Annand DSE5001 Final Exam

Joseph Annand

2023-06-28

# Libraries

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library("VennDiagram") # HW 5

## Loading required package: grid  
## Loading required package: futile.logger

library("statsr") # HW 7

## Warning: package 'statsr' was built under R version 4.3.1

## Loading required package: BayesFactor  
## Loading required package: coda

## Warning: package 'coda' was built under R version 4.3.1

## Loading required package: Matrix  
##   
## Attaching package: 'Matrix'  
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack  
##   
## \*\*\*\*\*\*\*\*\*\*\*\*  
## Welcome to BayesFactor 0.9.12-4.4. If you have questions, please contact Richard Morey (richarddmorey@gmail.com).  
##   
## Type BFManual() to open the manual.  
## \*\*\*\*\*\*\*\*\*\*\*\*

library("broom") # HW 7

## Warning: package 'broom' was built under R version 4.3.1

# Functions

# Check success-failure condition for Central Limit Theorem  
check\_clt <- function(n, p) {  
 return (n \* p >= 10 && n \* (1 - p) >= 10)  
}  
  
# Calculate standard error for point estimate  
find\_std\_error <- function(p, n) {  
 return (sqrt((p \* (1 - p)) / n))  
}  
  
# Perform hypothesis test using a p-value  
check\_hypothesis <- function(p, a) {  
 if (p < a) {  
 print("Reject null hypothesis")  
 } else {  
 print("Fail to reject null hypothesis")  
 }  
}

# Question 1

## Part A

The histogram shows us frequencies of the bins, or groups, of finishing times for male and female runners. The histogram also gives us insight to the mode, or the most common finishing times among winners. The box plot provides insight to where the outliers are in the data set of finishing times. The box plot also tells us where the median of the data set is.

## Part B

The bimodal distribution is the result of the data including finishing times for both male and female runners. The first prominent peak is likely the most common bin for male winner finishing times, which will tend to be faster than female winner times, and the second prominent peak is likely the most common bin for female winner finishing times.

## Part C

Men’s winner finishing times have a much narrower distribution than that of the women’s winner finishing times. Both may be right-skewed and have a long right tail due to the outliers that indicate times much slower than the rest of the data.

## Part D

The time series plot shows the year of each finishing time and allows us to see how finishing times for both men and women have changed or not changed from 1970-2000.

# Question 2

## Part A

The explanatory variable is the Percent with Bachelor’s degree and the response variable is the Per Capita Income.

## Part B

There is a strong, positive relationship between the two variables. It appears that the greater Percent of the population with a bachelor’s degree in a given county correlates with a higher per capita income in that county. On the left side of the graph, we see most of the data points form a large cluster, but on the upper right side of the graph, we see a few points that are far removed from the rest of the data. These points may be extreme cases or outliers.

## Part C

No, we cannot conclude that having a bachelor’s degree increases one’s income. This plot uses observational data, and a causal relationship cannot be determined without an experiment.

# Question 3

## Part A

slope <- (4.010-3.9983) / -(-0.0883)  
slope

## [1] 0.1325028

The slope is equal to 0.133.

## Part B

Yes, the data does provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive because the calculated slope is positive and the p-value for our slope is zero.

## Part C

The first diagnostic plot shows that there is constant variability across observations and that the variability of teaching evaluation does not change depending on the value of beauty, the explanatory variable. Also, the plot shows that points are fairly evenly distributed around the linear regression and no curvature is apparent, so the condition of linearity is met. The second diagnostic plot shows that the distribution of residuals is nearly normal, so the condition of nearly normal residuals is met. Finally, the third diagnostic plot shows that the order in which the data was collected does not affect the variability, so the condition of independent observations is met.

# Question 4

## Prepare

n\_children <- 194  
p\_hat <- 21 / 194  
alpha\_ <- 0.05

Null Hypothesis = Nearsightedness affects 8% of children. Alternative Hypothesis = Nearsightedness does not affect 8% of children.

## Check

p\_null = 0.08  
  
check\_clt(n\_children, p\_null)

## [1] TRUE

Central Limit Theorem holds because we pass success-failure condition and the observations are independent since they were collected through random sampling.

## Calculate

se\_ <- find\_std\_error(p\_null, n\_children)  
  
z\_score <- (p\_hat - p\_null) / se\_  
  
p\_value <- 2 \* pnorm(z\_score, lower.tail = FALSE)  
p\_value

## [1] 0.1469908

The p-value is 0.147.

## Conclude

check\_hypothesis(p\_value, alpha\_)

## [1] "Fail to reject null hypothesis"

Given that alpha is 0.05, the p-value is not less than the significance level, so we fail to reject our null hypothesis, meaning the data does not provide evidence that the 8% value is inaccurate.

# Question 5

## Part A

This distribution is called a sampling distribution.

## Part B

check\_clt(40, .16)

## [1] FALSE

# returns FALSE

We would expect the shape of the distribution to be right-skewed. We know that the distribution will not be symmetric because it does not meet for the success-failure condition, so either np or n(1-p) is small and the skew will be more noteworthy. Since the population proportion, 0.16 is close to zero, the distribution will be right skewed with a long, right tail.

## Part C

find\_std\_error(0.16, 40)

## [1] 0.05796551

The variability of the distribution is 0.058.

## Part D

The formal name for the value calculated in Part C is called the standard error.

## Part E

find\_std\_error(0.16, 90)

## [1] 0.03864367

The variability of the new distribution that collects 90 students per sample will be lower than that of when they collected 40 students per sample.

# Question 6

## Part A

(3/4) \* (3/4) \* (3/4) \* (3/4) \* (1/4)

## [1] 0.07910156

The probability the first question she gets right is the 5th question is 7.91%.

## Part B

(1/4) ^ 5

## [1] 0.0009765625

The probability that she gets all the questions right is 0.098%.

## Part C

1 - (3/4) ^ 5

## [1] 0.7626953

The probability that she gets at least one question right is 76.3%.

# Question 7

## Part A

qnorm(0.05, 4313, 583)

## [1] 3354.05

The cutoff for the the fastest 5% of athletes in the men’s group is 3354.05 seconds.

## Part B

qnorm(.90, 5261, 807)

## [1] 6295.212

The cutoff time for the slowest 10% of athletes in the women’s group is 6295.212 seconds.

# Question 8

## Part A

This study is a randomized experiment.

## Part B

Given the researchers reasonably include the four principles of randomized experiments; controlling, randomization, replication, and blocking; when necessary, the study can be used to conclude a causal relationship between stress and muscle cramps.

# Question 9

## Part A

pnorm(80, 72.6, 4.78)

## [1] 0.939203

93.9% of passenger vehicles travel slower than 80 mph.

## Part B

pnorm(80, 72.6, 4.78) - pnorm(60, 72.6, 4.78)

## [1] 0.9350083

93.5% of passenger vehicles travel between 60 and 80 mph.

## Part C

qnorm(0.95, 72.6, 4.78)

## [1] 80.4624

The fastest 5% of passenger vehicles travels 80.4624 mph or faster.

## Part D

pnorm(70, 72.6, 4.78, lower.tail = FALSE)

## [1] 0.7067562

70.68% of passenger vehicles drive over the speed limit on this stretch of the the I-5.

# Question 10

p\_b\_given\_a1 <- 0.50 # both girls given identical  
p\_a1 <- 0.30 # identical  
  
p\_b\_given\_a2 <- 0.25 # both girls given fraternal  
p\_a2 <- 0.70 # fraternal  
  
# Find P(identical | both girls) using Bayes Theorem  
(p\_b\_given\_a1 \* p\_a1) / ((p\_b\_given\_a1 \* p\_a1) + (p\_b\_given\_a2 \* p\_a2))

## [1] 0.4615385

Using Bayes Theorem, the probability that they are identical given that the twins are both girls is 46.15%.