Central Washington University

Computational Statistics

Winter 2019

Seminar 3 Report

Author:
Hermann Yepdjio

Professor:
Dr. Donald Davendra

February 27, 2019

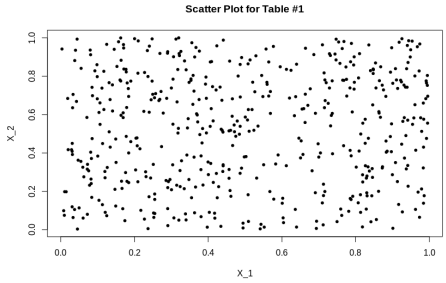
Contents

1 Introduction

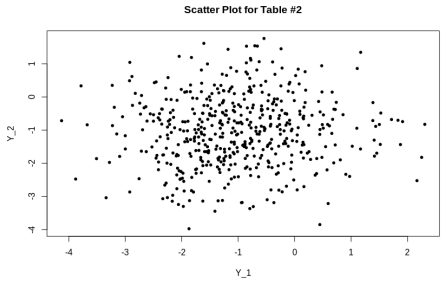
For seminar 3, we were given 6 columns of data containing 500 entries each. The goal was to analyze each 2 columns separately, for correlation using Pearson's R, Pearson's R^2 and Spearman's rho. In order to have a visual representation of the data, we started by making a scatter plot for each two columns. We then for each two columns, computed the covariance, Pearson's R and R^2, Spearman's rho and finally, we bootstrapped each two columns to obtain not only their standard errors, but also other important information. The details of the analysis we performed are described in the next sections as well as the results we obtained.

2 Scatter plots

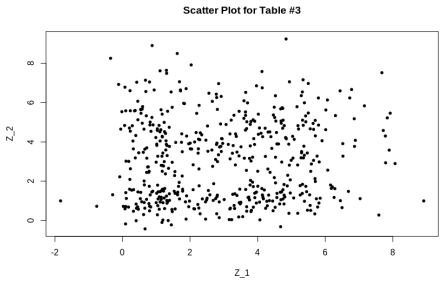
The goal of making the scatter plots was to see if we would be able to visually detect a trend in the repartition of the data.



(a) Scatter Plot for table 1



(b) Scatter Plot for table 2



(c) Scatter Plot for table 3

Figure 1: Scatter Plots

As we can see from Figure 1 above, none of the tables appear to have any type of correlation from a visual inspection.

3 Covariance

At this level of our analysis, we computed the covariance for each set of data and we obtained the following:

- Table1: cov(x, y) = 0.006
- Table2: cov(x, y) = 0.056
- Table3: cov(x, y) = 0.218

The covariance calculated above simply tells us for each set, by how much in average values of two variables differ from their respective means.

4 Pearson's R and R^2

Since the 3 data sets are relatively large (number of entries > 30), we assumed that they are normally distributed according to the Central Limit Theorem (CLT) and we applied Pearson's test to all of them and got the following:

- Table1:
  - R = 0.072
  - $R^2 = 0.051$
  - p-value = 0.110
- Table2:
  - R = 0.049
  - $R^2 = 0.002$
  - p-value = 0.273
- Table2:
  - R = 0.052
  - $R^2 = 0.003$
  - p-value = 0.243

Significance: As we can see from the results above, even though all the 3 sets have a non zero Pearson's R, the fact that their p-values are greater that 0.05 indicates that there is no significant correlation between their columns.

5 Spearman's Rho

Spearman's rho is a non-parametric correlation which is usually used when the data do not meet parametric assumptions. However, we computed the Spearman's correlation coefficient for each of the 3 sets just to confirm the results found in section 4 and we obtained the following:

- Table1:
  - R = 0.070
  - p-value = 0.117
- Table2:
  - R = 0.062
  - p-value = 0.163
- Table2:
  - R = 0.058
  - p-value = 0.196

Significance: Just like with Pearson's R the results above show that, even though all the 3 sets have a non zero Spearman's R, the fact that their p-values are greater that 0.05 indicates that there is no significant correlation between their columns.

6 Bootstrapping

Bootstrapping is another method that can be used to deal with data that do not meet parametric assumptions. However, we bootstrapped the 3 data sets first to obtained their standard errors based on bootstrap samples and second to confirm the results found in section 4. We obtained the following:

- Table1:
  - SE = 0.042
  - confidence interval: Normal = (-0.012, 0.154)
  - confidence interval: Basic = (-0.013, 0.151)
  - confidence interval: Percentile = (-0.008, 0.157)
  - confidence interval: BCa = (-0.011, 0.153)
- Table2:
  - SE = 0.041
  - confidence interval: Normal = (-0.032, 0.130)
  - confidence interval: Basic = (-0.032, 0.126)
  - confidence interval: Percentile = (-0.028, 0.130)
  - confidence interval: BCa = (-0.028, 0.130)
- Table2:
  - SE = 0.046
  - confidence interval: Normal = (-0.039, 0.143)
  - confidence interval: Basic = (-0.034, 0.146)
  - confidence interval: Percentile = (-0.042, 0.138)
  - confidence interval: BCa = (-0.044, 0.136)

Significance As we can see from the results above, we received 4 confidence intervals for each data set after bootstrapping. However, we can see that all those intervals cross zero. This means that the direction of the correlation in the original set is not necessarily the same as the direction of the correlation in the bootstrap samples. Which confirms that there is no significant correlation between the columns of the 3 original data sets.

7 conclusion

After performing a Pearson's test, Spearman's test and bootstrapping on the 3 data sets that were provided to us for seminar 3, we can conclude that there is no significant correlation between the 2 columns of each of those sets.