ADM 2303 - Assignment 1

Renad Gharz

06/05/2022

1. Online Shopping

The trend of online shopping is increasing significantly. Although the number of people who purchase various products and services online is increasing more than ever before, this trend is still relatively low among people older than 60. To learn more about this issue, a sample of 250 customers (refer to the Excel file) older than 60 who have bought an item online were selected. The number of hours spent related to online shopping during the past six months was recorded.

1.1 Mean, Median, and Mode

Calculate the mean, median, and mode of the number of hours spent related to online shopping.

```r
hours_online_mean <- mean(assign4_q1_data$Hours) # Mean of hours
hours_online_mean
```

```
## [1] 26.212
```

```r
hours_online_median <- median(assign4_q1_data$Hours) # Median of hours
hours_online_median
```

```
## [1] 26
```

```r
hours_online_mode <- Mode(assign4_q1_data$Hours) # Mode of hours
hours_online_mode
```

```
## [1] 21
## attr(,"freq")
## [1] 13
```

1.2 Variance and Standard Deviation

Calculate the variance and standard deviation of the number of hours spent related to online shopping.

```r
hours_online_var <- var(assign4_q1_data$Hours) # Variance of hours
hours_online_var
```

```
## [1] 89.87857
```

```r
hours_online_sd <- sd(assign4_q1_data$Hours) # Standard deviation of hours
hours_online_sd
```

```
## [1] 9.480431
```

1.3 Boxplot

Draw a box plot related to the number of hours spent related to online shopping.

```
hours_online_boxplot <- boxplot(assign4_q1_data$Hours,
                                main = "Hours Spent on Online Shopping",
                                sub = "Boxplot",
                                ylab = "Number of Hours")
```


```
summary(assign4_q1_data$Hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   19.25   26.00   26.21   33.00   54.00
```

1.4 Quartiles and IQR

Determine the first, second, third quartiles, and interquartile range of the number of hours spent related to online shopping.

```
hours_online_quart1 <- quantile(assign4_q1_data$Hours, 0.25) # First quartile
hours_online_quart1
```

```
##   25%
## 19.25
```

```
hours_online_quart2 <- quantile(assign4_q1_data$Hours, 0.5) # Second quartile
hours_online_quart2
```

```
## 50%
```

```
## 26
```

```r
hours_online_quart3 <- quantile(assign4_q1_data$Hours, 0.75) # Third quartile
hours_online_quart3
```

```
## 75%
## 33
```

```r
hours_online_IQR <- IQR(assign4_q1_data$Hours) # Interquartile range of the sample
hours_online_IQR
```

```
## [1] 13.75
```

1.5 Reporting Findings

Briefly describe what you have learned from the statistics you calculated.

The data is relatively symmetrical because the mean and the median are almost the same ($26.212 \approx 26$). There is a single outlier of one customer spending 54 hours. If that outlier is removed it brings the mean to 26.1004 which wouldn't change the shape as it was already almost perfectly symmetrical, thus the outlier does not have any real effect on the shape of the data's normal distribution so there is no need to remove the outlier. Because the outlier's impact is negligible and the mean and median are almost the same, either one of the averages can be used in this case. However, when outliers are present, it's usually best to use the median.

The minimum number of hours spent on online shopping by people over the age of 60 in the sample of 250 customers over the past 6 months is 2 hours and the maximum number of hours for online shopping was 50 hours. 25% of the sampled customers spent 19.5 hours or less on online shopping, 50% spend 26 hours or less, and 75% spend 33 hours or less.

2. Speed Limit

An insurance company claims that speed limits inside the city are low, and most drivers drive at speeds that they consider safe. The company believes that the city should set the speed limit at a "New" speed limit of the 85th percentile. They collected a random sample of 400 speeds where the limit is 60 km/hr was recorded. Compute the "New" speed limit that the company suggests.

```r
speed_quartile_85 <- quantile(assign4_q2_data$Speeds, 0.85) # 85% percentile of the data
speed_quartile_85
```

```
## 85%
## 75
```

The value of the 85th percentile and the new speed limit that the company suggests is 75 km/hr.

3. CO-OP Salary

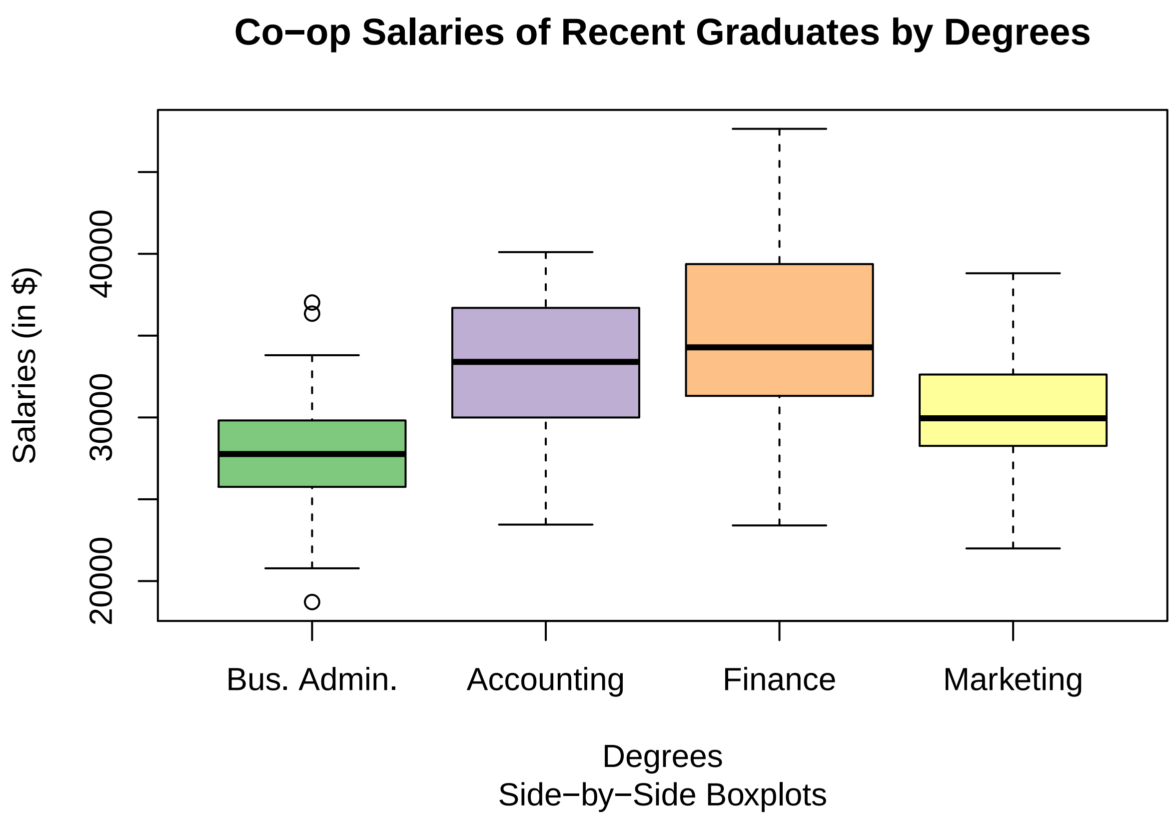
A business management student wants to learn more about the difference between co-op salaries related to her field of study. She uses a recent survey regarding the degree and co-op salary of four groups of graduates: (1) Business Administration, (2) Accounting, (3)Finance, and (4) Marketing.

3.1 Boxplots Comparison

Draw box plots to compare the four groups of starting salaries.

```r
coop_salaries_grouped_boxplots <-
  boxplot(assign4_q3_data$`Bus. Admin.`,
          assign4_q3_data$Accounting,
```

```r
            assign4_q3_data$Finance,
            assign4_q3_data$Marketing,
            main = "Co-op Salaries of Recent Graduates by Degrees",
            sub = "Side-by-Side Boxplots",
            xlab = "Degrees",
            ylab = "Salaries (in $)",
            col = brewer.pal(n = 4, name = "Accent"),
            names = colnames(assign4_q3_data))
```




```r
summary(coop_salaries_grouped_boxplots$stats)
```

```
##        V1              V2              V3              V4
##  Min.   :20780   Min.   :23451   Min.   :23401   Min.   :21994
##  1st Qu.:25757   1st Qu.:29998   1st Qu.:31318   1st Qu.:28262
##  Median :27765   Median :33397   Median :34284   Median :29951
##  Mean   :27585   Mean   :32730   Mean   :35203   Mean   :30329
##  3rd Qu.:29818   3rd Qu.:36698   3rd Qu.:39374   3rd Qu.:32625
##  Max.   :33804   Max.   :40105   Max.   :47639   Max.   :38812
```

3.2 Reporting Findings

Report your findings.

For business administration (BA) graduates, the maximum co-op salary is 33,804 and the minimum co-op salary is 20,780. Their median salary is 27,765. 75% of the BA graduates make 29,836 or less and 25% of them make 25,730 or less. Since the median is slightly greater than the mean ($M_d > \bar{y} \rightarrow 27,765 > 27,697$) then the salary distribution for the BA grads will be a little skewed to the left. There are 3 outliers for this degree with 3 graduates earning beyond the range of the data. One student earned $18,719 which is below the minimum salary of 20,780 and two students earned more than the maximum salary of 33,804 (36,345 and 37,025). Removing the 3 outliers would result in a new mean of 27,609 which is lower than the old mean which would make the distribution slightly more left skewed then the original. Because the outliers influence the shape of the distribution, they should be removed from the sample. Since the data contains outliers, it is best to use the median instead of the mean for the average.

Accounting graduates earn a minimum of 23,451 and a maximum of $40,105. 25% of them earn 29,297 or less and 75% earn 36,745 or less. The median salary for this degree is 33,397. Similarly, to the BA graduates the salary distribution for accounting grads is left skewed because the median is greater than the mean (33,397 > 33,148) by 249. However, the accounting distribution is more left skewed than the BA distribution because the difference between the median and mean is larger thus leading to more skew. Since there are no outliers present, it is best to use the mean for the average.

Finance grads earn a minimum 23,401 and a maximum of 47,639 while the median salary is 34,284. 25% of graduates with a finance degree earn $31,316 or less and 75% of them earn 39,551 or less. Contrary to the BA and accounting degrees, the mean is greater than the median (34,284 < 35,260) by 976 which means the distribution is right skewed and unlike the first two degrees, the skew for finance degrees is significantly more noticeable in the distribution. Since there are no outliers present, it is best to use the mean for the average.

Marketing graduates earn a minimum of 21,994 and a maximum of 38,812 during their co-op. Their median salary is 29,951 while 25% of these grads earned 28,254 or less and 75% earned 32,905 or less. Like the distribution for finance grads, this distribution's mean is greater than its median (29,951 < 30,474) by 523 meaning it is right skewed, although not as severely as the finance grads' distribution. Since there are no outliers present, it is best to use the mean for the average.

Looking at the four degrees together, BA grads had the lowest minimum salary of the group at 20,780 while accounting and finance grads have the highest minimum salary at 23,451 and 23,401 respectively. The BA graduates also have the lowest maximum salary of 33,804 excluding the two outliers while finance graduates have the highest max salary of 47,639 which is ~7,500 higher than the second highest max salary of 40,105 for the accounting grads.

Furthermore, The BA students have the smallest salary range of the group of 13,024, indicating that there isn't as much variation in salary for the BA grads compared to other degrees, while accounting has a range of 16,654 and marketing has a range of 16,818, indicating a higher degree of variation for graduates with those degrees compared to BA degree holders. Finance has by far the greatest range of variation of 24,238.

As for the IQR, finance graduates had the highest IQR of 8,235 indicating more variation where most of the data is clustered and BA had the smallest IQR of 4,106 indicating less variation where most of the data points are clustered.

4. Car Rental Company

A marketing manager of a car rental company wants to analyze the proportion of customers who are above 20 years old. He took a random sample of 500 customers and found out there were 375 customers above age 20. He knows that the true population proportion of customers above age 20 is 0.68.

4.1 Shape of Sampling Distribution

> Describe the shape of the sampling distribution of the proportion of customers who are above age 20.

First we verify that the success/failure condition is satisfied.

$$np = 500 \cdot 0.68 = 340 np \geq 10 \rightarrow 340 \geq 10$$

$$nq = 500 \cdot (1 - 0.68) = 500 \cdot 0.32 = 160 nq \geq 10 \rightarrow 160 \geq 10$$

Since the success/failure condition is met then the shape of the sampling distribution of the proportion of customers who are above age 20 is normal and is written as follows:

$$\hat{p} \rightarrow N(p, \sqrt{\frac{pq}{n}})$$

4.2 Mean and Standard Deviation

Find the mean and standard deviation of $\hat{p}$.

```
cust_above_20_mean <- 0.68 # Mean
cust_above_20_mean
```

```
## [1] 0.68
```

```
cust_above_20_sd <- sqrt((cust_above_20_mean * (1 - cust_above_20_mean)) / 500)
# Standard deviation
cust_above_20_sd
```

```
## [1] 0.02086145
```

4.3 Probability of Sample Proportion

What is the probability that the sample proportion $\hat{p}$ is within 0.03 of the true proportion of customers who are above age 20?

```
zscore_lower_bound <- ((cust_above_20_mean - 0.03) - cust_above_20_mean) / cust_above_20_sd
zscore_lower_bound <- round(zscore_lower_bound, digits = 2)

zscore_upper_bound <- ((cust_above_20_mean + 0.03) - cust_above_20_mean) / cust_above_20_sd
zscore_upper_bound <- round(zscore_upper_bound, digits = 2)

pnorm(zscore_upper_bound, lower.tail = TRUE) - pnorm(zscore_lower_bound, lower.tail = TRUE)
```

```
## [1] 0.8501326
```

5. Annual Salary

The annual salary of the new graduate students of a business school is normally distributed with a mean of 68,000 dollars and a standard deviation of 4,000 dollars.

5.1 Probability of Random Event

What is the probability that a randomly selected student receives more than $70,000?

```
annual_salary_new_grad_mean <- 68000
annual_salary_new_grad_sd <- 4000
zscore_greater_70k <- (70000 - annual_salary_new_grad_mean) / annual_salary_new_grad_sd
p_salary_greater_70k <- pnorm(zscore_greater_70k, lower.tail = FALSE)

pnorm(zscore_greater_70k, lower.tail = FALSE)
```

```
## [1] 0.3085375
```

5.2 Probability of Sample Mean

> A random sample of five students is selected. What is the probability that the sample mean is greater than $70,000?

```
new_grad_sample_size <- 5
zscore_mean_greater_70k <- (70000 - annual_salary_new_grad_mean) / (annual_salary_new_grad_sd / sqrt(5))

pnorm(zscore_mean_greater_70k, lower.tail = FALSE)
```

```
## [1] 0.1317762
```

5.3 Cumulative Probability of Random Event

> In a random sample of five graduate students, what is the probability that all receives more than $70,000?

```
dbinom(5,5, prob = p_salary_greater_70k)
```

```
## [1] 0.002796019
```