Does placing signs encouraging stairs over elevators increase stair use?

Darby Brown, Hanjin Jia, Thong Trinh, Victor Popp, Lena Reissinge

2024-10-15

Table of contents

Dedication Page

This analysis is dedicated to the only being that loves power more than we do.

Figure 1: Gollum when he sees our statistical analysis.

1 Experimental Power Analysis

1.1 Experiment Overview

In this experiment, we aim to examine whether the placement of a sign stating ‘Use the stairs, improve your health’ influences the choice between using stairs versus elevators. This document provides an overview of our power analysis to determine the necessary sample sizes to test our hypothesis.

1.2 Testing Procedure

- Metrics: Our metric of interest is the percentage of people choosing to use the elevator versus the stairs in both the control and treatment groups.
- Treatment: The treatment group is exposed to a sign with the message “Use the stairs, improve your health,” while the control group is not exposed to any sign.
- Statistical Power Approach: A series of t-tests are conducted to compare elevator use between the control and treatment groups (Null-hypothesis: treatment effect is zero). To ensure sufficient power, we simulate different sample sizes and calculate the proportion of experiments in which we would correctly reject the null hypothesis. The aim is to determine the minimum sample size required to detect significant differences at varying levels of treatment effects.
- Assumptions:

  1. Baseline Behavior: In the control group, 80% of participants use the elevator, with a standard deviation of 15%.
  2. Treatment Effect: We assume the sign may reduce elevator use by 2% to 20%, depending on its effectiveness.

```r
#Create dummy data
#assuming that:
    # control group has 80% of people taking elevator with a SD of 15%
    # treatment effect is a 2-20% reduction in people taking the elevator'
d <- data.table(
  id = 1:1100
)
```

```
d[ , D := sample(c('treatment', 'control'), size=.N, replace=T)]
d[D == "control",   Y_control   := rnorm(n=.N, mean=80, sd=15)]

#set up four possible treatment effect sizes [2, 8, 14, 20]
d[D == "treatment", Y2_treatment := rnorm(n=.N, mean=78, sd=15)]
d[D == "treatment", Y8_treatment := rnorm(n=.N, mean=72, sd=15)]
d[D == "treatment", Y14_treatment := rnorm(n=.N, mean=66, sd=15)]
d[D == "treatment", Y20_treatment := rnorm(n=.N, mean=60, sd=15)]
d[, Y2 := ifelse(D == "control", Y_control, Y2_treatment)]
d[, Y8 := ifelse(D == "control", Y_control, Y8_treatment)]
d[, Y14 := ifelse(D == "control", Y_control, Y14_treatment)]
d[, Y20 := ifelse(D == "control", Y_control, Y20_treatment)]
```

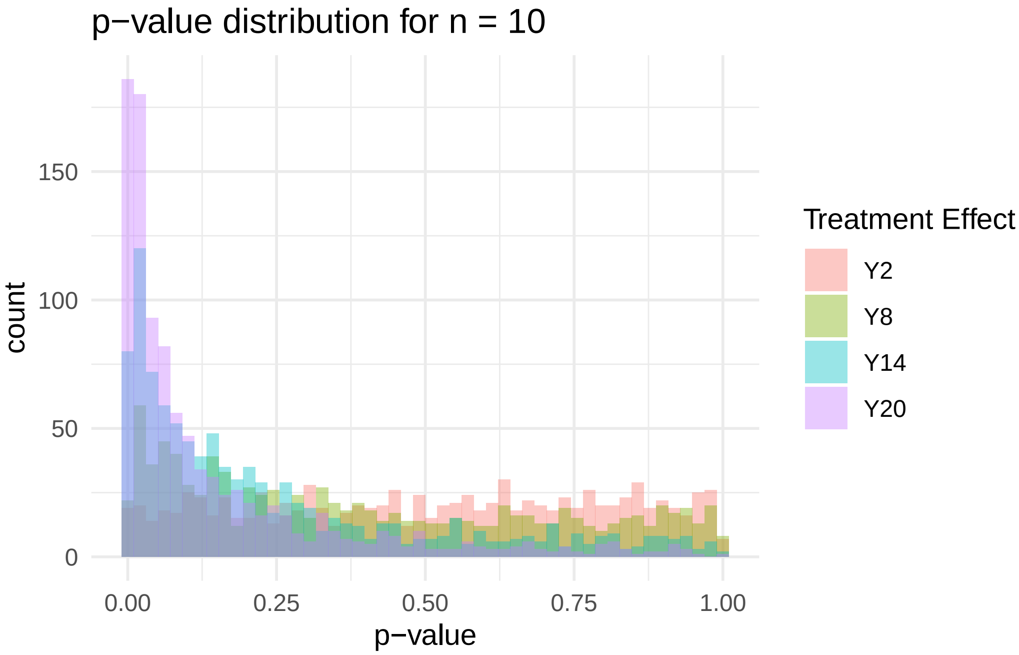
1.3 Simulation of 10 Subjects

In case we could only observe 10 people – 5 in treatment and another 5 in control - what sorts of p-values could we expect across the different treatment effect sizes we think could be possible? We sample 10 people from our dummy data and conduct t-tests four thousand times to plot the distribution of p-values observed at each sample size and ATE.

```
#list out hypothetical treatment sizes to run t test on
treatments_to_test <- c("Y2", "Y8", "Y14", "Y20")
#create a table to plot p-values per treatment effect
t_test_p_values <- data.table(
  id = 1:1000
)
#loop through the treatment effect sizes and simulate t tests
for (treatment_effect in treatments_to_test) {
  t_test_p_values_i <- c()
  for (i in 1:1000) {
    d10_i <- d[, .SD[sample(.N, 5)], keyby = D]

    t_test_ten_people_i <- t.test(d10_i[D == 'control', ..treatment_effect],
                                d10_i[D == 'treatment', ..treatment_effect])
    t_test_p_values_i <- c(t_test_p_values_i, t_test_ten_people_i$p.value)
  }
  t_test_p_values[, (treatment_effect) := t_test_p_values_i]
}
```

```
# Reshape the data from wide to long format
t_test_p_values_long <- melt(t_test_p_values, id.vars = "id",
                             variable.name = "treatment_effect",
                             value.name = "p_value")
# Plot the 4 series (columns) as histograms
ggplot(t_test_p_values_long, aes(x = p_value, fill = treatment_effect)) +
  geom_histogram(bins = 50, alpha = 0.4, position = "identity") +
  labs(
    title = 'p-value distribution for n = 10',
    x = 'p-value',
    fill = 'Treatment Effect'
  ) +
  theme_minimal()
```



From the plot, we observe flat distributions for Y2 and Y8 (which correlate to possible ATEs of 2% and 8%, respectively). Y14 (ATE 14%) and Y20 (ATE 20%) have distributions with a left skew, indicating that we would correctly reject the null hypothesis more often for those larger ATEs.

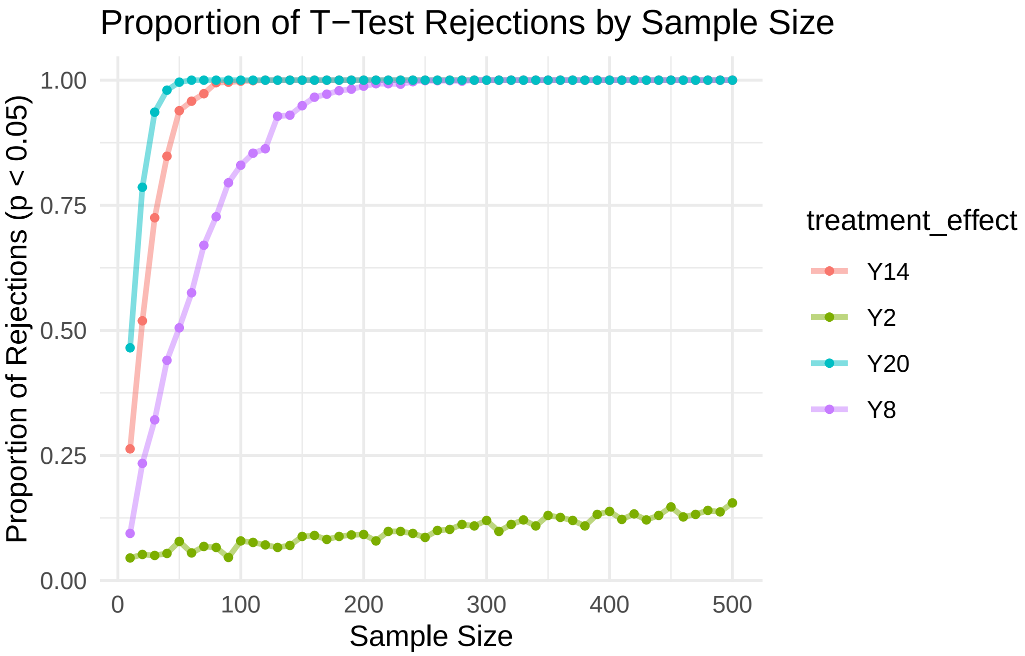
1.4 Determining Optimal Sample Size based on Estimated Treatment Effect

Here we simulate the percentage of times we would correctly reject the null across different effect sizes and sample sizes.

```r
#create a data table to record t-test rejects across varying sample sizes
experiment_simulations <- data.table(
  sample_size = integer(),
  t_test_rejects = numeric(),
  treatment_effect = character()
)
#make a vector to iterate through different numbers of people in the sample
people_in_sample <- c(seq(10,500, by = 10)) #sequence 10-1000 increment by 10

#three nested for loops this is embarassing...
#loop over the different possible treatment effects
for (treatment_effect in treatments_to_test) {
  #loop over the possible number of people we could have in our sample
  for (i in 1:(length(people_in_sample))) {
    t_test_p_values_i <- rep(NA, 1000)
    #simulate 100 randomized experiments with each setting for n (sample size)
    for (j in 1:1000) {
      d_i_j <- d[, .SD[sample(.N, people_in_sample[i]/2)], keyby = D]
      t_test_i_people_j <- t.test(d_i_j[D == 'control', ..treatment_effect],
                                  d_i_j[D == 'treatment', ..treatment_effect])
      t_test_p_values_i[j] <- t_test_i_people_j$p.value
    }
    t_test_rejects_i <- sum(t_test_p_values_i < 0.05) / length(t_test_p_values_i)
    new_row = data.table(sample_size = people_in_sample[i],
                         t_test_rejects = t_test_rejects_i,
                         treatment_effect = treatment_effect)
    experiment_simulations <- rbind(experiment_simulations, new_row)
  }
}
```

```r
#plot all simulated experiments
ggplot(experiment_simulations) +
  aes(x = sample_size, y = t_test_rejects, color = treatment_effect) +
  geom_line(linewidth = 1, alpha = 0.5) +
  geom_point(size = 1) +
  labs(
    title = "Proportion of T-Test Rejections by Sample Size",
    x = "Sample Size",
    y = "Proportion of Rejections (p < 0.05)"
    ) +
  theme_minimal()
```



Based on this power analysis, we can see that for any meaningful treatment effect size, we can expect very strong statistical power above approx. n = 200. If we wished to model a small treatment effect like 2% more people taking the stairs, we would need a very large experiment. Importantly, these conclusions rest on the assumption that the standard deviation of the percentage of people taking stairs each day is 15%. If we had a smaller standard deviation we may be able to measure smaller treatment effects with smaller sample sizes. The extent to which standard deviation impacts these results will be considered for future research.