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. We decided to do this because. . . . . NEED TO FINISH THIS JUSTIFICATION

```r
library(data.table)
library(magrittr)

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

Scenario 1: Different sample sizes

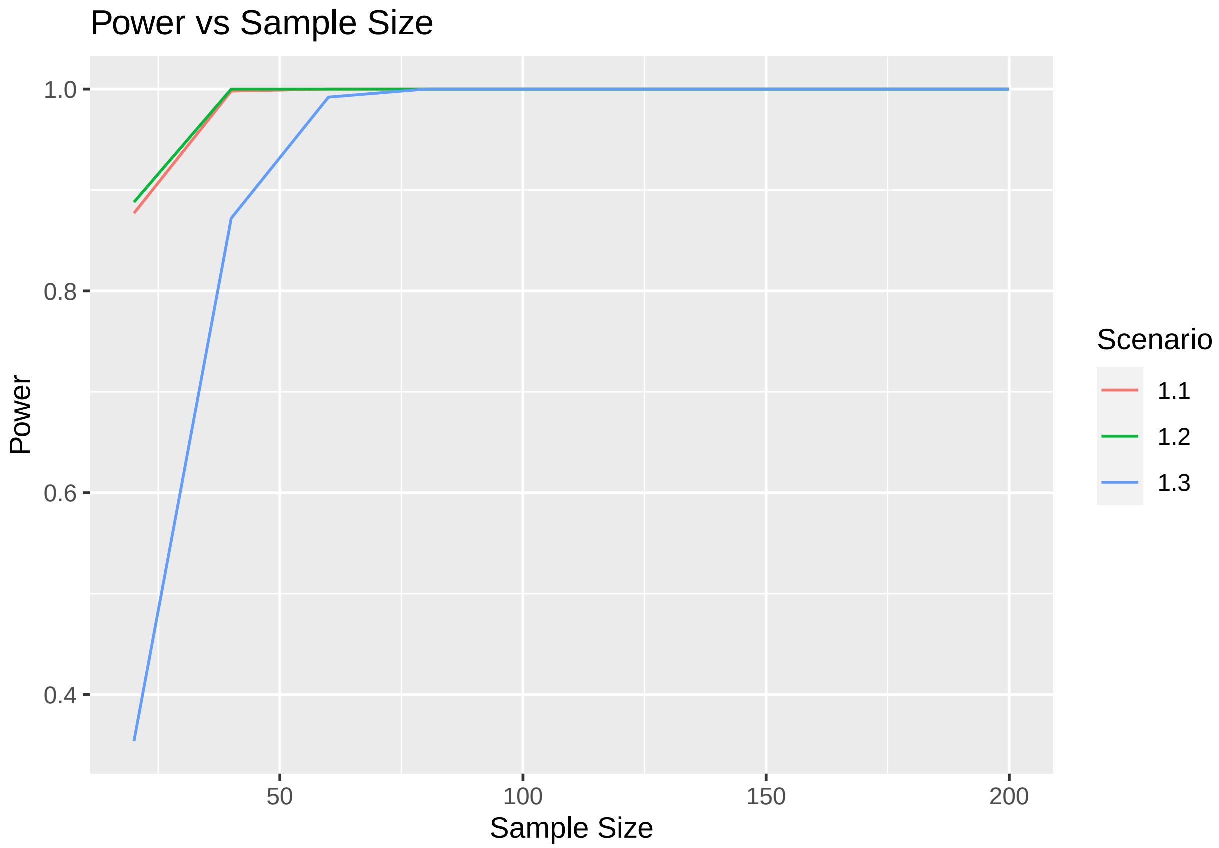
For scenario 1, we used different sample sizes to simulate our power. Our sample size ranged from 20 to 200, jumping by 20 subjects in each iteration. In scenario 1.1, we used the data as is with a 50/50 split for treatment and control. In scenario 1.2 we modified the data by decreasing the treatment 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.

```r
# 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 <- rep(NA, length(size_to_sample_11))

for (per in 1:length(size_to_sample_11)) {
    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 = size_to_sample_11[per]/2,
            replace = TRUE), ], d_cont[sample(nrow(d_cont), size = size_to_sample_11[per]/2,
            replace = TRUE), ])
        t_test_ten_people <- d_comb[, t.test(avg_score ~ treament_control)]

        t_test_p_values[i] = t_test_ten_people$p.value
    }
    power_values_11[per] <- mean(t_test_p_values < 0.05)
}
```

```
# Scenario #1.2: Different sample sizes, 50/50 split for
# control and treatment, decreased average of treatment
# effect

d$avg_score_dec = d$avg_score - 0.05

size_to_sample_12 <- seq(20, 200, by = 20)

power_values_12 <- rep(NA, length(size_to_sample_12))

for (per in 1:length(size_to_sample_12)) {
    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 = size_to_sample_12[per]/2,
            replace = TRUE), ], d_cont[sample(nrow(d_cont), size = size_to_sample_12[per]/2,
            replace = TRUE), ])

        t_test_ten_people <- d_comb[, t.test(avg_score_dec ~
            treament_control)]

        t_test_p_values[i] = t_test_ten_people$p.value
    }
    power_values_12[per] <- mean(t_test_p_values < 0.05)
}
```

```
# 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 <- rep(NA, length(size_to_sample_13))

for (per in 1:length(size_to_sample_13)) {
    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(size_to_sample_13[per] *
            0.8), replace = TRUE), ], d_cont[sample(nrow(d_cont),
            size = ceiling(size_to_sample_13[per] * 0.2), replace = TRUE),
            ])

        t_test_ten_people <- d_comb[, t.test(avg_score ~ treament_control)]
        t_test_p_values[i] = t_test_ten_people$p.value
    }
    power_values_13[per] <- mean(t_test_p_values < 0.05)
}
```

```
# 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.