Lab 2

Kristina Arevalo

Wed Sep 16 2020

Contents

Problem 1

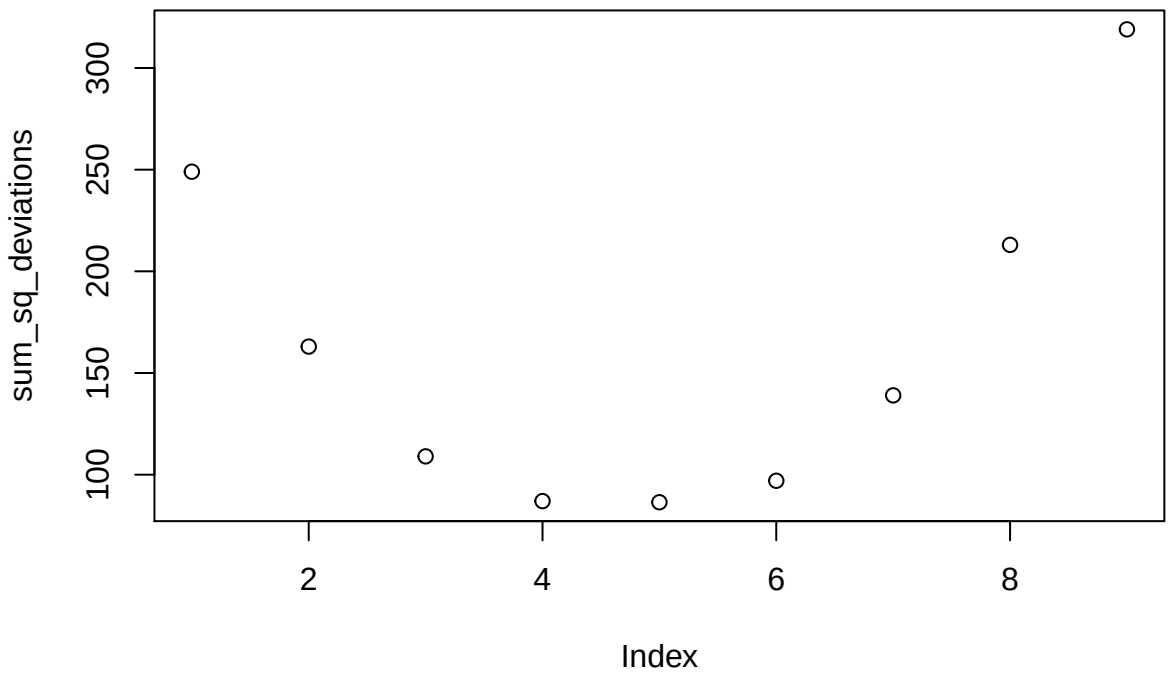
Use R to demonstrate that the mean minimizes the sum of the squared deviations from the mean. Accomplish the following steps:

a. Produce a sample of at least 10 or more different numbers
b. Produce a simulation following the example from the concepts section
c. Use your simulation to test a range of numbers smaller and larger than the mean to show that the mean minimizes the sum of the squared deviations from the mean.
d. Plot your results.

```
scores <- c(1,6,2,3,5,2,4,7,2,3,6,8,1,6,8,3)

numbers_to_test <- c(1,2,3,4,mean(scores),5,6,7,8)

sum_sq_deviations <- c()
for(i in 1:length(numbers_to_test)) {
  sum_sq_deviations[i] <- sum((scores-numbers_to_test[i])^2)
}

sum_sq_deviations
```

```
## [1] 249.0000 163.0000 109.0000  87.0000  86.4375  97.0000 139.0000 213.0000
## [9] 319.0000
```

```
plot(sum_sq_deviations)
```



For this problem I would rate it a 0, I wasn't sure what to do or what example you meant so I just followed the video solution

Problem 2

Write a custom R function for any one of the following descriptive statistics: median, mode, standard deviation, variance. Demonstrate that it produces the same value as the base R function given some set of numbers.

```
x <- rnorm(n=100, mean=12, sd=1)
n <- length(x)
standard_dev <- function(a){
  sqrt(sum((x - mean(x))^2) / (n - 1))
}

standard_dev(x)
```

```
## [1] 0.9906009
```

```
sd(x)
```

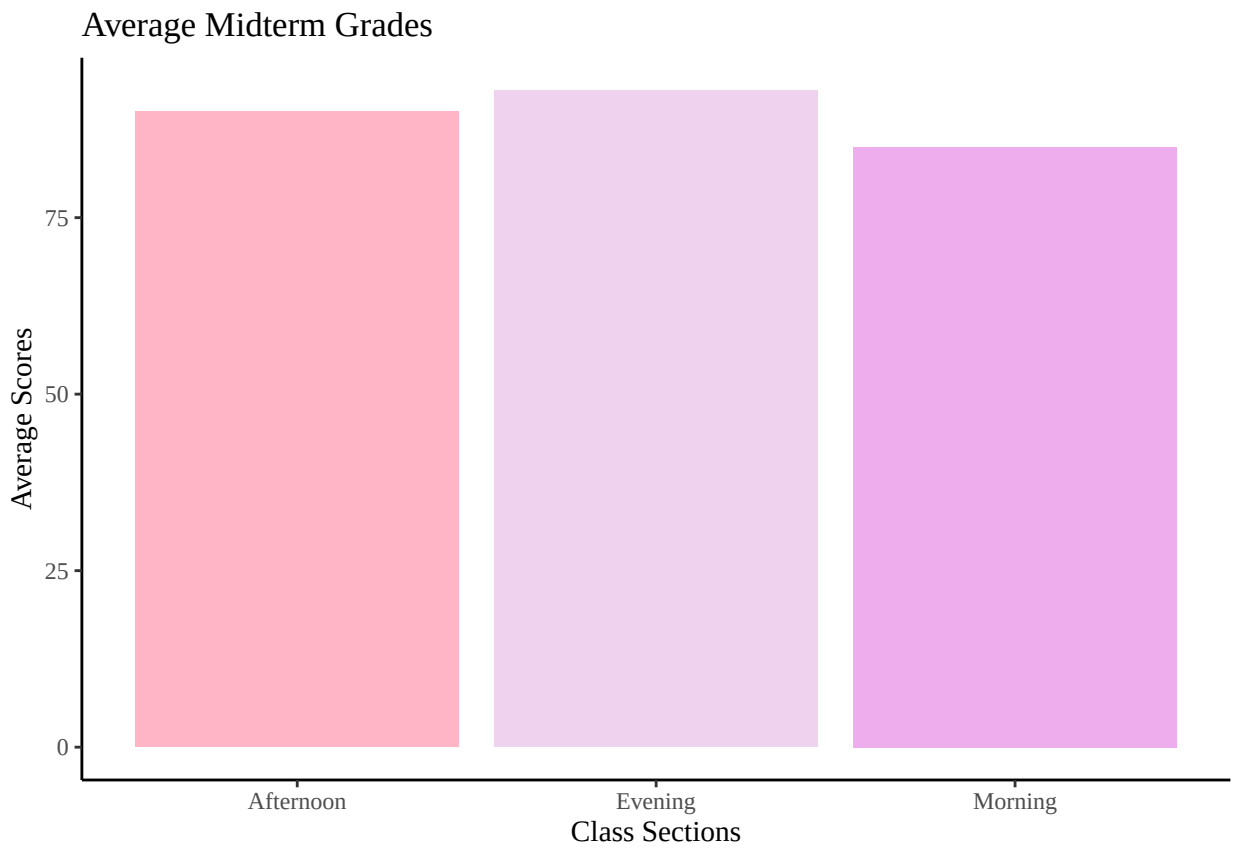
```
## [1] 0.9906009
```

Would rate this a 100 because it was something I had similiar notes on from Fahd's class.

Problem 3

Imagine the same instructor taught a morning, afternoon, and evening section of the same course. And, the average scores for each section on a midterm were 85% for the morning, 90% for the afternoon, and 93% for the evening sections. Create a data.frame representing these means for each section. Then, use ggplot2 to plot the means as bar graph. (hint you will need one vector for the class sections, and one vector for the means. Then you can combine them into a data.frame before plotting them)

```
class_sections <- c("Morning", "Afternoon", "Evening")
avg_scores <- c(85, 90, 93)
my_data_frame <- data.frame(class_sections,avg_scores)

avg_midterm_grades_graph <- ggplot(my_data_frame, aes(class_sections, avg_scores, fill= class_sections))
    geom_bar(stat="identity") +
    theme_classic()+
    theme(text = element_text(family = "Times"), legend.position = "none") +
    labs(x = "Class Sections", y = "Average Scores", title = "Average Midterm Grades") +
  scale_fill_manual(values= c("pink1", "thistle2", "plum2"))

avg_midterm_grades_graph
```

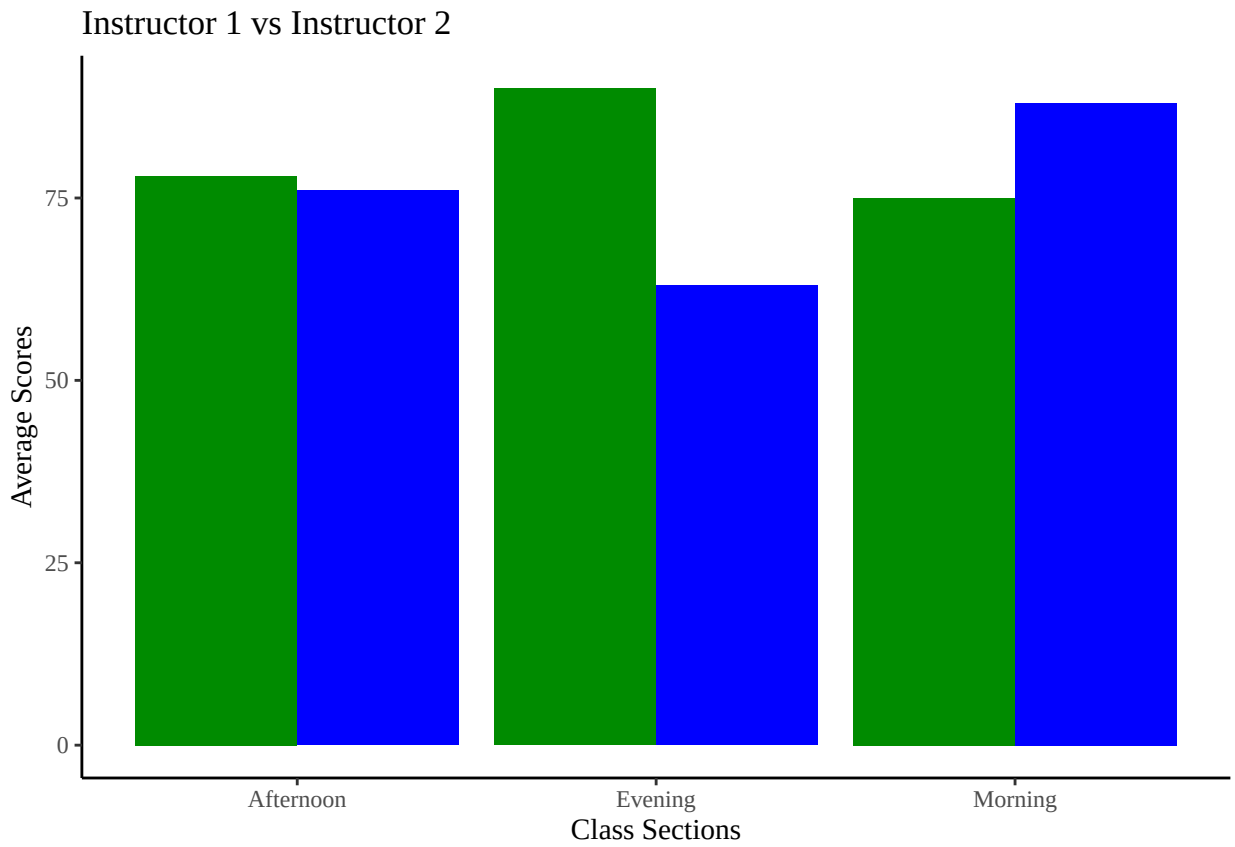


Would rate this as 100

Problem 4

Imagine there were two instructors, and they each taught different sections in the morning, afternoon and evening. The midterm averages for instructor 1 were 75%, 78%, and 80% in the morning, afternoon, and evening. The midterm averages for instructor 2 were 88%, 76%, and 63% for the morning, afternoon, and evening. Create a data.frame representing the means, the time of day, and the instructors (three columns). Then plot data.frame using ggplot2 as a bar graph.

```
instructor <- c("One", "One", "One", "Two", "Two", "Two")
class_section <- c("Morning", "Afternoon", "Evening", "Morning", "Afternoon", "Evening")
grades <- c(75,78,90,88,76,63)

my_data_frame2<- data.frame(instructor, class_section,grades)

ggplot(my_data_frame2, aes(class_section, grades, instructor, fill=instructor)) +
  geom_bar(stat= "identity", position= "dodge") +
  theme_classic()+
    theme(text = element_text(family = "Times"), legend.position = "none") +
    labs(x = "Class Sections", y = "Average Scores", title = "Instructor 1 vs Instructor 2") +
  scale_fill_manual(values= c("green4", "blue1"))
```



Would rate this a 50 because I couldn't figure out on my own how to separate the instructors grades because I kept making them separate vectors, but after that I made my own graph and such.

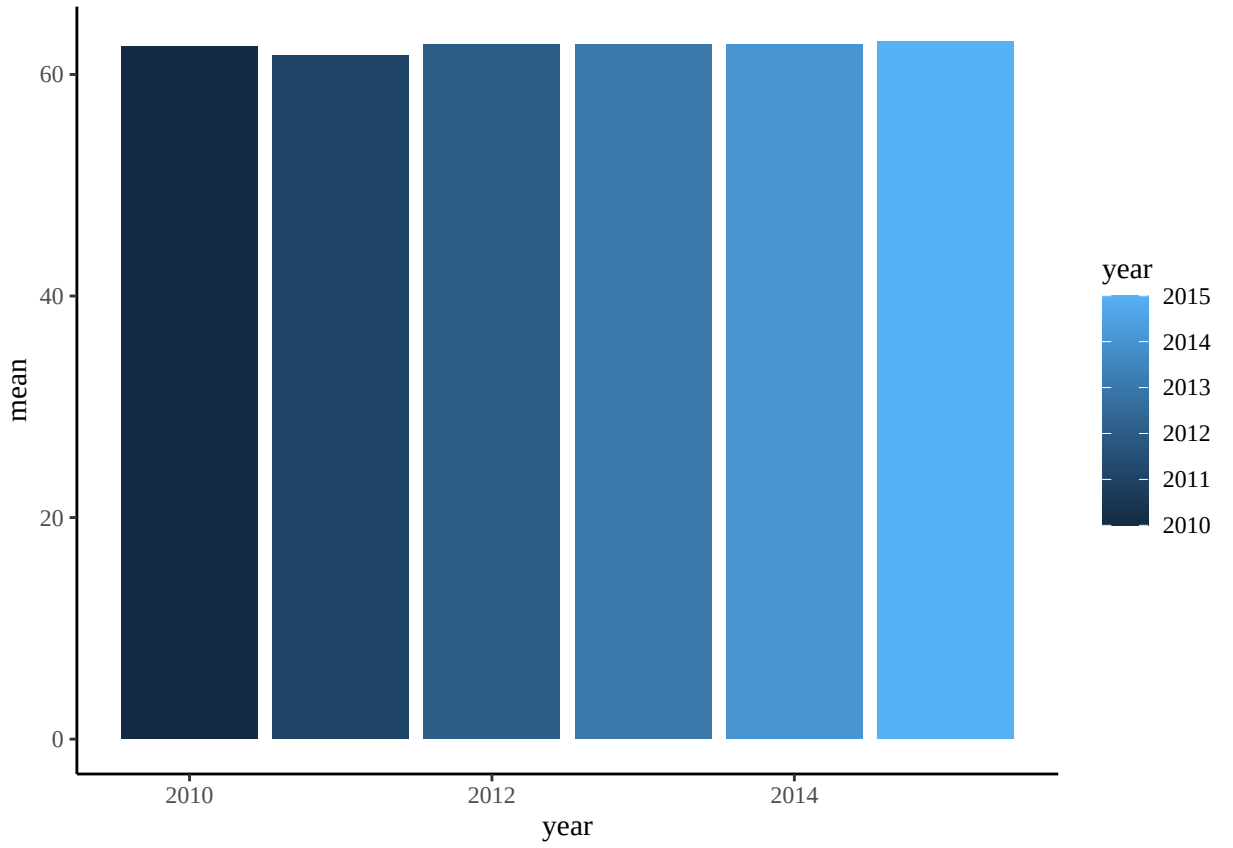
Problem 5

Import the WHR2018.csv data file, containing measure from the World Happiness report from 2018. For the years 2010 to 2015, what was the mean "healthy life expectancy at birth" for each year (find the mean for each year across countries). Show your results in a table and in a graph using ggplot.

```
WHR2018 <- read.csv(("open_data/WHR2018.csv"))

avg_HLE <- WHR2018 %>%
  filter(year >=2010, year <= 2015) %>%
  group_by(year) %>%
  summarize(mean = mean(Healthy.life.expectancy.at.birth, na.rm =TRUE))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
avg_HLE_graph <- ggplot(avg_HLE, aes(year, mean, fill= year)) +
  geom_bar(stat="identity") +
  theme_classic() +
  theme(text = element_text(family = "Times"))

avg_HLE_graph
```



Would rate this 100

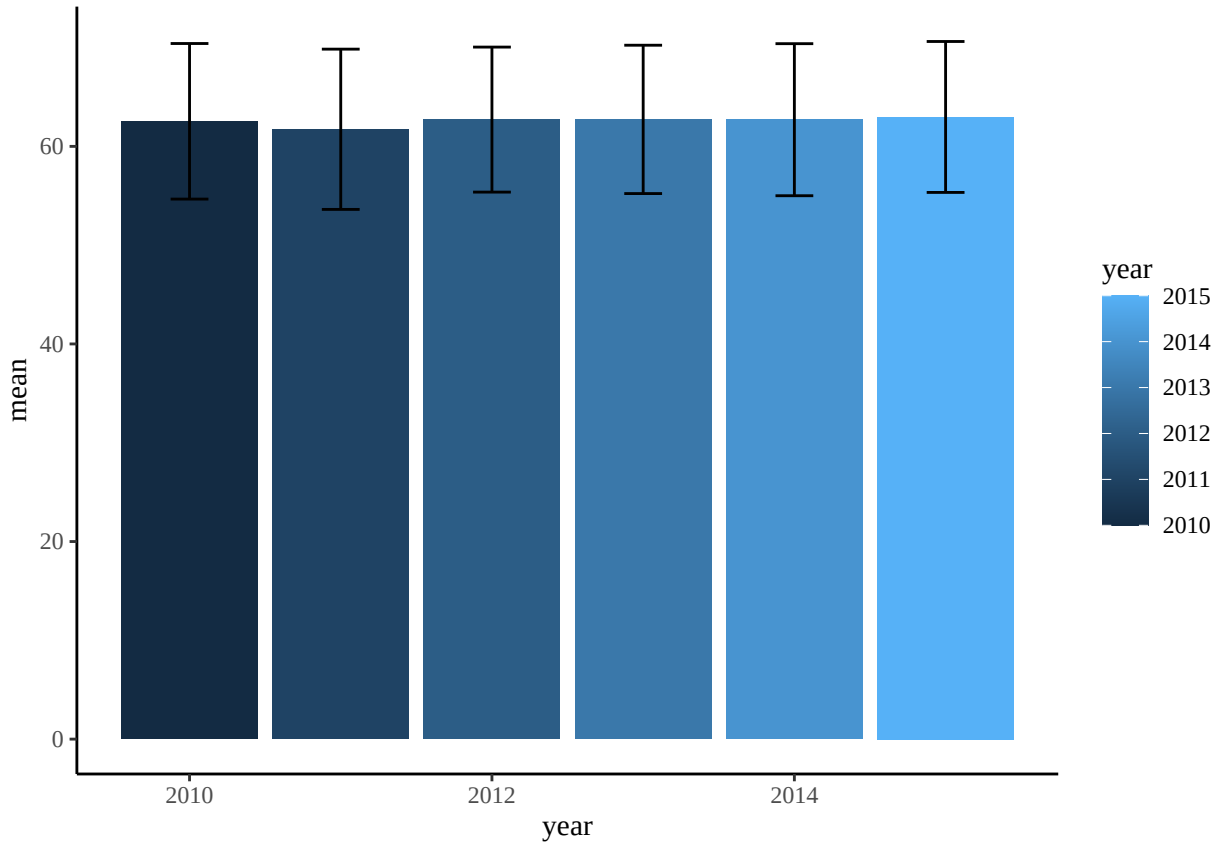
Problem 6

Repeat the above, except in addition to calculating the mean for each year, also calculate the standard deviation for "healthy life expectancy at birth" for each year. Then, add error bars to your graph using the +1 and -1 standard deviations from the means for each year.

```
avg_HLE_error <- WHR2018 %>%
  filter(year >=2010, year <= 2015) %>%
  group_by(year) %>%
  summarize(mean = mean(Healthy.life.expectancy.at.birth, na.rm =TRUE), sd = sd(Healthy.life.expectancy
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
avg_HLE_error_graph <- ggplot(avg_HLE_error, aes(year, mean, fill= year)) +
  geom_bar(stat="identity") +
  geom_errorbar(aes(ymin = mean - sd,
                    ymax = mean + sd),
                width = .25)+
  theme_classic() +
  theme(text = element_text(family = "Times"))

avg_HLE_error_graph
```



Would rate this a 100