Lab07

Jeffrey Wong

1/13/2020

```
setwd("/Users/jw-mba/Desktop/r-projects")
```

```
#source("GCP_local_IP_connection_setup.R")
#write.csv(first_300, "first_300.csv", row.names = F)
```

```
ag_data <- read.csv("ag_data.csv", stringsAsFactors = F)
airports <- read.csv("airports.csv", stringsAsFactors = F)
test_scores <- read.csv("test_scores.csv", stringsAsFactors = F)
first_100 <- read.csv("first_100.csv", stringsAsFactors = F)
first_300 <- read.csv("first_300.csv", stringsAsFactors = F)
```

1. Using the ag data dataset, calculate the year-to-year change in 'Food Availability per capita' for Total Grains and Cereals and Root Crops for Vietnam and Cambodia.

```
ag_data01 <- ag_data[(ag_data$Country == "Vietnam" | ag_data$Country == "Cambodia") &
                       ag_data$Commodity == "Total Grains/Cereals and Root Crops (R&T)"  &
                       ag_data$Item == "Food Availability per capita",
                     c(1,5,6)]
colnames(ag_data01)[3] <- "Food Availability per capita"

library(reshape2)
ag_data02 <- dcast(ag_data01, Year ~ Country, value.var = "Food Availability per capita")

df_yoy <- data.frame(year = ag_data02$Year[-1],
                     Cambodia = diff(ag_data02$Cambodia),
                     Vietnam = diff(ag_data02$Vietnam) )
```

   a. What is the effect size of the difference between the two countries?

```
pooled_sd <- sqrt((sd(df_yoy$Cambodia)^2 + sd(df_yoy$Vietnam)^2)/2)
h <- (mean(df_yoy$Cambodia) - mean(df_yoy$Vietnam))/pooled_sd
h
```

```
## [1] 0.1304192
```

   b. Interpret the results. How confident are you that there is an actual difference between the two countries and how their food availability per capita changed over time?

- an effect size of 0.13 means that whatever the distinction there is between these 2 countries, its effect is 0.13 standard deviation difference in outcome.

- However, the t-test of these 2 countries indiciates faily likely that there's no statistical difference b/w the two sets of data because the p-value is quite high, the confident level of an actual difference is only 40%.

```
t.test(df_yoy$Cambodia, df_yoy$Vietnam)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_yoy$Cambodia and df_yoy$Vietnam
## t = 0.52977, df = 54.701, p-value = 0.5984
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.892314  3.252048
## sample estimates:
## mean of x mean of y
##  2.245079  1.565212
```

c. To observe the same effect size at a 90% confidence level, how many *more* years of data would be needed for each country? (Assume alpha = 0.05 and power = 0.8.)

- in order to have 90% confidence level, we need to have 694 more years of data for each country

```
library(pwr)
pwr.2p.test(h = 0.1304192, power = 0.8, sig.level = 0.1, alternative = "two.sided")
```

```
##
##      Difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.1304192
##               n = 726.93
##       sig.level = 0.1
##           power = 0.8
##     alternative = two.sided
##
## NOTE: same sample sizes
```

```
727-33
```

```
## [1] 694
```

2. I am interested in conducting an experiment on 10% of the airports in the dataset. The expected effect size is 0.4.

a. What level of power can I expect from this experiment? Interpret this result.

- Power is 1, it means that we can definitely have true significant finding since there's no chnace to have false negative.

```
pwr.2p2n.test(h = 0.4, n1 = 7698, n2 = 770, alternative = "two.sided")
```

```
##
##      difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.4
##              n1 = 7698
##              n2 = 770
##       sig.level = 0.05
##           power = 1
##     alternative = two.sided
##
## NOTE: different sample sizes
```

b. Given that the US has far more airports than any other country, I would like the proportion of US/non-US airports to be the same in treatment group as the overall population. How will you construct your treatment group of airports?

- United States has 1512 airports in the datasets, which is 19.6% of the pool. If we conduct an experiment on 10% of the airports, we need to make sure it contains 151 airports in United States in the treatment group.

```
sum(airports$country == "United States")
```

```
## [1] 1512
```

```
sum(airports$country == "United States")/length(airports$country)
```

```
## [1] 0.1964147
```

```
770
```

```
## [1] 770
```

3. The school district is experimenting between two different student-level interventions for 5th grade math test scores: one at Wiggs, the other at Charles Middle. The 'Benchmark' test will be used to evaluate. The district needs a dashboard they can use to monitor the progress of the experiment. The dashboard needs to:

Plot the distribution of the two schools' scores and the mean of the scores Display the number of observations, the level of power of the experiment, and how many more observations from each school will be necessary to achieve 80% power, at which point, the experiment will be concluded. Note of all the 5th graders in the two schools, 64% attend Wiggs, 36% attend Charles Middle.

a. Given the results of your dashboard, what is your advice to the district at this point?

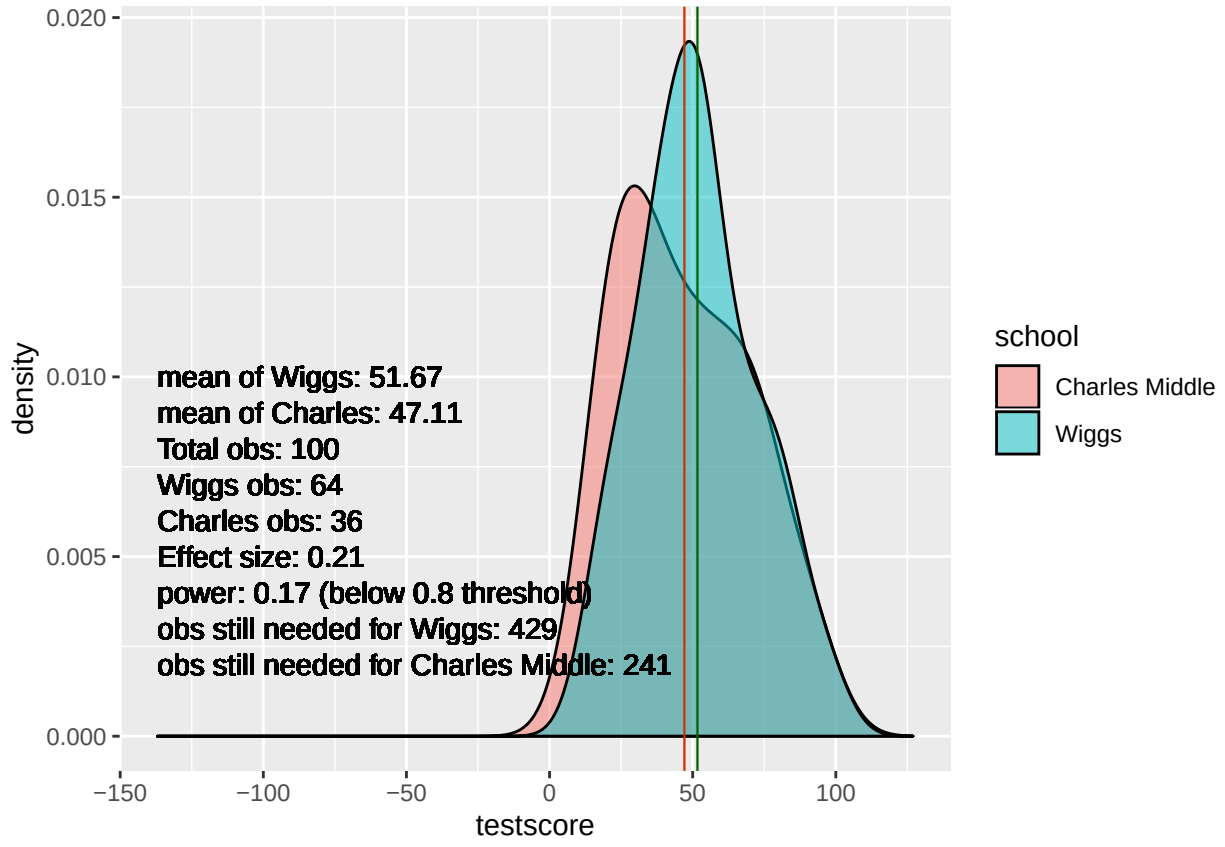
- with 100 obs of the 2 schools, the power is 0.17. In order to have 0.8 power, or 80% chance of finding true positive results between these school, we need to have a bigger sample of observations, 429 more from Wiggs and 241 more from Charles Middle. If no such amount of data can be obtained at the same period of time then a longer horizon time will be required to collect enough information to get a power of 0.8.

```r
total_size <- test_scores[test_scores$gradelevel == 5 &
                          test_scores$testname == "Benchmark" &
                          test_scores$subject == "Math" &
                          (test_scores$school == "Wiggs" | test_scores$school == "Charles Middle"),
                           c(5,10)]
table(total_size$school == "Wiggs")[2]/ length(total_size$school)
```

```
##      TRUE
## 0.6377778
```

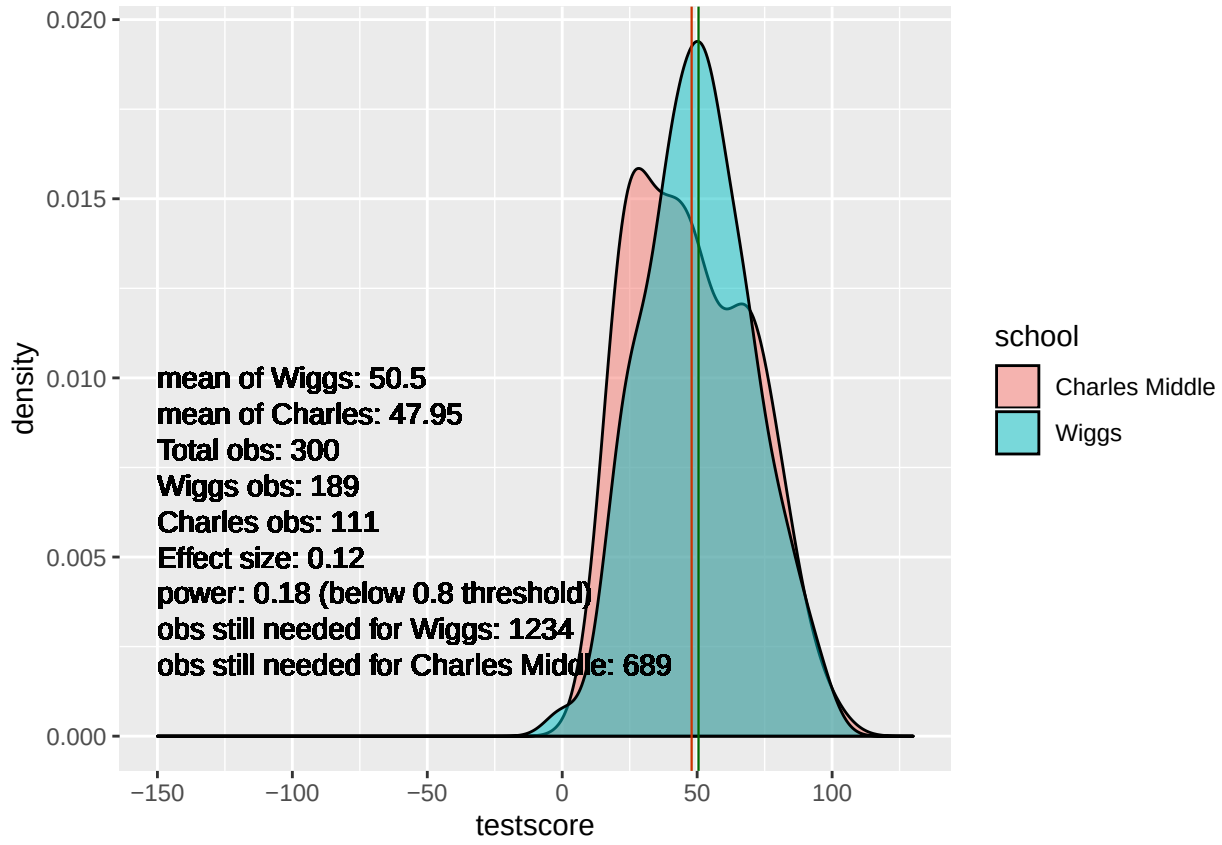
```r
experiment <- function(data) {
  w <- data[data$school == "Wiggs", 2]
  c <- data[data$school == "Charles Middle", 2]
  h <- abs((mean(w) - mean(c))/sqrt((sd(w)^2 + sd(c)^2)/2))
  n_w <- nrow(data[data$school == "Wiggs",])
  n_c <- nrow(data[data$school == "Charles Middle",])
  pwr <- pwr.2p2n.test(h = h, n1 = n_w, n2 = n_c)
  low <- min(data$testscore)
  high <- max(data$testscore)

  # labels
  w_mean_label <- paste("mean of Wiggs:", round(mean(w), digits = 2))
  c_mean_label <- paste("mean of Charles:", round(mean(c), digits = 2))
  power_label <- ifelse(pwr$power < 0.8,
                        paste("power:", round(pwr$power, digits = 2), "(below 0.8 threshold)"),
                        paste("power:", round(pwr$power, digits = 2), "(threshold met)"))
  #sample size required to achieve power = 0.8, given 64% attend Wiggs and 36% attend Charles
  i <- 2
  while(pwr.2p2n.test(h = h, n1 = i, n2 = i*0.64/0.36 ) $ power < .80) {i <- i + 1}
  req_pwr <- pwr.2p2n.test(h = h, n1 = i, n2 = i*0.64/0.36)
  obs_needed_w <- paste("obs still needed for Wiggs:", ceiling(req_pwr$n2 - n_w))
  obs_needed_c <- paste("obs still needed for Charles Middle:", ceiling(req_pwr$n1 - n_c))

  #plotting
  library(ggplot2)
  ggplot(data, aes(x = testscore, fill = school)) +
    geom_density(alpha = 0.5) +
    xlim(low - 150, high + 30) +
    geom_vline(xintercept = c(mean(w), mean(c)),
               color = c("#006600", "#CC3300"),
               size = 0.4) +
    geom_text(aes(low-150, 0.01, label = w_mean_label), hjust = 0) +
    geom_text(aes(low-150, 0.01-0.001, label = c_mean_label),  hjust = 0) +
    geom_text(aes(low-150, 0.01-0.002, label = paste("Total obs:", nrow(data))),  hjust = 0) +
    geom_text(aes(low-150, 0.01-0.003, label = paste("Wiggs obs:", n_w)),  hjust = 0) +
    geom_text(aes(low-150, 0.01-0.004, label = paste("Charles obs:", n_c)),  hjust = 0) +
    geom_text(aes(low-150, 0.01-0.005, label = paste("Effect size:", round(h, digits = 2))),
              hjust = 0) +
    geom_text(aes(low-150, 0.01-0.006, label = power_label),  hjust = 0) +
    geom_text(aes(low-150, 0.01-0.007, label = obs_needed_w),  hjust = 0) +
    geom_text(aes(low-150, 0.01-0.008, label = obs_needed_c),  hjust = 0)

}
```

```
experiment(first_100)
```



b. Provide an updated dashboard with more observations using this query:

```
experiment(first_300)
```


0.020
0.015
0.010
0.005
0.000
density
mean of Wiggs: 50.5
mean of Charles: 47.95
Total obs: 300
Wiggs obs: 189
Charles obs: 111
Effect size: 0.12
power: 0.18 (below 0.8 threshold)
obs still needed for Wiggs: 1234
obs still needed for Charles Middle: 689
−150
−100
−50
0
50
100
testscore
school
Charles Middle
Wiggs