

FREE Online Citizen Data Scientist Course

You Too Can be a **Citizen Data Scientist** –
No Matter Your Role, Skill or Job Function



Section 9 - Analytical Techniques - Part 3

Smarten ©Copyright 2022.

All rights Reserved.

Smarten is a trademark of Elegant MicroWeb.

All other trademarks are the property of their respective owners.

Instructor Notes: This supporting documentation includes a complete set of the slides used in the course material. Some sections also include expanded material, articles and documentation to further student understanding of more complex topics.

For your convenience, contact information is displayed at the end of the online course and at the conclusion of the supporting documentation in Section 12. We invite you to contact us with questions, requests or comments.

Section 9 - Analytical Algorithms and Techniques – Part 3

Clustering

Algorithm Techniques



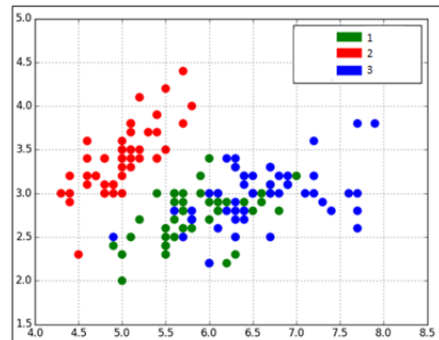
Clustering

Split data into groups when pre-assigned categories or classes are not available (as compared with "classification," where pre-assigned categories or classes are available).

Algorithms:

- K-Means Clustering
- Hierarchical Clustering

Example: Segmenting online customers into heavy/moderate/low purchaser groups based on purchasing frequency, average purchase amount, income, age, etc.



Clustering

What is it used for?

It's a process by which objects are classified into number of groups so that they are as much dissimilar as possible from one group to another group and as much similar as possible within each group

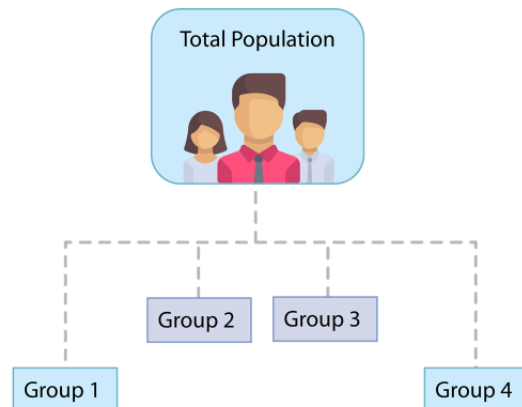
Thus it's simply a grouping of similar things/data points

For example, objects within group 1(cluster 1) shown in image above should be as similar as possible

But there should be much difference between an object in group 1 & group 2

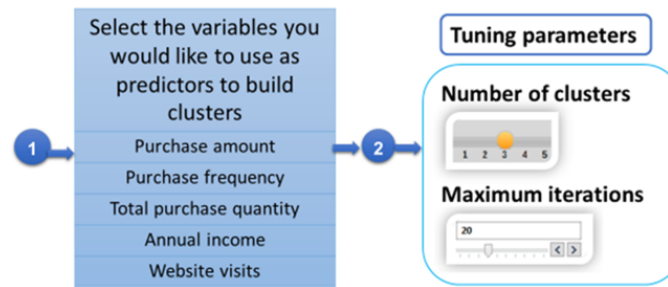
The attributes of objects decide which objects should be grouped together

Thus natural grouping of data points can be achieved



Clustering

Sample UI for Selecting Predictors And Applying Tuning Parameters: For Four Predictors



- The **silhouette score** is another useful criterion for assessing the natural and optimal number of clusters as well as for checking overall quality of partition
- The largest silhouette score, over different **K**, indicates the best number of clusters

Clustering

Limitations

- The number of clusters, k , must be determined before hand. Instead the algorithm should auto suggest this number for better user friendliness.
- It does not yield the same result with each run, since the resulting clusters depend on the initial random assignments for group centers.
- If it is inputted in a different order it may produce different cluster if the number of data points are few, hence number of data points must be large enough.
 - It has been suggested that $2m$ can be used (where m = number of clustering variables) as a rule to decide sample data size.
- K-means is suitable only for numeric data.
- Scale of data points influences Euclidean distance, so variable standardization becomes necessary.
- Empty clusters can be obtained if no points are allocated to a cluster during the assignment step.

Clustering

Some Examples

- It is used to find natural groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.
- Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the relevant group.

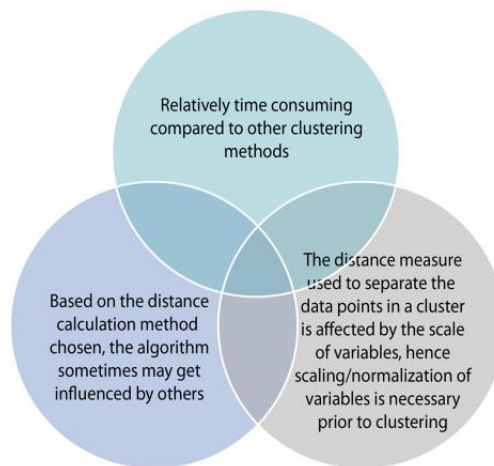
Let's look at a few examples

Movie tickets booking website users can be grouped into movie freaks/moderate watchers/rare watchers based on their past movie tickets purchase behavior, such as #days from last movie seen, average number of tickets booked each time, frequency of tickets booking per month, etc.

Retail customers can be grouped into loyal/infrequent/rare customer groups, based on retail outlet/website visits per month, purchase amount per month, purchase frequency per month, etc.

Clustering

Limitations



Clustering

Use case 2



Correlation

Algorithm Techniques



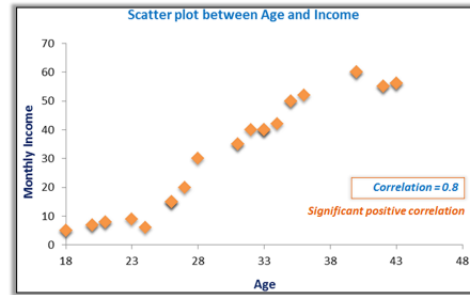
Correlation

Analyze how any two or more variables are associated.

Algorithms:

- Pearson Correlation
- Spearman Rank Correlation

Example: Analyze whether or not there is a strong positive association between age and online purchasing frequency.



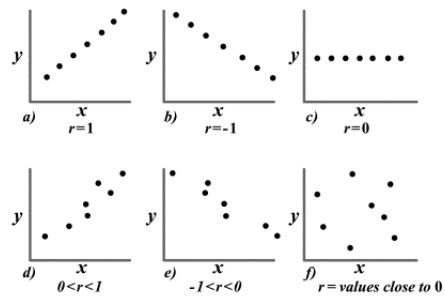
Correlation – Karl Pearson Correlation

Introduction : Karl Pearson’s correlation coefficient

- Karl Pearson’s correlation coefficient measures degree of linear relationship between two variables
- If the relationship between two variables X and Y is to be ascertained through Karl Pearson method, then the following formula is used:

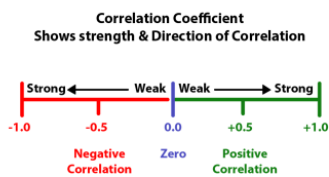
$$r = \frac{n\sum x_i y_i - \sum x_i \times \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \times \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

- The value of the coefficient of correlation always lies between ± 1



Correlation – Karl Pearson Correlation

Introduction : Interpretation of correlation coefficient



r value	Interpretation
+0.70 or higher	Very strong positive relationship
+0.40 to +0.69	Strong positive relationship
+0.30 to +0.39	Moderate positive relationship
+0.20 to +0.29	Weak positive relationship
+0.01 to +0.19	Negligible relationship
0	No relationship (zero order correlation)
-0.01 to -0.19	Negligible relationship
-0.20 to -0.29	Weak negative relationship
-0.30 to -0.39	Moderate negative relationship
-0.40 to -0.69	Strong negative relationship
-0.70 or higher	Very strong negative relationship

Correlation – Karl Pearson Correlation

Limitations

Karl Pearson correlation is affected by outliers

This correlation method measures the strength of relationship between only two variables, without taking into consideration the fact that both these variables may be influenced by a third variable

- For example, sale of ice cream and sale of cold drinks are related to weather conditions of the area. They may show a positive correlation but they are not related to each other

This method can handle only numeric data

Very Dissatisfied	Dissatisfied	Neither Satisfied nor Dissatisfied	Satisfied	Very Satisfied
(1)	(2)	(3)	(4)	(5)
0	0	0	0	0

In case of categorical ordinal (ranked) variables, they need to be converted into numeric ranks in order to proceed with Spearman's correlation

- For instance, as a survey response we may have variable values such as "Very dissatisfied", "dissatisfied", "neutral", "Satisfied", "very satisfied" etc. , these responses have to be converted into numeric ranks 1,2,3,4,5 respectively as shown in figure in right

Correlation – Spearman Rank

Introduction : Spearman’s Rank correlation coefficient

Spearman’s rank correlation is a measure of Correlation between two ranked (ordered) variables

It measures the strength and direction of association between two sets of data when ranked by each of their quantities

If the strength of association between two variables is to be ascertained through Spearman’s rank correlation method, then the following formula is used:

Spearman's coefficient: $r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$

Where n : Number of observations
 d : Difference between two ranks of each Observation

The value of r_s is:

$-1 < r_s < +1$

When there is complete disagreement among rankings

When there is complete agreement among rankings

Correlation – Spearman Rank

Example : Spearman’s Rank Correlation : Positive correlation

Let’s compute the Spearman’s correlation coefficient between two ranked variables X and Y :

X	Y	Rank of X	Rank of Y	d	d ²
5	9	1	1	0	0
6	10	2	2	0	0
7	11	3	3	0	0
$\sum d^2$					0

Closer this value to 0 , weaker the relationship /association between both variables

Closer this value to ± 1 , stronger the relationship between variables

Spearman's coefficient r_s -

$$1 - (6 \sum d^2 / n(n^2 - 1))$$

$$= 1 - (6 * 0 / 3(9 - 1))$$


$$= 1 - 0$$

$$= 1$$

= 1 ~ Perfect positive correlation

Correlation – Spearman Rank

Limitations



This correlation method measures the strength of relationship between only two variables, without taking into consideration the fact that both these variables may be influenced by a third variable

- For example, sale of ice cream and sale of cold drinks are related to weather conditions of the area. They may show a positive correlation but they are not related to each other

This method can handle only numeric data

Very Dissatisfied	Dissatisfied	Neither Satisfied nor Dissatisfied	Satisfied	Very Satisfied
(1)	(2)	(3)	(4)	(5)
0	0	0	0	0

In case of categorical ordinal (ranked) variables, they need to be converted into numeric ranks in order to proceed with Spearman's correlation

- For instance, as a survey response we may have variable values such as "Very dissatisfied", "dissatisfied", "neutral", "Satisfied", "very satisfied" etc. , these responses have to be converted into numeric ranks 1,2,3,4,5 respectively as shown in figure in right

Regression

Algorithm Techniques



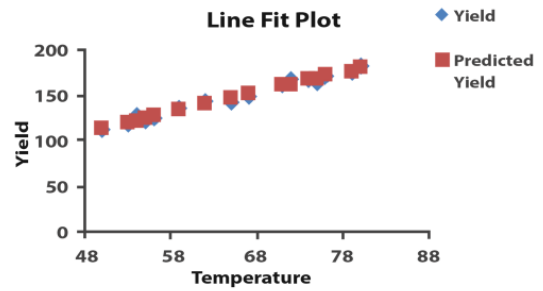
Regression

Predicts change in one variable based on change in one or more other variables

Algorithms:

- Simple Linear Regression
- Multiple Linear Regression

Example: eCommerce company can measure the sales impact of product price, product promotion, holidays, seasonality, etc.



Regression – Simple Linear Regression

Introduction : Simple linear regression

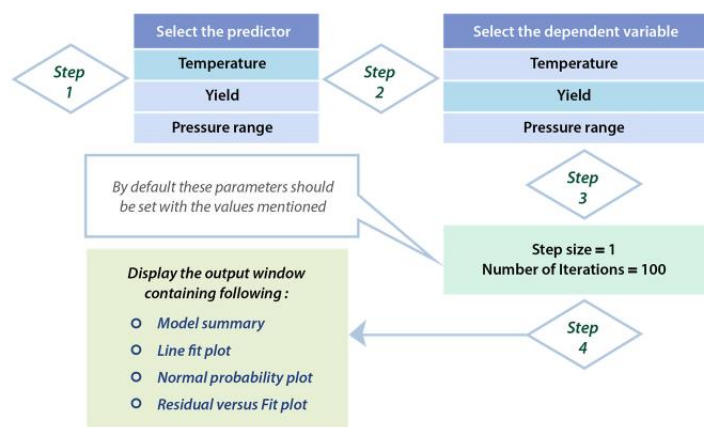
- **OBJECTIVE :**
 - It is a statistical technique that attempts to explore the relationship between one independent variable (X) and one dependent variable (Y)
- **BENEFIT :**
 - Regression model output helps identify whether independent variable/predictor X has any relationship with dependent variable Y and if yes then what is the nature/direction of relationship (i.e. positive/negative) between both
- **MODEL :**
 - Simple Linear regression model equation takes the form of $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ as shown in image in right :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The diagram shows the equation $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with callouts: Y_i is labeled 'Dependent variables', β_0 and β_1 are labeled 'Coefficients', X_i is labeled 'Predictor', and ϵ_i is labeled 'Error term'.

Regression – Simple Linear Regression

Standard input/tuning parameters & Sample UI



Note : Categorical predictors should be auto detected & converted to binary variables before applying regression

Regression – Simple Linear Regression

Limitations

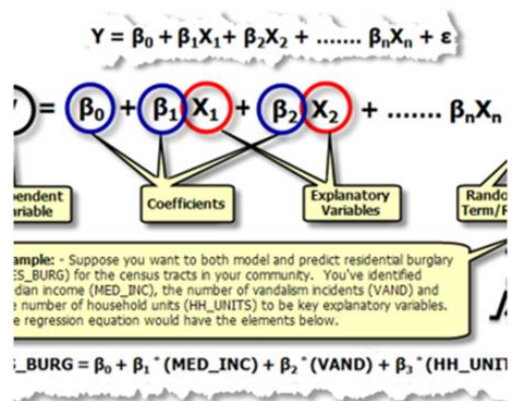
Simple linear regression is limited to predicting numeric output i.e. dependent variable has to be numeric in nature

- Minimum sample size should be $> 50+8m$ where m is number of predictors.
 - Hence in case of simple linear regression, minimum sample size should be $50+8(1) = 58$
- It handles only two variables : one predictor and one dependent variable but usually there are more than one predictor correlated with the dependent variable which can't be analyzed through simple linear regression

Regression – Multiple Linear Regression

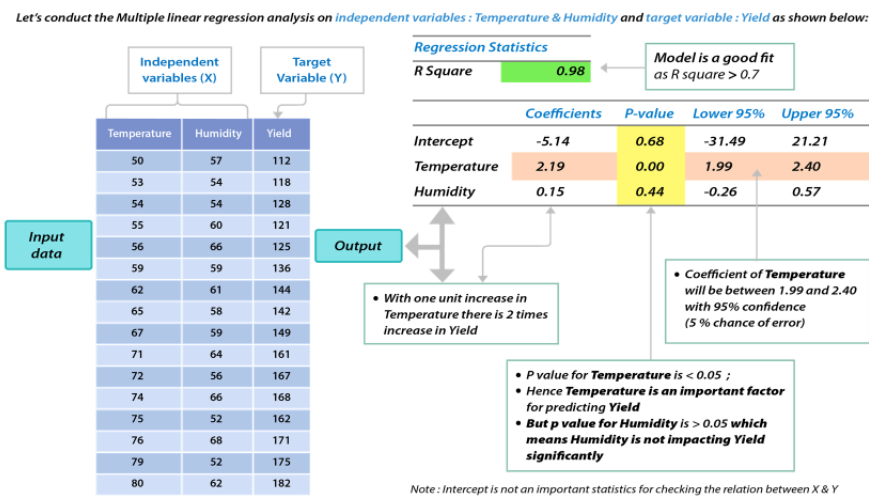
Introduction

- **OBJECTIVE :**
 - It is a statistical technique that attempts to explore the relationship between two or more variables (X ; and Y)
- **BENEFIT :**
 - Regression model output helps identify important factors (X), impacting the dependent variable (Y) and also the nature of relationship between each of these factors and dependent variable
- **MODEL :**
 - Linear regression model equation takes the form of $Y = \beta_0 + \beta_1 X_1 + \epsilon$ as shown in image in right :



Regression – Multiple Linear Regression

Example: Multiple linear regression



Regression – Multiple Linear Regression

Limitations

Linear regression is limited to predicting numeric output i.e. dependent variable has to be numeric in nature

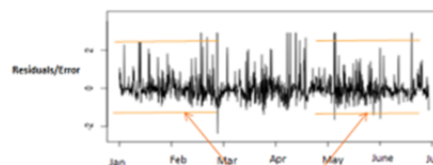
Minimum sample size should be at least 20 cases per independent variable

Multicollinearity among one or more predictors should be removed before running the model

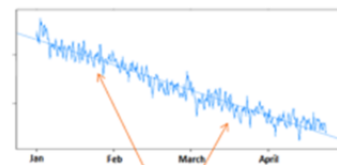
Multicollinearity is the situation in which two or more independent variables are highly correlated with one another

This method is applicable only when assumption of linearity between each X_i and Y is met which can be checked through the Line fit plot which is a scatter plot between each X_i and Y as described in the Interpretation section

Residuals should be time independent as described in the bottom image below



Time independent error (fairly constant over time & lying within certain range)



Time dependent error (decreasing with time)

Frequent Pattern Mining

Algorithm Techniques

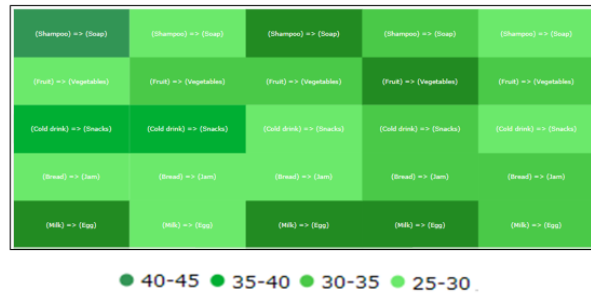


Frequent Pattern Mining
Finds frequent patterns from the data.

Algorithms:

- Frequent Pattern Mining

Example: A retail store can place bakery products, such as muffins, bread, and eggs, together if these products have a high frequency of being purchased together.



Frequent Pattern Mining

Example

TID	Milk	Bread	Butter	Beer
1	1	0	1	1
2	1	1	1	0
3	0	1	1	0
4	1	0	0	1
5	1	1	1	1

- Support (milk->bread) = 0.4 means milk & bread together occur in 40% of all transactions
- Confidence (milk->bread) = 0.5 means, if there are 100 transactions containing milk then there are 50 of them containing bread also

Here, support (Milk -> Bread):

$$= \text{Number of transactions containing milk \& bread} / \text{total transactions}$$

$$= 2/5 = 0.4$$

Confidence (Milk -> Bread):

$$= \text{support (milk -> bread)} / \text{support(milk)}$$

$$= 0.4 / [4/5]$$

$$= 0.4 / 0.8$$

$$= 0.5$$

Lift (Milk -> Bread):

$$= \text{support (milk-> bread)} / \text{support(milk)} * \text{support(bread)}$$

$$= 0.4 / [(4/5) * (3/5)]$$

$$= 0.4 / [0.8*0.6] = (0.4/0.48)$$

$$= 0.83$$

Frequent Pattern Mining

Limitations

- Processing time for running this algorithm is relatively high when compared to other algorithms due to millions of transaction level data in input
- The user must possess a certain amount of expertise in order to find the right settings for support and confidence to obtain the best association rules

Hypothesis Testing

Algorithm Techniques



Hypothesis Testing

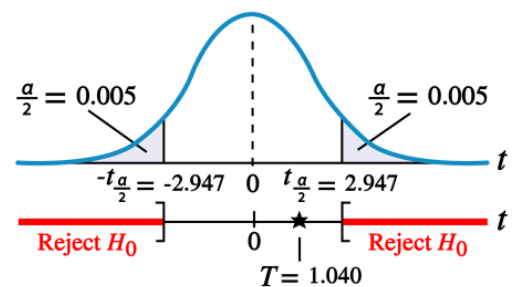
Answers such questions as the following:

- Are two samples significantly different?
- Is the treatment effective?
- Are two dimensions related or independent of each other?

Algorithms:

- Chi Square Test of Independence
- One Way Anova Test
- Independent Sample T-Test
- Paired Sample T-Test

Example: An eCommerce company can measure the regional influence on product category and gender influence on purchased product type.



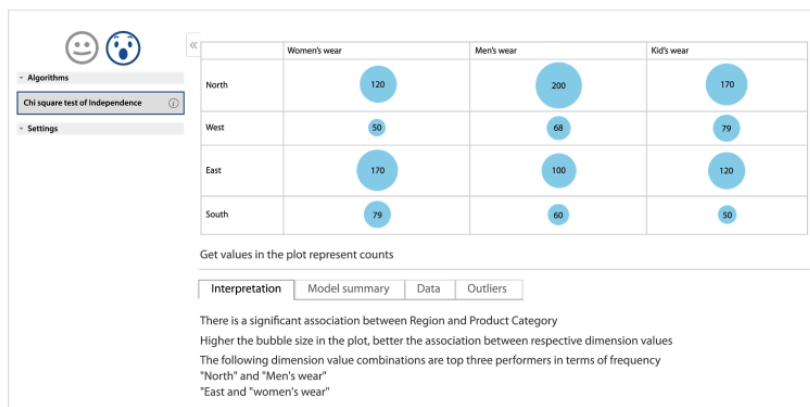
Hypothesis Testing

Introduction

- It is used to determine whether there is a statistically significant association between the two categorical variables
- Thus it finds out if the relationship exists between any two business parameters that are of categorical datatype
- Examples :
 - We could use a chi-square test for independence to determine whether gender is related to a voting preference
 - We could determine if region has any influence on product category purchased

Hypothesis Testing – Chi Square Test of Independence

Sample output 1 : Interpretation



Hypothesis Testing – Chi Square Test of Independence

Limitations

- Can be applied on only two categorical variables (two dimensions)
- Number of data points should be at least 50
- Frequency count for each dimension value combination, i.e. each cell value of the contingency table should be at least 5
- It tells the presence or, absence of an association between two parameters but doesn't measure the strength of association like correlation does

Hypothesis Testing – One Way Anova

Introduction

- It is used to determine whether there is a statistically significant difference among two or more group means
- Examples :
 - We could use One way Anova test to determine if out of three or more rivers, at least two of them differ significantly from each other in terms of pH, TDS, etc.
 - We could determine if at least two regions differ significantly in terms of average sales of a particular product category

Hypothesis Testing – One Way Anova

Example : Input

Let's conduct the One way Anova test on following two variables, one is a dimension and the other is a measure:

Date	River	pH level
1/7/17	Narmada	5.6
1/7/17	Sabarmati	6.8
1/7/17	Yamuna	6.8
1/8/17	Sabarmati	6.7
1/8/17	Yamuna	6.8
1/8/17	Narmada	5.6

This table displays pH measurements for three rivers taken on specific dates

Hypothesis Testing – One Way Anova

Limitations

- Total data points should be one more than total number of groups
 - For instance, if we have 15 rivers and their respective pH measurements then total data points must be at least $15+1=16$
- Outliers must be removed from the data before applying ANOVA as it is sensitive to outliers and may lead to incorrect results

Hypothesis Testing – Independent Sample T-Test

Introduction

- Independent sample t-test is a statistical test that determines whether there is a statistically significant difference between the means of two independent samples
 - For instance, checking if average value of a sedan car type is significantly different than the SUV car type
 - Here the hypothesis would be set as follows :
 - Null hypothesis : SUV and Sedan car types have insignificant difference in terms of value
 - Alternative hypothesis : Value of SUV and Sedan differ significantly

Hypothesis Testing – Independent Sample T-Test

Example : Input

Let's conduct the Independent t-test on following two variables, one is a dimension containing two values and the other is a measure:

Group	Value
A	90
A	95
A	80
B	78
B	75
B	70
B	65

Hypothesis Testing – Independent Sample T-Test

Limitations

- Can be applied on only two samples (one dimension with two values and one measure at a time)
- Observations within each group must be independent
- The values in each group must be normally distributed
- Number of data points should be at least 30

Hypothesis Testing – Paired T-Test

Introduction

- It is used to determine whether the mean of a dependent variable (e.g., weight, anxiety level, salary, reaction time, etc.) is the same in two related groups (e.g., two groups of participants that are measured at two different “time points” or who undergo two different “conditions”)
- Thus the classic use of the Paired t-Test is to evaluate the before and after of some treatment
- Examples :
 - Understand whether there was a difference in manager’s salaries before and after undertaking a PhD (i.e., your dependent variable would be “Salary”, and your two related groups would be the two different “time points”; that is, salaries “before” and “after” undertaking the PhD)
 - Measure the blood pressure of patient A, give him something (pharmaceutical, exercise, Tilapia) to reduce his blood pressure, then measure the blood pressure of patient A again. Repeat for patients B, C, D, ... In this case, the data of “Before” and “After” are paired by patient

Hypothesis Testing – Paired T-Test

Example : Input

Let's conduct the Paired sample t-test on following two variables, one is a time dimension containing months and other is a measure:

Month	Value
january	90
February	95
March	80
April	78
May	75
June	70

Let's say, measure values before April belong to 'before' or 'pre' sample and from April belong to 'After' or 'post' sample

Hypothesis Testing – Paired T-Test

Limitations

Can only be applied to two samples (One measure and one time dimension or a sequence ID to decide the cut point for division of measure values into pre and post samples)

Number of data points should be at least 30

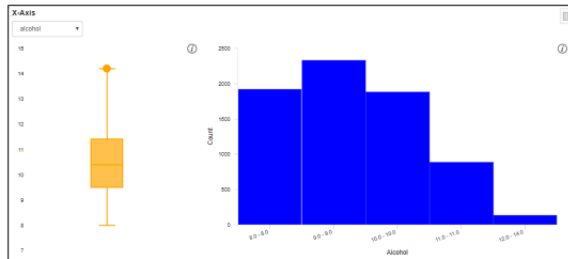
Descriptive Statistics

Algorithm Techniques



Descriptive Statistics

Provides basic statistics, such as mean, median, mode, standard deviation, variance, skewness, and kurtosis.



Descriptive Statistics

Introduction with example

Mean:

- It is simply the average of all the data values
- This measure can be biased in case of significant number of outliers present in data
- **Descriptive statistics** help describe and understand the features of a specific dataset, by giving short summaries about the measures of the data. The most recognized types of descriptive statistics are listed and explained below
- Mean, median, and mode are different ways to figure out an average

Median:

- It is the value in the middle when the data items are arranged in ascending order
- This measure is relatively robust in case of significant number of outliers present in data making it more appropriate measure of average in case of presence of outliers in data
- For instance, when profiling customers based on various attributes such as income or balances, their median age/income/balance etc. can be looked at instead of mean to avoid bias due to outliers

Mode:

- It is the most frequently occurring value in a series of data
- In case of no repeating values, there would be no mode
- For example, in satisfaction survey analysis, mode can be used to find what is the most common rating provided by responders to a particular service/product
- The second most popular use of mode is while imputing missing values of a character variable; when we have number of missing values in say, region variable then it's general tendency to replace these missing values with most frequently occurring region i.e. mode of region

Mean

Add all the numbers then divide by the amount of numbers

9, 3, 1, 8, 3, 6

$$9 + 3 + 1 + 8 + 3 + 6 = 30$$

$$30 \div 6 = 5$$

The mean is 5

Median

Order the set of numbers, the median is the middle number

9, 3, 1, 8, 3, 6

1, 3, 3, 6, 8, 9

The median is 4.5

Mode

The most common number

9, 3, 1, 8, 3, 6

The mode is 3

Range

The difference between the highest and lowest number

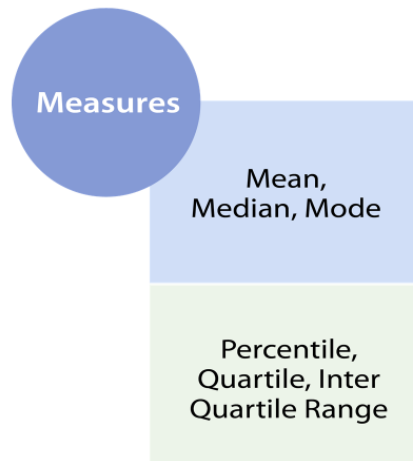
9, 3, 1, 8, 3, 6

$$9 - 1 = 8$$

The range is 8

Descriptive Statistics

Statistical summary measures



Descriptive Statistics

Business User Cases – Standard Deviation/Variance

Let's compute the standard deviation for stock prices over a 10 day period.

- Calculate the average (mean) price for the number of periods or observations
- Determine the deviation (price-mean) for each period
- Square the deviations for each period
- Divide the sum by the number of observations (this is the variance)
- The standard deviation is equal to the square root of this number

The lower the number, the less volatility in price. This calculation makes it easier to estimate future stock prices.

	Date	QQQ Price	10-period Average (mean)	Deviation	Deviation Squared	10-period Average of Deviation Squared	Standard Deviation
1	2-Dec-10	53.73	54.09	-0.36	0.13		
2	3-Dec-10	53.87	54.09	-0.22	0.06		
3	6-Dec-10	53.85	54.09	-0.24	0.06		
4	7-Dec-10	53.85	54.09	-0.21	0.04		
5	8-Dec-10	54.08	54.09	-0.01	0.00		
6	9-Dec-10	54.14	54.09	0.05	0.00		
7	10-Dec-10	54.50	54.09	0.41	0.16		
8	13-Dec-10	54.30	54.09	0.21	0.04		
9	14-Dec-10	54.40	54.09	0.31	0.09		
10	15-Dec-10	54.16	54.09	0.07	0.01	0.06	0.24

Descriptive Statistics

Business use cases – In general

- Descriptive Statistics as the name implies describes or summarizes the raw data and makes it interpretable by humans
- Common examples of descriptive analytics are reports that provide historical insights regarding the company's production, financials, operations, sales, finance, inventory and customers
- Thus, Descriptive Statistics is to be used when you need to understand at an aggregate level what is going on in your company, and when you want to summarize and describe different aspects of your business



This article summarizes our recent article series on the definition, meaning and use of the various algorithms and analytical methods and techniques used in predictive analytics for business users, and in augmented data preparation and augmented data discovery tools.

The article series is designed to help business users better understand the analytical techniques so that the average user can feel more confident in adopting, embracing and sharing these tools.

This twenty-four (24) article series includes:

Naïve Bayes Classification:

What is Naïve Bayes Classification and How is it Used for Enterprise Analysis?

<https://www.smartent.com/blog/what-is-naive-bayes-classification-and-how-is-it-used-for-enterprise-analysis>

Use Case(s): Weather Forecasting, Fraud Analysis and more.

Frequent Pattern Mining (Association):

What is Frequent Pattern Mining (Association) and How Does it Support Business Analysis?

<https://www.smartent.com/blog/what-is-frequent-pattern-mining-association-and-how-does-it-support-business-analysis>

Use Case(s): Market Basket Analysis, Frequently Bundled Products and more.

<p>KNN Classification:</p>	<p><u>What is KNN Classification and How Can This Analysis Help an Enterprise?</u> (https://www.smartent.com/blog/what-is-knn-classification-and-how-can-this-analysis-help-an-enterprise) Use Case(s): Predicting Loan Default, Predicting Success of Medical Treatment and more.</p>
<p>Multiple Linear Regression:</p>	<p><u>What is Multiple Linear Regression and How Can it be Helpful for Business Analysis?</u> (https://www.smartent.com/blog/what-is-multiple-linear-regression-and-how-can-it-be-helpful-for-business-analysis) Use Case(s): Impact of Product Pricing, Promotion on Sales, Impact of rainfall, humidity on crop yield an more.</p>
<p>Independent Samples T Test:</p>	<p><u>What is the Independent Samples T Test Method of Analysis and How Can it Benefit an Organization?</u> (https://www.smartent.com/blog/what-is-the-independent-samples-t-test-method-of-analysis-and-how-can-it-benefit-an-organization) Use Case(s): Are men more satisfied with their jobs than women? Does customer group A spend more on products than customer group B, and more.</p>
<p>Simple Random Sampling and Stratified Random Sampling:</p>	<p><u>What Are Simple Random Sampling and Stratified Random Sampling Analytical Techniques?</u> (https://www.smartent.com/blog/what-are-simple-random-sampling-and-stratified-random-sampling-analytical-techniques) Use Case(s): Average value of all cars in U.S. based on sample, sampling by age, gender, religion, race, educational attainment, socioeconomic status, and nationality and more.</p>
<p>Spearman's Rank Correlation:</p>	<p><u>What is Spearman's Rank Correlation and How is it Useful for Business Analysis?</u> (https://www.smartent.com/blog/what-is-spearman-s-rank-correlation-and-how-is-it-useful-for-business-analysis) Use Case(s): Cluster various survey responders into groups, based on rank correlation, assess student rating by department chairs and by the faculty members and more.</p>
<p>Binary Logistic Regression Classification:</p>	<p><u>What is Binary Logistic Regression Classification and How is it Used in Analysis?</u> (https://www.smartent.com/blog/what-is-binary-logistic-regression-classification-and-how-is-it-used-in-analysis) Use Case(s): Predict if loan default based on attributes of applicant; predict likelihood of successful treatment of new patient based on patient attributes and more.</p>

<p>Paired Sample T Test:</p>	<p><u>What is the Paired Sample T Test and How is it Beneficial to Business Analysis?</u> https://www.smartent.com/blog/what-is-the-paired-sample-t-test-and-how-is-it-beneficial-to-business-analysis Use Case(s): Manufacturing unit manager analyzes statistical significance of cycle time difference, pre and post process change, determine whether sales increased following a particular campaign and more.</p>
<p>Simple Linear Regression:</p>	<p><u>What is Simple Linear Regression and How Can an Enterprise Use this Technique to Analyze Data?</u> https://www.smartent.com/blog/what-is-simple-linear-regression-and-how-can-an-enterprise-use-this-technique-to-analyze-data Use Case(s): Measure the impact of product price on product sales, measure the impact of temperature on crop yield an more.</p>
<p>ARIMAX Forecasting:</p>	<p><u>What is ARIMAX Forecasting and How is it Used for Enterprise Analysis?</u> https://www.smartent.com/blog/what-is-arimax-forecasting-and-how-is-it-used-for-enterprise-analysis Use Case(s): Forecast product line growth based on data from the past 30 years based on yearly consumer inflation rate, yearly GDP data, target variables for user-specified time periods to clearly illustrate results for planning, production, sales and other factors and more.</p>
<p>Karl Pearson Correlation Analysis:</p>	<p><u>What is Karl Pearson Correlation Analysis and How Can it be Used for Enterprise Analysis Needs?</u> https://www.smartent.com/blog/what-is-karl-pearson-correlation-analysis-and-how-can-it-be-used-for-enterprise-analysis-needs Use Case(s): Correlation between income and credit card delinquency rate, identify negative, positive and neutral correlations between the age of a consumer and the color of shirt they might purchase and more.</p>
<p>Hierarchical Clustering:</p>	<p><u>What is Hierarchical Clustering and How Can an Organization Use it to Analyze Data?</u> https://www.smartent.com/blog/what-is-hierarchical-clustering-and-how-can-an-organization-use-it-to-analyze-data Use Case(s): Group loan applicants into high/medium/low risk based on attributes such as loan amount, installments, or employment tenure, organize customers into groups/segments based on similar traits, product preferences and expectations and more.</p>
<p>SVM Classification Analysis:</p>	<p><u>What is SVM Classification Analysis and How Can It Benefit Business Analytics?</u> https://www.smartent.com/blog/what-is-svm-classification-analysis-and-how-can-it-benefit-business-analytics Use Case(s): Predict success of treatment success based on attributes of a patient, improve weather forecasting results and more.</p>

<p>Outlier Analysis:</p>	<p><u>What is Outlier Analysis and How Can It Improve Analysis?</u> (https://www.smartent.com/blog/what-is-outlier-analysis-and-how-can-it-improve-analysis) Use Case(s): Outliers are sometimes discounted, or in other cases, they will indicate that the organization should focus solely on those outliers; identify when a person recovered from a particular disease in spite of the fact that most other patients did not survive, and more.</p>
<p>Decision Tree Analysis:</p>	<p><u>What is the Decision Tree Analysis and How Does it Help a Business to Analyze Data?</u> (https://www.smartent.com/blog/what-is-the-decision-tree-analysis-and-how-does-it-help-a-business-to-analyze-data) Use Case(s): Classify customers into those that will default and those that will not default. And assess the characteristics of customers that are likely to default, based on customer attributes and past online shopping behavioral data, one can predict the future purchases of customers and more.</p>
<p>Chi Square Test of Association:</p>	<p><u>What is the Chi Square Test of Association and How Can it be Used for Analysis?</u> (https://www.smartent.com/blog/what-is-the-chi-square-test-of-association-and-how-can-it-be-used-for-analysis) Use Case(s): Determine if a product sells better in certain locations, verify if gender has an influence on purchasing decisions, Identify if demographic factors influence banking channel/product/service preference or selection of a type of term insurance plan and more.</p>
<p>FP Growth Analysis:</p>	<p><u>What is FP Growth Analysis and How Can a Business Use Frequent Pattern Mining to Analyze Data?</u> (https://www.smartent.com/blog/what-is-fp-growth-analysis-and-how-can-a-business-use-frequent-pattern-mining-to-analyze-data) Use Case(s): Select items in a business catalog to complement each other so that buying one item will lead to buying another, analyze the association of purchased items in a single basket or single purchase and more.</p>
<p>ARIMA Forecasting:</p>	<p><u>What is ARIMA Forecasting and How Can it Be Used for Enterprise Analysis?</u> (https://www.smartent.com/blog/what-is-arima-forecasting-and-how-can-it-be-used-for-enterprise-analysis) Use Case(s): Predict sales of a drug for the next 2 months, based on drug sales from the past 12 months, suitable for forecasting when data is stationary or non-stationary, will produce accurate, dependable forecasts, when planning for short-term business results and more.</p>

<p>Multinomial-Logistic Regression Classification:</p>	<p><u>What is the Multinomial-Logistic Regression Classification Algorithm and How Does One Use it for Analysis?</u> https://www.smartent.com/blog/what-is-the-multinomial-logistic-regression-classification-algorithm-and-how-does-one-use-it-for-analysis) Use Case(s): Based on the attributes of a respondent e.g., demographics, marital status, gender, income, age, qualification etc., analysis can check the level of likely satisfaction with life/job/product/services, given a list of symptoms, one can predict if a patient is likely to be diagnosed with initial/intermediate/serious stages of a particular disease and more.</p>
<p>KMeans Clustering Algorithm:</p>	<p><u>What is the KMeans Clustering Algorithm and How Does an Enterprise Use it to Analyze Data?</u> https://www.smartent.com/blog/what-is-the-kmeans-clustering-algorithm-and-how-does-an-enterprise-use-it-to-analyze-data) Use Case(s): Loan applicants grouped as low, medium, and high risk based on applicant age, annual income, employment tenure, a movie ticket booking website can group users into frequent ticket buyers, moderate ticket buyers and occasional ticket buyers, based on past movie ticket purchases, and more.</p>
<p>Descriptive Statistics:</p>	<p><u>What is Descriptive Statistics and How Do You Choose the Right One for Enterprise Analysis?</u> https://www.smartent.com/blog/what-is-descriptive-statistics-and-how-do-you-choose-the-right-one-for-enterprise-analysis) Use Case(s): Average age and income for a particular type of product category purchased, Identify the most popular dish served in the restaurant or find out the most frequent rating given by customers for a given movie/restaurant or most frequent size or category of a sold product and more.</p>
<p>Holt-Winters Forecasting:</p>	<p><u>What is the Holt-Winters Forecasting Algorithm and How Can it be Used for Enterprise Analysis?</u> https://www.smartent.com/blog/what-is-the-holt-winters-forecasting-algorithm-and-how-can-it-be-used-for-enterprise-analysis) Use Case(s): Forecasting number of viewers by day for a particular game show for next two months. Input data: Last six months daily viewer count data, insurance claim manager can forecast policy sales for next month based on past 12 months data and more.</p>
<p>Trends and Patterns:</p>	<p><u>What Are Data Trends and Patterns, and How Do They Impact Business Decisions?</u> https://www.smartent.com/blog/what-are-data-trends-and-patterns-and-how-do-they-impact-business-decisions) Use Case(s): identify seasonality pattern when fluctuations repeat over fixed periods of time and where patterns do not extend beyond 1 year, analyze a stationary time series with statistical properties, where variances are all</p>

	constant over time, or cyclical when fluctuations do not repeat over fixed periods of time, are unpredictable and extend beyond a year, and more.
--	---

Each of these techniques, methods and algorithms has a unique value in advanced analytics. Augmented Data Discovery tools allow business users to gather and analyze data using these techniques within a sophisticated, intuitive navigation that is designed to guide users through the processing of selecting the appropriate algorithm or analytical technique based on the type of data selected.