Lab 8

Code

AUTHOR
Maxim Dokukin

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.

```
knitr::opts_chunk$set(message = FALSE, warning = FALSE) # I tri
library(palmerpenguins)
library(tidyverse)
```

```
── Attaching core tidyverse packages ────────────────────────
tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.4.4     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ purrr     1.0.2     
── 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
```

```
set.seed(66)
```

A. Bootstrapping the sampling distribution of the median

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()`.

```
data(penguins)

fem_adelie_data <- penguins |>
             filter(!is.na(body_mass_g), species == 'Adelie', s

norm_confint <- function(data, field, alpha){

  column_data <- data[[field]]
  column_data <- na.omit(column_data)

  mean_data  <- mean(column_data)
  n <- length(column_data)
  se <- sd(column_data) / sqrt(n)

  z_score <- qnorm(1 - alpha/2)
  me <- z_score * se

  paste('[', mean_data - me, ',', mean_data + me, ']')
}

cat(norm_confint(fem_adelie_data, 'body_mass_g', 0.05))
```

```
[ 3307.04078189664 , 3430.63045098007 ]
```

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

```
calculate_mean <- function(data) {

  sample_data <- sample(data$body_mass_g, size = length(data$bo
                        replace = T)
  mean(sample_data)
}

set.seed(66)
bootstrap_means <- replicate(10000, calculate_mean(fem_adelie_d

cat('Mean interval: [', quantile(bootstrap_means, probs = 0.025
    quantile(bootstrap_means, probs = 0.975), ']')
```

```
Mean interval: [ 3305.822 , 3429.795 ]
```

---

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

```
calculate_median <- function(data) {

  sample_data <- sample(data$body_mass_g, size = length(data$boo
                        replace = T)
  median(sample_data)
}

set.seed(66)
bootstrap_medians <- replicate(10000, calculate_median(fem_adel

cat('Median interval: [', quantile(bootstrap_medians, probs = 0.
    ',', quantile(bootstrap_medians, probs = 0.975), ']')
```

```
Median interval: [ 3300 , 3450 ]
```

B. Simulations

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

```
n <- 10000
lambdas <- rexp(n, rate = 1)
y <- rpois(n, lambdas)

cat("Est E(Y):", mean(y))
```

```
Est E(Y): 1.0097
```

```
cat("Est var(Y):", var(y))
```

```
Est var(Y): 2.049011
```

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:

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

```r
binom_single_sim <- function(n, p){

  X <- rbinom(1, n, prob=p)
  ph <- X / n
  se <- sqrt(ph*(1-ph)/n)
  z <- qnorm(0.975)

  return((p >= ph - z * se) & (p <= ph + z * se))
}
```

```
b. Next, write a second function that takes three inputs: `n`,
`p`, and `n_runs`, representing the number of times to run your
simulation. This function should use your function from (a) to
simulate `n_runs` binomial random variables and return the
proportion of the `n_runs` for which $p$ is contained in the
confidence interval.
```

```r
binom_mult_sim <- function(n, p, n_runs){

  return(sum(replicate(n_runs, binom_single_sim(n, p))) / n_runs
}
```

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

```
binom_mult_sim(20, 0.5, 1000)
```

```
[1] 0.958
```

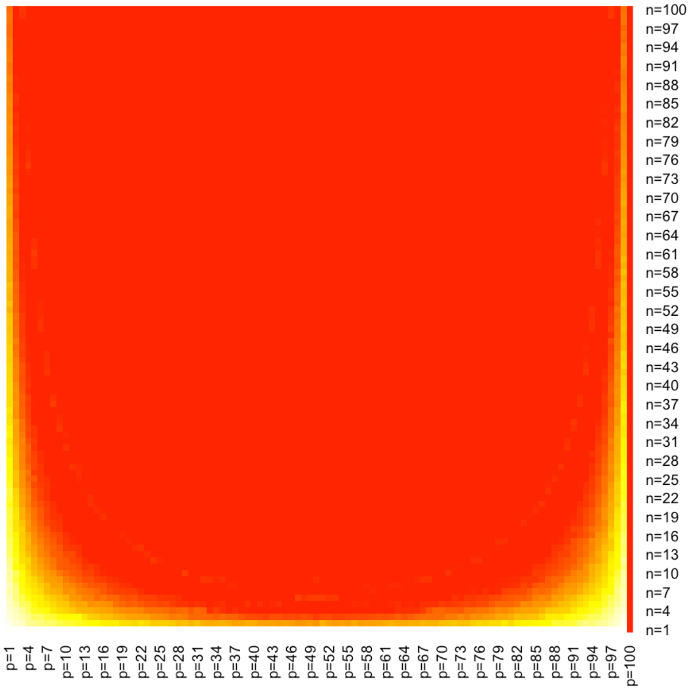
d. Use your simulation code to investigate the following questions: For what values of `n` and `p` is the frequentist coverage close to the expected 95\% value? For what values of `n` and `p` is the frequentist coverage very different to the expected 95\% value?

```
mat <- matrix(NA, nrow = 100, ncol = 100)

for(i in 1:100){
  for(j in 1:100){
    mat[i, j] <- binom_mult_sim(i, j/100, 1000)
  }
}
```

```
mat_dist <- abs(mat - 0.95)
rownames(mat_dist) <- paste("n=", 1:nrow(mat_dist), sep = "")
colnames(mat_dist) <- paste("p=", 1:ncol(mat_dist), sep = "")

heatmap(mat_dist, Rowv = NA, Colv = NA, scale = "none",
        margins = c(5,5), col = heat.colors(256))
```



The relationship between n and p is quite complex. The more red, the closest it is to 0.95.

—————————————————-

C. Hypothesis Testing

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

```
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.

```
flights_as <- flights_sample |>
              filter(carrier == "AS") |>
```

```r
              drop_na()

cat('Alaska Airlines arrival delay CI95', norm_confint(flights_a
```

```
Alaska Airlines arrival delay CI95 [ -12.6163019459496 ,
-7.24547520496716 ]
```

The CI95 for Alaska Airlines arrival time is 7-13 min ahead of the schedule.

---

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

```r
flights_ha <- flights_sample |>
              filter(carrier == "HA") |>
              drop_na()
cat('Hawaiian Airlines arrival delay CI95', norm_confint(flight
```

```
Hawaiian Airlines arrival delay CI95 [ -14.8776245271162 ,
1.04721517039107 ]
```

The CI95 for the average arrival delay for Hawaiian Airlines flights ranges from 1 min late to 15 minutes ahead of schedule.

---

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

```r
flights_ha_late <- flights_ha$arr_delay > 0

ph <- mean(flights_ha_late)
n <- length(flights_ha_late)
se <- sqrt(ph * (1 - ph) / n)
z <- qnorm(0.975)
me <- z * se

ci_lower <- ph - me
ci_upper <- ph + me
```

```r
cat('Hawaiian Airlines delayed proportion CI95: [', ph - me, ',
```

```
Hawaiian Airlines delayed proportion CI95: [ 0.2358532 ,
0.3313982 ]
```

We are 95% confident that 23-33% of Hawaiian Airlines arrive delayed.

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.

```r
t.test(flights_as$arr_delay, flights_ha$arr_delay)
```

```

    Welch Two Sample t-test

data:  flights_as$arr_delay and flights_ha$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
```

p-value = 0.4822 > 0.05; There is no statistically significant difference in mean arrival delays between the two airlines at the 95% confidence level.

D. Linear Regression

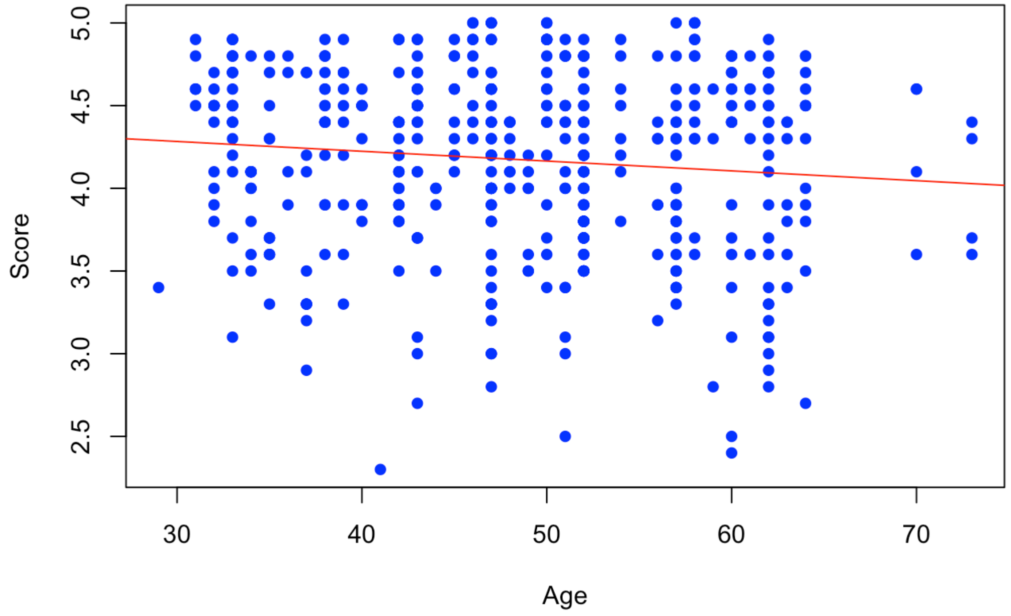
Researchers at the University of Texas in Austin, Texas tried to figure out what causes differences in instructor teaching evaluation scores. Use the following code to load data on 463 courses. A full description of the data can be found here.

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

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

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

plot(score ~ age, data = evals,
     xlab = "Age",
     ylab = "Score",
     pch  = 16, col = "blue")
abline(lm_res,  col = "red")
```



```
summary(lm_res)
```

```
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
```

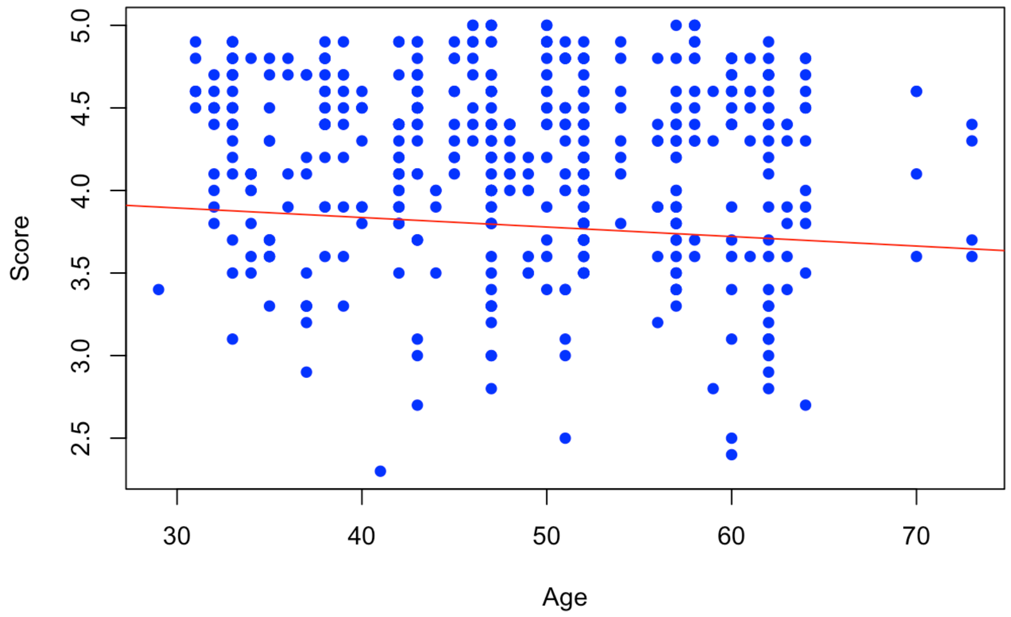
There is statistically significant slight negative correlation (p=0.02125) For every one-year increase in age the score decreases by 0.005938 points.

---

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

```r
lm_res = lm(score ~ age + bty_avg + gender + language, data = ev

plot(score ~ age, data = evals,
     xlab = "Age",
     ylab = "Score",
     pch  = 16, col = "blue")
abline(lm_res,  col = "red")
```



```
summary(lm_res)
```

```
Call:
lm(formula = score ~ age + bty_avg + gender + language, data =
evals)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8648 -0.3538  0.1154  0.4105  0.9180 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          4.066822   0.166480  24.428  < 2e-16 ***
age                 -0.005764   0.002702  -2.133 0.033432 *  
bty_avg              0.064691   0.016775   3.856 0.000132 ***
gendermale           0.200574   0.051507   3.894 0.000113 ***
languagenon-english -0.251754   0.102133  -2.465 0.014069 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5238 on 458 degrees of freedom
Multiple R-squared:  0.08052,	Adjusted R-squared:  0.07249 
F-statistic: 10.03 on 4 and 458 DF,  p-value: 8.7e-08
```

There is a positive relationship between the beauty average and the score, with a statistically significance (p=0.000132) Male gender correlates with better score; statistically significant (p=0.000113) Non English language has negative correlation to the score, at (p=0.014069)

————–-