Hannah Roach
2/8/2019

Lab Exercise 3: Basic Statistics, Visualization, and Hypothesis Tests

Purpose:	The lab introduces you to the analysis of data using the R statistical package within the Data Science and Big Data Analytics environment. After completing the tasks in this lab you should able to: • Perform summary (descriptive) statistics on the data sets • Create basic visualizations using R both to support investigation of the data as well as exploration of the data • Create plot visualizations of the data using a graphics package • Test a hypothesis about the data

Tasks:	Tasks you will complete in this lab include: • Reload data sets into the R statistical package • Perform summary statistics on the data • Remove outliers from the data • Plot the data using R • Plot the data using lattice and ggplot • Test a hypothesis about the data

References:	References used in this lab are located in your ***Student Resource Guide Appendix***. See the Appendix for: • R Commands – Quick Reference • Surviving LINUX – Quick Reference

Part 1 – Basic Statistics and Visualization Using R

Workflow Overview

1. Prepare working environment for the Lab and load data files
2. Obtain summary statistics for Household Income and visualize data
3. Obtain summary statistics for number of rooms and visualize data
4. Remove Outliers
5. Stratify Variable – Household Income and plot the results
6. Plot Histogram and Distributions
7. Compute Correlation between income and number of rooms
8. Create a Boxplot – Distribution of income as a factor of number of rooms
9. Exit R

LAB Instructions

Step	Action

1	**<u>Prepare working environment for the Lab and load data files</u>**
	1. Set the working directory to LAB01 where we have stored the data. On the console window type:
	`setwd("~/LAB01")`
	
	2. In the script window, open the script called "Module3Lab2.R". (Click on "File", "Open File" and Navigate to directory LAB03 and click on file "Module3Lab2.R").
	Start R and Read the Data Set Back Into Your Workspace:
	
	3. Execute the following commands from the script window:
	`options(digits=3)`
	`options(width=68)`
	
	`ls()`
	`load(file="Labs.Rdata")`
	`ls()`
	
	`rm(lab2)`
	
	`ds <- lab1`
	`colnames(ds) <- c("income", "rooms")`

2	**<u>Obtain summary statistics for Household Income and visualize data:</u>**
	
	1. Execute the following commands from the script window:
	
	`summary(ds$income)`
	`range(ds$income)`
	`sd(ds$income)`
	`var(ds$income)`
	
	`plot(density(ds$income))   # left skewed`
	
	2. What is the mean? 67200
	3. What is the median? 50300
	4. What is the standard deviation? 68178

Step	Action

3	**<u>Obtain summary statistics for Number of rooms and visualize data:</u>**
	
	Execute the following commands from the script window:
	`summary(ds$rooms)`
	`range(ds$rooms)`
	`sd(ds$rooms)`
	`plot(as.factor(ds$rooms))`
	
	What is the mean? 5.63
	
	What is the median? 6.00
	
	What is the standard deviation? 1.99

Step	Action
4	**<u>Remove Outliers</u>** In a previous lab, you recorded the range of income. You observed that the minimum household income is 4, and the maximum is 1,620,560. 1. Does this make sense to you? Why? Yes, this make sense because the median is 50,300 and the 3rd quarter is 84200 and the plot appears to be left skewed. 2. What happens if you throw out the top and bottom 10%? Execute the following line from the script window **`(m <- mean(ds$income, trim=0.10) )`** `The new mean becomes 55,347` 3. How does this compare to the previous mean of this variable? The new mean with the outliers removed is lower than the original mean. 4. Execute the following commands from the script window: **`ds <- subset(ds, ds$income >= 10000 & ds$income < 1000000)`** **`summary(ds)`** **`quantile(ds$income, seq(from=0, to=1, length=11))`** 5. How do these values vary from the values in the original data set? These values only include incomes between $10,000 and $1,000,000. 6. Do they make more sense? Yes 7. Which data set would you prefer to use? The second dataset because it excludes the outliers. ________________________________ *We might consider the high and low value as outliers, and get rid of them. On the other hand, as we will discover, income is best described via a lognormal distribution, and hence these values are in the extreme ends +- 3 sds from the mean.

Step	Action

5	**<u>Stratify Variable – Household Income and plot the results:</u>**
	
	Stratify breaks that occur close to U.S. Guidelines for Poverty, Median Income, Wealth, and Rich (> $250k @ year)
	
	1. Execute the following code (listed under comment heading "step 5" in the script file):
	`breaks <- c(0, 23000, 52000, 82000, 250000, 999999)`
	`labels <- c("Poverty", "LowerMid", "UpperMid", "Wealthy", "Rich")`
	`wealth <- cut(ds$income, breaks, labels)`
	`# add wealth as a column to ds`
	`ds <- cbind(ds, wealth)`
	`# show the 1st few lines.`
	`head(ds)`
	
	2. Continue to execute the remaining part of the code in Step 5
	`wt <- table(wealth)`
	`percent <- wt/sum(wt)*100`
	`wt <- rbind(wt, percent)`
	`wt`
	`plot(wt)`
	
	3. Take another look at the relationship between wealth and income. Execute the following lines:
	`# take another look -- wealth by rooms`
	
	`nt <- table(wealth, ds$rooms)`
	`print(nt)`
	`plot(nt)          # nice mosaic plot`
	
	4. Execute this code from the script file. These lines will remove the variables wealth, breaks and labels, and then save the variables data set and write into a file named "Census.Rdata".
	`rm(wealth,breaks,labels)`
	`save(ds, wt, nt, file="Census.Rdata")`

Step	Action

6	**Plot Histogram and Distributions:**
	
	Problem: How do you represent income given the range of values? Given this rage of values, you could illustrate the data as a histogram.
	
	1. Select and execute the code under Step 6 Histograms and distributions in the script file.
	`library(MASS)`
	
	`with(ds, {`
	`  hist(income, main="Distribution of Household Income", freq=FALSE)`
	`  lines(density(income), lty=2, lwd=2)`
	`# line type (lty) 2 is dashed`
	`  xvals = seq(from=min(income), to=max(income), length=100)`
	`  param = fitdistr(income, "lognormal")`
	`  lines(xvals, dlnorm(xvals, meanlog=param$estimate[1], sdlog=param$estimate[2]), col="blue")`
	`})`
	
	2. Now try the same thing with log10(income)
	
	`logincome = log10(ds$income)`
	`hist(logincome, main="Distribution of Household Income", freq=FALSE)`
	`# line type lty(2) is a dashed line`
	`lines(density(logincome), lty=2, lwd=2)`
	`xvals = seq(from=min(logincome), to=max(logincome), length=100)`
	`param = fitdistr(logincome, "normal")`
	`lines(xvals, dnorm(xvals, param$estimate[1], param$estimate[2]), lwd=2, col="blue")`

<table>
<tr><th>Step</th><th>Action</th></tr>
<tr><td>7</td><td>
<u>Compute Correlation between income and number of rooms:</u>

1. You need to consider your hypothesis.

• Your hypothesis is that the number of rooms in a house is predicted by household income (the rich can buy bigger houses), e.g. lm(rooms ~ income)
• Therefore, our null hypothesis: no correlation between income and number of rooms.
• Alternate hypothesis: there is a correlation between income and the number of rooms.
4. Execute the following code (listed after the comment line “Step7 in the script file).
<code>with(ds, cor(income, rooms))</code>

<code>with(ds, cor(log(income), rooms))) # this will give a better correlation</code>

5. For comparison, correlate rooms with a completely unrelated variable.
<code>n = length(ds$income)</code>
<code>with(ds, cor(runif(n), rooms))</code>
</td></tr>
<tr><td>8</td><td>
<u>Create a Boxplot - Distribution of income as a factor of number of rooms:</u>

1. Select and execute the code (Listed after the comment line “Step 8”) in the script window.
2. Plot the distribution of income as a factor of # of rooms. ‘log=”y”’ plots income on log scale. We will suppress the outlier points and let the whiskers cover the full range of the data.
<code>boxplot(income ~ as.factor(rooms), data=ds, range=0, outline=F, log=”y”,xlab="# rooms", ylab="Income")</code>

3. Plot the # of rooms as a function of wealth level.
<code>boxplot(rooms ~ wealth, data = ds,main="Room by Wealth", Xlab="Category", ylab="# rooms")</code>

<code># we’ll keep the outlier points in this one</code>
</td></tr>
</table>

Step	Action
9	**<u>Exit R:</u>** 1. Type the following command into the RStudio command window: **`q()`** 2. R will ask you if you want to save your workspace. Answer "**`no`**."

End of Lab Exercise