Research Methodology: Bivariate Analysis

Yesoda Bhargava

January 12, 2021

Discussion Points

- Two categorical variables are generally analysed via a contingency table.
- One categorical and one continuous variable are analysed using box plot.
- Two continuous variables are analysed using scatter plot, correlation.
- Variables must be cleaned for any missing/blank values before bivariate analysis.

Contingency table

- Refer to the previous dataset used. Reference here.
- We want to see how the number of diabetes cases lie across the gender.
- Clearn variable sex, diabetes for any invalid/blank values.
- Sex: Male, Female. Diabetes: Yes/No.

```
> table(data$diabetes)

    1     2     3     4     7     9
57401  3782 347091  9149   605   231
> data_new=subset(data, data$diabetes==1 | data$diabetes==3)
> table(data_new$diabetes)

    1     3
57401 347091
```

Contingency Table: Diabetes Vs. Sex

```
> tb=table(data_new$diabetes,data_new$sex)
> tb

        1      2
  1  27170  30231
  3 158141 188950
> percent =function(a,b){
+   pr=a/(a+b)
+   pr=pr*100
+   pr=round(pr,2)
+   return(pr)
+ }
> percent(tb[1],tb[2])
[1] 14.66
> percent(tb[3],tb[4])
[1] 13.79
```

Homework

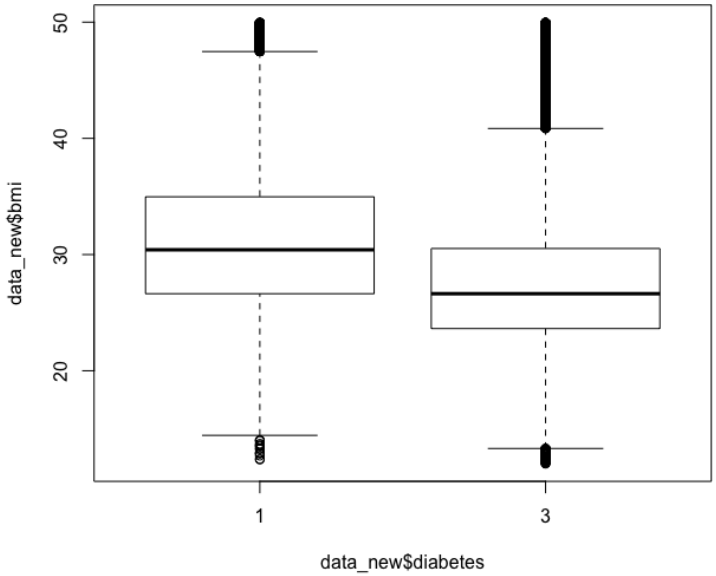
- Tabulate Sex Vs. Cholesterol, Sex Vs. Blood Pressure, Sex Vs. Exercise, Sex. Vs. Genhlth.
- Be careful in analysing each variable before tabulating it with other variable.
- Cleaning of data is important.

One Categorical and One Continuous Variable

- Suppose we want to see the BMI distribution among diabetic and non-diabetic people.
- In previous code we already cleaned data for diabetes values (Yes, No).
- Let us also clean it for the BMI values in the new data set **data_new**.

```
> summary(data_new$bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  12.00   23.91   27.28   28.27   31.32   99.84   34729
> data_new<-data_new[!is.na(data_new$bmi),]
> hist(data_new$bmi)
> data_new=subset(data_new, data_new$bmi<50)
> hist(data_new$bmi)
> boxplot(data_new$bmi~data_new$sex)
> boxplot(data_new$bmi~data_new$diabetes)
```

Side-by-side boxplot



Observation

- The median for diabetic is higher than that of non-diabetic.
- The inter-quartile range of diabetic is higher than that of non-diabetic.
- The maximum value of BMI is much less in non-diabetics as compared to diabetics.
- **Homework**: Try to create side-by-side box plot for Diabetes age and Sex.

Two continuous variables

- Scatter plot.
- Correlation.
- Analyse the diabetes age and BMI.
- Correlation: **cor(data_new$bmi, data_new$diabage)** = -0.145. Interpretation?

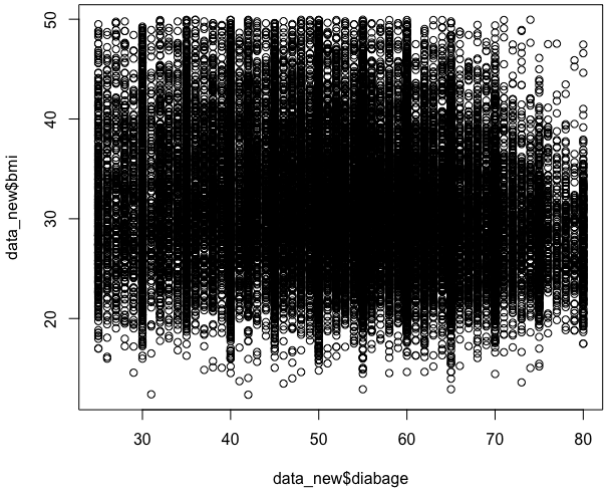


Figure 1: Scatter plot for Diabetes age vs. BMI in the dataset.

Words of Caution

- Just because patterns are seen graphically may not mean a significant relationship between the variables.
- Additional statistical analysis is required : **Inferential Statistics**.
- An elementary course book of statistics could help in this regard.
- Suggestion: Stanford Probability and Statistics Course.

- Next Lecture: Reference Managers.