STA 2260

Homework 2 - R Portion

Kevin Bailey

Due 9/11/2020

Instructions

- Create a new `R` Script
  - When saving this file, use the format `LastName_FirstName_hw2`, this should be a `.R` file when saved.
- When doing separate parts, create clear breaks in the code separated by a comment about which part will follow. Something like:

```
code is here

# part (b)
more code is here
```

This assignment will have you get the standard deviation and variance, as well as create boxplots, in `R`. We will also learn what the **coefficient of variation** is through this assignment. The goal is not only to be able to create this all in `R`, but be able to analyze the data from the plots themselves.

This time, we have random samples of people who play osu! (ew) and people who play Beatmania IIDX. We recorded reaction times (in ms) of players from each of these samples, and I provide you this data below. You need to copy and paste this code into your `R` script as it will be used for this entire assignment. It may also be helpful to remember that `boxplot(...)` has an argument for `horizontal` to make the boxplots horizontally oriented rather than vertical.

```
osu_data <- c(450, 420, 469, 360, 450, 390, 250, 415, 410, 480, 444, 461, 260, 440,
   345, 435, 449)

iidx_data <- c(350, 369, 275, 215, 249, 210, 360, 320, 215, 233, 280, 274, 290, 310,
   320, 290, 304)
```

Part (a)

For the osu! data, you should:

- Get the variance and standard deviation of the data.
- Create a box plot of the data.
  - Do you think the data is skewed or symmetric? If skewed, what kind? Explain what made you come to your conclusion.
  - It may be helpful to store this boxplot into a variable so you can access things to help with other parts.
- What is the spread of fourths for this data?
  - This is where you can use your variable for the boxplot with the `$` in `R` to help get exact values.

Part (b)

For the IIDX data, you should do the same as you did for the osu! data:

- Get the variance and standard deviation of the data.
- Create a box plot of the data.
  - Do you think the data is skewed or symmetric? If skewed, what kind? Explain what made you come to your conclusion.
  - It may be helpful to store this boxplot into a variable so you can access things to help with other parts.
- What is the spread of fourths for this data?
  - This is where you can use your variable for the boxplot with the `$` in `R` to help get exact values.

Part (c)

- Based on the standard deviations from parts (a) and (b), which data set do you think is more spread?

Construct side-by-side boxplots (multiple boxplots in the same window) in `R` using the `boxplot(...)` command. You can throw multiple datasets in there, give it a shot and see what happens.

- Do you think these side-by-side boxplots agree with your answer about which one is more spread?
  - Note that the standard deviation is sensitive to outliers since it relies on the mean for its calculation, which is also sensitive to outliers.

There is actually something called the **coefficient of variation** (CV), which expresses the standard deviation of a data set as a percentage of the mean.

$$CV = \frac{s}{\bar{x}}$$

This is a *unitless* measurement and can be used to compare spreads of data from entirely different measurements if needbe.

- Now, calculate the CV for both the osu! data as well as the IIDX data. Do these results still agree with what you said earlier about which dataset may be more spread?

Part (d)

You may have noticed there were outliers in the osu! data. Let's see what happens if we just remove them.

- Create a new variable that removes the outliers from the osu! data. You can just copy your code above and manually remove them.
- Get the standard deviation, the variance, and the coefficient of variation of this new dataset. Is it more or less spread than the other two datasets?
- Now, create side-by-side boxplots using all three of the data sets. Comment on the shape of the new dataset.
  - Notice anything off about the new data set? Comment on anything that stuck out to you.

This shows that we can't really handle outliers by just simply removing them, especially with a small sample size (n). Removing outliers can still result in more outliers, and removing those can result in *more* outliers, so they must be handled with care. This is why most of the time outliers are still included, but handled in different ways (covered in other classes) or just kept in the back of our mind when analyzing the data.

FEEL FREE TO CONTACT ME ON ANY OF THESE PARTS AT ANY TIME! I want to help you understand! As always, you can reference the lecture videos and/or the `R` script in Week 2, and anything else really. Please communicate!