PES University, Bangalore

Established under Karnataka Act No. 16 of 2013

UE20CS312 - Data Analytics - Worksheet 1b - Correlation Analysis

Designed by Vibha Masti, Dept. of CSE - vibha@pesu.pes.edu

Correlation

Correlation is a measure of the strength and direction of relationship that exists between two random variables and is measured using correlation coefficient. Correlation can assist data scientists to choose the variables for model building that is used for solving an analytics problem.

There are different types of correlation coefficients, based on the nature of the data being compared:

- Between two continuous (interval, ratio) random variables - *Pearson's Product Moment Correlation Coefficient*
- Between two ordinal random variables - *Spearman-Rank Correlation Coefficient*
- Between a continuous RV and a dichotomous RV - *Point Bi-Serial Correlation Coefficient*
- Between two binary random variables - *Phi Coefficient*

Road Accidents

India is the world's second-most populous country with a population of around 1.2 billion people (as of July 2022). Roads are a very important mode of transport in India, spanning over 6.2 million kilometers of length, making it the country with the second-largest road network, after the United States of America. (Source: Wikipedia). With India trying to modernize its road infrastructure, there is still the problem of frequent road accidents.

Road accidents in India is a major cause of death and injury. The NCRB (National Crime Records Bureau) of India collects detailed data on traffic accidents and collisions annually. Please download the dataset from the GitHub repository that contains road accident data in India from 2016. The data was obtained from this kaggle dataset.

Data Dictionary

```
S. No.: Serial number
State/ UT: name of state/union terrirory in India
Fine/Clear - Total Accidents: total accidents per state/UT in Fine/Clear weather conditions
Fine/Clear - Persons Killed: total fatalities per state/UT in Fine/Clear weather conditions
Fine/Clear - Persons Injured: total injured people per state/UT in Fine/Clear weather conditions
Mist/ Foggy - Total Accidents: total accidents per state/UT in Mist/Foggy weather conditions
Mist/ Foggy - Persons Killed: total fatalities per state/UT in Mist/Foggy weather conditions
Mist/ Foggy - Persons Injured: total injured people per state/UT in Mist/Foggy weather conditions
Cloudy - Total Accidents: total accidents per state/UT in Cloudy weather conditions
Cloudy - Persons Killed: total fatalities per state/UT in Cloudy weather conditions
Cloudy - Persons Injured: total injured people per state/UT in Cloudy weather conditions
Rainy - Total Accidents: total accidents per state/UT in Rainy weather conditions
Rainy - Persons Killed: total fatalities per state/UT in Rainy weather conditions
Rainy - Persons Injured: total injured people per state/UT in Rainy weather conditions
Snowfall - Total Accidents: total accidents per state/UT in Snowfall weather conditions
```

```
Snowfall - Persons Killed: total fatalities per state/UT in Snowfall weather conditions
Snowfall - Persons Injured: total injured people per state/UT in Snowfall weather conditions
Hail/Sleet - Total Accidents: total accidents per state/UT in Hail/Sleet weather conditions
Hail/Sleet - Persons Killed: total fatalities per state/UT in Hail/Sleet weather conditions
Hail/Sleet - Persons Injured: total injured people per state/UT in Hail/Sleet weather conditions
Dust Storm - Total Accidents: total accidents per state/UT in Dust Storm weather conditions
Dust Storm - Persons Killed: total fatalities per state/UT in Dust Storm weather conditions
Dust Storm - Persons Injured: total injured people per state/UT in Dust Storm weather conditions
Others - Total Accidents: total accidents per state/UT in Other weather conditions
Others - Persons Killed: total fatalities per state/UT in Other weather conditions
Others - Persons Injured: total injured people per state/UT in Other weather conditions
```

Points

The problems in this worksheet are for a total of 10 points with each problem having a different weightage.

- *Problem 1*: 2 points
- *Problem 2*: 2 points
- *Problem 3*: 3 points
- *Problem 4*: 1.5 points
- *Problem 5*: 1.5 points

Problem 1 (2 points)

Find the total number of accidents in each state for the year 2016 and display your results. Make sure to display all rows while printing the dataframe. Print only the necessary columns. (Hint: use the grep command to help filter out column names).

```
library(ggpubr)
library(dplyr)
df <- read.csv('road_accidents_india_2016.csv', row.names=1)
```

Problem 2 (2 points)

Find the (fatality rate $= \dfrac{\text{total number of deaths}}{\text{total number of accidents}}$) in each state. Find out if there is a significant linear correlation at a significance of $\alpha = 0.05$ between the *fatality rate* of a state and the *mist/foggy rate* (fraction of total accidents that happen in mist/foggy conditions).

Correlation between two continuous RVs: Pearson's correlation coefficient. Pearson's correlation coefficient between two RVs x and y is given by:

$$\rho = \frac{\text{Covariance}(x, y)}{\sigma_x \sigma_y}$$

where σ is the standard deviation of a variable.

Plot the fatality rate against the mist/foggy rate. (Hint: use the `ggscatter` library to plot a scatterplot with the confidence interval of the correlation coefficient).

Plot the fatality rate and mist/foggy rate (see this and this for R plot customization).

Problem 3 (3 points)

Rank the states based on total accidents and total fatalities (give a rank of 1 to the state that has the highest value of a property). You are free to use any tie-breaking method for assigning ranks.

Find the Spearman-Rank correlation coefficient between the two rank columns and determine if there is any statistical significance at a significance level of $\alpha = 0.05$. Also test the hypothesis that the correlation coefficient is at least 0.2.

The t statistic is given by

$$t = \frac{r_s - \rho_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}}$$

Where r_s is the calculated Spearman-Rank correlation coefficient and ρ_s is the value of the population correlation coefficient being tested against.

Problem 4 (1.5 points)

Convert the column `Hail.Sleet...Total.Accidents` to a binary column as follows. If a hail/sleet accident has occurred in a state, give that state a value of 1. Otherwise, give it a value of 0. Once converted, find out if there is a significant correlation between the `hail_accident_occcur` binary column created and the number of rainy total accidents for every state.

Calculate the point bi-serial correlation coefficient between the two columns. (Hint: it is equivalent to calculating the Pearson correlation between a continuous and a dichotomous variable. You could also use the `ltm` package's `biserial.cor` function).

Problem 5 (1.5 points)

Similar to in Problem 4, create a binary column to represent whether a dust storm accident has occurred in a state (1 = occurred, 0 = not occurred). Convert the two columns into a contingency table.

Calculate the phi coefficient of the two tables. (Hint: use the `psych` package).