Lab 8

Jack Rellamas

Remember, **follow the instructions below and use R Markdown to create a pdf document with your code and answers to the following questions on Gradescope.** You may find a template file by clicking "Code" in the top right corner of this page.

A. Bootstrapping the sampling distribution of the median

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
penguins <- palmerpenguins::penguins
```

1. Using the `penguins` dataset in the `palmerpenguins` package, construct a confidence interval for the mean `body_mass_g` for female Adelie penguins based on using a normal distribution based on the central limit theorem. You should compute the confidence interval without using `confint()`.

```
f_adelie <- filter(penguins, species == "Adelie", sex == "female")
mean_body_mass <- mean(f_adelie$body_mass_g)
z <- 1.96

# get sample standard dev.
sum <- 0
for (i in f_adelie$body_mass_g) {
  sum = sum + (i - mean_body_mass)**2
}

s <- (sum / (length(f_adelie$body_mass_g) - 1))**0.5
r <- z * s / ((length(f_adelie$body_mass_g)))**0.5
CI <- c(mean_body_mass - r,
        mean_body_mass + r)

print(CI)
```

```
## [1] 3307.040 3430.632
```

2. Construct a bootstrap confidence interval for the mean `body_mass_g` for female Adelie penguins using 10000 resamples.

```
set.seed(7)
peng_resample <- replicate(10000, mean(sample(f_adelie$body_mass_g, 73, replace = T)))
t.test(peng_resample)$conf.int
```

```
## [1] 3368.015 3369.248
## attr(,"conf.level")
## [1] 0.95
```

3. Construct a bootstrap confidence interval for the median `body_mass_g` for female Adelie penguins using 10000 resamples.

```
set.seed(7)
peng_resample <- replicate(10000, median(sample(f_adelie$body_mass_g, 73, replace = T)))
t.test(peng_resample)$conf.int
```

```
## [1] 3377.98 3379.81
## attr(,"conf.level")
## [1] 0.95
```

B. Simulations

4. Suppose that $Y \sim \text{Poisson}(X)$ where $X \sim \text{Exponential}(1)$. Use simulation to estimate $E(Y)$ and $\text{Var}(Y)$.

```
set.seed(20)
y <- replicate(10000, rpois(n = 1, lambda = rexp(1,1)))
e_y <- mean(y)

mystr <- paste("E(X):", e_y)
print(mystr)
```

```
## [1] "E(X): 0.9849"
```

```
e2 <- mean(y**2)
var_y <- e2 - e_y
mystr <- paste("Var(Y):", var_y)
print(mystr)
```

```
## [1] "Var(Y): 1.9396"
```

5. For this question, you will write a simulation to test the frequentist coverage of a 95% confidence interval for a proportion based on the normal approximation.

a. First, write a function that takes two inputs: `n` and `p`. Your function should randomly generate some $X \sim \text{Binomial}(n, p)$, compute $\widehat{p} = X/n$, and then compute the corresponding normal distribution-based confidence interval for p **based on your sample** $\widehat{p}$. Your function should return `TRUE` if p is in the confidence interval. You may use the following formula for the confidence interval:

$$\widehat{p} \pm z_{.975}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

```
in_CI <- function(n, p) {
  X <- rbinom(n = n, size = 1, p = p)
  S <- sum(X)
  phat <- S / n

  CI <- c(phat - qnorm(0.975) * ((phat * (1 - phat)) / n)**0.5,
          phat + qnorm(0.975) * ((phat * (1 - phat)) / n)**0.5)

  if (p >= CI[1] & p <= CI[2]) {
    return(T)
  }
  return(F)
}
```

b. Next, write a second function that takes three inputs: ‘n‘, ‘p‘, and ‘n_runs‘, representing the numb

```
proportion_in_CI <- function(n, p, n_runs) {
  runs <- replicate(n_runs, in_CI(n, p))
  return(sum(runs) / n_runs)
}
```

c. Test your function from (b) with ‘n = 20‘, ‘p = .5‘, and ‘n_runs = 1000‘.

```
set.seed(2017)
proportion_in_CI(n = 20, p = .5, n_runs = 1000)
```

```
## [1] 0.953
```

d. Use your simulation code to investigate the following questions: For what values of ‘n‘ and ‘p‘ is t

For what values of `n` and `p` is the frequentist coverage close to the expected 95% value?}

The frequentist coverage is close the the expected 0.95 value when n >= 5 and when p = 0.3, 0.5, 0.7.

For what values of `n` and `p` is the frequentist coverage very different to the expected 95% value?

It is very different when n < 5 and when p < 0.1 or p > 0.9.

C. Hypothesis Testing

Use the following code to obtain the Hawaiian Airlines and Alaska Airlines flights from the `nycflights13` package.

```r
library(tidyverse)
library(nycflights13)
data("flights")
flights_sample <- flights |>
  filter(carrier %in% c("HA", "AS"))
```

6. Compute a 95% confidence interval for the mean `arr_delay` for Alaska Airlines flights. Interpret your results.

```r
compute_CI <- function(vect) {
  xbar <- mean(vect)
  z <- 1.96

  # get sample standard dev.
  sum <- 0
  for (i in vect) {
    sum = sum + (i - xbar)**2
  }

  s <- (sum / (length(vect) - 1))**0.5
  r <- z * s / ((length(vect)))**0.5
  CI <- c(xbar - r,
          xbar + r)
  return(CI)
}
```

```r
alaska <- flights_sample %>%
  filter(carrier == "AS") %>%
  drop_na(arr_delay)
alaska_CI <- compute_CI(alaska$arr_delay)
alaska_CI
```

```
## [1] -12.616351  -7.245426
```

```r
# t.test(alaska$arr_delay)$conf.int
```

We can be 95% certain that the true mean delay of Alaska Airlines flights lies between -12.616351 and -7.245426.

7. Compute a 95% confidence interval for the mean `arr_delay` for Hawaiian Airlines flights. Interpret your results.

```r
hawaiian <- flights_sample %>%
  filter(carrier == "HA") %>%
  drop_na(arr_delay)
hawaii_CI <- compute_CI(hawaiian$arr_delay)
hawaii_CI
```

```
## [1] -14.877771   1.047361
```

We can be 95% certain that the true mean delay of Hawaiian Airlines flights lies between -14.877771 and 1.047361.

8. Compute a 95% confidence interval for the proportion of flights for which `arr_delay > 0` for Hawaiian Airlines flights. Interpret your results.

```
h_flights <- as.numeric(hawaiian$arr_delay > 0)
compute_CI(h_flights)
```

```
## [1] 0.2357824 0.3314691
```

We can be 95% certain that the true proportion of Hawaiian Airline flights with a delay lies between 0.2357824 and 0.3314691.

9. Consider the null hypothesis that the mean `arr_delay` for Alaska is equal to the mean `arr_delay` for Hawaiian and the alternative hypothesis that the mean `arr_delay` values are different for the two airlines. Perform an appropriate hypothesis test and interpret your results.

```
t.test(x = alaska$arr_delay, y = hawaiian$arr_delay)
```

```
##
##  Welch Two Sample t-test
##
## data:  alaska$arr_delay and hawaiian$arr_delay
## t = -0.70339, df = 420.37, p-value = 0.4822
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.443017   5.411649
## sample estimates:
## mean of x mean of y
## -9.930889 -6.915205
```

We do not have enough evidence to dismiss the null hypothesis that the mean arrival delay of Alaska Airlines is equal to the mean arrival delay of Hawaiian Airlines.

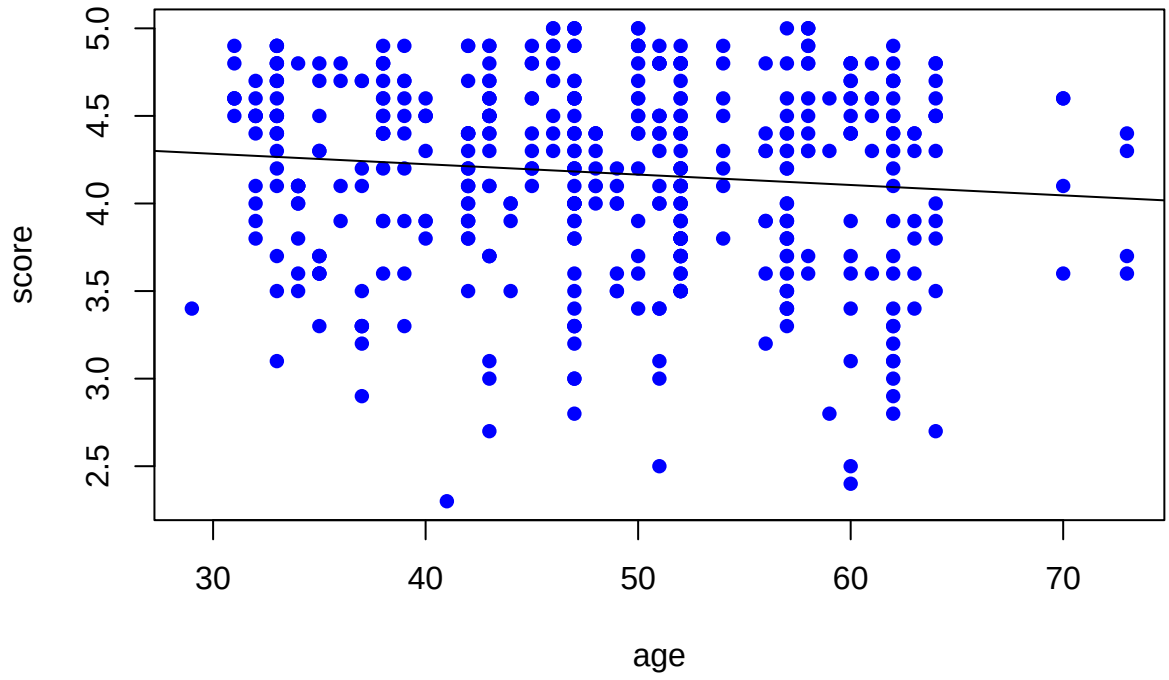
D. Linear Regression

Researchers at the University of Texas in Austin, Texas tried to figure out what causes differences in instructor teaching evaluation scores. Use the following code to load data on 463 courses. A full description of the data can be found here.

```
evals <- readr::read_csv("https://www.openintro.org/book/statdata/evals.csv")
```

10. Carry out a linear regression with `score` as the response variable and `age` as the single explanatory variable. Interpret your results.

```
plot(score ~ age, data = evals,
     pch  = 16, col = "blue") +
  abline(lm(evals$score ~ evals$age))
```




```
## integer(0)
```

```
summary(lm(score ~ age, data = evals))
```

```
##
## Call:
## lm(formula = score ~ age, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9185 -0.3531  0.1172  0.4172  0.8825
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.461932   0.126778  35.195   <2e-16 ***
## age         -0.005938   0.002569  -2.311   0.0213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5413 on 461 degrees of freedom
## Multiple R-squared:  0.01146,    Adjusted R-squared:  0.009311
## F-statistic: 5.342 on 1 and 461 DF,  p-value: 0.02125
```

With a slope of -0.005938 and a significance level of 0.0213, there there does seem to be a slight decrease in a teacher's score as age increases.

10. Extend your regression model by adding an additional explanatory variable. What happens to your results? Are the new p-values appropriate to use?

```
lm_res <- lm(score ~ age + bty_avg, data = evals)
summary(lm_res)
```

```
##
## Call:
## lm(formula = score ~ age + bty_avg, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9427 -0.3474  0.1293  0.3957  0.9478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.054732   0.169865  23.870  < 2e-16 ***
## age         -0.003059   0.002664  -1.148 0.251396
## bty_avg      0.060656   0.017098   3.548 0.000429 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5347 on 460 degrees of freedom
## Multiple R-squared:  0.03778,    Adjusted R-squared:  0.0336
## F-statistic: 9.031 on 2 and 460 DF,  p-value: 0.0001422
```

There is a strong positive correlation between the teacher's beauty score and their evaluation score. Beauty is a stronger predictor of a score than age is.

The p-value in this multiple linear regression analysis for age is much greater than it was when it was the only independent factor we were looking at.

The new p-values are appropriate to use since. We can reject the null hypothesis for beauty, that beauty does not affect evaluation score, since the p-value is so low.