Project Power Analysis

Oscar Garcia, Alejandro Reskala, Jessica Stockham

02/21/2023

```
library(data.table)
library(magrittr)
library(ggplot2)

knitr::opts_chunk$set(dpi = 300)

set.seed(3)
```

The project is "A Field Experiment on eBay: item prices versus shipping and handling: are they equivalent in bidders' minds?"

Data

We start with data from the referenced paper (https://www.researchgate.net/publication/4985612_Plus_Shipping_and_Handling_Revenue_Non_Equivalence_in_Field_Experiments_on_eBay)for exploration purposes.

```
d_lowr <- fread('./data_AB.csv')
d_lowr[, differences:= RevB - RevA]
d_lowr
```

```
##     Type Item  RevA  RevB differences
##  1:   CD    1  5.50  7.24        1.74
##  2:   CD    2  6.50  7.74        1.24
##  3:   CD    3  8.50 10.49        1.99
##  4:   CD    4 12.50 11.99       -0.51
##  5:   CD    5 11.00 15.99        4.99
##  6:   CD    6 13.50 14.99        1.49
##  7:   CD    7  0.00  9.99        9.99
##  8:   CD    8  7.28  9.49        2.21
##  9:   CD    9  6.07  8.25        2.18
## 10:   CD   10  4.50  5.24        0.74
## 11:   XB   11 34.05 41.24        7.19
## 12:   XB   12 44.01 33.99      -10.02
## 13:   XB   13 40.99 39.99       -1.00
## 14:   XB   14 36.01 36.99        0.98
## 15:   XB   15 41.00 32.99       -8.01
## 16:   XB   16 37.00 38.12        1.12
## 17:   XB   17 42.12 42.99        0.87
## 18:   XB   18 26.00 33.99        7.99
```

```
## 19:   XB   19 36.00 37.00        1.00
## 20:   XB   29 33.99 40.99        7.00
```

```
d_lowr[ , .(group_mean = mean(RevA), group_sd = sd(RevA)), keyby = .(Type)]
```

```
##    Type group_mean group_sd
## 1:   CD      7.535 4.035689
## 2:   XB     37.117 5.248073
```

```
d_lowr[ , .(group_mean = mean(RevB), group_sd = sd(RevB)), keyby = .(Type)]
```

```
##    Type group_mean group_sd
## 1:   CD     10.141 3.390953
## 2:   XB     37.829 3.452862
```

```
d_lowr[ , .(group_mean = mean(differences), group_sd = sd(differences)), keyby = .(Type)]
```

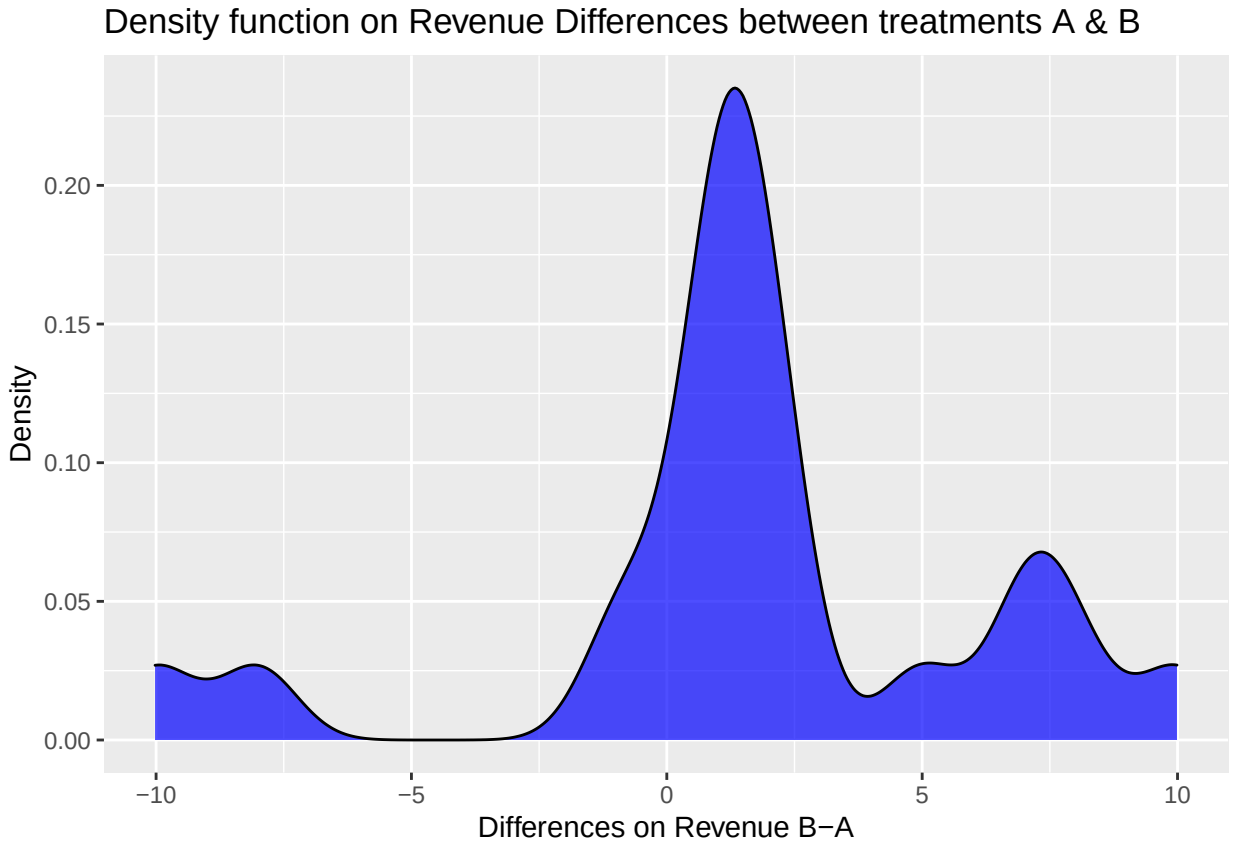
```
##    Type group_mean group_sd
## 1:   CD      2.606 2.943495
## 2:   XB      0.712 6.046109
```

Assumptions and preliminary tests

We have paired dependent samples and metric scale (revenue), so applicable tests are: - Paired t-test - Wilcoxon Signed rank test

In order to choose one, we test difference between Treatments A and B. If the difference is approximately symmetric around the mean, we choose Wilcoxon. If the difference looks approximately normal, we can use paired t-test.

```
p <- ggplot(d_lowr, aes(x=differences)) +
  geom_density(alpha = 0.7, fill = "blue") +
  ggtitle("Density function on Revenue Differences between treatments A & B") +
  ylab("Density") +
  xlab("Differences on Revenue B-A")
p
```



Form the plot above, which looks reasonably symmetrical around the mean but not quite normal, we decide to use Wilcoxon test.

```
Wilcox <- d_lowr[ , wilcox.test(d_lowr$RevA, d_lowr$RevB, paired=TRUE)]
```

```
## Warning in wilcox.test.default(d_lowr$RevA, d_lowr$RevB, paired = TRUE): cannot
## compute exact p-value with ties
```

```
str(Wilcox)
```

```
## List of 7
##  $ statistic  : Named num 44.5
##   ..- attr(*, "names")= chr "V"
##  $ parameter  : NULL
##  $ p.value    : num 0.0251
##  $ null.value : Named num 0
##   ..- attr(*, "names")= chr "location shift"
##  $ alternative: chr "two.sided"
##  $ method     : chr "Wilcoxon signed rank test with continuity correction"
##  $ data.name  : chr "d_lowr$RevA and d_lowr$RevB"
##  - attr(*, "class")= chr "htest"
```

```
Wilcox
```

```
##
```

```
##  Wilcoxon signed rank test with continuity correction
##
## data:  d_lowr$RevA and d_lowr$RevB
## V = 44.5, p-value = 0.02508
## alternative hypothesis: true location shift is not equal to 0
```

p-value 0.0250808 is lower than 5% so we reject the null hypothesis that the mean difference in revenues between paired treatments is zero, for the data in the paper.

Let's use a binomial statistic to test the null hypothesis against the one-sided alternative that B outperforms A.

```
bcount <- length(which(d_lowr$RevB>d_lowr$RevA))
btest <- binom.test(bcount, nrow(d_lowr), 0.5, alternative='greater')
btest
```

```
##
##  Exact binomial test
##
## data:  bcount and nrow(d_lowr)
## number of successes = 16, number of trials = 20, p-value = 0.005909
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5989719 1.0000000
## sample estimates:
## probability of success
##                    0.8
```

In this example, B outperforms A 16 of 20 times.

Since the p-value that we get is **0.005909**, we **reject the null hypothesis**. We do not have enough evidence to say that the revenue of treatment A is the same as the revenue of treatment B under same total reserve price for the low reserve prices group.

Power Analysis simulation

We have two separate populations under treatments A and B - one lower-priced item, and a higher-priced item (in the case of the paper, they used music CDs and Xbox games, respectively). In our case we plan to replace music CDs with Blue Rays (movies and TV shows), because they are more adapted to the current market in 2023. We will still use video games. In the following, we will calculate the estimated power for each of these two items.

Item type 1: Blue Rays

We assume mean revenues under treatment A of 20 dollars (we picked this value after looking at current eBay auctions for popular new blue ray auctions) and sd of 4 dollars (similar to the paper); under treatment B, mean revenues are approximately 20% higher (percentage taken from the music CDs experiment in the paper referenced, excluding unsold copies), with approximately same sd of 4, similar to the paper.

```
power_wilcox_test_1 <- function(
  mean_A = 20,
  mean_B = 24,
```

```
  sd_A = 4,
  sd_B = 4,
  number_per_condition = 10,
  power_loops = 100,
  ri_loops = 100,
  verbose = TRUE) {

    p_values <- NA
    ri <- NA
    d <- data.table()
    NUM_ITEMS <- 10
    d[ , Item := seq(1:NUM_ITEMS)]

    for(power_loop in 1:power_loops) {
      if(verbose == TRUE) {
        if(power_loop %% 10 == 0) {
          cat(sprintf('Loop Number: %.0f\n', power_loop))
        }
      }

      # d[, RevA := rnorm(.N, mean = mean_A, sd = sd_A)]
      # d[, RevB := rnorm(.N, mean = mean_B, sd = sd_B)]
      #
      # ate <- mean(d$RevB) - mean(d$RevA)
      #
      # for(ri_loop in 1:100) {
      #   d[, RevB := rnorm(.N, mean = mean_A, sd = sd_A)]
      #   ri[ri_loop] <- mean(d$RevB) - mean(d$RevA)
      # }
      #
      # p_values[power_loop] <- mean(abs(ri) > abs(ate))

      p_values[power_loop] <- wilcox.test(
        x = rnorm(number_per_condition, mean = mean_A, sd = sd_A),
        y = rnorm(number_per_condition, mean = mean_B, sd = sd_B),
        paired=TRUE
      )$p.value
    }

    return(list(
      'p_values' = p_values,
      'power' = mean(p_values < 0.05)
      ))
}
```

```
power_wilcox_test_1()$power
```

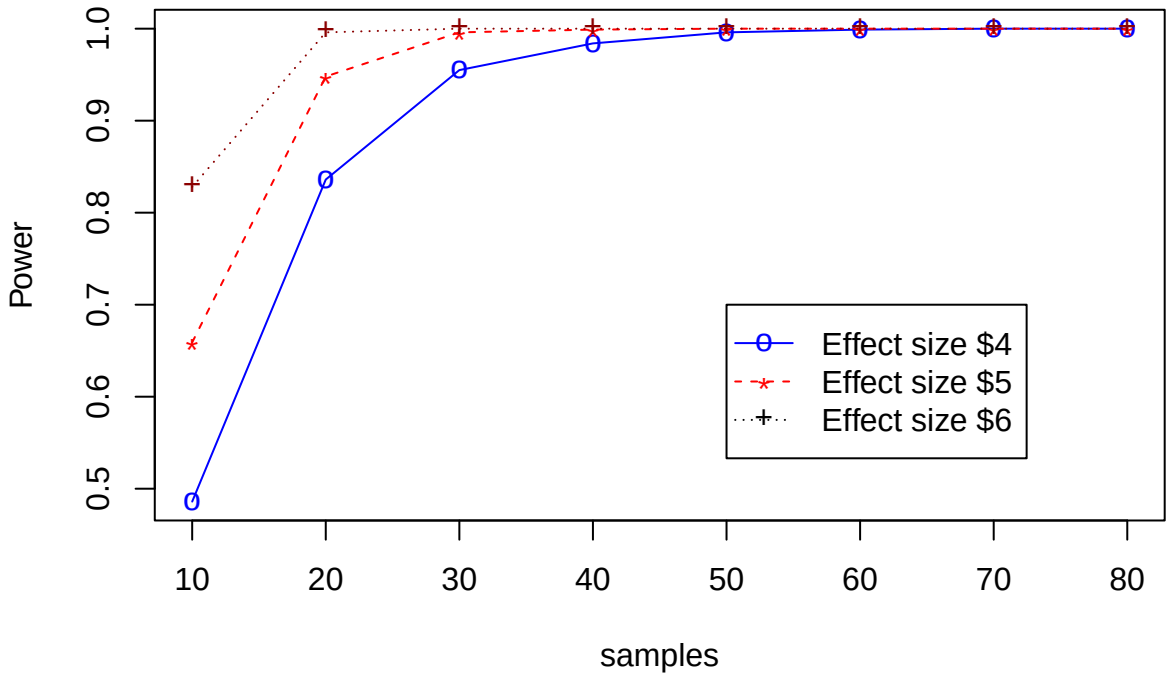
```
## Loop Number: 10
## Loop Number: 20
## Loop Number: 30
## Loop Number: 40
## Loop Number: 50
## Loop Number: 60
```

```
## Loop Number: 70
## Loop Number: 80
## Loop Number: 90
## Loop Number: 100
```

```
## [1] 0.52
```

Scenarios: increase assumed treatment effect size Treatment effect size starts at 20% (similar to the paper for music CDs as explained above), and we increase it from there.

```
# Increasing sample size
samples <- c(10, 20, 30, 40, 50, 60, 70, 80)

size_power_1 <- NA
size_power_2 <- NA
size_power_3 <- NA

for(i in 1:length(samples)) {
  size_power_1[i] <- power_wilcox_test_1(
    mean_A = 20, mean_B = 24,
    power_loops = 1000, verbose = FALSE,
    number_per_condition = samples[i]
    )$power

  size_power_2[i] <- power_wilcox_test_1(
  mean_A = 20, mean_B = 25,
  power_loops = 1000, verbose = FALSE,
  number_per_condition = samples[i]
  )$power

  size_power_3[i] <- power_wilcox_test_1(
  mean_A = 20, mean_B = 26,
  power_loops = 1000, verbose = FALSE,
  number_per_condition = samples[i]
  )$power

}

plot(x = samples, y = size_power_1, type="o", col="blue", pch="o", lty=1, ylab="Power")

points(x = samples, y = size_power_2, col="red", pch="*")
lines(x = samples, y = size_power_2, col="red",lty=2)

points(x = samples, y = size_power_3, col="dark red",pch="+")
lines(x = samples, y = size_power_3, col="dark red", lty=3)

legend(50,0.7,legend=c("Effect size $4","Effect size $5","Effect size $6"),
       col=c("blue","red","black"),
       pch=c("o","*","+"),lty=c(1,2,3), ncol=1)
```



Item type 2: Video Games

We assume mean revenues under treatment A of 50 dollars (we picked this value after looking at current eBay auctions for popular video games auctions) and sd of 4 dollars (similar to the paper); under treatment B, mean revenues are approximately 2% higher (percentage taken from the Xbox games experiment in the paper referenced, with approximately same sd of 4, similar to the paper.

```
power_wilcox_test_2 <- function(
  mean_A = 50,
  mean_B = 51,
  sd_A = 4,
  sd_B = 4,
  number_per_condition = 10,
  power_loops = 100,
  ri_loops = 100,
  verbose = TRUE) {

    p_values <- NA
    ri <- NA
    d <- data.table()
    NUM_ITEMS <- 10
    d[ , Item := seq(1:NUM_ITEMS)]

    for(power_loop in 1:power_loops) {
      if(verbose == TRUE) {
        if(power_loop %% 10 == 0) {
          cat(sprintf('Loop Number: %.0f\n', power_loop))
        }
      }

      # d[, RevA := rnorm(.N, mean = mean_A, sd = sd_A)]
```

```
      # d[, RevB := rnorm(.N, mean = mean_B, sd = sd_B)]
      #
      # ate <- mean(d$RevB) - mean(d$RevA)
      #
      # for(ri_loop in 1:100) {
      #   d[, RevB := rnorm(.N, mean = mean_A, sd = sd_A)]
      #   ri[ri_loop] <- mean(d$RevB) - mean(d$RevA)
      # }
      #
      # p_values[power_loop] <- mean(abs(ri) > abs(ate))

      p_values[power_loop] <- wilcox.test(
        x = rnorm(number_per_condition, mean = mean_A, sd = sd_A),
        y = rnorm(number_per_condition, mean = mean_B, sd = sd_B),
        paired=TRUE
      )$p.value
    }

    return(list(
      'p_values' = p_values,
      'power' = mean(p_values < 0.05)
      ))
}
```

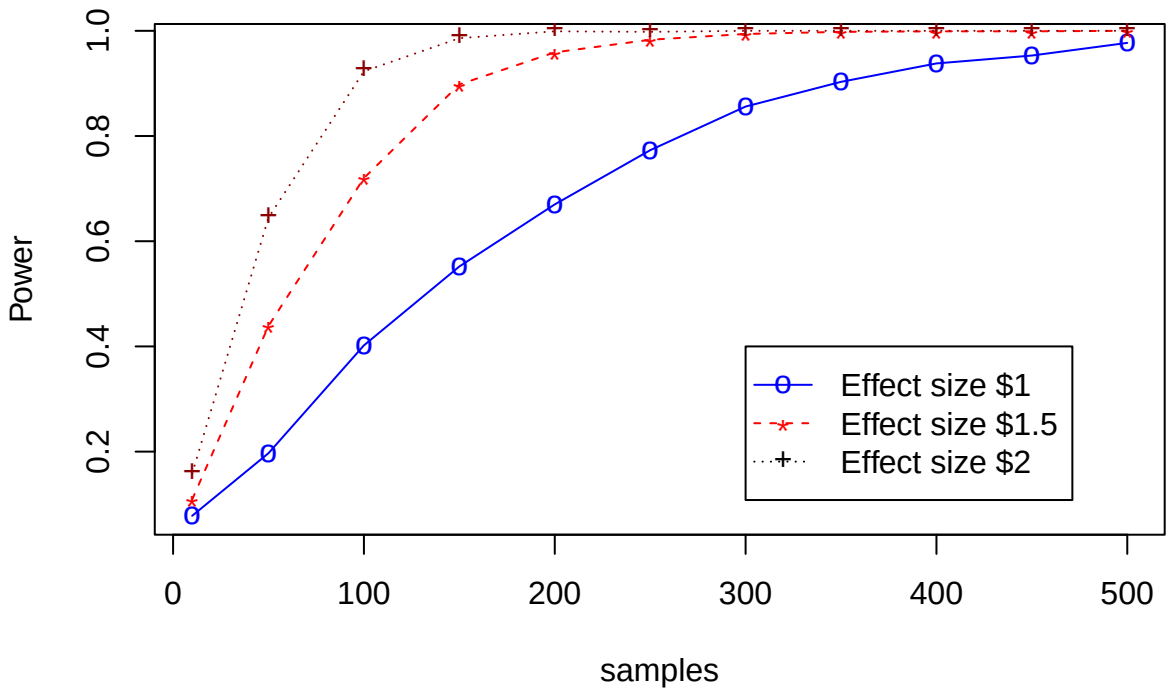
```
power_wilcox_test_2()$power
```

```
## Loop Number: 10
## Loop Number: 20
## Loop Number: 30
## Loop Number: 40
## Loop Number: 50
## Loop Number: 60
## Loop Number: 70
## Loop Number: 80
## Loop Number: 90
## Loop Number: 100

## [1] 0.02
```

Scenarios: increase assumed treatment effect size Treatment effect size starts at 2% (similar to the paper as explained above), and we increase it from there.

```
# Increasing sample size
samples <- c(10, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500)

size_power_1 <- NA
size_power_2 <- NA
size_power_3 <- NA

for(i in 1:length(samples)) {
  size_power_1[i] <- power_wilcox_test_2(
    mean_A = 50, mean_B = 51,
```

```
    power_loops = 1000, verbose = FALSE,
    number_per_condition = samples[i]
    )$power

  size_power_2[i] <- power_wilcox_test_2(
  mean_A = 50, mean_B = 51.5,
  power_loops = 1000, verbose = FALSE,
  number_per_condition = samples[i]
  )$power

  size_power_3[i] <- power_wilcox_test_2(
  mean_A = 50, mean_B = 52,
  power_loops = 1000, verbose = FALSE,
  number_per_condition = samples[i]
  )$power

}

plot(x = samples, y = size_power_1, type="o", col="blue", pch="o", lty=1, ylab="Power")

points(x = samples, y = size_power_2, col="red", pch="*")
lines(x = samples, y = size_power_2, col="red",lty=2)

points(x = samples, y = size_power_3, col="dark red",pch="+")
lines(x = samples, y = size_power_3, col="dark red", lty=3)

legend(300,0.4,legend=c("Effect size $1","Effect size $1.5","Effect size $2"),
      col=c("blue","red","black"),
      pch=c("o","*","+"),lty=c(1,2,3), ncol=1)
```



Conclusions

For the Blue Rays, we probably need a sample size of 20 to get a power close to 80% with the assumed treatment effect (the lowest scenario). For video games, even with the highest assumed treatment effect we would need a sample of around 70-80 samples, which is likely challenging to manage for our purposes. This leads us to re-evaluate the assumptions for this type of item to decide if we should find another type.