Training on Data Management and Analysis Using R Software

The East Africa Community (EAC) and Institut National de la Statistique du Burundi (INSBU)

Royal Palace Hotel (4th to 16th September 2023)

Bujumbura, Burundi

Day-5 : September, 2023

Basic Level : Statistical Analysis

National Training Workshop on R Software for INSBU Staff

Leguma Bakari
Email: leguma.bakari@eastc.ac.tz
Phone:+255 762 760 095

September 07,2023

Physical Address
Eastern Africa Statistical Training Center(EASTC)
Changanyikeni Area
P. O. Box 35103
Dar es Salaam, Tanzania

Outline

Data Analysis Introduction

Scales of Measurement

Descriptive Analysis

Inferential Analysis

Bivariate Analysis

t-tests

Data Analysis

- Data analysis process of extract useful information, draw conclusions, and support decision-making by using data.
- In statistics data analysis can be categorized in two groups
  - Descriptive analysis, and
  - Inferential analysis.
- The goal of data analysis is to turn raw data into meaningful insights that can inform business decisions, scientific research, or other areas of inquiry.
- Data analysis can be applied to a wide range of fields, including finance, healthcare, marketing, social science, and any other filed as long as data is available.
- We need data and analysis for planning phase, for monitoring and evaluation, and for informed decision marking.

Outline

Data Analysis Introduction

Scales of Measurement

Descriptive Analysis

Inferential Analysis

Bivariate Analysis

t-tests

Scales of Measurement

- Scales of measurement refer to ways in which variables/numbers are defined and categorized.
- Each scale of measurement has certain properties which in turn determines the appropriateness for use of certain statistical analyses.
- There are four scales of measurement are
  - Nominal scale,
  - Ordinal scale,
  - Interval scale, and
  - Ratio scale.
- Mathematical operation are important criteria on naming and identifying a scale of measurement.
- The mathematical operations are addition(+),subtraction(-), multiplication(×), division(÷) and inequalities($<$ or $>$).

Nominal Scale

- A nominal scale is a scale, in which numbers only serve as tags or labels for the purpose of categorizing responses.
- No any mathematical operation which can be applied in nominal variable.
- Nominal responses can not be ranked.
- Examples of nominal variables includes gender, any binary (ie yes-no) response, race, marital status e.t.c
- In nominal scale you can only compute mode value, but not median and any form of mean.

Nominal Scale Arithmetic Application

- Let us use marital status nominal variable for demonstration.
- Numbers in nominal variables are just assigned number (not original).
- The possible values for marital status variable are
  1. Single
  2. Married
  3. Widowed
  4. Divorced
- Addition (+)
  - From mathematics; 1+2=3,
  - For nominal variable this is equivalent to
  - Single+Married=Widowed; this is impossible,
  - Therefore addition operation can not be applied in nominal data.
  - Since addition can not be applied, and the mean computation involves addition in it's computation,
  - Therefore mean is an invalid statistic for nominal data.
  - Mean = $\frac{x_1+x_2+\cdots+x_n}{n}$

- Subtraction (−)
    - From mathematics; 4-2=2,
    - For nominal variable this is equivalent to
    - Divorced-Married=Married; this is impossible
    - Therefore subtraction operation can not be applied in nominal data.
- Multiplication (×)
    - From mathematics; $2 \times 2 = 4$,
    - For nominal variable this is equivalent to
    - Married × Married = Divorced; this is impossible
    - Therefore multiplication operation can not be applied in nominal data.
- Division (÷)
    - From mathematics; $3 \div 1 = 3$,
    - For nominal variable this is equivalent to
    - Widowed ÷ Single = Single; this is impossible
    - Therefore division operation can not be applied in nominal data.

- Inequality ($<$ or $>$)
  - From mathematics; $2 > 1$,
  - For nominal variable this is equivalent to
  - Married > Single; this is impossible
  - Inequality in other words implies ranking,
  - Since inequality can not be applied, and the median computation involves ranking (ascending or descending),
  - Therefore median is an invalid statistic for nominal data.
- Among three common measure of central tendency, mean median and mode,
- Mode is the only valid statistic since in involves counts (frequency)
- This conclude that, no any mathematical operation can be applied in nominal data.

Ordinal Scale

- An ordinal scale is a scale, in which numbers only serve as tags or labels for the purpose of categorizing an ordered responses.
- An ordinal scale implies that the categories must be put into an order such that each category in one class is considered greater (or less) than every category in another class.
- For ordinal variable inequality mathematical operations are only mathematical operations which can be applied.
- Example of ordinal variables includes education level, GPA class, e.t.c
- In ordinal scale you can compute mode and median values but not and any form of mean.

Ordinal Scale Arithmetic Application

- Let us use education level ordinal variable for demonstration.
- Numbers in nominal variables are just assigned number (not original).
- The possible values for education level variable are
  1. Primary
  2. Secondary
  3. College
  4. University
- Addition (+)
  - From mathematics; 1+2=3,
  - For nominal variable this is equivalent to
  - Primary+Secondary=University; this is impossible,
  - Therefore addition operation can not be applied in ordinal data.
  - Since addition can not be applied, and the mean computation involves addition in it’s computation,
  - Therefore mean is an invalid statistic for ordinal data.

- Subtraction (−)
  - From mathematics; 4-2=2,
  - For nominal variable this is equivalent to
  - University-Secondary=Secondary; this is impossible
  - Therefore subtraction operation can not be applied in ordinal data.
- Multiplication (×)
  - From mathematics; $2 \times 2 = 4$,
  - For nominal variable this is equivalent to
  - Secondary × Secondary = University; this is impossible
  - Therefore multiplication operation can not be applied in ordinal data.
- Division (÷)
  - From mathematics; $3 \div 1 = 3$,
  - For nominal variable this is equivalent to
  - College ÷ Primary = College; this is impossible
  - Therefore division operation can not be applied in ordinal data.

- Inequality ($<$ or $>$)
    - From mathematics; $2 > 1$,
    - For nominal variable this is equivalent to
    - Secondary > Primary; this is possible
    - Inequality in other words implies ranking,
    - Since inequality can be applied, and the median computation involves ranking (ascending or descending),
    - Therefore median is an valid statistic for ordinal data.
- Among three common measure of central tendency, mean median and mode,
- Mode and median are the only valid statistic for summarization.
- This conclude that, only inequality mathematical operation can be applied in ordinal data.

Interval Scale

- An interval scale is a scale of measurement whereby the presence of zero (0) does not indicate the absence of measurement.
- Zero value which do not indicate absence of measurement is also called false zero.
- For interval variable three mathematical operations are applicable which are.
    - Inequalities,
    - Addition, and
    - Subtraction.
- Example of interval variable is temperature and deviations.
- In interval scale you can compute mode, median and arithmetic mean values, but not geometric mean.

Interval Scale Arithmetic Application

- Let us use degree centigrade interval variable for demonstration.
- Numbers in interval variable are original number (from measurement).
- The possible values for education level variable are
- $15^o C, -9^o C, 0^o C, 5^o C, -15^o C, 25^o C$
- Addition ($+$)
  - From mathematics; $15^o C + 5^o C = 30^o C$, which is possible and valid
  - Since addition can be applied, and the mean computation involves addition in it's computation,
  - Therefore mean is a valid statistic for interval data.
  - Mean = $\frac{x_1+x_2+\cdots+x_n}{n}$

- Subtraction ($-$)
  - From mathematics; $15^oC - 5^oC = 10^oC$, which is possible and valid
  - Any computation which involves subtraction will become valid due to this property.
- Multiplication ($\times$)
  - From mathematics; $15^oC \times 0^oC = ??$, which is impossible and not valid
  - Any computation which involves multiplication (ie geometric mean) will become invalid due to this property.
  - Geometric mean = $\sqrt[n]{y_1 \times y_2 \times \cdots \times y_n}$
- Division ($\div$)
  - From mathematics; $15^oC \div -5^oC = ??$, which is impossible and not valid
  - Any computation which involves division (ie harmonic mean) will become invalid due to this property.
  - Harmonic mean = $\frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}$

- Inequality ($<$ or $>$)
  - From mathematics; $15^oC > 5^oC$, which is possible and valid
  - Inequality in other words implies ranking,
  - Since inequality can be applied, and the median computation involves ranking (ascending or descending),
  - Therefore median is an valid statistic for interval data.
- All three common measure of central tendency, arithmetic mean median and mode, are valid statistic for summarization.
- Geometric mean and harmonic mean can not be applied.
- This conclude that only inequality, addition and subtraction mathematical operation can be applied in nominal data.

Ratio Scale

- A ratio scale is a of measurement whereby the presence of zero (0) indicates the absence of measurement.
- Zero value which indicates absence of measurement is also called true zero.
- For ratio variable all five forms mathematical operations can be applicable.
- Examples of ratio variables are height, weight, age e.t.c
- In ratio scale you can compute many statistics include mode, mean and arithmetic mean values, geometric mean, etc.

Ratio Scale Arithmetic Application

- Let us use length ratio variable for demonstration.
- Numbers in interval variable are original number (from measurement).
- The possible values for length variable are
- $150\ cm, 80\ cm, 130\ cm, 50\ cm$
- Addition ($+$)
  - From mathematics; $150\ cm + 50\ cm = 200\ cm$, which is possible and valid
  - Since addition can be applied, and the mean computation involves addition in it's computation,
  - Therefore mean is a valid statistic for ratio data.
  - Mean = $\frac{x_1+x_2+\cdots+x_n}{n}$

- Subtraction ($-$)
    - From mathematics; $150\ cm - 50\ cm = 100\ cm$, which is possible and valid
    - Any computation which involves subtraction will become valid due to this property.
- Multiplication ($\times$)
    - From mathematics; $50\ cm \times 5\ cm = 250\ cm^2$, which is possible and valid
    - Any computation which involves multiplication (ie geometric mean) will become valid due to this property.
    - Geometric mean = $\sqrt[n]{y_1 \times y_2 \times \cdots \times y_n}$
- Division ($\div$)
    - From mathematics; $50\ cm \div 5\ cm = 10$, which is possible and valid
    - Any computation which involves division (ie harmonic mean) will become valid due to this property.
    - Harmonic mean = $\frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}$

- Inequality ($<$ or $>$)
  - From mathematics; $50\ cm > 5\ cm$, which is possible and valid
  - Inequality in other words implies ranking,
  - Since inequality can be applied, and the median computation involves ranking (ascending or descending),
  - Therefore median is an valid statistic for ratio data.
- All three common measure of central tendency, arithmetic mean median and mode, are valid statistic for summarization.
- Geometric mean and harmonic mean can also be applied.
- This conclude that, all forms mathematical operation can be applied in nominal data.

Scale of Measurement Summary

- You can first classify scale of measurements into two groups, namely categorical and measured data.
- Categorical data is a variable with a fixed number of responses, this can either be in nominal or ordinal scale.
  - In nominal scale the categories can not be ranked, eg male and female in gender variable .
  - In ordinal scale the categories can be ranked,eg agree, neutral, disagree in likert scale variables.
- Measured data is any data with a unit of measurement i.e cm,kg, etc, which can either be interval or ratio scale.
  - Interval data is a measured data with false zero such as temperature.
  - Ratio data is a data with true zero such as height, weight etc.
- All data which are in ratio or interval scale they can be converted to categorical data.
- When they are converted they become ordinal data but not nominal, eg GPA to GPA class, age to age group etc.

Outline

Data Analysis Introduction

Scales of Measurement

Descriptive Analysis

Inferential Analysis

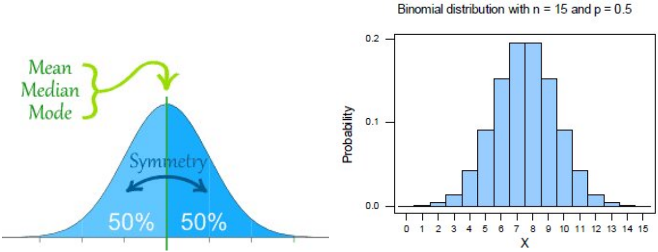
Bivariate Analysis

t-tests

Descriptive Analysis

- Descriptive data analysis is a method of summarizing and describing the main characteristics of a dataset.
- It is used to understand the basic features of the data, such as
  - measure of central tendency (mean, median, mode),
  - measure of dispersion (standard deviation, range, interquartile range), and
  - distribution of the data.
- Descriptive data analysis also involves using visualizations such as histograms, bar charts, and scatter plots to represent the data and help identify patterns and trends.
- Descriptive data analysis can be done with the use of summary statistics and visualizations,
- It is the first step in data analysis (EDA) and it provides a general understanding of the data.
- Descriptive data analysis can be performed by using various software like Excel, SPSS,Stata, R, Python, and many more.

Symmetric Graph for Data

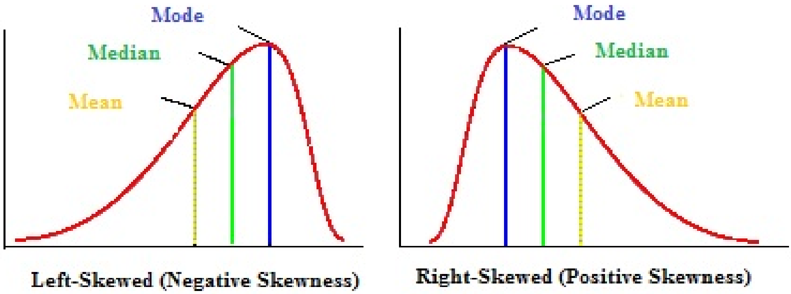


(a) Continuous Data

(b) Discrete Data

Figure 1: Symmetric Graphs

Skewned Graphs for Continuous Data

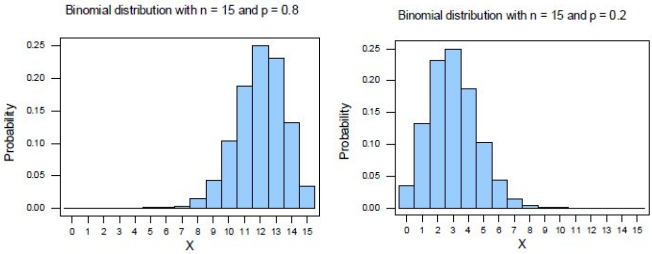


Left-Skewed (Negative Skewness)

Right-Skewed (Positive Skewness)

Figure 2: Skewness Graph for Continuous Data

Skewed Graphs for Discrete Data



(a) Left Skewed

(b) Right Skewed

Figure 3: Skewness Graph for Discrete Data

Measure of central tendency

- Central tendency is the name for measurements that look at the typical central values within a dataset.
- This does not just refer to the central value within an entire dataset, which is called the median.
- Rather, it is a general term used to describe a variety of central measurements.
- For instance, it might include central measurements from different quartiles of a larger dataset.
- Common measures of central tendency include:
  - The mean: The average value of all the data points.
  - The median: The central or middle value in the dataset.
  - The mode: The value that appears most often in the dataset.

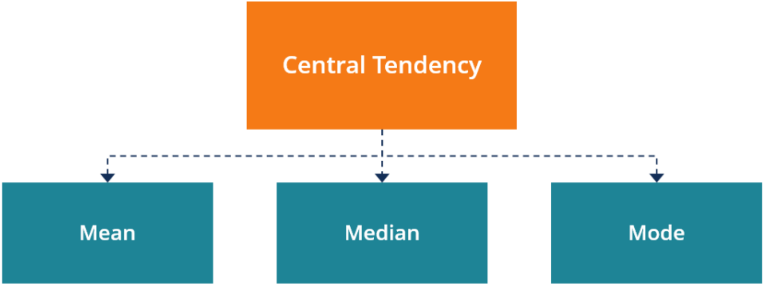


Figure 4: Measures of Central Tendency

Caution on the Use of Measure of Central Tendency

- Even though the measures above are the most commonly used to define central tendency.
- This includes but not limited to, geometric mean, harmonic mean, midrange, and geometric median.
- The selection of a central tendency measure depends on the properties of a dataset.
- Although the mean is regarded as the best measure of central tendency for quantitative data, that is not always the case.
- For example, the mean may not work well with quantitative datasets that contain
  - extremely large or small values (outliers),or
  - Skewed data.
- The extreme (outliers) values may distort the mean.

Measure of variability (or Variations, Dispersions)

- These are values which shows the scatterings (spreadness) of the data.
- It tells the variation of the data from one another and gives a clear idea about the distribution of the data.
- The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.
- However, like central tendency, variability is not just one measure.
- The spread of a data set can be described by
    - Variance and standard deviation,
    - Range and Inter quartile range (IQR)
- The spreadness can also be shown in graphs like scatter plots, boxplots, stem and leaf plots an etc.
- Identifying variability relies on understanding the central tendency measurements of a dataset.

Variation by Graph

- The sketch at the right shows that, although the mean three data sets (A,B and C) are the same but they differ in terms of variation.
- Variation in data C is greater than variation in data B and variation in data B is greater than variation in data A.

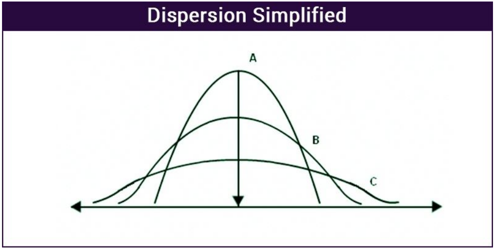


Figure 5: Variation

Hypothetical Data

Table 1: Hypothetical Data

	Hypothetical Data					**Mean**	**Variance**
Data1	20	20	20	20	20	**20**	**0**
Data2	18	19	20	21	22	**20**	**2.5**
Data3	16	18	20	22	24	**20**	**10**
Data4	10	15	20	25	30	**20**	**62.5**
Data5	12	16	20	24	28	**20**	**40**
Data6	36	38	40	42	44	**40**	**10**
Data7	23	24	25	26	27	**25**	**2.5**
Data8	41	43	45	47	49	**45**	**10**

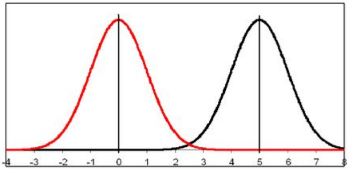
Questions:

Pictured at the right are two different normal distributions. Which is different between the two distributions?

[✓ A. Mean]

B. Standard deviation

C. Both

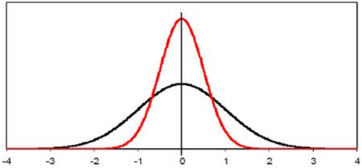


Which is different between the two normal distributions to the right?

A. Mean

[✓ B. Standard deviation]

C. Both

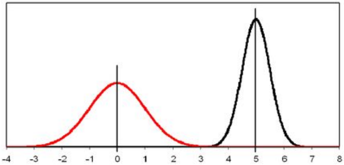


Which is different between the two normal distributions to the right?

A. Mean

B. Standard deviation

[✓ C. Both]



Suitable Analysis for Descriptive Analysis

- Just because descriptive analysis is used for summarization it doesn’t mean it can be used just as you wish,
- There is a number of criteria which make one technique to be suitable for that particular summarization.
- For nominal data, the suitable measure is mode and frequency distribution in general.
- For ordinal data the suitable measure is mode, median and frequency distribution in general.
- For interval and ratio data which are normally distributed the suitable measure are mean, median, mode, variance and standard deviation.
- For interval and ratio data which are not normally distributed (skewed) the suitable measure are median and interquartile range.

Outline

Data Analysis Introduction

Scales of Measurement

Descriptive Analysis

Inferential Analysis

Bivariate Analysis

t-tests

Inferential Analysis

- Inferential analysis is a method of making predictions or inferring information about a population based on a sample of data.
- It uses statistical techniques to draw conclusions about a larger group of individuals or data based on observations of a smaller subset.
- The goal of inferential analysis is to make generalizations about a population based on a sample, and to estimate population parameters such as means, proportions, and variances.
- Inferential statistics includes a variety of techniques which includes
  1. Estimation (like maximum likelihood and least square methods),
  2. Hypothesis testing (like t-test, chi-square test, etc), and
  3. Statistical model fitting for prediction and forecasting.
- Model fitting is used to develop a statistical model that describes the relationship between the variables in the data (linear and logistic regressions).
- Inferential data analysis can be performed by using various software like Excel, SPSS,Stata, R, Python, and many more.

Outline

Data Analysis Introduction

Scales of Measurement

Descriptive Analysis

Inferential Analysis

Bivariate Analysis

t-tests

Bivariate Analysis

- Bivariate analysis is an analysis that is performed to determine the relationship between 2 variables.
- In this analysis, two measurements were made for each observation.
- The samples used could be pairs or each independent with different treatments.
- In general, there are 3 types of analysis based on number of variables
    i. Univariate analysis (1 variable)
    ii. Bivariate analysis (2 variables), and
    iii. Multivariate analysis (more than 2 variables)

Chi-square (χ^2) Measure of Association

- chi-square measure of association is a statistical test that is used to determine whether there is a significant association between two categorical variables.
- It is a non-parametric test, which means that it does not make any assumptions about the distribution of the data.

- chi-square measure of association is a statistical test that is used to determine whether there is a significant association between two categorical variables.
- It is a non-parametric test, which means that it does not make any assumptions about the distribution of the data.

- $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$
- Where
  - O is for observed count in cell, and E is for expected count in cell.
  - r is for total number of rows, and c is for total number of columns.

When to use Chi-square (χ^2) Measure of Association

For chi-square measure of association to be useful under the following scenarios

1. Both variables should be in nominal scale, or
2. One variable is nominal and another is ordinal.
3. Cell counts must be at least 5.

Chi-square (χ^2) Measure of Association by R

- Hypothesis for chi-square test
  - Null (H_0): There is no association between var1 and var2.
  - Alternative (H_1): There is an association between var1 and var2.
- Conclusion: Reject Null (H_0) is p_value < α;
- Most common values for α are 0.01 (1%), 0.05 (5%) and 0.1 (10%)
- In R chi-square test is performed by using the chisq.test().

Chi-square test by R

```
chisq.test(vector1,vector2)
chisq.test(contingency_table)
```

Example

```
chisq.test(mtcars$vs,mtcars$am)
chisq.test(table(mtcars$vs,mtcars$am))
```

- NB: For chi-square to be efficient, all observed cell value must have at least 5 counts.

When to use Fisher’s exact Measure of Association

For Fisher’s exact measure of association to be useful under the following scenarios

1. Both variables should be in nominal scale, or
2. One variable is nominal and another is ordinal.
3. Cell counts can be at less than 5.

Fisher's exact Measure of Association by R

- Fisher's exact is suitable substitute for chi-square test when at least one observed cell value is less than 5 counts.
- The only difference here is the test, all other things like hypothesis are exactly the same as in chi-square test.
- In R Fisher's exact test is performed by using the fisher.test().

Chi-square test by R

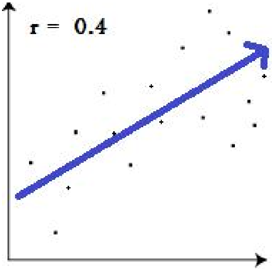
```
fisher.test(vector1,vector2)
fisher.test(contingency_table)
```

Example

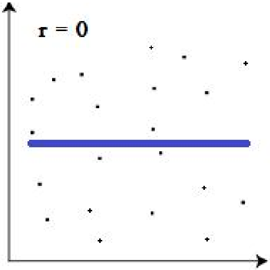
```
fisher.test(mtcars$vs,mtcars$am)
fisher.test(table(mtcars$vs,mtcars$am))
```

Correlation analysis

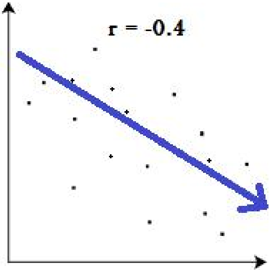
- Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two variables measured with at least ordinal scale.
- Useful for checking if if there are possible connections between variables.
- Correlation between variables ranges between $-1 \leq r \leq +1$.
- Correlations between variables will results in one of the three types
   i. Positive correlation which indicate the variables increases simultaneously with the other and vice-versa.
   ii. Negative correlation which indicate the variables are moving in opposite directions,ie one variable decreases when the other increases.
   iii. Zero correlation indicate no relationship between two variables.



Positive Correlation



No correlation



Negative

Types of correlation coefficients

- There are several correlation coefficients based on the linearity of the relationship, the level of measurement of your variables, and the distribution of your data.
- For high statistical power and accuracy, it's best to use the correlation coefficient that's most appropriate for your data.
- The most commonly used correlation coefficient are
    i. Pearson's correlation,
    ii. Spearman correlation,
    iii. Kendall's tau, and
    iv. Point-biserial.

Pearson’s correlation

- Pearson’s correlation describes the linear relationship between two quantitative variables.
- These are the assumptions your data must meet if you want to use Pearson’s r:
    i. Both variables are on an interval or ratio level of measurement
    ii. Data from both variables follow normal distributions
    iii. Your data have no outliers
- The Pearson’s r is a parametric test, so it has high power.
- But it’s not a good measure of correlation if your variables have a non-linear relationship, or if your data have outliers.
- The same goes if the distribution is skewed , or come from categorical variables.
- If any of these assumptions are violated, you should consider a rank correlation measure.

Computing Pearson’s correlation

- The Pearson’s correlation have got different styles of writing the same formula
- $r = \frac{n\sum XY - \sum X \sum Y}{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}$
- $r = \frac{\sum XY - n\bar{X}\bar{Y}}{[\sum X^2 - n\bar{X}^2][\sum Y^2 - n\bar{Y}^2]}$
- $r = \frac{cov(X,Y)}{sd(X)sd(Y)}$
- Where
  - r is the Pearson’s correlation coefficient.
  - n is the sample size.
  - $\sum$ is the summation symbol
  - X and Y are two variables
  - cov is covariance, and
  - sd is standard deviation

When to use Pearson Correlation test

Pearson correlation will be useful under the following scenarios

1. Both variables should be in ration or interval scale
2. Both variables should be approximately normal (symmetric)

Pearson Correlation test by R

- Pearson correlation is conduction by using cor.test() with active method as "pearson"
- Pearson correlation is also a default cor.test() function

Spearman Rank Correlation test by R

```
cor.test(var1,var2,alternative = , conf.level = )
cor.test(var1,var2,alternative = ,method = "pearson", conf.level = )
```

Example

```
cor.test(mtcars$hp,mtcars$disp)
cor.test(mtcars$hp,mtcars$disp,method = "pearson")
```

Spearman rank correlation

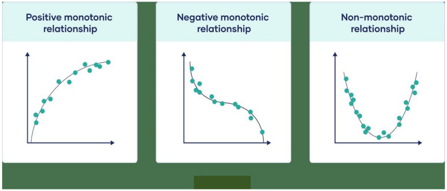
- Spearman’s rank correlation coefficient, is the most common alternative to Pearson’s r.
- It’s a rank correlation coefficient because it uses the rankings of data from each variable (e.g., from lowest to highest) rather than the raw data itself.
- You should use Spearman’s rho when your data fail to meet the assumptions of Pearson’s r.
- This happens when at least one of your variables is on an ordinal level of measurement or when the data from one or both variables do not follow normal distributions.
- While the Pearson correlation coefficient measures the linearity of relationships, the Spearman correlation coefficient measures the monotonicity of relationships.

Computing Spearman rank correlation

- Spearman’s rank correlation coefficient formula is
- $r = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$
- where
  - r is the Spearman rank correlation coefficient.
  - n is the sample size.
  - d is the difference between the x-variable rank and the y-variable rank for each pair of data.
  - $\sum$ is the summation symbol
- If you have a correlation coefficient of 1, all of the rankings for each variable match up for every data pair.
- If you have a correlation coefficient of -1, the rankings for one variable are the exact opposite of the ranking of the other variable.
- A correlation coefficient near zero means that there’s no monotonic relationship between the variable rankings.

Linear vs Monotonic relationship

- In a linear relationship, each variable changes in one direction at the same rate throughout the data range.
- In a monotonic relationship, each variable also always changes in only one direction but not necessarily at the same rate.
  - Positive monotonic: when one variable increases, the other also increases.
  - Negative monotonic: when one variable increases, the other decreases.
- Monotonic relationships are less restrictive than linear relationships.



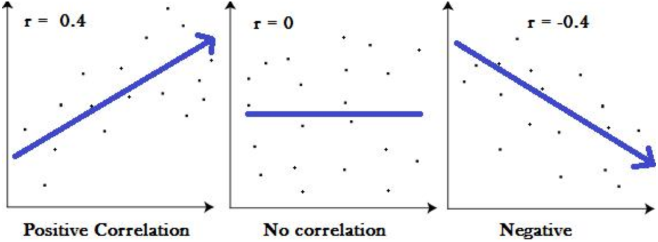


Figure 6: Linear Relationship Measured by Pearson's Correlation

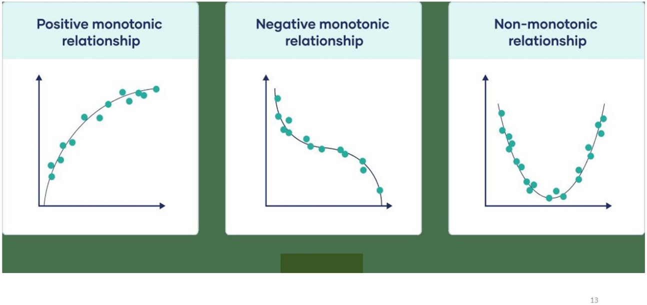


Figure 7: Monotonic Relationship Measured by Spearman's Rank

When to use Spearman Rank Correlation test

Spearman Rank correlation will be useful under the following scenarios

1. Both variables be in ordinal scale
2. One variable is ordinal and other is ratio or interval.
3. Both variables are in ratio or interval but skewed.
4. Both variables are in ratio or interval but one is skewed and the other is normal.

Spearman Rank Correlation test by R

- Spearman rank correlation is conduction by using cor.test() with active method as "spearman"

Spearman Rank Correlation test by R

```
cor.test(var1,var2,alternative = ,method = "spearman", conf.level = )
```

Example

```
cor.test(mtcars$hp,mtcars$disp,method = "spearman")
```

Outline

Data Analysis Introduction

Scales of Measurement

Descriptive Analysis

Inferential Analysis

Bivariate Analysis

t-tests

One sample t-test for mean

- The test on one sample mean
- Hypothesis:
  - $H_0; \mu = \mu_0$ vs $H_1; \mu \neq \mu_0$ (tow tailed (sided))
  - $H_0; \mu = \mu_0$ vs $H_1; \mu > \mu_0$ (right tailed)
  - $H_0; \mu = \mu_0$ vs $H_1; \mu < \mu_0$ (left tailed)
- Level of significance (Type-I error): alpha(α) usually
  - α= 0.01(1%) or
  - α= 0.05(5%) or
  - α= 0.1(10%)
- Test statistic: $t = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim t_{n-1}$
- Decision: Reject H_0 if $p - value < \alpha$
  - If null hypothesis is rejected ($p - value < \alpha$) conclusion will be in favour of alternative hypothesis, and
  - If null hypothesis is not rejected ($p - value \geq \alpha$) conclusion will be in favour of null hypothesis.
- NB: For mean test the data must be either in interval or ratio scale

One sample t-test by R

One sample t-test

```
t.test(x,mu=#)
t.test(x) for zero μ
t.test(x, mu = #, alternative = "greater")
t.test(x, mu = #, alternative = "less")
t.test(x, mu = #, alternative = "two.sided") default
t.test(x,mu=#,conf.level = 0.95) default
t.test(x,mu=#,conf.level = 0.90)
t.test(x,mu=#,conf.level = 0.99)
```

Assume μ_0=120

Example

```
t.test(mtcars$hp,mu=120)
t.test(mtcars$hp,mu=120,conf.level = 0.99)
t.test(mtcars$hp,mu=120, alternative = "greater")
```

Two samples t-test

- The two-sample independent t-test is a statistical test that is used to compare the means of two independent groups.
- This means are measured on the same variable but for two different groups.
- Consider $y_1 \sim N(\mu_1, \sigma_1^2)$ and $y_2 \sim N(\mu_2, \sigma_2^2)$
- We assume that the two samples are independent and that $\sigma_1^2 = \sigma_2^2 = \sigma^2$
- The assumptions of independence and equal variances are necessary in order for the t-statistic.
- The estimate pooled variance is computed as $s_{pl}^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
- $E(s_{pl}^2) = \sigma^2$,unbiased estimator.

- The test on two samples mean
- Hypothesis:
  - $H_0; \mu_1 = \mu_2$ vs $H_1; \mu_1 \neq \mu_2$
  - $H_0; \mu_1 = \mu_2$ vs $H_1; \mu_1 > \mu_2$
  - $H_0; \mu_1 = \mu_2$ vs $H_1; \mu_1 < \mu_2$
- Level of significance (Type-I error): alpha(α) usually
  - α= 0.01(1%) or
  - α= 0.05(5%) or
  - α= 0.1(10%)
- Test statistic:$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$
- Decision: Reject H_0 if $p - value < \alpha$
  - If null hypothesis is rejected ($p - value < \alpha$) conclusion will be in favour of alternative hypothesis, and
  - If null hypothesis is not rejected ($p - value \geq \alpha$) conclusion will be in favour of null hypothesis.
- NB: For mean test the data must be either in interval or ratio scale

Two samples t-test by R

- For independent t-test by R use the following syntax
- Here var1 is numeric and var2 is binary categorical

Independent t-test by R

```
t.test(var1 ~ var2, data = , var.equal = TRUE)
t.test(var1 ~ var2, data = , var.equal = TRUE,alternative = "greater")
t.test(var1 ~ var2, data = , var.equal = TRUE, alternative = "less")
```

Example

```
t.test(hp~vs,var.equal = TRUE,data=mtcars)
```

Here

- hp is a numeric variable, and
- vs is binary variable which define the two groups of hp

Paired samples by t-test

- The paired t-test is a method used to test whether the mean difference between pairs of measurements is significant different from zero or not.
- You can use the test when your data values are paired measurements.
- For example, you might have before-and-after measurements for a group of people.
- The paired t-test is also known as the dependent samples t-test, the paired-difference t-test, and the matched pairs t-test.
- The paired sample t-test requires the sample data to be numeric and continuous, as it is based on the normal distribution.
- The variables should not contain any outliers.

Paired t-test computation (Dependent Sample)

- The first step is to compute the differences between the values of a single pair d
- Let x and y be the two paired variables, then
- $d_i = x_i - y_i$
- $H_0; \mu_x = \mu_y$
- The formula for the paired t-test is given by
- $t = \frac{\overline{d}}{s/\sqrt{n}} \sim t_{n-1}$
- where
    - $\overline{d} = \frac{\sum d_i}{n}$
    - $s = \sqrt{\frac{\sum (d_i - \overline{d})^2}{n-1}}$
    - n is the sample size

- The test on paired samples
- Hypothesis:
    - $H_0; \mu_x = \mu_y$ vs $H_1; \mu_x \neq \mu_y$
    - $H_0; \mu_x = \mu_y$ vs $H_1; \mu_x > \mu_y$
    - $H_0; \mu_x = \mu_y$ vs $H_1; \mu_x < \mu_y$
- Level of significance (Type-I error): alpha(α) usually
    - α= 0.01(1%) or
    - α= 0.05(5%) or
    - α= 0.1(10%)
- Test statistic:
- $t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}$
- Decision: Reject H_0 if $p - value < \alpha$
    - If null hypothesis is rejected ($p - value < \alpha$) conclusion will be in favour of alternative hypothesis, and
    - If null hypothesis is not rejected ($p - value \geq \alpha$) conclusion will be in favour of null hypothesis.
- NB: For mean test the data must be either in interval or ratio scale

Paired t-test by R

- For paired t-test by R use the following syntax
- Here var1 and var2 are both numeric variables

Independent t-test by R

```
t.test(var1,var2,paired=TRUE)
```

Example

```
t.test(mtcars$hp,mtcars$disp,paired=TRUE)
```

Here

- hp is a numeric variable, and
- disp is also numeric.