post2 – Reproducible Data Visualization on Hypothesis Tesing through Shiny App

Ashley Gu

11/30/2017

Reproducible resources:

This post focuses on the reproducibility of the data analysis, thus I included the libraries I used for this data analysis below:

- ggplot2

The dataset is obtained through a school project and it is a public dataset that was used for a research paper http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.1152&rep=rep1&type=pdf.

The dataset is included in the data/cleandata/ folder and all the paths in this doc and the shiny app file are relative paths so if you download the entire post2 folder and set working directory to app/, you will be able to reproduce the analysis correctly.

How to correctly reproduce the shiny app display?

- Download "post2" file folder through this google drive link. https://drive.google.com/drive/folders/1jUJJNrrq5J5xA8QhfE1PKi9m6HPIWYUA?usp=sharing
- Open "hypotest.R" in post2/app/ using RStudio. Set "post2/app" as working directory.
- Make sure you install the shiny app and ggplot2 package.
- On the top right corner, you should be able to see a "Run App" button. Click this button and you will be able to see the shiny app page.
- Play around with it as much as you want.

Introduction on dataset and hypothesis testing:

This post explores the topic of hypothesis testing and shiny app. I want to build a hypothesis testing visualizer to help me have a more direct and intuitive understanding of the dataset when I conduct the hypothesis tests.

Recently I saw a dataset called 'oliveoil.txt' which contains the percentages (mulitplied by 100) of each of eight fatty acids comprising olive oil across three regions of Italy which are further partitioned into a total of nine subregions.

Here is the first few rows of our dataset:

```
#import ggplot library
library(ggplot2)

# Load the data files and scale the percentage
olive_oil = read.table('../data/cleandata/oliveoil.txt', sep = ',', header = T)
olive_oil[,c(5:12)] = olive_oil[,c(5:12)]/100

#show the head of dataset
head(olive_oil)
```

```
##   Obs        Label Region Area palmitic palmitoleic stearic oleic linoleic
## 1   1 North-Apulia      1    1    10.75        0.75    2.26 78.23     6.72
## 2   2 North-Apulia      1    1    10.88        0.73    2.24 77.09     7.81
## 3   3 North-Apulia      1    1     9.11        0.54    2.46 81.13     5.49
## 4   4 North-Apulia      1    1     9.66        0.57    2.40 79.52     6.19
## 5   5 North-Apulia      1    1    10.51        0.67    2.59 77.71     6.72
## 6   6 North-Apulia      1    1     9.11        0.49    2.68 79.24     6.78
##   linolenic arachidic eicosenoic
## 1      0.36      0.60       0.29
## 2      0.31      0.61       0.29
## 3      0.31      0.63       0.29
## 4      0.50      0.78       0.35
## 5      0.50      0.80       0.46
## 6      0.51      0.70       0.44
```

I want to understand for the same acid type, whether the regions of olive oil is correlated with the percentages of that type of acid. In other words, for the same acid type, does the pattern of percentages vary by region?

The regions are devided as follows:

- South Italy: North-Apulia; Calabria; South-Apulia; Sicily
- Sardinia: Inland-Sardinia; Coast-Sardinia
- North Italy: Umbria; East-Liguria; West-Liguria;

I would use hypothesis testing to answer this question, specifically, I will use ANOVA test to test for the homogeneity of percentages across various regions for the same acid type. To explain ANOVA test in a more intuitive way, let's take a look at a subset of our dataset.

```
olive_oil[c(1,2,3,4,5,403,413,423,505,506,507),c(2,3,5,6)]
```

```
##              Label Region palmitic palmitoleic
## 1      North-Apulia      1    10.75        0.75
## 2      North-Apulia      1    10.88        0.73
## 3      North-Apulia      1     9.11        0.54
## 4      North-Apulia      1     9.66        0.57
## 5      North-Apulia      1    10.51        0.67
## 403 Inland-Sardinia      2    10.40        1.03
## 413  Coast-Sardinia      2    11.20        0.90
## 423          Umbria      3    10.85        0.70
## 505    East-Liguria      3    10.90        0.60
## 506    East-Liguria      3    11.50        0.90
## 507    East-Liguria      3    12.40        0.90
```

Under null hypothesis, I assume that for the same type of acid, the acid percentage pattern is the same across different regions.

In the example above, it would mean that for column “palmitic”, the percentages should be the same for acorss regions, it is just due to chance that we observe a difference among the region 1,2,3.

Under this null assumption, we will generate a null model and have a distribution of the percentages if our null hypothesis is true.

We then compare our observed percentages with the distribution under null assumption to see how likely/unlikely it is for us to observe this data if null hypothesis is true. P-value describes this likelihood/probability. Normally, if p-value is less than 5%, we will reject the hypothesis.

Combine shiny app for interactive visualization

Why shiny app?

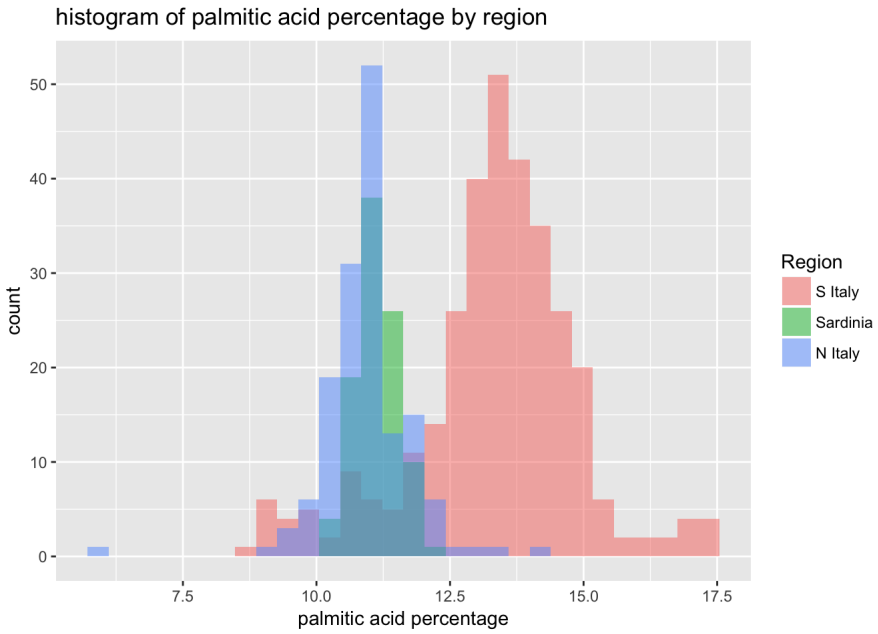
Before we start simulating null hypothesis and calculating p-values, it would be really helpful to do some data visualization and look at ,for each acid type, the distribution of percentages across region.

One example is listed below:

```
#change Region into factor for plotting

olive_oil$Region = factor(olive_oil$Region)
ggplot(olive_oil) + geom_histogram(aes(x = olive_oil$palmitic,  fill = Region), alpha = 0.5, position = "identity"
) + scale_fill_discrete(name="Region",
                        breaks=c(1, 2, 3),
                        labels=c("S Italy", "Sardinia", "N Italy")) + xlab(paste("palmitic", "acid percentage"))
+ ggtitle(paste("histogram of" , "palmitic", "acid percentage by region"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



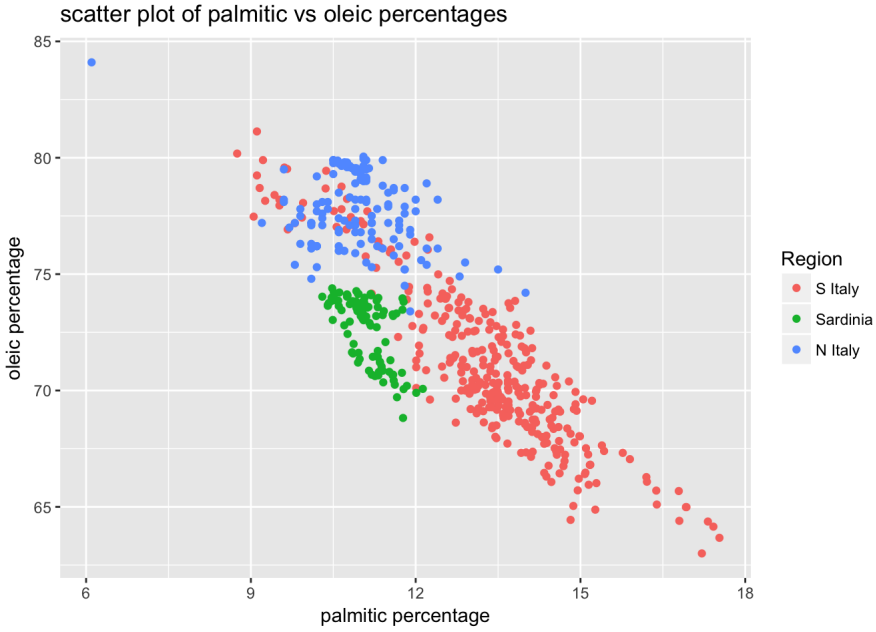
However, we have nine different acids so we want to generalize this process. In addition to that, for convenience, when we are trying to interpret the result of our hypothesis testing, we want to be able to jump between acid types quite easily and don’t spend too much time going back and forth to run and scroll through pages of histograms to find the one with the right acid type.

In this situation, Shiny App comes in very helpful since it could display visualizations interactively and researchers could easily change the acid type and jump between histograms.

What’s in My Shiny App?

In addition to histograms, I also included a scatter plot that shows the percentages of acid A versus acid B, colored by regions. Below is one example:

```
#palmitic vs oleic
ggplot(data = olive_oil, aes(x = olive_oil$palmitic, y = olive_oil$oleic)) + geom_point(aes(color = Region)) + sca
le_color_discrete(name="Region",
                        breaks=c(1, 2, 3),
                        labels=c("S Italy", "Sardinia", "N Italy")) + xlab("palmitic percentage") + ylab("oleic p
ercentage") + ggtitle("scatter plot of palmitic vs oleic percentages")
```



As we can see from the scattered plot above, it's very easy to see how the percentages are clustered according to regions, which helps us to better interpret the relationship between region and pattern of acid percentages.

To make it easy to compare with our test results, I included a test statistics table on the "Histogram" page. I also include the test statistics of chi square test. Different from ANOVA test, chi square tests compares the pattern for all acid types at once. Therefore, the chi square test tells us in general, the acid percentages are related with regions. And ANOVA test tells us that for each region, whether this relationship is random or it is statistically significant.

Key Takeaway

My key takeaway for this post is that shiny app makes data visualization more user friendly and flexible. Users can quickly generate and navigate through plots without having to worry about how the code works.

Reference

- https://support.rstudio.com/hc/en-us/articles/218294727-Why-would-I-use-Shiny-instead-of-Tableau-Spotfire-Qlikview-or-similar-BI-tools-
- https://www.analyticsvidhya.com/blog/2016/10/creating-interactive-data-visualization-using-shiny-app-in-r-with-examples/
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.1152&rep=rep1&type=pdf
- https://medium.com/ibm-data-science-experience/shiny-a-data-scientists-best-friend-883274c9d047
- http://shiny.stat.calpoly.edu/t_Test/
- https://interestingyhu.wordpress.com/2015/03/06/create-a-t-test-shiny-app/
- http://hselab.org/first-shiny-app.html