INFO-241: Power Analysis

Aarushi Somani, Austin Tao, Eleanor Kim, Grayson Meckfessel, Lauren Murai

February 26th, 2024

Experiment Background

We hope to investigate and quantify the effect that labels and quality information have on people's taste perception of foods. In particular, we use a difference-in-difference model to estimate the effects of quality information on taste ratings of cookies.

The treatment group will be informed about which off-brand oreo they are tasting before they eat it. They will be explicitly told when they are eating a "high quality" cookie versus a standard off brand cookie. On the other hand, the control group will taste both types of cookies without knowing any information about the brand/ingredients. They will be blinded to the brand identity to observe the outcome of pure taste preference without the influence of preconceived notions. The cookies will be packaged in the same manner (on a plate) for both groups and one group will be given the information and the other will not. They will eat both cookies and then give a rating on a discrete scale of 1-10 for both the cookies they eat.

Let's say we get high-quality, organic, gourmet off-brand oreos and poorer quality off-brand oreos. The treatment group eats both and gives a rating of both (we assume for now that the order in which they eat the cookies is independent to their ratings) and the control group also does the same. The difference in the rating of the two oreos is calculated for each group, and we calculate the difference between the groups.

Power Analysis

To get an idea of the amount of data we may need to run such an experiment, we simulate a power analysis below to see if the difference between treatment group and control group is significant or not. For both cases, our null hypothesis is that on average, there is no difference between the ratings of the treatment and control groups, and the alternative is that there is a difference between the groups.

In order to inform the parameters we use in our simulation, we conducted a small test survey from 20 individuals: 10 of whom were in treatment providing ratings for the "good" cookie and the "bad" cookie after we told them about the quality of their cookies, and 10 of whom were in control providing ratings for the "good" cookie and the "bad" cookie without being given information about their cookies. We take the mean ratings from each of the four subgroups and their standard deviations for our simulated data to get the following results:

```
# Data
good_cookie_control <- c(9, 7, 6, 7, 6, 5, 8, 7, 9, 7)
bad_cookie_control <- c(8, 7, 4, 6, 8, 5, 5, 6, 8, 7)
good_cookie_treatment <- c(8, 9, 7, 7, 6, 8, 8, 7, 9, 7)
bad_cookie_treatment <- c(7, 5, 4, 5, 5, 6, 7, 5, 8, 6)

# Calculations for Control Group
control_stats <- data.frame(
  Type = c("Mean", "SD"),
  Good_Cookie = c(mean(good_cookie_control), sd(good_cookie_control)),
```

```
  Bad_Cookie = c(mean(bad_cookie_control), sd(bad_cookie_control))
)

# Calculations for Treatment Group
treatment_stats <- data.frame(
  Type = c("Mean", "SD"),
  Good_Cookie = c(mean(good_cookie_treatment), sd(good_cookie_treatment)),
  Bad_Cookie = c(mean(bad_cookie_treatment), sd(bad_cookie_treatment))
)

# Printing Tables
knitr::kable(control_stats, caption = "Control Group Statistics")
```

Table 1: Control Group Statistics

Type	Good_Cookie	Bad_Cookie
Mean	7.100000	6.400000
SD	1.286684	1.429841

```
knitr::kable(treatment_stats, caption = "Treatment Group Statistics")
```

Table 2: Treatment Group Statistics

Type	Good_Cookie	Bad_Cookie
Mean	7.6000000	5.800000
SD	0.9660918	1.229273

These results seem reasonable compared to research we've found online, which demonstrate an approximate 10% increase in rating on their respective rating scales (see Schouteten et al., 2015 and Sutterlin and Siegrist). Both of these studies have sample sizes of around n = 150.

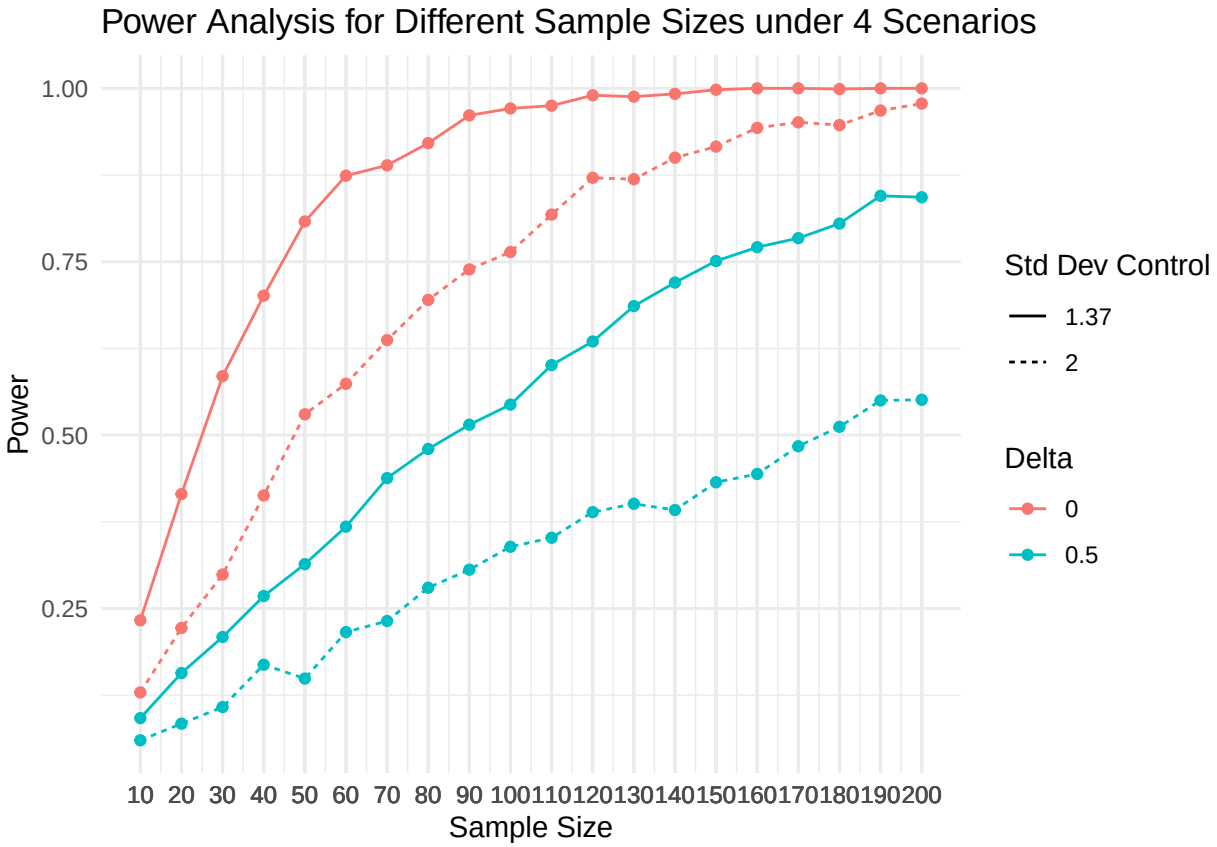
To modify our simulations to produce four scenarios, we consider what happens if we have 2 different delta values (0 and 0.5) and 2 different sets of standard deviations for the treatment in control, where one is a constant conservative (larger) estimate for both the control and treatment and the other are the standard deviation values obtained from our sample. This gives us 4 different scenarios to simulate, which we do below, using a two-sample t-test. In two of our scenarios, we modify the treatment group ratings by scaling down the mean score of the "good" cookies (we call this quantity delta) to produce a more conservative difference in means than what we observed in our survey. Our 4 different scenarios are the following:

1. Using both the sample means and the sample standard deviations from our survey
2. Using a smaller treatment mean for the "good" cookie by subtracting 0.5, and then using the standard deviations from our survey
3. Using the sample means from our survey, but using a constant larger standard deviation
4. Using a smaller treatment mean for the "good" cookie by subtracting 0.5, but using a constant larger standard deviation

```
# Function to simulate the experiment and perform a t-test
simulate_experiment <- function(n, delta, std_dev_t, std_dev_c) {

  # Simulate ratings for real and off-brand oreos for the treatment group
  good_oreo_treatment <- rnorm(n, mean = mean(good_cookie_treatment) - delta,
```

```
                                sd = std_dev_t)
  bad_oreo_treatment <- rnorm(n, mean = mean(bad_cookie_treatment) ,
                              sd = std_dev_t)

  # Simulate ratings for real and off-brand oreos for the control group
  good_oreo_control <- rnorm(n, mean = mean(good_cookie_control), sd = std_dev_c)
  bad_oreo_control <- rnorm(n, mean = mean(bad_cookie_control), sd = std_dev_c)

  # Calculate the differences in ratings
  diff_treatment <- good_oreo_treatment - bad_oreo_treatment
  diff_control <- good_oreo_control - bad_oreo_control


  # Perform a two-sample t-test and return the p-value
  t_test <- t.test(diff_treatment, diff_control)
  return(t_test$p.value)
}

# Power analysis function across different sample sizes and scenarios
power_analysis <- function(initial_n, delta_values, std_dev_values, alpha,
                           step_size, max_multiplier) {
  # Create a grid of scenarios
  scenarios <- expand.grid(delta = delta_values, std_dev = std_dev_values)

  # Initialize vectors to store results
  sample_sizes <- seq(initial_n * step_size, initial_n * max_multiplier,
                      by = initial_n * step_size)
  powers <- numeric(length(sample_sizes) * nrow(scenarios))

  # Loop over different sample sizes and scenarios
  for (i in 1:length(sample_sizes)) {
    n <- sample_sizes[i]

    for (j in 1:nrow(scenarios)) {
      delta <- scenarios$delta[j]
      std_dev_t <- scenarios$std_dev[[j]][1]
      std_dev_c <- scenarios$std_dev[[j]][2]
      num_simulations <- 1000
      p_values <- replicate(num_simulations, simulate_experiment(n, delta,
                                                                 std_dev_t, std_dev_c))

      # Calculate the power for this sample size and scenario
      powers[(i - 1) * nrow(scenarios) + j] <- mean(p_values < alpha)
    }
  }


  # Create a data frame of results
  results <- data.frame(sample_size = rep(sample_sizes, each = nrow(scenarios)),
                        power = powers,
                        delta = rep(scenarios$delta, times = length(sample_sizes)),
                        std_dev_t = rep(unlist(lapply(scenarios$std_dev,
                                        function(sublist) sublist[[1]])),
```

```
                                        times = length(sample_sizes)),
                        std_dev_c = rep(unlist(lapply(scenarios$std_dev,
                                        function(sublist) sublist[[2]])),
                                        times = length(sample_sizes)))

  return(results)
}

# Scenario 1: Use the mean and standard deviations we found from our sample

# Scenario 2: Use a more conservative mean by modifying with delta
#             standard deviations we found from our sample

# Scenario 3: Use the mean we found from our sample, more conservative & constant sd's

# Scenario 4: Use a more conservative mean by modifying with delta,
#.            more conservative and constant sd's

# Parameters
initial_n <- 100
delta_values <- c(0, .5) # Different delta values for different scenarios
std_dev_values <- list(c(1.42,1.37), c(2,2)) # Diff std_dev values for diff scenarios
alpha <- 0.05 # Significance level
step_size <- 0.1 # Starting from 10% of the initial sample size
max_multiplier <- 2 # Up to 200% of the initial sample size

# Conduct the power analysis for each combination of delta and std_dev
results <- power_analysis(initial_n, delta_values, std_dev_values, alpha,
                          step_size, max_multiplier)

# Plot the results
ggplot(results, aes(x = sample_size, y = power, color = factor(delta),
                    linetype = factor(std_dev_c))) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = results$sample_size) +
  labs(title = "Power Analysis for Different Sample Sizes under 4 Scenarios",
       x = "Sample Size",
       y = "Power",
       color = "Delta",
       linetype = "Std Dev Control") +
  theme_minimal()
```



From the plots above, we find that our test's power exceeds 0.5 for all combinations of parameters after around n = 170. At that sample size, every parameter combination except for delta = 0.5, std = 2 achieves a power greater than 0.8. This informs us that as long as we can get a sample size of around n = 170, we expect our test to have high power. In fact, this value is very similar to those used in the studies mentioned above, confirming the validity of our power simulation.