Problem Set 1

Chase Bookin

June 24, 2020

Question 1

```
sample_1 <- c(3, 4, 5, 6, 7)
sample_2 <- c(67, 68, 69, 70, 71)

mean_samp1 <- sum(sample_1) / length(sample_1)

sd_samp1 <- round(sqrt(((3-mean_samp1)^2 + (4-mean_samp1)^2 + (5-mean_samp1)^2
                  + (6-mean_samp1)^2 + (7-mean_samp1)^2) / (length(sample_1) - 1)), 3)

mean_samp2 <- sum(sample_2) / length(sample_2)

sd_samp2 <- round(sqrt(((67-mean_samp2)^2 + (68-mean_samp2)^2 + (69-mean_samp2)^2
                  + (70-mean_samp2)^2 + (71-mean_samp2)^2) / (length(sample_2) - 1)), 3)
```

Sample 1:
- Mean: 5
- Standard Deviation: 1.581

Sample 2:
-Mean: 69
-Standard Deviation: 1.581

The standard deviations of sample 1 and sample 2 are equal. This shows that although the samples have different means, they are similarly dispersed around their center.

Question 2

```
z_tokyo <- (380000 - 420000) / 20000

z_germany <- (3100 - 3200) / 57
```

Relative to their peers, the worker in Germany is earning more than the worker in Tokyo. This is demonstrated by the z-score of each worker's salary. The z-score of the workers' salaries tells the relative position of their salary to their peers using the mean and standard deviation. In this case, the German worker's z-score of roughly -1.75 is greater than the Tokyo worker's z-score of -2, demonstrating that the German worker is earning comparatively more than the Tokyo worker.

Question 3

a)

```
z_prob_keane <- 1-.192
z_keane <- 0.87
sd_keane <- (25000 - 21000) / 0.87
```

Standard Deviation: 4597.7

b)

```r
z_42nd <- -0.2
percentile_42 <- (sd_keane * z_42nd) + 21000
```

42nd Percentile: 20080.46

c)

```r
# -1.55 <= z <= 1.55
percentile_94 <- (sd_keane * 1.55) + 21000
percentile_06 <- (sd_keane * -1.55) + 21000
```

Middle 88% values: (13873.56, 28126.44)

Question 4

a)

```r
# finding P(1502 <= sample mean <= 1,748)
pop_mean <- 1573
pop_var <- 952021
n <- 85
std_error_a <- sqrt(pop_var / n)

z_1502 <- (1502 - 1573) / std_error_a
z_1748 <- (1748 - 1537) / std_error_a

CDF_z_1502 <- 0.25143
CDF_z_1748 <- 0.97670

prob_btwn <- CDF_z_1748 - CDF_z_1502
```

The probability the sample average lies between 1502 and 1748 is 0.725

b)

```r
# 92% middle pack boundaries; sample size of 63

std_error_b <- sqrt(pop_var / 63)
z_0.96 <- 1.75
z_0.04 <- -1.75

upper_bound <- round((z_0.96 * std_error_b) + pop_mean, 3)
lower_bound <- round((z_0.04 * std_error_b) + pop_mean, 3)
```

The boundaries of the middle 92% of the sample average estimator with n = 63 are (1357.875, 1788.125).

Question 5

a)

```r
accidents <- c(12, 7, 17, 11, 9, 8, 19, 22, 12, 17, 15, 9, 12, 21, 15)
n <- length(accidents)
sample_avg <- sum(accidents) / n
```

The sample average is 13.7333333.

b)

```
SD_accidents <- tibble(accidents = accidents) %>%
  summarize(SD_accidents = sqrt(sum((accidents - sample_avg)^2) / (n - 1))) %>%
  pull(SD_accidents)

SE_accidents <- SD_accidents / sqrt(n)
```

The standard error of the sample average is 1.22.

c)

```
null <- 11.2
t_critical <- 2.05
t_sample <- (sample_avg - null) / SE_accidents

# p_value to right of null hypothesisis 2*(1-CDF(t-stat))
p_value <- 2 * (1 - 0.98124)
```

Using the t-stat method to test the friend's claim, we have enough evidence to reject the null hypothesis that the population average for daily car crashes is 11.2. The t-stat of the sample mean is 2.075, which is greater than the t-critical value of 2.05 using a significance level of 4%.

Using the p_value method, we see the p-value is approximately 0.038, which is less than the alpha value of 0.04. Therefore, we reject the null hypothesis.