Correlation Analysis

INTRODUCTION

A statistical metric known as **correlation** may be used to characterize the relationship, or how two variables are related to one another. It explains how changing one variable affects the other.

Correlation coefficients offer valuable insights into the relationship between variables, with Karl Pearson's and Spearman's Rank correlation coefficients being prominent measures.

Karl Pearson's Correlation Coefficient (Pearson's r):

This coefficient assesses the strength and direction of linear relationships between quantitative variables.

Represented by $\rho\ (x, y)$, it ranges from -1 to 1, indicating perfect positive, perfect negative, or no linear relationships.

Computation involves covariance and standard deviations, using the formula:

$$\rho(x,y) = \frac{\text{Covariance}(x,y)}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sqrt{\sum (y_i - \bar{y})^2}}}$$

Spearman's Rank Correlation Coefficient:

Utilized for ordinal or non-parametric data, where variables are ranked instead of measured.

Denoted by $\rho\ (x, y)$, it also ranges from -1 to 1, signifying perfect positive, perfect negative, or no monotonic relationships.

Calculation involves comparing the ranks of corresponding variables and utilizing the formula:

$$\rho(x,y) = 1 - \frac{6 \sum d^2}{n(n^2-1)}$$

Here, d represents the difference between ranks, and n is the number of observations.

These coefficients serve as vital statistical tools, aiding researchers in comprehending the interplay between variables within their datasets.

The different Types of correlation are:

- Positive Correlation: If two variables are either increasing or decreasing in parallel.
- Negative Correlation: If there is an increase and the other decreases or vice versa.
- Zero Correlation: If change of a variable does not have any effect on another.
- Perfect Correlation: When the change of one variable has the same change in the other. These are two types of perfect correlation which are perfectly positive and

perfectly negative. The former is when the change of increase in both variables are same whereas the latter is when the change of decrease in both variables are same.

UNIVARIATE ANALYSIS

1) LakeHuron Data

DATA DESCRIPTION

Annual measurements of the level, in feet, of Lake Huron 1875–1972.

Data set with a time series of length 98.

2) Nile Data

DATA DESCRIPTION

Measurements of the annual flow of the river Nile at Aswan (formerly `Assuan`), 1871–1970, in $10^8 m^3$, "with apparent changepoint near 1898" (Cobb(1978), Table 1, p.249).

Data Set with a time series of length 100.

OBJECTIVE

Conduct a correlation analysis by examining the goodness of fit to the normal distribution and draw conclusions based on the outcome of the test.

METHODOLOGY

To assess normality numerically, various statistical tests are employed due to limitations in graphical interpretation. These tests include the Shapiro-Wilk, Jarque-Bera, Anderson-Darling, and Kolmogorov-Smirnov tests. Hypothesis testing is then applied to ascertain the normality of the distribution.

H0: The variable's distribution is not significantly different from a normal distribution.

H1: The variable's distribution is significantly different from a normal distribution.

Employing a significance level of 5%, we utilize these tests to draw conclusions regarding the normality of the data distribution.

(i) R-CODE AND INTERPRETATION OF RESULTS: LakeHuron data Set

Input

Understanding the data set

head(LakeHuron)

attach(LakeHuron)

Checking for normality Graphically

hist(LakeHuron)

Shapiro-Wilk test

shapiro.test(LakeHuron)

Jarque-Bera test

library(tseries)

jarque.bera.test(LakeHuron)

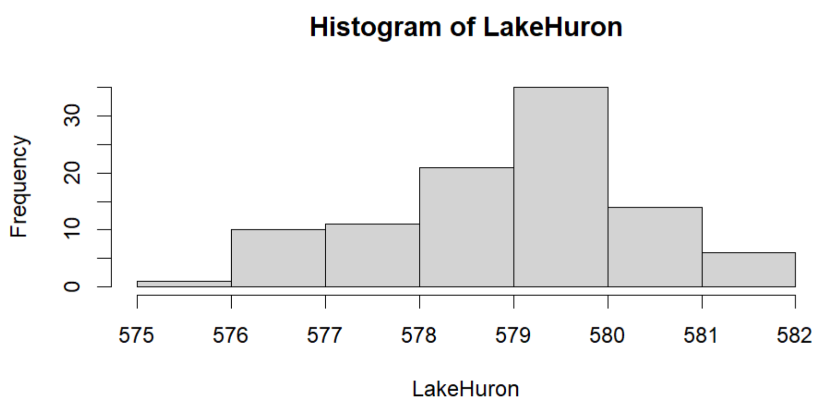
Anderson-Darling test

library(nortest)

ad.test(LakeHuron)

Output

```
> head(LakeHuron)
[1] 580.38 581.86 580.97 580.80 579.79 580.39
```



Interpretation of result: While the histogram displays a bell-shaped curve, suggesting a degree of normality in the dataset, it is insufficient to conclusively confirm normality. Thus, numerical evidence is necessary to substantiate the assumption of normality.

```
> shapiro.test(LakeHuron)

Shapiro-Wilk normality test
```

```
data:  LakeHuron
W = 0.98492, p-value = 0.3271
```

Interpretation of result:

Using the p-value we test for significance.

Since p-value > 0.05, We do not Reject H0

Thus, there is no significance identified. Hence The data set follows normal distribution.

```
> jarque.bera.test(LakeHuron)

Jarque Bera Test

data:  LakeHuron
X-squared = 1.3433, df = 2, p-value = 0.5109
```

Interpretation of result:

Using the p-value we test for significance.

Since p-value > 0.05, We do not Reject H0

Thus, there is no significance identified. Hence The data set follows normal distribution.

```
> ad.test(LakeHuron)

Anderson-Darling normality test

data:  LakeHuron
A = 0.43831, p-value = 0.2888
```

Interpretation of result:

Using the p-value we test for significance.

Since p-value > 0.05, We do not Reject H0

Thus, there is no significance identified. Hence The data set follows normal distribution.

CONCLUSION

Based on the results of the Shapiro-Wilk, Jarque-Bera, and Anderson-Darling tests, it can be inferred that the dataset LakeHuron adheres to a normal distribution, suggesting that the LakeHuron dataset exhibits normality.

ii) R-CODE AND INTERPRETATION OF RESULTS: Nile data Set

Input

Understanding the data set

head(Nile)

attach(Nile)

Checking for normality Graphically

hist(Nile)

Shapiro-Wilk test

shapiro.test(Nile)

Jarque-Bera test

library(tseries)

jarque.bera.test(Nile)

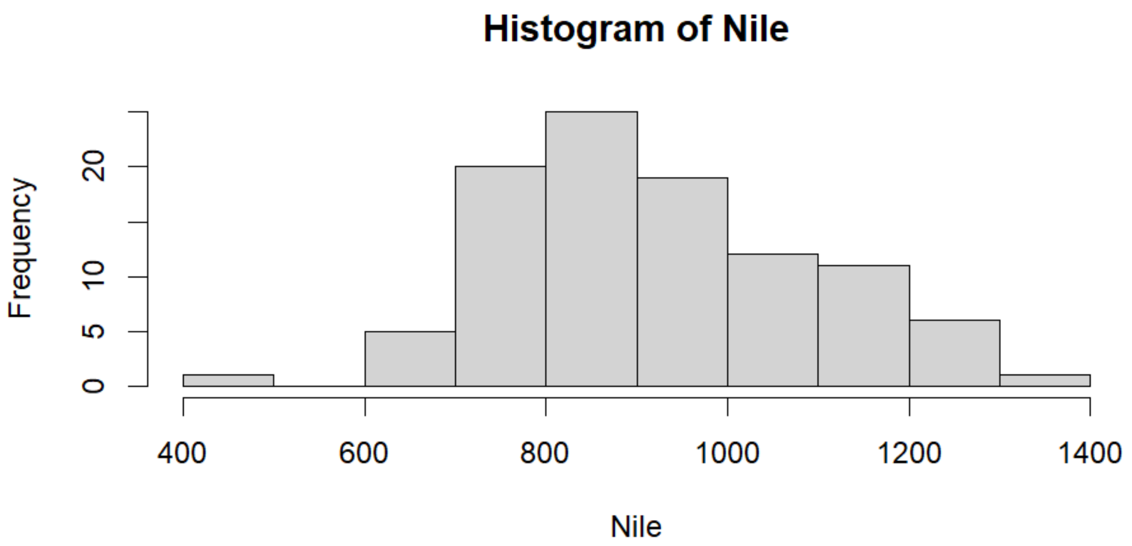
Anderson-Darling test

library(nortest)

ad.test(Nile)

Output

```
> head(Nile)
[1] 1120 1160  963 1210 1160 1160
```



Interpretation of result: While the histogram displays a bell-shaped curve, suggesting a degree of normality in the dataset, it is insufficient to conclusively confirm normality. Thus, numerical evidence is necessary to substantiate the assumption of normality.

```
> shapiro.test(Nile)

Shapiro-Wilk normality test

data:  Nile
W = 0.97343, p-value = 0.04072
```

Interpretation of result:

Using the p-value we test for significance.

Since p-value < 0.05, We Reject H0

Thus, there is significance identified. Hence The data set does not follow normal distribution.

```
> jarque.bera.test(Nile)

Jarque Bera Test

data:  Nile
X-squared = 2.1194, df = 2, p-value = 0.3466
```

Interpretation of result:

Using the p-value we test for significance.

Since p-value > 0.05, We do not Reject H0

Thus, there is no significance identified. Hence The data set follows normal distribution.

```
> ad.test(Nile)

Anderson-Darling normality test

data:  Nile
A = 1.032, p-value = 0.009821
```

Interpretation of result:

Using the p-value we test for significance.

Since p-value < 0.05, We Reject H0

Thus, there is significance identified. Hence The data set does not follow normal distribution.

CONCLUSION

Based on the results of the Shapiro-Wilk and Anderson-Darling tests, it can be inferred that the dataset Nile does not adhere to a normal distribution, suggesting that the Nile dataset does not exhibit normality.

Multivariate Analysis

DATA DESCRIPTION: LifeCycleSavings

Data on the savings ratio 1960–1970.

A data frame with 50 observations on 5 variables.

[,1] sr numeric aggregate personal savings
[,2] pop15 numeric % of population under 15
[,3] pop75 numeric % of population over 75
[,4] dpi numeric real per-capita disposable income
[,5] ddpi numeric % growth rate of dpi

Details

Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio (aggregate personal saving divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old. The data are averaged over the decade 1960–1970 to remove the business cycle or other short-term fluctuations.

OBJECTIVE

Evaluate the multivariate normality of a dataset using statistical tests and recommend additional tests based on the outcomes.

METHODOLOGY

To evaluate multivariate normality, we often resort to 3D plots for bivariate data. However, as the number of features increases, visualization becomes challenging, and we typically rely on matrix scatter diagrams instead.

To numerically assess multivariate normality, several statistical tests are available, including the Mardia, Royston, Hinze-Zirkler, Doorkin-Hansen, and Energy tests.

Employing hypothesis testing

H0: There is no significant difference from a multivariate normal distribution.

H1: There is significant difference from a multivariate normal distribution.

we conduct our analysis with a significance level set at 5%, these tests guide our inference regarding the distribution of the data.

R-CODE AND INTERPRETATION OF RESULTS: EuStockMarkets data Set

Input

```
# Understanding the Dataset

library(MVN)
```

head(LifeCycleSavings)

Checking for multivariate normality Graphically

c=cor(LifeCycleSavings,method="pearson")#Correlation matrix will be given

round(c,2)

#pairs.panels -> to graphically show the correlation

library(psych)

pairs.panels(LifeCycleSavings)

corrplot.mixed(c,upper="number",lower="circle")

#Mardia test

mvn(LifeCycleSavings,mvnTest="mardia")

Hz test

mvn(LifeCycleSavings,mvnTest="hz")

Energy test

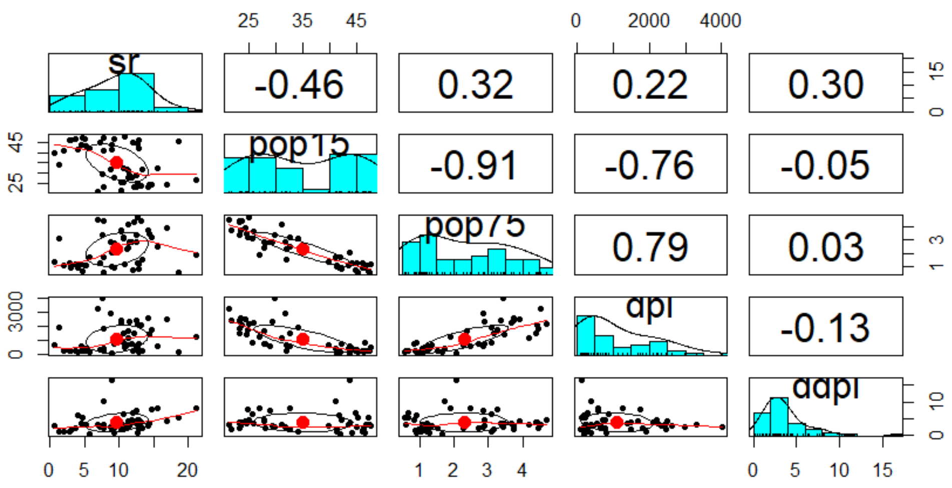
mvn(LifeCycleSavings,mvnTest="energy")

Output

```
> head(LifeCycleSavings)
             sr pop15 pop75     dpi ddpi
Australia 11.43 29.35  2.87 2329.68 2.87
Austria   12.07 23.32  4.41 1507.99 3.93
Belgium   13.17 23.80  4.43 2108.47 3.82
Bolivia    5.75 41.89  1.67  189.13 0.22
Brazil    12.88 42.19  0.83  728.47 4.56
Canada     8.79 31.72  2.85 2982.88 2.43
```

```
> round(c,2)
         sr pop15 pop75   dpi  ddpi
sr     1.00 -0.46  0.32  0.22  0.30
pop15 -0.46  1.00 -0.91 -0.76 -0.05
pop75  0.32 -0.91  1.00  0.79  0.03
dpi    0.22 -0.76  0.79  1.00 -0.13
ddpi   0.30 -0.05  0.03 -0.13  1.00
```

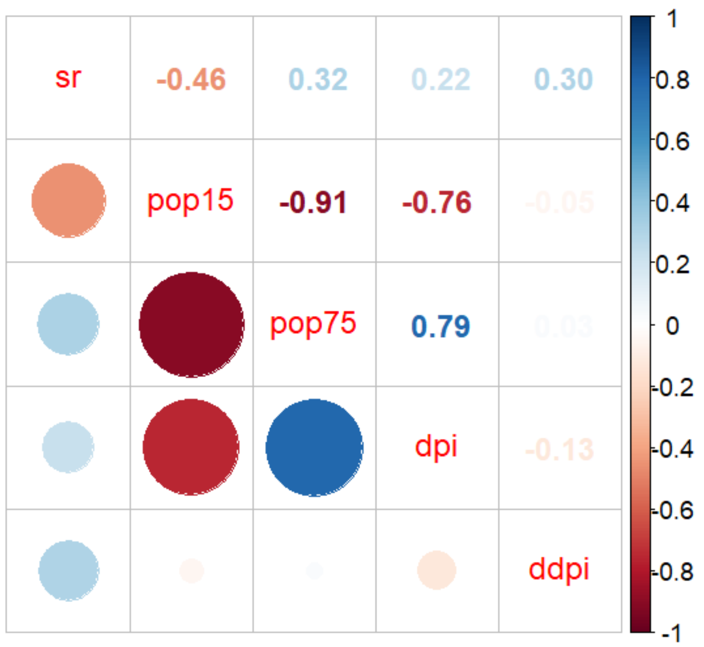
```
> pairs.panels(EuStockMarkets)
```



Interpretation of graph:

Scatter plots representing each pair of variables are displayed below the diagonal, allowing v isual inspection of their relationships.
Above the diagonal, correlation coefficients are computed to assess the correlation between pairs of variables.
Along the diagonal, histograms illustrate the distribution of each variable, aiding in the evalu ation of normality. Univariate normality tests, such as the Anderson-Darling test, are conduct ed to further validate the normality assumption for each variable.
Among all pairs Pop15 and pop75 are highly negatively correlated and ddpi and pop75 are th e least correlated.

```
> corrplot.mixed(cr,upper="number",lower="circle")
```



Interpretation of graph:

The heatmap below the diagonal visually represents varying correlation values between diffe rent pairs of variables, indicated by differences in color intensity. Similarly, above the diagon al, correlation coefficient values for each pair are depicted, akin to the previous graph.

```
> mvn(LifeCycleSavings,mvnTest="mardia")
$multivariateNormality
             Test          Statistic              p value Result
1 Mardia Skewness 114.224217857025 2.47688793880192e-10     NO
2 Mardia Kurtosis  2.8063692862095  0.00501032357416742     NO
3             MVN             <NA>                 <NA>     NO

$univariateNormality
              Test  Variable Statistic   p value Normality
1 Anderson-Darling    sr        0.3985  0.3531      YES
2 Anderson-Darling   pop15      2.2978  <0.001      NO
3 Anderson-Darling   pop75      1.4713   7e-04      NO
4 Anderson-Darling    dpi       2.5155  <0.001      NO
5 Anderson-Darling   ddpi       2.3033  <0.001      NO

$Descriptives
       n      Mean    Std.Dev  Median   Min     Max     25th       75th 
Skew     Kurtosis
sr    50    9.6710   4.480407  10.510  0.60   21.10   6.9700   12.6175 -0.
005569743 -0.32369517
pop15 50   35.0896   9.151727  32.575 21.44   47.64  26.2150   44.0650 -0.
001188000 -1.68025919
pop75 50    2.2930   1.290771   2.175  0.56    4.70   1.1250    3.3250  0.
305162641 -1.33181496
dpi   50 1106.7584 990.868889 695.665 88.94 4001.89 288.2075 1795.6225  0.
949629305 -0.09116257
ddpi  50    3.7576   2.869871   3.000  0.22   16.71   2.0025    4.4775  2.
140592209  6.39547229
```

Interpretation of result:

i) For Mardia Skewness

Using the p-value we test for significance.

Since p-value < 0.05, We Reject H0

Thus, there is significance identified. Hence The data set does not follow normal distribution.

ii) For mardia Kurtosis

Using the p-value we test for significance.

Since p-value < 0.05, We Reject H0

Thus, there is significance identified. Hence The data set does not follow normal distribution.

To confirm our results, we may further choose to perform other such tests such as Hinze-Zirkler test and Energy test.

```
> mvn(LifeCycleSavings,mvnTest="hz")
$multivariateNormality
           Test       HZ      p value MVN
1 Henze-Zirkler 1.353345 3.723682e-08  NO
```

```
$univariateNormality
              Test  Variable Statistic   p value Normality
1 Anderson-Darling    sr        0.3985  0.3531      YES
2 Anderson-Darling   pop15      2.2978  <0.001      NO
3 Anderson-Darling   pop75      1.4713   7e-04      NO
4 Anderson-Darling    dpi       2.5155  <0.001      NO
5 Anderson-Darling   ddpi       2.3033  <0.001      NO

$Descriptives
       n      Mean    Std.Dev  Median   Min     Max     25th       75th
Skew    Kurtosis
sr     50    9.6710   4.480407  10.510  0.60   21.10   6.9700   12.6175 -0.
005569743 -0.32369517
pop15  50   35.0896   9.151727  32.575 21.44   47.64  26.2150   44.0650 -0.
001188000 -1.68025919
pop75  50    2.2930   1.290771   2.175  0.56    4.70   1.1250    3.3250  0.
305162641 -1.33181496
dpi    50 1106.7584 990.868889 695.665 88.94 4001.89 288.2075 1795.6225  0.
949629305 -0.09116257
ddpi   50    3.7576   2.869871   3.000  0.22   16.71   2.0025    4.4775  2.
140592209  6.39547229
```

```
> mvn(LifeCycleSavings,mvnTest="energy")
$multivariateNormality
         Test Statistic p value MVN
1 E-statistic  1.733879       0  NO

$univariateNormality
              Test  Variable Statistic   p value Normality
1 Anderson-Darling    sr        0.3985  0.3531      YES
2 Anderson-Darling   pop15      2.2978  <0.001      NO
3 Anderson-Darling   pop75      1.4713   7e-04      NO
4 Anderson-Darling    dpi       2.5155  <0.001      NO
5 Anderson-Darling   ddpi       2.3033  <0.001      NO

$Descriptives
       n      Mean    Std.Dev  Median   Min     Max     25th       75th
Skew    Kurtosis
sr     50    9.6710   4.480407  10.510  0.60   21.10   6.9700   12.6175 -0.
005569743 -0.32369517
pop15  50   35.0896   9.151727  32.575 21.44   47.64  26.2150   44.0650 -0.
001188000 -1.68025919
pop75  50    2.2930   1.290771   2.175  0.56    4.70   1.1250    3.3250  0.
305162641 -1.33181496
dpi    50 1106.7584 990.868889 695.665 88.94 4001.89 288.2075 1795.6225  0.
949629305 -0.09116257
ddpi   50    3.7576   2.869871   3.000  0.22   16.71   2.0025    4.4775  2.
140592209  6.39547229
```

Interpretation of result:

Note that for both the Hinze-Zirkler test and the Energy test, the p-value is less than 0.05, indicating rejection of the null hypothesis. This suggests significant evidence that the distribution of the data deviates from multivariate normality.

CONCLUSION

The Mardia test revealed that the data does not adhere to multivariate normality, and subsequent tests such as Hinze-Zirkler and Energy corroborated this finding. Given the consistent results across all tests indicating significant deviation from multivariate normality, it can be concluded that the LifeCycleSavings dataset does not conform to a multivariate normal distribution.