Adding data analysis to a mathematical statistics course

Johanna Franklin

Hofstra University

January 17, 2020

Situation

I inherited a very theoretical upper-division prob/stats sequence and discovered that the students

- *hadn't been learning to use statistical software,*
- *hadn't been required do any kind of project, and*
- *generally didn't use actual data in a meaningful way.*

Problem

How do you introduce data analysis into a "pure" prob/stats sequence without losing coverage of the theoretical topics?

Mathematical Probability and Statistics 1 & 2

First semester:

- univariate probability
- brief introduction to basic descriptive statistics, regression, confidence intervals, and hypothesis testing

Second semester:

- multivariate probability
- the rest of statistics: CIs, hypothesis testing, and the theory behind it all (maximum likelihood estimation, sufficient statistics, best critical regions, etc.)

Student population

Required for students in:

- math ed (only the 1st semester)
- actuarial science
- mathematical finance

Popular among students in the sciences/economics/CS.

Across the sequence.

First semester:

- Get them used to the idea that the data they will analyze is real and not a toy data set (and cite sources!)
- Use Excel

Second semester:

- Give them toy data sets only for calculations by hand, and then only when necessary
- Use R
- Assign a project

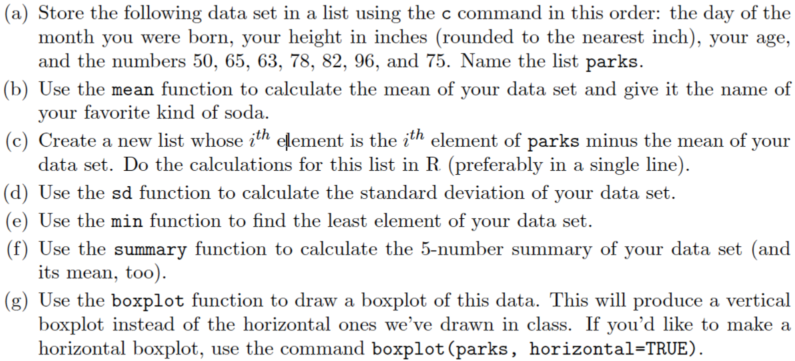
During the probability phase

Goal

The students will learn how to utilize data and calculate basic descriptive statistics in R via homework with minimal class demonstration.

Week 1	—
Week 2	create and manipulate lists use `summary` and `boxplot`
Week 3	import data work with a single column of a table use `hist` and `plot`
Week 4	clean up data sets create subtables
Week 5	—
Week 6	create histograms and a q-q plot to estimate normality

Sample homework problem: Week 2



(a) Store the following data set in a list using the `c` command in this order: the day of the month you were born, your height in inches (rounded to the nearest inch), your age, and the numbers 50, 65, 63, 78, 82, 96, and 75. Name the list `parks`.

(b) Use the `mean` function to calculate the mean of your data set and give it the name of your favorite kind of soda.

(c) Create a new list whose i^{th} element is the i^{th} element of `parks` minus the mean of your data set. Do the calculations for this list in R (preferably in a single line).

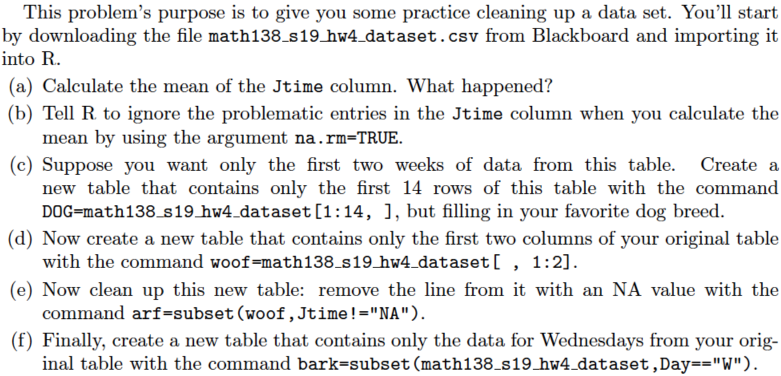
(d) Use the `sd` function to calculate the standard deviation of your data set.

(e) Use the `min` function to find the least element of your data set.

(f) Use the `summary` function to calculate the 5-number summary of your data set (and its mean, too).

(g) Use the `boxplot` function to draw a boxplot of this data. This will produce a vertical boxplot instead of the horizontal ones we've drawn in class. If you'd like to make a horizontal boxplot, use the command `boxplot(parks, horizontal=TRUE)`.

Sample homework problem: Week 4



This problem's purpose is to give you some practice cleaning up a data set. You'll start by downloading the file `math138_s19_hw4_dataset.csv` from Blackboard and importing it into R.

(a) Calculate the mean of the `Jtime` column. What happened?

(b) Tell R to ignore the problematic entries in the `Jtime` column when you calculate the mean by using the argument `na.rm=TRUE`.

(c) Suppose you want only the first two weeks of data from this table. Create a new table that contains only the first 14 rows of this table with the command `DOG=math138_s19_hw4_dataset[1:14, ]`, but filling in your favorite dog breed.

(d) Now create a new table that contains only the first two columns of your original table with the command `woof=math138_s19_hw4_dataset[ , 1:2]`.

(e) Now clean up this new table: remove the line from it with an NA value with the command `arf=subset(woof,Jtime!="NA")`.

(f) Finally, create a new table that contains only the data for Wednesdays from your original table with the command `bark=subset(math138_s19_hw4_dataset,Day=="W")`.

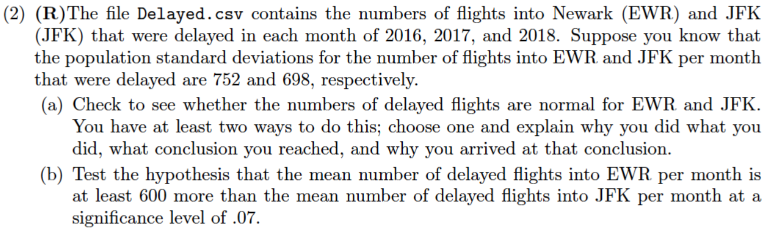
During the statistics phase

Goal

The students will learn to carry out statistical tests/regression analyses in R as demonstrated in class.

Split class examples: first one by hand, then one in R—and post screenshots to the class website afterwards!

Sample homework problem: Week 8



(2) **(R)**The file `Delayed.csv` contains the numbers of flights into Newark (EWR) and JFK (JFK) that were delayed in each month of 2016, 2017, and 2018. Suppose you know that the population standard deviations for the number of flights into EWR and JFK per month that were delayed are 752 and 698, respectively.

(a) Check to see whether the numbers of delayed flights are normal for EWR and JFK. You have at least two ways to do this; choose one and explain why you did what you did, what conclusion you reached, and why you arrived at that conclusion.

(b) Test the hypothesis that the mean number of delayed flights into EWR per month is at least 600 more than the mean number of delayed flights into JFK per month at a significance level of .07.

Now possible: projects!

Goal

Each student will design and carry out their own final project on a topic that interests them that they can talk about in an interview or carry over into their workplace.

Week 7: Proposal due, identifying the research question, statistical method, background sources, and potential data source

Week 11: Data set due

Week 15 (last day): Poster session and reports due

summary(talk)

To add data analysis to a theoretical statistics course (or a financial derivatives course, or a stochastic processes course, or...):

- don't be afraid to delegate basic descriptive stats to the homework sets,
- do regression/hypothesis testing/etc. examples in R during class as well as examples by hand,
- don't believe you have to give up teaching the theory to do it, and
- do have faith in your students!

For more details, contact me:

`johanna.n.franklin@hofstra.edu`

`@JohannaF_Math` (Twitter)