INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: M.Sc.

Subject: Research Methodology

Chapter: Unit 3 Chapter 1

Chapter Name: Data Analysis and Interpretation

1 Data Preprocessing

Data preprocessing is the first step of a data analysis process. This method involves preparing data so that it can be made ready for research, analysis and modeling. You must prepare and transform the raw data in a format that is easy to interpret and work with.

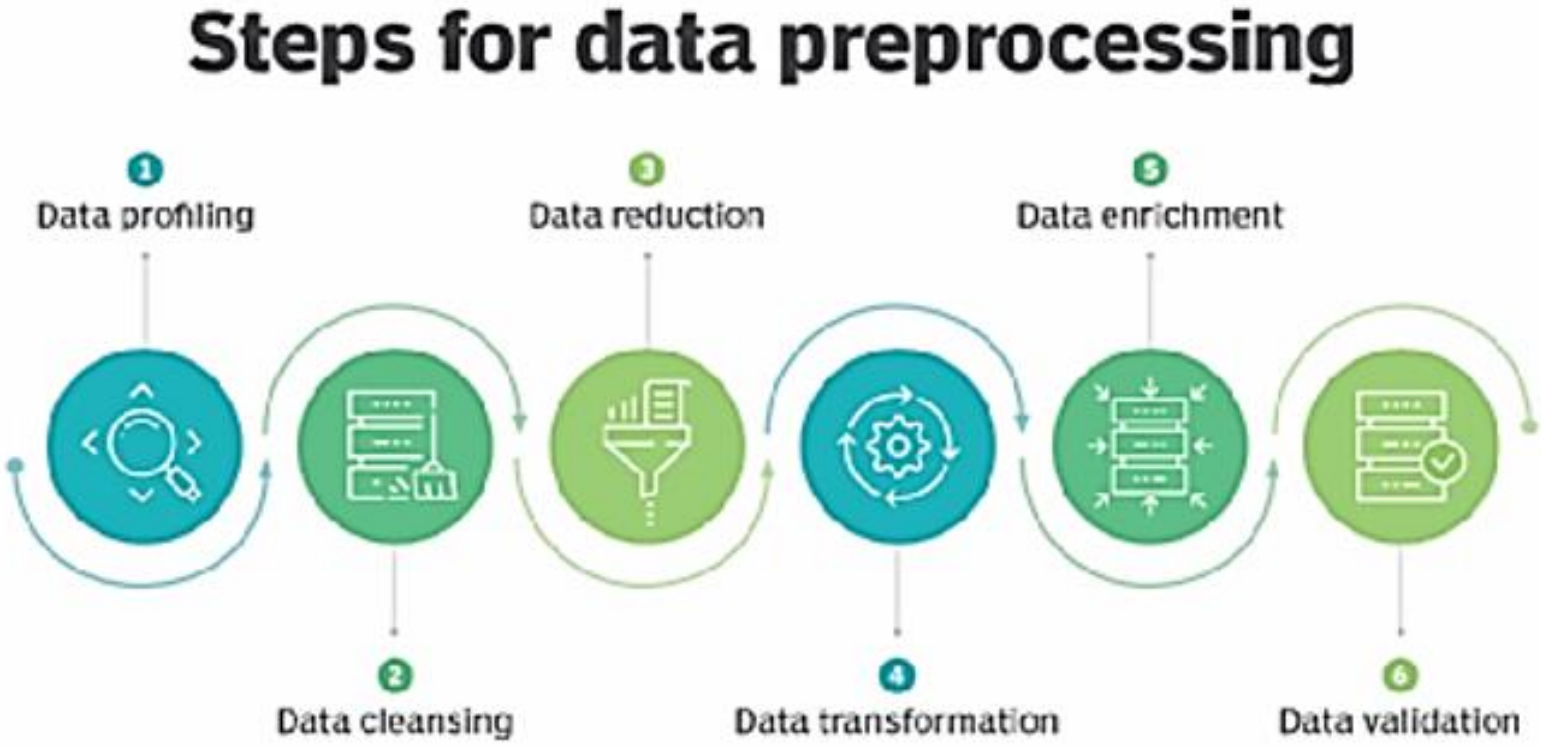
Need for Data Processing:

- **Noise Reduction**: Data preprocessing eliminates errors in the dataset, reducing the noise produced by inconsistencies.

- **Normalization of Data**: Data preprocessing helps normalize the data so that the data can be converted into equalized scale values.

- **Structural errors** usually refer to some typos and inconsistencies in the values of the data.

Data Preprocessing



1 Data Analysis

Data analysis refers to the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, drawing conclusions, and supporting decision-making. It involves applying various statistical and computational techniques to interpret and derive insights from large datasets.

The ultimate aim of data analysis is to convert raw data into actionable insights that can inform business decisions, scientific research, and other endeavors.

The primary purposes of data analysis can be summarized as follows:

To gain insights: Data analysis allows you to identify patterns and trends in data, which can provide valuable insights into the underlying factors that influence a particular phenomenon or process.
To inform decision-making: Data analysis can help you make informed decisions based on the information that is available.
To improve performance: Data analysis can help you optimize processes, products, or services by identifying areas for improvement and potential opportunities for growth.
To measure progress: Data analysis can help you measure progress towards a specific goal or objective, allowing you to track performance over time and adjust your strategies accordingly.
To identify new opportunities: Data analysis can help you identify new opportunities for growth and innovation by identifying patterns and trends that may not have been visible before.

1 Qualitative Data Analysis

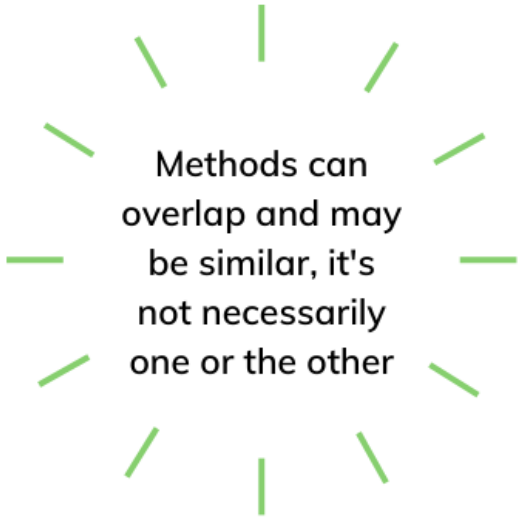
Qualitative data analysis is a process of gathering, structuring and interpreting qualitative data to understand what it represents.

Qualitative data is non-numerical and unstructured. Qualitative data generally refers to text, such as open-ended responses to survey questions or user interviews, but also includes audio, photos and video.

Businesses often perform qualitative data analysis on customer feedback. And within this context, qualitative data generally refers to verbatim text data collected from sources such as reviews, complaints, chat messages, support centre interactions, customer interviews, case notes or social media comments.

5 common methods of Qualitative Data Analysis:

- Content Analysis
- Narrative Analysis
- Discourse Analysis
- Thematic Analysis
- Grounded Theory



Qualitative Data Analysis Techniques

1 Qualitative Data Analysis Techniques

Content Analysis

This is a popular approach to qualitative data analysis. Other qualitative analysis techniques may fit within the broad scope of content analysis.
Thematic analysis is a part of the content analysis.

Content analysis is used to identify the patterns that emerge from text, by grouping content into words, concepts, and themes.

Content analysis is useful to quantify the relationship between all of the grouped content.

1 Qualitative Data Analysis Techniques

Narrative Analysis

Narrative analysis focuses on the stories people tell and the language they use to make sense of them.

It is particularly useful in qualitative research methods where customer stories are used to get a deep understanding of customers' perspectives on a specific issue.

A narrative analysis might enable us to summarize the outcomes of a focused case study.

1 Qualitative Data Analysis Techniques

Discourse Analysis

Discourse analysis is used to get a thorough understanding of the political, cultural and power dynamics that exist in specific situations.

The focus of discourse analysis here is on the way people express themselves in different social contexts.

Discourse analysis is commonly used by brand strategists who hope to understand why a group of people feel the way they do about a brand or product.

Discourse Analysis: A Powerful Tool That Ensures Your Strategy's Success (emotivebrand.com)

1 Qualitative Data Analysis Techniques

Thematic Analysis

Thematic analysis is used to deduce the meaning behind the words people use. This is accomplished by discovering repeating themes in text. These meaningful themes reveal key insights into data and can be quantified, particularly when paired with sentiment analysis.

Often, the outcome of thematic analysis is a code frame that captures themes in terms of codes, also called categories. So, the process of thematic analysis is also referred to as "coding".

A common use-case for thematic analysis in companies is analysis of customer feedback.

https://getthematic.com/insights/thematic-analysis-software/

1 Qualitative Data Analysis Techniques

Grounded Theory

Grounded theory is a useful approach when little is known about a subject. Grounded theory starts by formulating a theory around a single data case. This means that the theory is “grounded”.

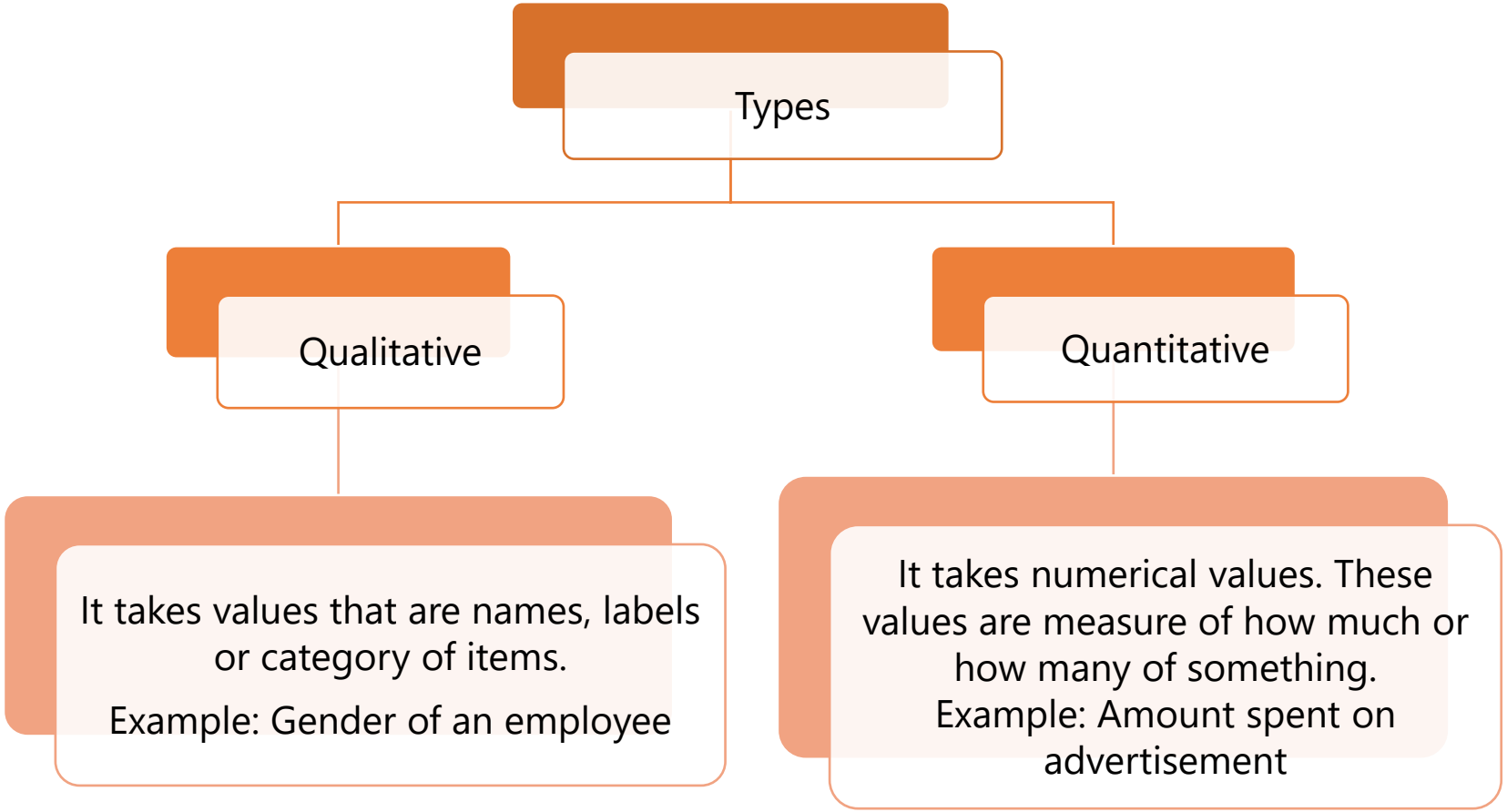
Grounded theory analysis is based on actual data, and not entirely speculative.
Then additional cases can be examined to see if they are relevant and can add to the original grounded theory.

1 Quantitative Data Analysis

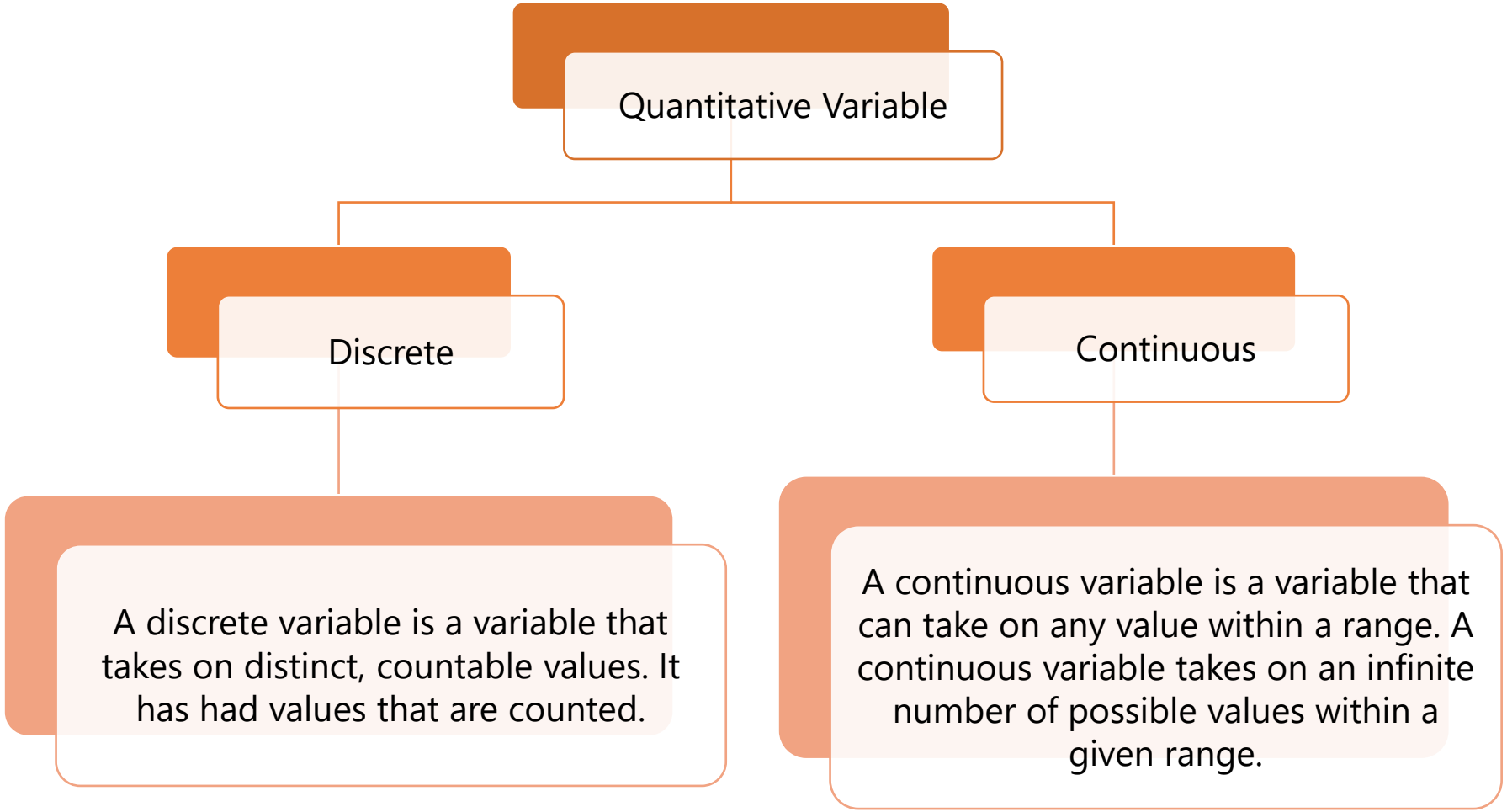
Variable

Measurable, quantifiable, countable or classifiable attribute or characteristic which varies from one entity to another in a group is termed as variable.

Example: - Number of units sold in a month - Blood pressure of an employee



1 Quantitative Data Analysis



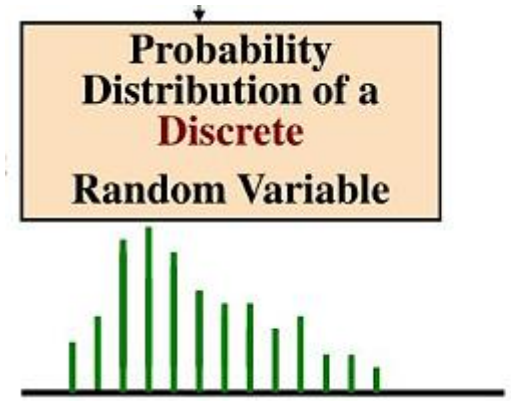
1 Variable

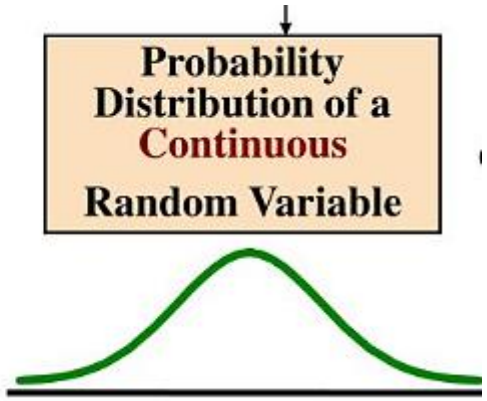
Cause-Effect relationship

In quantitative research, researcher often studies effect of one variable on another, to establish cause and effect relationship.

Dependent Variable – effect, variable being predicted. Its value depends upon or is a consequence of the change in value of another variable.

Independent Variable – cause, variable being used to predict the most likely value of dependent variable. Its value is independent of values of other variables.





1 Types of Quantitative Data

Univariate Data

- When data pertains to only one characteristic or variable of each entity in a category of like items under study, the data is called Univariate data.

  Examples: Number of units sold in a month.

Bivariate Data

- When two variables are observed simultaneously to study each entity in a category of like items, Data obtained is called Bivariate data.

  Examples: Number of units sold in a month and amount spent on advertisement.

Multivariate Data

- When more than two variables are observed simultaneously to study each individual or entity in a certain population, Data obtained is called Multivariate Data.

1 Analysis of Quantitative Data

Analysis of Univariate Data

Measures in Descriptive Statistics

The measures used to summarize the univariate data –

- The distribution concerns the frequency of each value observed.
- The central tendency concerns the averages of the values.
- The variability or dispersion concerns how spread out the values are.

Analysis of Bivariate Data

Identifying correlation between two variables if any and then establishing cause and effect relationship between them and describing the nature of it.

Correlation is the study that involves - knowing existence of relationship between two quantitative variables, - and then knowing its magnitude and direction.

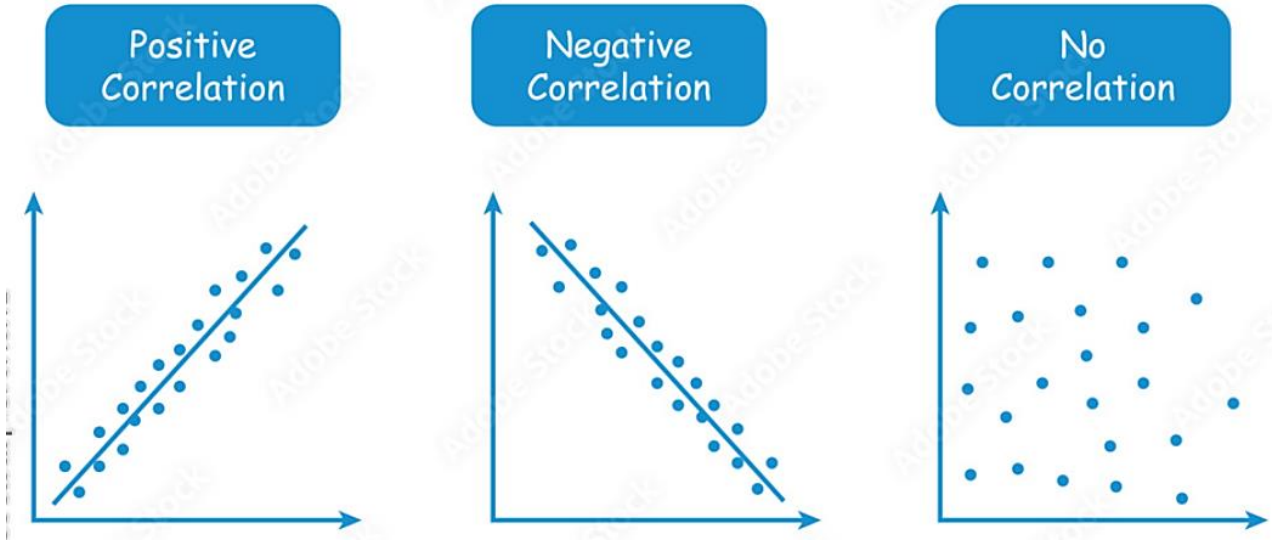
1 Analysis of Bivariate Data

Correlation and Cause-effect
When two variables are correlated, one of them is cause and other is the effect.

Examples - Minimum temperature of town and sale of woolen garments.
- Rain fall and production of paddy

The relation can be positive or negative.

- When one variable increases as the other increases, the correlation is positive.
- When one variable decreases as the other increases, the correlation is negative.



1 Methods of studying linear correlation

The methods are as follows

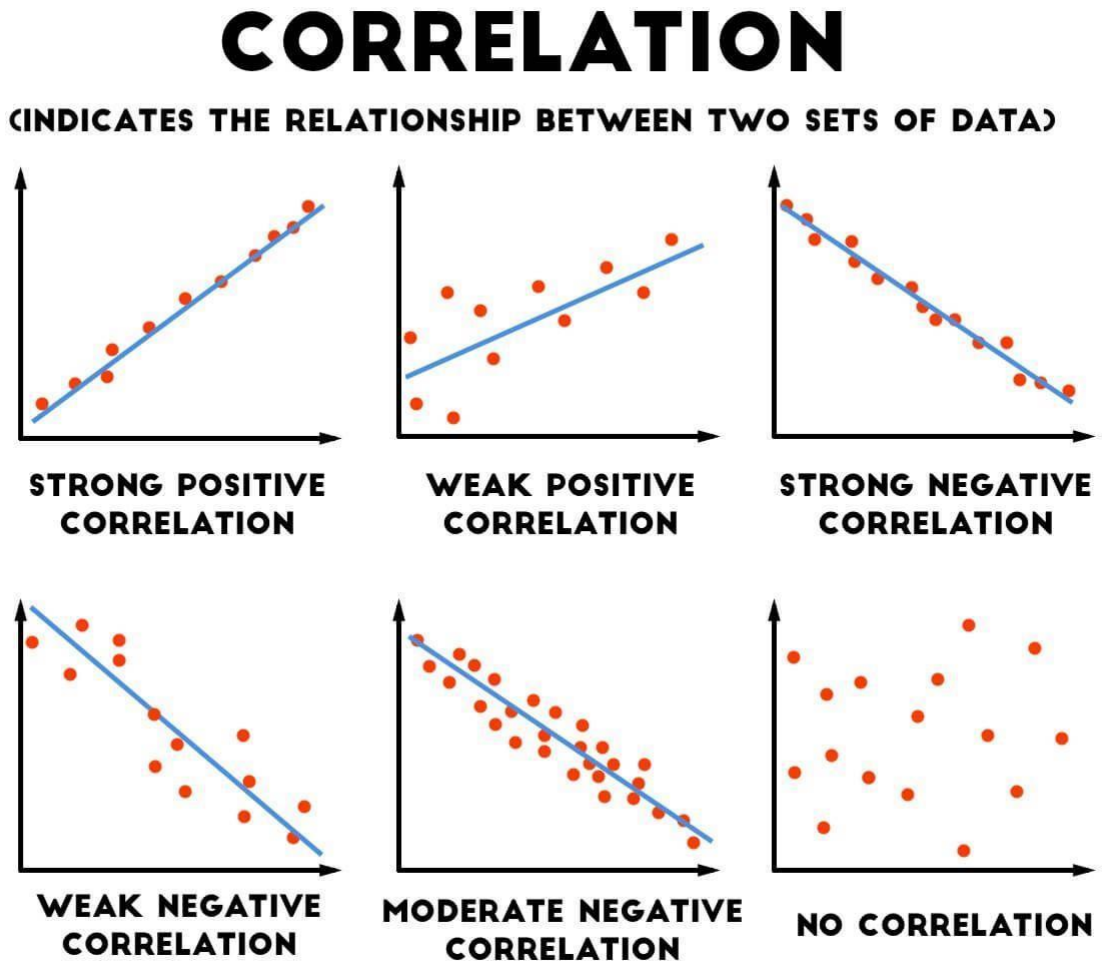
- Scatter plot
- Covariance
- Coefficient of correlation
- Spearman's Rank correlation coefficient

1 Methods of studying linear correlation

Scatter plot

Scatter plot displays the bi-variate data in a graphical form to reveal the relationship between two variables.

A scatter plot of two variables shows the values of one variable on the X axis (cause) and the values of the other variable on the Y axis (effect) .



1 Methods of studying linear correlation

Coefficient of Correlation

It measures the degree of association between two quantitative variables.

Karl Pearson's Coefficient of Correlation

$$r = \frac{\sum[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Interpreting Karl Pearson's Correlation Coefficient

- Value of Karl Pearson's correlation coefficient varies from +1 through 0 to -1.
- The greater the absolute value, the stronger the linear relationship.
- Complete linear correlation between two variables is expressed by either +1 or -1.
- Complete absence of linear correlation is represented by 0.

1 Methods of studying linear correlation

Spearman's Rank Correlation Coefficient

Sometimes the observations are expressed in comparative terms or ranks.

In such case, correlation coefficient is given by

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

d – difference between ranks of ith observation
n – number of observations

It varies from +1 through 0 to -1.

1 Regression

Research inferences and managerial decisions are often based on the nature of relationship between the two or more variables.
Example – Relation between advertisement expenditure and sales

Regression is a technique used to develop an equation showing how are the two or more variables are related to each other.

Regression equation

- Independent Variable – variable taking observed values (cause).
- Dependent Variable – variable value of which is to be predicted (effect).
- Regression equation expresses independent variable as a function of dependent variables on the basis of data collected.
- Then, it is used to predicts the most likely value of dependent variable for the given value of independent variables.

1 Regression

Mathematical Model for Simple Linear Regression

If $\hat{y}$ is the estimated or predicted value of dependent variable and x is the observed value independent variable, then the regression of Y on X relationship is described as follows:

$$\hat{y} = \alpha + \beta x + \varepsilon$$

α and β are population parameters which are not known.
ε is error in estimation i.e. difference in actual value and estimated value of dependent variable.

Least square method is used to find the regression equation of y on x that best represents bivariate sample data,
$\hat{y} = a + b x$

a and b denote sample estimates of α and β.

1 Regression

Calculating Values of a and b

$\hat{y} = a + b\,x$

$$b = \frac{\sum[\,(x_i - \bar{x})\,(y_i - \bar{y})\,]}{\sum[\,(x_i - \bar{x})^2]}$$

$a = \bar{y} - b\,\bar{x}$

where b is the constant in the regression equation representing slope of the line.

where a is the y intercept of the regression line.

1 Regression Analysis

- For equation $\hat{y} = a + b\,x$, the estimate of y derived from equation may not be equal to the actual observed value of y.
- The difference between estimated value and corresponding observed value depends up on the extent of scatter of various points around the line of best fit.
- The closer the various paired sample points clustered around the line of best fit, the smaller the difference between the estimated value and observed value.

1 Coefficient of determination

- It is a key output of regression analysis.
- It is equal to square of coefficient of correlation and so denoted as R^2.
- It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable

- It shows numerical measure about how good is the "fit" between actual observations and predicted value.
- Higher the R^2 value, data points are less scattered, so it is a good model. Lesser the R^2 value is more scattered the data points.

- It ranges from 0 to 1 indicating the extent to which the dependent variable is predictable.
- R^2 = 0 means that the dependent variable cannot be predicted from the independent variable.
  R^2 = 1 means the dependent variable can be predicted without error from the independent variable.
- R^2 = 0.9 means that 90 percent of the variance in value of dependent variable, y, is explained by value of independent variable, x. Remaining 10% is unexplained and can be due to sampling error or other variables.

1 Inferential Statistics

While descriptive statistics summarize the characteristics of a data set, **inferential statistics help you come to conclusions and make predictions based on your data.**

When you have collected data from a sample, you can use inferential statistics to understand the larger population from which the sample is taken.

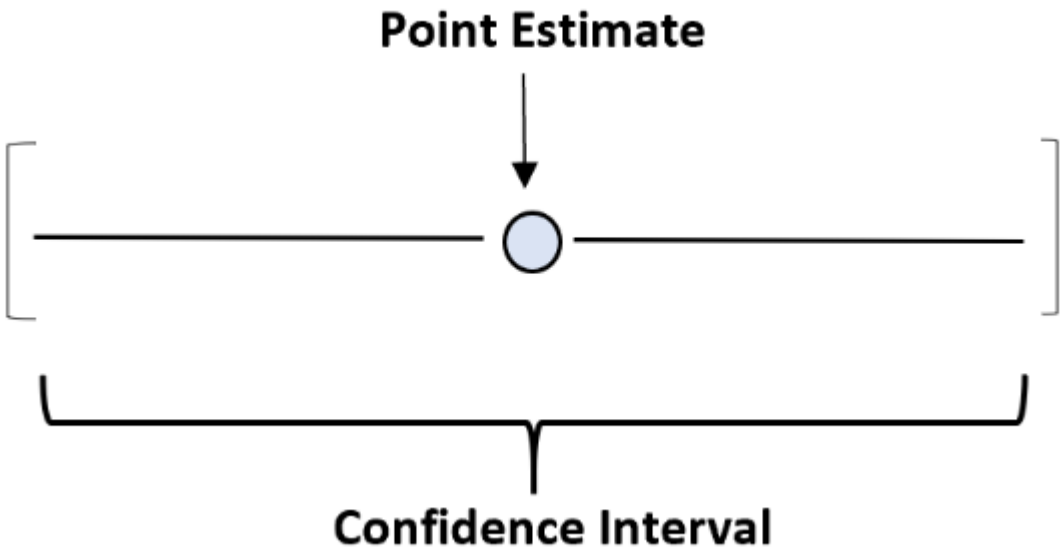
Inferential statistics have two main uses:

- **making estimates about populations** (for example, the mean SAT score of all 11th graders in the US).
  - Point Estimates
  - Interval Estimates or Confidence Interval
- **testing hypotheses to draw conclusions about populations** (for example, the relationship between SAT scores and family income).

1 Inferential Statistics

Point Estimate

A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.



1 Inferential Statistics

Point Estimate

1. The method of moments

The one-parameter case

This is the simplest case: to equate population mean, E X( ) , to sample mean, x, and solve for the parameter, ie:

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

1 Inferential Statistics

Point Estimate

1. The method of moments

The two-parameter case

With two unknown parameters, we will require two equations.
This involves equating the first and second-order moments of the population and the sample and solving the resulting pair of equations.

Moments about the origin can be used but the solution is the same (and often more easily obtained) using moments about the mean – apart from the first-order moment being the mean itself.

The second-order equation is:

$$E\left[X^2\right] = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

or equivalently:

$$E\left[(X-\mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

or: $$\text{var}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

1 Inferential Statistics

Point Estimate

1. The method of maximum likelihood

The method of maximum likelihood is widely regarded as the best general method of finding estimators. In particular, maximum likelihood estimators have excellent and usually easily determined asymptotic properties and so are especially good in the large-sample situation. 'Asymptotic' here just means when the samples are very large.

It seems reasonable that a good estimate of the unknown parameter θ would be the value of θ that **maximizes** the probability, errrr... that is, the **likelihood**... of getting the data we observed. So, that is, in a nutshell, the idea behind the method of maximum likelihood estimation.

In light of the basic idea of maximum likelihood estimation, one reasonable way to proceed is to treat the "**likelihood function**" $L(\theta)$ as a function of θ and find the value of θ that maximizes it.

1 Inferential Statistics

Point Estimate

1. The method of maximum likelihood (Example of exponential distribution)

Given a random sample of size n $(ie\ x_1, \ldots, x_n)$ from the exponential population with density $f(x) = \lambda e^{-\lambda x}$, $x > 0$, the MLE, $\hat{\lambda}$, is found as follows:

$$L(\lambda) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$

$$\therefore \log L(\lambda) = n \log \lambda - \lambda \sum x_i$$

$$\frac{\partial}{\partial \lambda} \log L(\lambda) = \frac{n}{\lambda} - \sum x_i$$

equating to zero:

$$\frac{n}{\lambda} - \sum x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$$

$$\therefore \text{MLE is } \hat{\lambda} = \frac{1}{\bar{X}}$$

1 Inferential Statistics

Confidence Intervals

A **confidence interval** uses the variability around a statistic to come up with an interval estimate for a parameter. Confidence intervals are useful for estimating parameters because they take sampling error into account.

Each confidence interval is associated with a confidence level. A confidence level tells you the probability (in percentage) of the interval containing the parameter estimate if you repeat the study again.

Example: A 95% confidence interval means that if you repeat your study with a new sample in exactly the same way 100 times, you can expect your estimate to lie within the specified range of values 95 times.

1 Inferential Statistics

Confidence Intervals

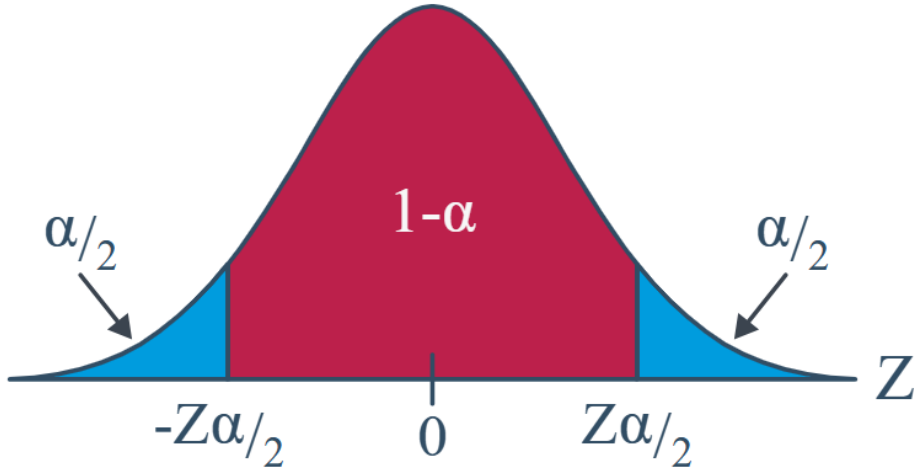
$X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2. So that:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ and } Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

The population variance σ^2 is known.

Then, $(1-\alpha)100\%$ confidence interval for the mean μ is:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$



1 Inferential Statistics

Confidence interval for the mean of normally-distributed data

Normally-distributed data forms a bell shape when plotted on a graph, with the sample mean in the middle and the rest of the data distributed fairly evenly on either side of the mean.

The confidence interval for data which follows a standard normal distribution is:

$$CI = \bar{X} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

Where:

CI = the confidence interval
$\bar{X}$ = the population mean
Z* = the critical value of the z distribution
σ = the population standard deviation
√n = the square root of the population size

The confidence interval for the t distribution follows the same formula but replaces the Z* with the t*.

1 Inferential Statistics

Confidence interval for the mean of normally-distributed data

In real life, you never know the true values for the population (unless you can do a complete census). Instead, we replace the population values with the values from our sample data, so the formula becomes:

$$CI = \hat{x} \pm Z^* \frac{s}{\sqrt{n}}$$

Where:

$\hat{x}$ = the sample mean
s = the sample standard deviation

1 Inferential Statistics

Confidence interval for proportions

The confidence interval for a proportion follows the same pattern as the confidence interval for means, but place of the standard deviation you use the sample proportion times one minus the proportion:

$$CI = \hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where:
$\hat{p}$ = the proportion in your sample (e.g. the proportion of respondents who said they watched any television at all)
Z*= the critical value of the *z* distribution
n = the sample size

1 Inferential Statistics

Hypothesis Testing

In many research areas, such as medicine, education, advertising and insurance, it is necessary to carry out statistical tests. These tests enable researchers to use the results of their experiments to answer questions such as:

- Is drug A a more effective treatment for AIDS than drug B?
- Does training programme T lead to improved staff efficiency?
- Are the severities of large individual private motor insurance claims consistent with a lognormal distribution?

A hypothesis is where we make a statement about something; for example the mean lifetime of smokers is less than that of non-smokers.

A hypothesis test is where we collect a representative sample and examine it to see if our hypothesis holds true.

1 Inferential Statistics

The testing procedure

The standard approach to carrying out a statistical test involves the following steps:

- specify the hypothesis to be tested
- select a suitable statistical model
- design and carry out an experiment/study
- calculate a test statistic
- calculate the probability value
- determine the conclusion of the test.

1 Inferential Statistics

The Hypotheses

The basic hypothesis being tested is the null hypothesis, denoted H0 – it can sometimes be regarded as representing the current state of knowledge or belief about the value of the parameter being tested (the 'status quo' hypothesis). In many situations a difference between two populations is being tested and the null hypothesis is that there is no difference.

In a test, the null hypothesis is contrasted with the alternative hypothesis, denoted H1.

1 Inferential Statistics

One-sided and two-sided tests

In a test of whether smoking reduces life expectancies, the hypotheses would be:
H0 : smoking makes no difference to life expectancy
H1 : smoking reduces life expectancy

This is an example of a one-sided test, since we are only considering the possibility of a reduction in life expectancy ie a change in one direction.

However, we could have specified the hypotheses:
H0 : smoking makes no difference to life expectancy
H1 : smoking affects life expectancy

This is a two-sided test, since the alternative hypothesis considers the possibility of a change in either direction, ie an increase or a decrease.

1 Inferential Statistics

Test statistics

The actual decision is based on the value of a suitable function of the data, the test statistic.

The set of possible values of the test statistic itself divides into two subsets, a region in which the value of the test statistic is consistent with H0 , and its complement, the critical region (or rejection region), in which the value of the test statistic is inconsistent with H0 . If the test statistic has a value in the critical region, H0 is rejected.

Errors

The level of significance of the test, denoted α , is the probability of committing a Type I error, ie it is the probability of rejecting H0 when it is in fact true.
The probability of committing a Type II error, denoted β , is the probability of accepting H0 when it is false.
An ideal test would be one which simultaneously minimises α and β – this ideal however is not attainable in practice.

1 Inferential Statistics

P-values

Merely comparing the observed test statistic with some critical value and concluding eg ‘using a 5% test, reject H0 ’ or ‘reject H0 with significance level 5%’ or ‘result significant at 5%’ (all equivalent statements) does not provide the recipient of the results with clear detailed information on the strength of the evidence against H0 .

A more informative approach is to calculate and quote the probability value (p-value) of the observed test statistic. This is the observed significance level of the test statistic – the probability, assuming H0 is true, of observing a test statistic at least as ‘extreme’ (inconsistent with H0) as the value observed.

The p-value is the lowest level at which H0 can be rejected.

The smaller the p-value, the stronger is the evidence against the null hypothesis.

1 Inferential Statistics

Testing the value of a population mean

Situation: random sample, size n, from $N(\mu, \sigma^2)$ - sample mean $\bar{X}$

Testing: $H_0: \mu = \mu_0$

(a) σ known: test statistic is $\bar{X}$, and $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ under H_0

(b) σ unknown: test statistic is $\frac{\bar{x}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$ under H_0

1 Inferential Statistics

Testing the value of a population proportion

Situation: $\boldsymbol{n}$ binomial trials with P (success) $= p$; we observe x successes.

Testing: $H_0: p = p_0$.

Test statistic is $X \sim \text{Bin}(n, p_0)$ under H_0.
For large $\boldsymbol{n}$, use the normal approximation to the binomial (with continuity correction), ie use:

$$\frac{\frac{X \pm \frac{1}{2}}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

or:

$$\frac{X \pm \frac{1}{2} - np}{\sqrt{np(1-p)}} \dot{\sim} N(0,1)$$

1 Inferential Statistics

Testing for Paired Data

In testing for a difference between two population means, the use of independent samples can have a major drawback. Even if a real difference does exist, the variability among the responses within each sample can be large enough to mask it. The random variation within the samples will mask the real difference between the populations from which they come. One way to control this variability external to the issue in question is to use a pair of responses from each subject, and then work with the differences within the pairs. The aim is to remove as far as possible the subject-to-subject variation from the analysis, and thus to 'home in' on any real difference between the populations.

Assumption: differences constitute a random sample from a normal distribution.

Testing: $H_0: \mu_D (= \mu_1 - \mu_2) = \delta$

Test statistic is $\frac{\bar{D}-\delta}{S_D/\sqrt{n}} \sim t_{n-1}$ under H_0.

We can use $N(0,1)$ for t, and do not require the 'normal' assumption, if n is large.

1 Inferential Statistics

Decide whether to reject or fail to reject your null hypothesis

Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.

In most cases you will use the p-value generated by your statistical test to guide your decision. And in most cases, your predetermined level of significance for rejecting the null hypothesis will be 0.05 – that is, when there is a less than 5% chance that you would see these results if the null hypothesis were true.

In some cases, researchers choose a more conservative level of significance, such as 0.01 (1%). This minimizes the risk of incorrectly rejecting the null hypothesis

Question

The average IQ of a sample of 50 university students was found to be 105.

Carry out a statistical test to determine whether the average IQ of university students is greater than 100, assuming that IQs are normally distributed.

It is known from previous studies that the standard deviation of IQs among students is approximately 20.

Solution

We are testing:

$$H_0 : \mu = 100 \quad vs \quad H_1 : \mu > 100 \ (\sigma \text{ known})$$

Under H_0, $\dfrac{\bar{X} - 100}{\sigma/\sqrt{n}} \sim N(0,1)$.

The test statistic is $\dfrac{105 - 100}{20/\sqrt{50}} = 1.768$.

Calculate the probability of getting a result as extreme as the test statistic (*ie* the *p*-value). If $Z \sim N(0,1)$:

$$P(Z > 1.768) = 1 - 0.96147 = 0.03853$$

We are carrying out a 5% one-tailed test. The probability we have obtained is less than 5%, so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the average IQ of university students is greater than 100.