# Inference using Randomization and Bootstrapping

Mini-Assignment - MTH 361 A/B - Spring 2023

**Instructions:**

- Please provide complete solutions for each problem. If it involves mathematical computations, explanations, or analysis, please provide your reasoning or detailed solutions.
- Note that some problems have multiple solutions or ways to solve it. Make sure that your solutions are clear enough to showcase your work and understanding of the material.
- Creativity and collaborations are encouraged. Use all of the resources you have and what you need to complete the mini-assignment. Each student must take personal responsibility and submit their work individually. Please abide by the University of Portland Academic Honor Principle.
- **Please save your work as one pdf file, don't put your name in any part of the document, and submit it to the Teams Assignments for this course. Your document upload will correspond to your name automatically in Teams.**
- If you have questions or concerns, please feel free to ask the instructor.

**R Packages:**

```
library(tidyverse)
library(openintro)
library(infer)
```

# I. Randomization and Simulations for Inference

### Materials

- Randomization for Inference for One Proportion

  Given $p_0$ (proportion null value), $\hat{p}$ (sample proportion), $N$ (trials). You can simulate the null and sampling distributions using the *rbinom* function and compute the p-value and confidence interval. Below is an example code for bootstrapping for one proportion.

```
# variables
p_0 <- 0.50
p_hat <- 0.60
N <- 100
samples <- 1000

# the null and sampling distributions
oneprop_null_distribution <- rbinom(samples,N,p_0)
oneprop_sampling_distribution <- rbinom(samples,N,p_hat)
oneprop_df <- tibble(counts = c(oneprop_null_distribution,
                                oneprop_sampling_distribution),
                     proportions = counts/N,
                     label = rep(c("null","sampling"),each=samples))

# computing the p-value
oneprop_p_value <- oneprop_df %>%
  filter(label == "null",
         proportions > p_hat) %>%
  summarise(p_value = n()/samples)
oneprop_p_value
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.027
```

```
# computing the confidence interval
alpha <- 0.05
cl <- 1-alpha
z_star <- qnorm(cl + (alpha/2),0,1)
oneprop_confidence_interval <- oneprop_df %>%
  filter(label == "sampling") %>%
  summarise(standard_error = sd(proportions),
            lowerbound = p_hat - z_star*standard_error,
            upperbound = p_hat + z_star*standard_error)
oneprop_confidence_interval
```

```
## # A tibble: 1 x 3
##   standard_error lowerbound upperbound
##            <dbl>      <dbl>      <dbl>
## 1         0.0474      0.507      0.693
```

- Randomization for Two-Way Tables

  Given two categorical variables with multiple levels. You can compute the chi-squared statistic by using the method of randomization. Below is an example code using the `ask` dataset.

```
# summarized table of the data
table(ask %>% select(question_class,response))
```

```
##                 response
## question_class   disclose hide
##   general               2   71
##   neg_assumption       36   37
##   pos_assumption       23   50
```

```
set.seed(4747)

samples <- 1000

# compute the chi-square statistic using the data
ask_rand_obs <- ask %>%
  specify(response ~ question_class) %>%
  calculate(stat = "Chisq") %>%
  pull()

# perform randomization assuming the null hypothesis is true
ask_rand_dist <- ask %>%
  specify(response ~ question_class) %>%
  hypothesise(null = "independence") %>%
  generate(reps = samples, type = "permute") %>%
  calculate(stat = "Chisq")

# compute the pvalue
ask_pvalue = ask_rand_dist %>%
  filter(stat > ask_rand_obs) %>%
  summarise(p_value = n()/samples)
ask_pvalue
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

**Exercises**

1. **Malaria Vaccine Trial.** Perform an inference of two proportions using simulations. Use the `malaria` dataset to determine if there is a statistically significance difference between the proportion of infected individuals between the placebo and vaccine groups. Modify the example code given above. In addition, perform an inference for two-way tables using randomization using the same dataset.
2. (Outstanding Question) **Sex Discrimination on Promotions.** Perform an inference for two-way tables using randomization. Use the `sex_discrimination` dataset to determine if there is a statistically significant association between who gets promoted based on sex. Modify the example code given above.