Power Analysis

Ange, Srila, Iris

10/03/2022

Install and load the necessary packages for this analysis.

```
library(data.table)
install.packages("ggplot2")
library(ggplot2)
install.packages("sandwich")
library(sandwich)
install.packages("lmtest")
library(lmtest)
install.packages("reshape2")
library(reshape2)
```

These are the sample sizes we will use to evaluate power.

```
n_sizes <- c(1:8)*25
```

First Scenario: Large Effect

Previous research shows that having participants actively interact with cybersecurity trainings, via games, helps improve their habits. In "A Game or Notes? The Use of a Customized Mobile Game to Improve Teenagers' Phishing Knowledge, Case of Tanzania" (2022)[1], the authors find that teenagers who interacted with a game designed to teach them about phishing performed better on a test about their phishing knowledge than the control group which only read the notes (45.1 percentage points performed better). Thus, we simulate our data to show that those in the treatment group, where they use active learning and interact with the training like they would a game, identify the messages more accurately than the control group does..

```
set.seed(1)

power_calc1 <- function(n){
  data <- data.table(id=1:n)
  data <- data[, group:=sample(0:1, size=.N, replace=T)]
  data <- data[group==0,correct:= as.integer(runif(.N, min = 1, max = 15))]
  data <- data[group==1,correct:= as.integer(runif(.N, min = 5, max = 15))]
  data <- data[,measurement:=correct/20]
  mod1 <- data[,lm(measurement~group)]
  mod1$vcovHC_ <- vcovHC(mod1)
  mod1_p_value <- coeftest(mod1, vcov. = mod1$vcovHC_)[8]
  return(mod1_p_value)
}

power_values1 <- c()
```

[1] https://doi.org/10.3390/jcp2030024

```
for (n in seq_along(n_sizes)){
  mod1_p_values <- replicate(n = 1000,
                             expr = power_calc1(n_sizes[n]))
  power <- mean(mod1_p_values<0.05)
  power_values1 <- c(power_values1,power)
}
```

Second Scenario: Blocking by Age

We suspect age may play a factor in the experimental results. In "Is This Phishing? Older Age Is Associated With Greater Difficulty Discriminating Between Safe and Malicious Emails" (2020) [2], they find that it seems like older people have some trouble when discriminating between safe and malicious emails; however, other research find that within a company, it is the younger people (18-24 years old) who are less likely to follow the cybersecurity policy [3]. Thus, we simulate our data to show that younger people who did not receive the treatment ("active" training) will identify the emails correctly the fewest, while those who were older and received the treatment will identify the emails correctly the most.

```
set.seed(1)

power_calc_scenario_2 <- function(n){
  data <- data.table(id = 1 : n)
  ### group 0 : control and group 1 : treatment
  data <- data[, group := sample(0:1, size = .N, replace = T)]
  treatment <- data[group==1,]
  control <- data[group==0,]
  #50% of treatment will be assigned 18-22
  treatment[sample(1:nrow(treatment),size=round(nrow(treatment)/2),0),
            age:=sample(18:22,size=.N, replace = T)]
  #the rest of treatment will be assigned 23-99 for age
  treatment[is.na(age),
            age:=sample(23:99,size=.N, replace = T)]
  # repeat for control
  #50% of control will be assigned 18-22
  control[sample(1:nrow(control),size=round(nrow(control)/2),0),
          age:=sample(18:22,size=.N, replace = T)]
  #the rest of control will be assigned 23-99 for age
  control[is.na(age),
          age:=sample(23:99,size=.N, replace = T)]
  data <- rbind(treatment,control)

  ### Scenarios:-
  ### young, control, age between 18-22
  ### old, control, age not between 18-22
  ### young, treatment, age between 18-22
  ### old, treatment, age not between 18-22
  young_control_correct_ans <- c(12, 14)
  young_treatment_correct_ans <- c(12, 18)
  old_control_correct_ans <- c(15, 18)
  old_treatment_correct_ans <- c(16, 20)

  ### Old and Treatment
  data <- data[group == 1 & age %in% c(23:99),
```

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8557838/
[3] https://www.techrepublic.com/article/why-genx-and-baby-boomers-are-the-most-cybersecure-employees/

```
              correct := as.integer(runif(.N,
                                          old_treatment_correct_ans[1],
                                          old_treatment_correct_ans[2]
                                          ))]
  ### Old and control
  data <- data[group == 0 & age %in% c(23:99),
              correct := as.integer(runif(.N,
                                          old_control_correct_ans[1],
                                          old_control_correct_ans[2]
                                          ))]
  ### Young and treatment
  data <- data[group == 1 & age %in% c(18:22),
              correct := as.integer(runif(.N,
                                          young_treatment_correct_ans[1],
                                          young_treatment_correct_ans[2]
                                          ))]
  ### Young and control
  data <- data[group == 0 & age %in% c(18:22),
              correct := as.integer(runif(.N,
                                          young_control_correct_ans[1],
                                          young_control_correct_ans[2]
                                          ))]

  ###Calculating measurement value
  data <- data[, measurement := correct / 20 ]
  #view(data)
  model_scenario_2 <- data[, lm(measurement ~ group + age)]
  #summary(model_scenario_2)
  model_scenario_2$vcovHC_ <- vcovHC(model_scenario_2)
  model_scenario_2_p_value <- coeftest(model_scenario_2,
                                       vcov. = model_scenario_2$vcovHC_)[11]
  # model_scenario_2_p_value
  return(model_scenario_2_p_value)
}

power_values_scenario_2_vector <- c()

for (sample_size in n_sizes){
  print(c("sample_size : ", sample_size))
  sceario_2_p_values <- replicate(n = 1000,
                                  expr = power_calc_scenario_2(sample_size)
                                  )
  calculated_power <- mean(sceario_2_p_values < 0.05)
  print(c("calculated_power : ",calculated_power))
  power_values_scenario_2_vector <- c(power_values_scenario_2_vector,
                                      calculated_power)
  }
```

```
## [1] "sample_size : " "25"
## [1] "calculated_power : " "0.746"
## [1] "sample_size : " "50"
## [1] "calculated_power : " "0.98"
## [1] "sample_size : " "75"
## [1] "calculated_power : " "0.999"
```

```
## [1] "sample_size : " "100"
## [1] "calculated_power : " "1"
## [1] "sample_size : " "125"
## [1] "calculated_power : " "1"
## [1] "sample_size : " "150"
## [1] "calculated_power : " "1"
## [1] "sample_size : " "175"
## [1] "calculated_power : " "1"
## [1] "sample_size : " "200"
## [1] "calculated_power : " "1"
```

Third Scenario: Almost Everyone is Correct

If people have good cybersecurity hygiene and are familiar with how to identify threats, then they would be able to identify which emails are legitimate or phishing. In this scenario, we are providing both control and treatment groups with the cybersecurity training, so it is possible that there will be a heavy skew to the left, with most people correctly identifying 20/20 of the emails as phishing. We simulate our data with this in mind, but people in the treatment group will have slightly higher accuracy in identifying phishing emails from legitimate emails.

```
set.seed(1)

power_calc3 <- function(n){
  data <- data.table(id=1:n)
  data <- data[, group:=sample(0:1, size=.N, replace=T)]
  data <- data[group==0,correct:=round(runif(.N, min = 14, max = 23),0)]
  data <- data[group==1,correct:=round(runif(.N, min = 17, max = 23),0)]
  data <- data[correct>20,correct:=20]
  data <- data[,measurement:=correct/20]
  mod3 <- data[,lm(measurement~group)]
  mod3$vcovHC_ <- vcovHC(mod3)
  mod3_p_value <- coeftest(mod3, vcov. = mod3$vcovHC_)[8]
  return(mod3_p_value)
}

power_values3 <- c()
for (n in seq_along(n_sizes)){
  mod3_p_values <- replicate(n = 1000,
                             expr = power_calc3(n_sizes[n]))
  power <- mean(mod3_p_values<0.05)
  power_values3 <- c(power_values3,power)
}
```

Plot

```
plotdf <- data.frame(n_sizes,
                     power_values1,
                     power_values_scenario_2_vector,
                     power_values3)
melted <- reshape2::melt(plotdf,id=c("n_sizes"))

ggplot(data = melted,
       aes(x=n_sizes,
           y=value,
```

```
          group=variable,
          colour=variable)) +
geom_line() +
geom_point() +
theme_minimal() +
labs(title="Power Analysis",
     x="Sample Sizes",
     y="Power",
     color = "Scenario") +
theme(legend.position = "right") +
scale_color_manual(labels = c("1: Large Effect",
                              "2: Blocking by Age",
                              "3: Almost Everyone is Correct"),
                   values = c("red","blue","black"))
```