might incorrectly claim significance. That's basically p-hacking 101!"

Power Calculations With Clustering

Scene 2: A (Cluster-)Randomized Trial

Now we come to the **correct** approach: our **unit of randomization** is the **clinic**, not the individual. That means each clinic either receives the intervention or does not, and all patients in a clinic share the same treatment assignment.

Step 2: Incorporating Clustering

We'll model:

- **`num_clusters`** = number of clinics.
- **`cluster_size`** = number of patients per clinic.
- **`icc`** (*Intraclass Correlation Coefficient*): A measure of how strongly patients in the same clinic resemble each other. High ICC means patients within a clinic are more correlated.
- We'll generate clinic-level "random effects" and individual-level error terms to reflect **both** measurement error and the variability in potential outcomes across individuals.

❗ Task 2: Correcting the Simulation for Clustering

Fill in the code below to acheive a power of 0.8:

1. Define the **number of clinics** and the **number of patients per clinic**.
2. Assign each entire clinic to treatment or control.
3. Incorporate **cluster-level** random effects.
4. Fit the model but adjust standard errors for clustering.

```r
library(multiwayvcov)  # for cluster-robust standard errors

# Define cluster parameters
num_clusters <-     # e.g. 50
```

```r
cluster_size <-     # e.g. 10
total_sample <- num_clusters * cluster_size
icc <-              # e.g. 0.4

# Variances
ind_err_var <- 1
# cluster_err_var is derived from the intraclass correlation coefficient
cluster_err_var <- (icc * ind_err_var) / (1 - icc)

sim.size <-         # e.g. 2000
effect <-           # e.g. 0.2
prop <-             # e.g. 0.5
t_alpha <-          # e.g. 0.05

reject_t <- numeric(sim.size)

set.seed(654321)
for (i in 1:sim.size) {

  # Create a data table with one row per cluster, repeated for each patient
  clusters <- data.table(
    cluster = rep(1:num_clusters, each = cluster_size)
  )

  # Generate a cluster-level random effect
  clusters[, cluster_error := rep(
    rnorm(num_clusters, mean = 0, sd = sqrt(cluster_err_var)),
    each = cluster_size
  )]

  # Randomize at the clinic level
  clusters[, treatment := rep(
    rbinom(num_clusters, 1, prop),
    each = cluster_size
  )]

  # Add an individual-level error
  clusters[, individual_error := rnorm(.N, mean = 0, sd = sqrt(ind_err_var))]

  # Final outcome for each individual
  clusters[, outcome := 10 + effect * treatment + cluster_error + individual_error]

  # Fit a naive linear model (ignoring clustering in standard errors)
  fit <- lm(outcome ~ treatment, data = clusters)

  # Adjust standard errors for clustering
  robust_SE <- cluster.vcov(fit, clusters$cluster, df_correction = TRUE)
  robust_coef <- coeftest(fit, robust_SE)

  # Get the p-value for the treatment coefficient
  p_value <- robust_coef[2,4]
  reject_t[i] <- ifelse(p_value < t_alpha, 1, 0)
}

# Compute estimated power with clustering
power_clustered <- mean(reject_t)
cat("Estimated Power (Clustered):", power_clustered, "\n")
```

Discussion of Step 2

Dr. Doub R. Obust looks at the new power estimate and remarks:

> "As you can see, once we account for clustering, the power is (usually) **lower** than in the flawed approach. That's because those cluster-level similarities effectively reduce the amount of independent information we have. Our standard errors are bigger, so it's harder to find significance unless the effect size or sample size is larger."

CEO Barnaby Beta shakes his head: "But that means our study might be underpowered if we stick to our current budget!"

Nurse Random replies with a grin: "Don't worry, we can plan more carefully. Otherwise, if we run a smaller study, we risk a **Type II error**—failing to reject the null hypothesis when there really is an effect. You know, like leaving the gold standard intervention on the shelf because we didn't gather enough data to prove it works."

Dr. P-Hacker interjects:

> "And there's always `p < 0.10`, right? We can move our threshold for significance to get more 'positive' results!"
> **Nurse Random (scolding):** "That's *literally* p-hacking. Please step away from the analysis, Doctor."

The Truth About p-Values

A quick comedic break for a lesson on p-values:

- **Dr. P-Hacker** keeps shouting that $p < 0.05$ means there's a 5% chance the null hypothesis is true.

- **Nurse Random** corrects him: "No, no, no. A p-value is the probability of observing a result *at least as extreme as ours* if the null is true. It's NOT the probability the null is true. You can't interpret it that way."

  Dr. P-Hacker: "I was so sure it was the probability that *I* was right."
  Dr. Doub R. Obust: "We live and learn, Doctor."

(Optional) Plotting the Minimum Detectable Effect (MDE)

Because **CEO Barnaby Beta** wants to know how large an effect must be to have a reasonable chance of detection, you might want to simulate across a range of effect sizes and/or sample sizes. You can then plot the resulting power curves to see what the **Minimum Detectable Effect** (MDE) is for a given power requirement.

Task 3: Create an MDE Plot

1. Loop over a grid of *effect sizes* (or sample sizes).

2. For each value, compute power using the cluster-based simulation approach.

3. Plot the *effect size* on the x-axis and the corresponding *power* on the y-axis.

```r
# Example code snippet (feel free to modify)
library(ggplot2)

effect_sizes <- seq(0, 0.5, by = 0.05)
results <- data.table(effect = effect_sizes, power = NA_real_)

for (e in seq_along(effect_sizes)) {
  # Repeat your cluster simulation steps but with effect = effect_sizes[e]
  # ...
  # store the average of reject_t in results$power[e]
  # ...
}

ggplot(results, aes(x = effect, y = power)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Power vs. Effect Size",
    x = "Effect Size",
    y = "Power"
  ) +
  theme_minimal()
```

Final Words from the Hospital Staff

CEO Barnaby Beta: "Alright, so if we need to ensure we have enough clinics and participants to achieve our desired power, we may need a bigger budget than expected. Let's not forget that ignoring clustering would have given us a false sense of security in our power. Now we know better."

Nurse Random: "And no more Type I or Type II error confusion, Dr. P-Hacker. We must keep our definitions straight if we're to have any credibility around here!"

Dr. P-Hacker (sighing): "Fine, fine. Guess I'll tone down the p-value hype. But I'm keeping my neon sign."

Dr. Doub R. Obust (with a grin): "We'll allow the neon sign, as long as you promise to interpret it *correctly*."

Submission Instructions

1. Make sure your `.qmd` file knits successfully (to `.pdf` or `.html`).

2. Upload your compiled document to Gradescope:

3. **Include** your discussion of the results, your code, and your responses to the callout sections.

Remember, the moral of the story: **Always check who (or what) is being randomized, and account for clustering when necessary!** Otherwise, your study conclusions might be as random as Dr. P-Hacker's next dinner choice.

Good luck!