Assignment 1

Introduction to Data Science

5/26/2021

Introduction

Solve the questions below and report your solutions and findings using RMarkdown. The final pdf should be submitted via Canvas. The deadline for this assignment is **June 12, 2021, 11.55pm**.

This assignment will ask you to use some useful R commands. Figures should be made using the pacakge ggplot. Pay attention to the layout of the plot.

The sample mean

If we have a data set given with n observations denoted by $x_1, x_2, \ldots, x_n$. Then we can always determine the sample mean, denoted by $\hat{\mu}$. Moreover, the sample mean will always be finite since we have a finite sample.

However, we have to be careful with blindly using the sample mean to help to summarize a data set. In this assignment, we will show situations where the sample mean does not provide useful information about the data set. Moreover, using the sample mean can be dangerous, since it does not reflect the true nature of the data.

Question 1

Run the following code.

```
library(Pareto) # if necessary, first install the package. Use install.pacakges("Pareto)
set.seed(100)
Data=data.frame(x.n=rnorm(50000),x.p=rPareto(50000,t=1,alpha=2))
```

Assume the observations in the data frame `Data` represent observations of two variables that you have to investigate.

1. Use `ggplot` to make a histogram and a boxplot of the variable `x.n`.
2. Determine the sample mean and sample standard deviation of the variable `x.n`.
3. Explain how the sample mean and standard deviation that you calculated in the previous question can be used to summarize the variable.
4. Consider the following statement: *'The mean and the standard deviation of the observations of the variable `x.p` cannot be used to summarize the data. Moreover, the mean is a bad predictor for new observations because it neglects possible very extreme realizations.'* Provide an analysis to support this statement. Make useful plots and tables.
   Tip: Start by determining the mean and standard deviation of the data set. Make a histogram and boxplot. You can use the function `filter` to determine a subset of a data frame.

Question 2

1. Load the data set `DataAssignment1.txt`. Transform the data to the log scale and make a histogram and boxplot.

2. Are there outliers in this data set that should be deleted before we start our exploratory analysis?

3. Determine the mean and median of the data set. Explain what you see?

4. This data set contains the daily claim amounts of a large insurance company is receiving. The most recent claim amounts are at the bottom of the data set. Determine at each day the sample mean and median using all past, but no future, observations (rolling mean and median). Use these figures to justify which measure should we use in this example, the mean or the median. Or can we use both? *Use the functions* `cumsum`, `sapply` *and* `head` *to determine the rolling mean and median without using a for loop.*