Essential Skills: Assignment Statistics (1)
2016-2017

In lab sessions of the course ES related to the lectures about Statistics, we have decided to ask you to practice doing statics with R. We have selected two themes, namely:

- Examining Data
- Summary Statistics With Aggregate

The two themes are proposed by the web site Rexercises [1] and provide a progressive introduction, which allow learning how to use R functions to get some useful statistics about a given data set.

Examining Data

One of the first steps of data analysis is the *descriptive analysis*; this helps to understand: (1) how the ***data is distributed*** and (2) provides important information for further steps. Examining Data tutorial helps you to practice with functions useful for one variable descriptive analysis, including graphs.

NOTE: Before proceeding, it might be helpful to look over the help pages for the length, range, median, IQR, hist, quantile, boxplot, and stem functions. For this Tutorial you will use a dataset called islands, an R dataset that contains the areas of the world's major landmasses expressed in squared miles.

To load the dataset run the following instruction: data(islands).

1. Load the islands dataset and obtain the total number of observations.

2. **Measures of central tendency**. Obtain the following statistics of islands
   - Mean, Median

3. Using the function range, obtain the following values:
   - Size of the biggest island, Size of the smallest island

4. **Measures of dispersion**. Find the following values for islands:
   - Standard deviation
   - The range of the islands size using the function range.

5. **Quantiles**. Using the function quantile obtain a vector including the following quantiles:
   - 0%, 25%, 50%, 75%, 100%
   - .05%, 95%

[1] R exercise http://r-exercises.com/

6. Interquartile range. Find the interquartile range of islands.

7. Create a histogram of islands with the following properties.
   - Showing the frequency of each group.
   - Showing the proportion of each group

8. Create box-plots with the following conditions
   - Including outliers
   - Without outliers

9. Using the function boxplot find the outliers of islands. Hint: use the argument prob=F.

10. Create a stem and leaf plot of islands

Summary Statistics With Aggregate

Using R's built-in time series dataset, "AirPassengers", compute the average annual standard deviation.

1. Aggregate the "airquality" data by "airquality$Month", returning means on each of the numeric variables. Also, remove "NA" values.

2. Aggregate the "airquality" data by the variable "Day", remove "NA" values, and return means on each of the numeric variables.

3. Aggregate "airquality$Solar.R" by "Month", returning means of "Solar.R". The header of column 1 should be "Month". Remove "not available" values.

4. Apply the standard deviation function to the data aggregation from Exercise

5. The structure of the aggregate() formula interface is `aggregate(formula, data, FUN).`
   - The structure of the formula is y ~ x. The "y" variables are numeric data. The "x" variables, usually factors, are grouping variables, that subset the "y" variables.
   - aggregate.formula allows for one-to-one, one-to-many, many-to-one, and many-to-many aggregation. Therefore, use aggregate.formula for a one-to-one aggregation of "airquality" by the mean of "Ozone" to the grouping variable "Day".

6. Use aggregate.formula for a many-to-one aggregation of "airquality" by the mean of "Solar.R" and "Ozone" by grouping variable, "Month".

7. Dot notation can replace the "y" or "x" variables **inaggregate.formula**. Therefore, use "." dot notation to find the means of the numeric variables in airquality", with the grouping variable of "Month"

8. Use dot notation to find the means of the “airquality” variables, with the grouping variables of “Day” and “Month“. Display only the first 6 resulting observations.

9. Use dot notation to find the means of “Temp“, with the remaining “airquality” variables as grouping variables.

References

[1] Exercise with Answers in R: http://r-exercises.com/2016/09/06/examining-data-solutions/
[2] Elementary statistics with R: http://www.r-tutor.com/elementary-statistics
[3] R: A self-learn tutorial https://www.nceas.ucsb.edu/files/scicomp/Dloads/RProgramming/BestFirstRTutorial.pdf
[4] Introduction to Probability and Statistics Using R https://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf