Final 241 Project: Power Analysis

Arisa Nguyen, Annie Vellon, Samuel Omosuyi

10/4/2022

Synthetic Data

The synthetic data contained columns for subject ID, treatment/control indicator, and the a average score on the subject's opinion of working from the office. The average score was prescribed based on a 1 - 5 Likert scale, with 1 being a negative opinion of working from the office and 5 being a positive opinion of working from the office. The average score was assigned using a random distribution. Treated subjects were randomly assigned a average score between 3 to 5, while control subjects were randomly assigned a average score between 1 - 4. In industry, it has been noticed that when office ammenities are open (like gyms and cafes), show up rates are higher (show up rate defiend as the percentage of people badging into the office on a given day). In at least one of our messages, we plan to call out office ammenities, and therefore assume that this will increase the average survey score (just like it did with show up rates).

```
d <- fread("practice_data/power_analysis_data.csv")
head(d, 3)
```

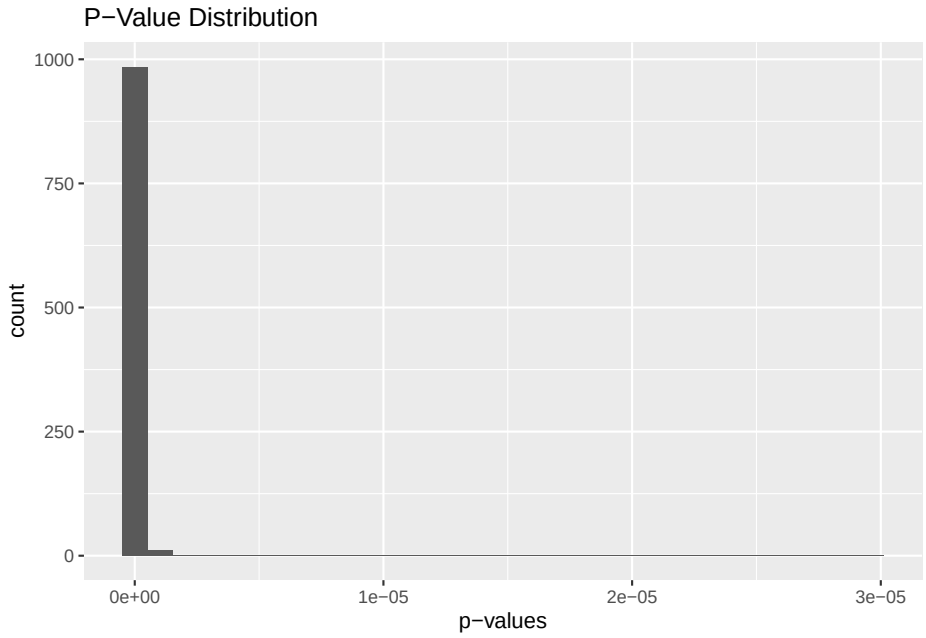
```
##    subject_id treament_control avg_score
## 1:          1                0         2
## 2:          2                1         4
## 3:          3                0         3
```

Initial Analysis

Our initial analysis shows that for 1000 iterations of 50 subjects, the p-values peak closer to 0. This indicates that the probability for a false positive is low.

```
experiment <- function(data, num_subjects) {
    exp_data <- data[, .(avg_score = sample(avg_score, num_subjects,
        replace = TRUE)), by = treament_control]
    ten_t_test <- exp_data[, t.test(avg_score ~ treament_control)]
    p_value = ten_t_test$p.value
    return(p_value)
}
p_values <- replicate(n = 1000, expr = experiment(data = d, num_subjects = 50))
```

```
ggplot() + aes(p_values) + geom_histogram() + labs(title = "P-Value Distribution",
    x = "p-values")
```



Scenario 1.1 - 1.3: Different sample sizes

For scenario 1.1, we used different sample sizes to simulate our power. Our sample size ranged from 20 to 200, jumping by 20 subjects in each iteration. In scenario 1.1, we used the data as is with a 50/50 split for treatment and control. In scenario 1.2 we modified the data by decreasing the treatment effect for treated subjects so that the mean was shifted from 4 to 3.5. This is to simulate a scenario in which the expected treatment effect is not as much as we anticipated it would be. In scenario 1.3, we changed the treatment and control split from 50/50 to 80/20. In this scenario, we ended up with many more treated individuals responding to the survey compared to control subjects.

```
# Function to calculate power using t-test
power <- function(d, sample_size, t_split) {
    power_values <- rep(NA, length(sample_size))

    for (per in 1:length(sample_size)) {
        t_test_p_values <- rep(NA, 1000)
        for (i in 1:1000) {
            d_treat = d[d$treament_control == 1]
            d_cont = d[d$treament_control == 0]

            d_comb = rbind(d_treat[sample(nrow(d_treat), size = ceiling(sample_size[per] *
                (t_split)), replace = TRUE), ], d_cont[sample(nrow(d_cont),
                size = ceiling(size_to_sample_11[per] * (1 -
                  t_split)), replace = TRUE), ])

            t_test <- d_comb[, t.test(avg_score ~ treament_control)]

            t_test_p_values[i] = t_test$p.value
        }
        power_values[per] <- mean(t_test_p_values < 0.05)
    }
```

```
    return(power_values)
}
```

```
# Scenario #1.1: Different sample sizes, 50/50 split for
# control and treatment
size_to_sample_11 <- seq(20, 200, by = 20)
power_values_11 <- power(d, size_to_sample_11, 0.5)
```

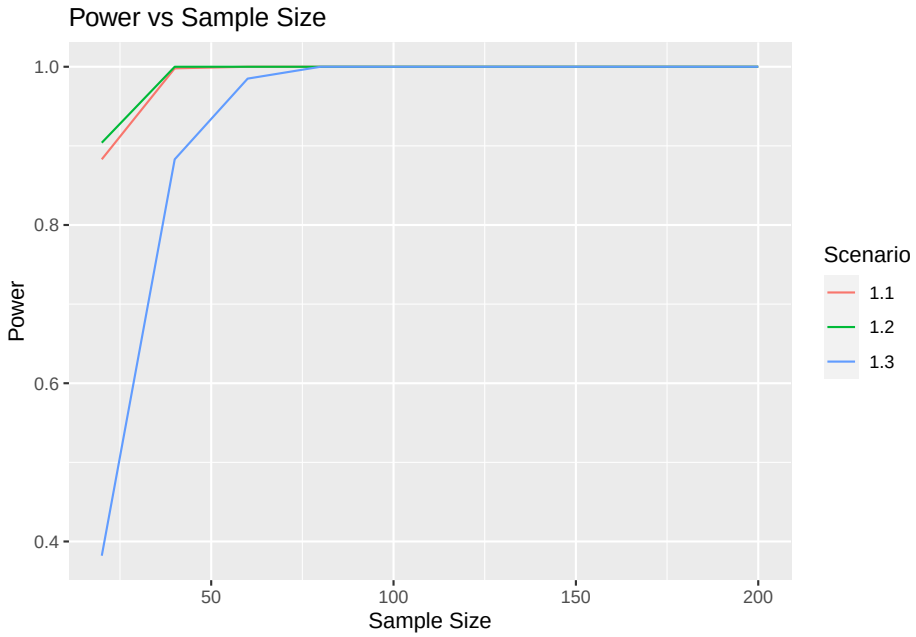
```
# Scenario #1.2: Different sample sizes, 50/50 split for
# control and treatment, decreased average of treatment
# effect
d2 <- d
d2$avg_score = d$avg_score - 0.05
size_to_sample_12 <- seq(20, 200, by = 20)
power_values_12 <- power(d2, size_to_sample_12, 0.5)
```

```
# Scenario #1.3: Different sample sizes, 20/80 split for
# control and treatment
size_to_sample_13 <- seq(20, 200, by = 20)
power_values_13 <- power(d, size_to_sample_13, 0.8)
```

```
# Creating data frames for each scenario for plotting
# purposes
df1 = data.frame(size_to_sample_11, power_values_11)
df2 = data.frame(size_to_sample_12, power_values_12)
df3 = data.frame(size_to_sample_13, power_values_13)
names(df1) <- c("sample_size", "power")
names(df2) <- c("sample_size", "power")
names(df3) <- c("sample_size", "power")

# Plot data
dat <- rbind(df1, df2, df3)
dat$grp <- rep(factor(1:3), times = c(nrow(df1), nrow(df2), nrow(df3)))

ggplot(data = dat, aes(x = sample_size, y = power, colour = grp)) +
    geom_line() + ggtitle("Power vs Sample Size") + labs(y = "Power",
    x = "Sample Size") + scale_color_discrete(name = "Scenario",
    labels = c(1.1, 1.2, 1.3))
```



According to our figure above, all three of the scenarios was able to achieve high power by sample size ~ 150. As expected, scenario 1.3 took the longest to achieve power because of small control group size.

Scenario 2

We further explored how varying split between control and treatment affected power when total number of subjects is held constant at 20. According to our figure below, power continues to increase with % in treatment until about 50%, at which it starts to drop off. Thus, we will aim for 50/50 split in responses from treatment and control in our experiment.

```
# Set Up Test function to return the P-Value Same sample
# sizes, different control and treatment split
sample_d_split <- function(data, num_subjects, split) {
    d_treat = data[data$treament_control == 1]
    d_cont = data[data$treament_control == 0]
    d_comb = rbind(d_treat[sample(nrow(d_treat), size = ceiling(num_subjects *
        split), replace = TRUE), ], d_cont[sample(nrow(d_cont),
        size = ceiling(num_subjects * (1 - split)), replace = TRUE),
        ])
    exp_data <- d_comb[, .(avg_score = sample(avg_score, num_subjects,
        replace = TRUE)), by = treament_control]
    t_test <- exp_data[, t.test(avg_score ~ treament_control)]
    return(t_test$p.value)
}
```

```
percentages_to_sample_v2 <- c()
num_split_v2 <- c()
sample_power_v2 <- c()
for (split in seq(from = 0.1, to = 0.9, by = 0.1)) {
    num_split_v2 <- c(num_split_v2, split)
    p_values_v2 <- replicate(1000, sample_d_split(d, 20, split))
    power_v2 <- mean(p_values_v2 < 0.05)
```

```
    sample_power_v2 <- c(sample_power_v2, power_v2)
}
```

```
ggplot() + aes(x=num_split_v2, y=sample_power_v2) +
  labs( x = "Control / Treatment Split", y = "Power", title ="Power Analysis based on varying split in c
  geom_line(color = 'red') +
  theme_bw() + # has to be before axis text manipulations because disables their effect otherwise
  theme(axis.text.x = element_text(angle = 0, hjust=1),text = element_text(size=8))+
  scale_x_continuous("% in Treatment", labels = as.character(num_split_v2), breaks = num_split_v2)
```