Worksheet 3

Author:

Discussants:

==========

The purpose of this worksheet is to gain more experience examining quantitative data using histograms, boxplots, 5 number summaries and z-scores. Please submit a compiled pdf with your answers *by 11:59pm on Sunday September 30th.* Be sure to answer all the questions that are posed in the different exercise steps, and always label your axes!

Some useful functions for completing this worksheet are: dim(), length(), hist(), mean(), median(), sd(), min(), fivenum(), boxplot() and quantile().

Exercise 1a: For the first set of exercises we will examine information about the heights of major league baseball players. The code below loads information about every major league baseball player who has played until the 2016 season. What does a case correspond to in this data frame and how many cases are there?

```
# load a data frame called 'Master' with baseball player
# information
library(Lahman)

# A data frame with information about baseball players
Master <- Master
```

Answers:

Exercise 1b: The code below extracts a vector of heights of all the baseball players in inches. Create a histogram of their heights, and be sure to use an appropriate number of bins and to put appropriate labels on your axes. Describe the shape of the histogram. Are there any noticeably small or large outliers?

```
# extract a vector with baseball player heights
player_heights <- Master$height


# create a histogram of the baseball player's heights
```

Answers:

Exercise 1c: Now create a boxplot of the heights, and report the 5 number summary. Do you notice any large outliers now? Describe what you should do when you you see an extreme outlier, and then do it!

Answers:

Exercise 1d: Next calculate the mean and standard deviation of the heights of baseball players. Note, you will need to use the na.rm = TRUE option since there is missing data. Then report the range of heights that one would expect for the middle 95% of the heights to be in if it was the case that the heights were normally distributed (hint: for a normal distribution, the middle 95% of the data is within ± two standard deviations of the mean). Then calculate the actual 2.5 and 97.5th percentile values for the heights which correspond to the end points for the middle 95% of the data. Are these end points close to what you calculated for the end points that were based on assumption that the heights had a normal distribution?

Answers:

Exercise 1e: Finally calculate the z-score for the minimum value in the data set (again use the na.rm = TRUE argument when calculating the mean, the standard deviation, and the minimum value in the data). Report what this value is, and what this value means.

Answers:

Exercises 2a: For the second set of exercies let's analyze data from 5 different brands (Toyota Corollas, Subaru Imprezas, Honda Civics, Hyundai Elantras and the Mazda 3s) in the car_data data frame. The code below extracts vectors with the prices of Mazda 3s and Toyota Corollas. Start by reporting the number of elements in these two vectors and create histograms of the prices of these two cars. Be sure to choose an appropriate number of bins (n) to most clearly see the shape of the underlying distribution and be sure to label the axes appropriately. Describe the shapes of these distributions below.

```
# do no change the lines below
load("/home/shared/intro_stats/cs206_data/cars_small.Rda")


# do no change the lines below
mazda_prices <- subset(car_data$price, car_data$brand == "Mazda")
toyota_prices <- subset(car_data$price, car_data$brand == "Toyota")


# calculate the number of elements in these vectors and
# create histograms of their distributions
```

Answer:

Exercise 1b: Now create side-by-side boxplots comparing the prices of Mazdas and Toyotas, and as always label the axes appropriatley. Which brand of car has the higher median price and which type of car has more variability as measured by the inter-quartile range?

Answers:

Exercise 2c: Finally calculate the mean and standard deviation of the prices of both brands and report which brand has the higher mean and standard deviation. Also report the range of values that approximately 95% of the Mazdas will be if it was the case that the Mazda prices were normally (bell-shaped) distributed.

Then calculate the actual 5th and 95th percentile for the Mazda prices. Are these close to what you calculated based on assuming the distribution of prices had a normal distribution?

Answers: