WORKSHOP ON SPSS

SHIJINA.A.S

ASSISTANT PROFESSOR

MANNANIYA COLLEGE OF ARTS AND SCIENCE, PANGODE

Statistical Packages

- Computer programmes written for statistical analysis
- Simplified the work of researchers who have to do statistical analysis in their research work
- SPSS, MSTAT, INDOSTAT, MINITAB, EXCEL

SPSS

- Relatively comprehensive package for use in Economics, Business and Research
- Software package used for logical batched and non-batched statistical analysis.
- First launched in 1968
- Acquired by IBM in 2009
- Officially known as IBM SPSS
- Software for editing and analyzing all sorts of data
- Can open all file formats that are commonly used for structured data- spreadsheets, plain text files, relational databases, etc

Measures of Scaling

1. Nominal Scale

Divides the data into two or more mutually exclusive and exhaustive categories

2. Ordinal Scale

places events in order, but the interval of the scale need not be equal in terms of any rule

3. Interval Scale

All the characteristics of nominal and ordinal scales

It has the property of equality of interval

4. Ratio Scale

measures the actual amount of variables

It is more flexible and most powerful

Nominal	Ordinal	Interval	Ratio
Classification data: e.g. Male / Female No ordering: e.g. it makes no sense to state that M > F Arbitrary labels: e.g., M/F, 0/1, etc	Ordered but differences between values are not important e.g., Political parties on left to right spectrum given labels 0, 1, 2 e.g., Likert scales, rank on a scale of 1..5 your degree of satisfaction e.g., Restaurant ratings	Ordered, constant scale, but no natural zero Differences make sense, but ratios do not e.g. Temperature (C,F), Dates	Ordered, constant scale, natural zero e.g., Height, Weight, Age, Length

BASIC CONCEPTS

- **Population**

  Collection of all individuals or objects or items under study and denoted by N

- **Sample**

  A part of a population and denoted by n

- **Variable**

  Characteristic of an individual or object.

  - Qualitative and Quantitative variables

- **Parameter**

  Characteristic of the population

- **Statistic**

  Characteristic of the sample

The Concept of P Value

- Given the observed data set, the P value is the smallest level for which the null hypothesis is rejected (and the alternative is accepted)

- P value 0.000 to 0.010 then reject NH at 1% (Denoted by **)

  i.e. Highly Significant

- P value 0.011 to 0.050 then reject NH at 5% (Denoted by *)

  i.e. Significant

- P value 0.051 to 1.000 then accept NH at 5% (Do not put star)

  i.e. Not Significant

- **0.000 => < 0.001****

Non-Parametric Tests

- In some situations, the practical data may be non-normal and/or it may not be possible to estimate the parameter(s) of the data
- The test which are used for such situations are called non-parametric tests
- Since these tests are based on the data which are free from distribution and parameter, these tests are known as non-parametric(NP) test or Distribution Free tests
- NP test can be used even for nominal data (qualitative data like greater or less, etc.) and ordinal data, like ranked data.
- NP test required less calculation, because there is no need to compute parameters.

List of Non-Parametric Tests

1. One-sample test
   - One sample sign test
   - Chi-square one sample test
   - Kolmogorov-Smirnov test
2. Two related samples tests
   - Two samples sign test
   - Wilcoxon Matched-pairs signed –rank test
3. Two independent samples test
   - Chi-Square test for two independent samples
   - Mann-Whitney U test
   - Kolmogorov-Smirnov two sample test

List of Non-Parametric Tests

4 K Related Samples test

- Friedman Two way Analysis of Variance by Ranks
- The Coehran Q test

5. K Independent samples

- Chi-Square test for k Independent samples
- The extension of the Median test
- Kruskal-Wallis one-way Analysis of Variance by Rank

One sample sign test

- This test is applied to a situation where a sample is taken from a population which has a continuous symmetrical distribution and known to be non-normal such that the probability of having a sample values less than the mean value as well as probability of having a sample values more than the mean value(p) is ½.

- Classified into four categories
  1. One-tailed one-sample sign tests for small sample
  2. Two-tailed one-sample sign tests for small sample
  3. One-tailed one-sample sign tests for large sample
  4. Two-tailed one-sample sign tests for large sample

Kolmogorov-smirnov test

- It is similar to the chi-square test to do goodness of fit of a given set of data to an assumed distribution
- This test is more powerful for small samples whereas the chi-square test is suited for large sample

- H_0: The given data follow an assumed distribution

  H_1: The given data do not follow an assumed distribution

- K-S test is an one-tailed test. Hence if the calculated value of D is more than the theoretical value of D for a given significance level, then reject H_0 ; otherwise accept H_0

Two samples sign test

- Two samples sign test is applied to a situation, where two samples are taken from two populations which have continuous symmetrical distributions and known to be non-normal
- Modified sample value, $Z_i = +$ if $X_i > Y_i$

  $= -$ if $X_i < Y_i$

  $= 0$ if $X_i = Y_i$
- Classified into four categories
  1. One-tailed two-sample sign tests with binomial distribution
  2. Two-tailed two-sample sign tests with binomial distribution
  3. One-tailed two-sample sign tests with normal distribution
  4. Two-tailed two-sample sign tests with normal distribution

The Wilcoxon Matched-pairs signed-ranks test

- The Wilcoxon test is a most useful test for behavioral scientist
- Let d_i = the difference score for any matched pair
- Rank all the d_i without regard to sign
- T = Sum of rank with less frequent sign
- Compute $Z = [T - E(T)]/SD(T)$

Mann-Whitney U Test

- Mann-Whitney U test is an alternate to the two sample t-test
- This test is based on the ranks of the observations of two samples put together
- Alternate name for this test is **Rank-Sum Test**
- Let R_1 = The sum of the ranks of the observations of the first sample
- Let R_2 = The sum of the ranks of the observations of the second sample
- Objective: To check whether the two samples are drawn from different populations having the same distribution
- Compute $Z = [U - E(U)]/SD(U)$

  where $U = n_1n_2 + [n_1(n_1 + 1)/2] - R_1$
- or $U = n_1n_2 + [n_2(n_2 + 1)/2] - R_2$

Correlation and Regression Analysis

- The Chi-square test measures the association between two or more variables.This test is applicable only when data is on nominal scale.
- Correlation and Regression analysis is used for measuring the relationship between two variables measured on interval or ratio scale.

Correlation Analysis

- Correlation analysis is a statistical technique used to measure the magnitude of linear relationship between two variables.
- Correlation analysis cannot be used in isolation to describe the relationship between variables.
- It can be used along with regression analysis to determine the nature of the relationship between two variables.
- Thus correlation analysis can be used for further analysis
- Two prominent types of correlation Coefficient are
  - Pearson Product Moment correlation coefficient
  - Spearman's Rank correlation coefficient
- Testing the significance of correlation coefficient
- Type I $H_0: \rho = 0$ and $H_1: \rho \neq 0$

  Type II $H_0: \rho = r$ and $H_1: \rho \neq r$

  Type III $H_0: r_1 = r_2$ and $H_1: r_1 \neq r_2$

Regression Analysis

- Regression analysis is used to predict the nature and closeness of relationships between two or more variables
- It evaluate the causal effect of one variable on another variable
- It used to predict the variability in the dependent (or criterion) variable based on the information about one or more independent (or predictor) variables.
- Two variables : Simple or Linear Regression Analysis
- More than two variables : Multiple Regression Analysis

Linear Regression Analysis

- Linear regression : $Y = \alpha + \beta X$

  Where Y : Dependent variable

  X : Independent variable

  α and β : Two constants are called regression coefficients

  β : Slope coefficient i.e. the change in the value of Y with the corresponding change in one unit of X

  α : Y intercept when X = 0

- R^2 : The strength of association i.e. to what degree that the variation in Y can be explained by X.
- $R^2 = 0.10$ then only 10% of the total variation in Y can be explained by the variation in X variables

THANK YOU