Indian Statistical Institute

Foundation Course

on

Predictive Modeling

using

Python

1

CONTENTS

Indian Statistical Institute

2

Introduction *to* Predictive Modeling

PREDICTIVE MODELING

Introduction

A set of methods to arrive at quantitative solutions to problems of business interest

Part of data science or statistical learning

Has assumed tremendous importance in the recent past as the data availability is constantly on the rise

There is widespread belief that the existing data may be fruitfully analyzed to arrive at hitherto unknown insights

PREDICTIVE MODELING

Examples

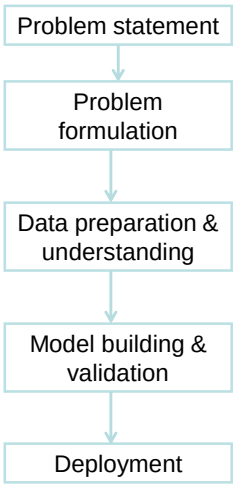
1. An automobile manufacturer wants to understand how the fault and failure related data captured through the sensors may be used to classify the condition of vehicles so that preventive maintenance may be carried out optimally
2. An insurance company may wish to classify drivers as very risky, risky, safe etc. on the basis of their driving habits so that insurance premium may be fixed intelligently.
3. A company engaged in oil exploration may need to estimate the time and expenses of drilling under different geological conditions before taking up a drilling assignment.
5. Credit card as well as health insurance companies may wish to identify fraudulent transactions so that appropriate actions may be initiated
6. An email service provider may wish to develop a method to classify spam mail from usual mail on the basis of the content of the mail

5

Indian Statistical Institute

PREDICTIVE MODELING

Predictive Modeling Process



6

Indian Statistical Institute

PREDICTIVE MODELING

Supervised learning

Understanding the behavior of a target (response / dependent / *y*) variable as a set of input or process varaibles (independent / explanatory / *x*) change

Typically attempts are made to develop a function or model to estimate the target

Often called dependency analyses

7

Indian Statistical Institute

PREDICTIVE MODELING

Supervised learning: Examples

1. Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient
2. Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data
3. Predict whether a particular credit card transaction is fraudulent. The prediction is to be based on past transaction history, transaction type, reputation of the merchants involved and other similar variables
4. Identify the impact of different variables like price, relative brand position, general economic condition, level of competition, and product type (luxury / necessity, etc) on the demand of a particular product during a given period

8

Indian Statistical Institute

PREDICTIVE MODELING

Unsupervised learning: Examples

1. Identification of typical profile of employees who quit quickly
2. Identification of products that are usually sold together
3. Grouping of cities or geographies with respect to their characteristics
4. Develop a scale to measure brand position

9

Indian Statistical Institute

PREDICTIVE MODELING

Predictive modeling tasks

1. Hypotheses testing
2. Classification and class probability estimation
3. Value estimation, explanatory and causal models

10

Indian Statistical Institute

PREDICTIVE MODELING

Hypothesis testing

Hypotheses are statements about a given phenomenon

Hypothesis testing consists of determining the plausibility of the statements on the basis of data

Examples

1. Increasing number of years of education increases earning potential
2. Design A produces a lower defect rate compared to design B
3. A particular design of a web page leads to more conversion compared to another

11

Indian Statistical Institute

PREDICTIVE MODELING

Classification & class probability estimation

Used in situations where the target is to be classified

The problem is to allocate the target variable to one of the classes based on the value of some explanatory variable(s)

In most cases the probability that the target will belong to different classes is first estimated

Allocation to a particular class is made on the basis of the estimated probabilities

Examples

1. Classification of credit card transaction as fraudulent or not
2. Classification of whether a customer will renew her contract or not
3. Classification of whether a sales bid will be won, lost or abandoned by the customer
4. Classification of a loan application as low, high or medium risk

12

PREDICTIVE MODELING

Value estimation

Used to estimate or predict the value of a target variable rather than classifying the same

The value needs to be estimated based on certain explanatory variables

Examples

1. Finding the lifetime value of a customer
2. Estimating the effort required to complete a software development project
3. Finding the total number of cheques that may arrive for processing

PREDICTIVE MODELING

Fundamental tasks and techniques: Relationship

SL No	Fundamental task	Statistical / data mining technique
1	Phenomenon Understanding	Descriptive Statistics, hypothesis testing, graphical analysis and data visualization, contingency tables
2	Classification	Logistic regression, Discriminant analysis, Decision trees, Neural networks, Support vector machine, Naive Bayes classification, etc
3	Value Estimation	Table lookup, k nearest neighbor, Regression models – Multiple linear regression and its variants including shrinkage methods, Survival Analysis, Neural networks, non-parametric methods, etc

DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

Importance of Measurement

We don't know what we don't know

If we can't express what we know in the form of numbers, we really don't know much about it

If we don't know much about it, we can't control it

If we can't control it, we are at the mercy of chance

Successful organizations have a common language to communicate

Common language promotes objectivity in decision-making process

Does the problem really exist? Measurement will answer that question

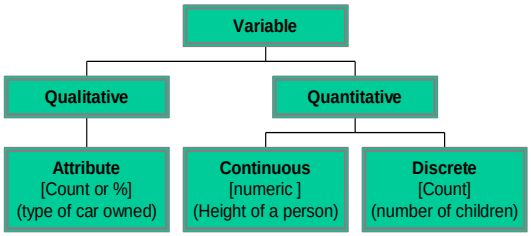
Improvement can happen only if we understand where we are and where we should go

Have we reached where we intended to? -- *only data answers that question*

A good data collection simplifies the problem solving effort

DESCRIPTIVE STATISTICS

Variable Types



Qualitative variable examples : Programming Style**,** Gender, religious affiliation, type of automobile owned, eye color

Quantitative variable examples : Size of a Program, Balance in bank account, minutes remaining in class, number of children in a family.

17

DESCRIPTIVE STATISTICS

Types of Data

Quantitative characteristics are of two principal types – Discrete and continuous.

Discrete: A characteristic may take only some isolated or discrete values, e.g., the number of defects on an item, the number of breakdown of machines in a shop, etc. This is often obtained by counting.

Continuous: A characteristic may theoretically take any value within a definite range, e.g., the body weight, length of a pencil, service response time, temperature in a room, etc. This is recorded with the help of a measuring equipment of defined accuracy.

Qualitative characteristic is typically represented by

Attribute Data: A characteristic that is judged “Good” or “Bad” by comparing it to a referenced standard

Example: Defective / Non Defective, Delivered on time / Not on time

18

DESCRIPTIVE STATISTICS

Types of Data

Discrete data

Data that can take a limited number of values. For example:

- Days in a week
- Number of defects

Attribute Data

Data that can be classified in binary form. For example:

- Items passed / rejected
- Responses to a customer satisfaction survey in 2 categories" Satisfied" and "Not Satisfied"

Continuous Data

Data that can take any value within a specified range. For example:

- Temperature in this room
- Exchange rate of a currency
- Yield of a process
- Height of a person

DESCRIPTIVE STATISTICS

Quiz

1. Give 3 examples of each type of data
   - Quantitative Continuous
   - Qualitative Attribute
   - Quantitative Discrete

2. What guidelines help you distinguish between Quantitative Continuous and Quantitative Discrete data?

3. What guidelines help you distinguish between Qualitative Attribute and Quantitative Discrete data?

DESCRIPTIVE STATISTICS

Answers

1.

Quantitative Continuous	Qualitative Attribute	Quantitative Discrete
Cycle time	Late deliveries	No. of Complaints
Cost	Accurate PO	No. of Errors
Effort	Defective programs	No. of Bugs

2.

–Continuous data are measurable, usually on a continuous scale, such as time, amount (money), volume, length, or temperature

–Discrete data are countable, on an integer scale, such as items with a characteristic (attribute) or number of occurrences

21

DESCRIPTIVE STATISTICS

Answers

3.

Qualitative Attribute	Quantitative Discrete
You are interested in counting items with an attribute (Ex: projects delivered late)	You are interested in counting occurrences for a given opportunity (Ex: complaints per week)
You can also count items without the attribute (Ex: orders delivered “not-late” = on time)	You *cannot* count a “non occurrence” (Ex: impossible to count “non -complaints”)
You can determine the proportion of items with the attribute (Ex: % late deliveries)	There are no physical limits to the number of occurrences (no limit to the number of complaints possible; this is often referred to as “countable infinity”)

22

DESCRIPTIVE STATISTICS

Role of Statistics

"There are three kind of lies: Lies, Damn Lies and Statistics….." Benjamin Disraeli
"Figures don't lie: liars figure…" Anonymous

What is Statistics?

Statistics is the science concerned with collection, organizing, presenting and analyzing data so that intelligent and more effective judgment may be formed upon them.

"Statistics may be regarded as the technique of drawing valid conclusions from a limited body of experimental or observational data."

"Statistics is the name for the body of scientific methods (statistical methods) which are meant for the collection, analysis and interpretation of numerical data."

DESCRIPTIVE STATISTICS

Population and Sample

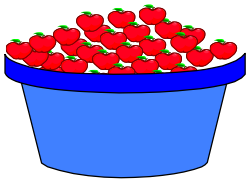
Population: An aggregate of living or non-living things whose characteristic(s) is under study.

Sample: A sample is part of the population. A sample is used as a basis for making estimates and inferences about the population. By size of the sample, we mean the number of elements or groups of elements of the population that constitute the sample.

DESCRIPTIVE STATISTICS

Population and Sample

- The entire set of items is called the Population.
- The small number of items taken from the population to make a judgment of the population is called a Sample.
- The numbers of samples taken to make this judgment is called Sample size.

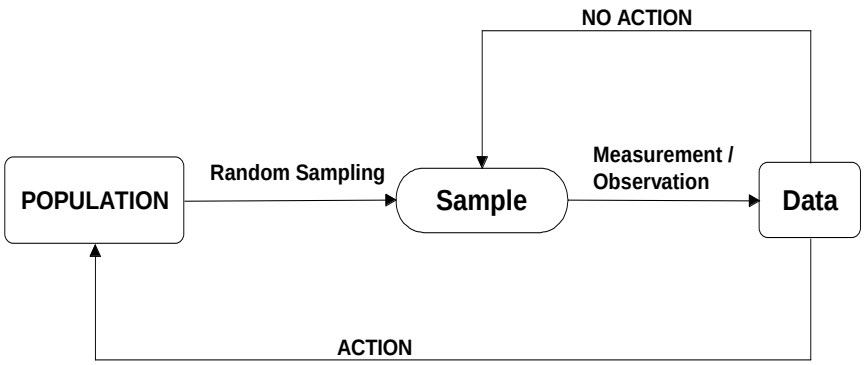


Population

Sample

DESCRIPTIVE STATISTICS

Population and Sample



DESCRIPTIVE STATISTICS

Advantages of sampling over complete enumeration

- Reduced cost
- Greater speed
- Greater scope
- Desired accuracy

27

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Frequency distribution

Frequency distribution is a form of data summarization. It is obtained by first dividing the range of (sample) values into a number of class-intervals, and then classifying the values into these intervals. A frequency distribution thus consists of a set of class-intervals and the frequency of values in each of them. It helps comprehension of both location and dispersion of data.

28

DESCRIPTIVE STATISTICS

Example: Response Time (in minutes) to 100 customer complaints

93.7	98.8	100.5	75	87.6	102.3	83.4	98.1	98.3	147
100.4	116.1	82.9	71.3	94.6	117.8	104.3	109.9	109	106.7
109.2	91.1	89.7	107.2	65.9	96.8	112.9	86.2	112.3	112.2
101.2	116.1	80.2	105.2	90.6	130.1	108.1	98	94.1	107.6
105.3	81.8	93.3	99.4	109.5	87.4	142.1	87.2	101.5	101.6
143.9	96.3	84.9	109.3	94.5	79	115.9	93.1	110.3	121.7
63.9	87.2	107.2	132.7	126.2	131.4	125.2	109	104.3	106.9
79.3	89.8	88.9	103.7	119.6	77.4	76.5	94.5	98.5	80.9
111.5	88.3	100	99.7	127.5	121.5	91.8	74.6	90.1	110.5
76.3	87.9	98.5	82.8	100.2	114.4	92.9	110.2	97.5	100.3

29

DESCRIPTIVE STATISTICS

The steps involved in the construction of a frequency table are:

1. Collect around 100 observations(measurements) to form a frequency distribution. Each observation should be recorded to the same degree of accuracy.
2. Obtain the range of measurements as $R = X_{Max} - X_{Min}$
3. Decide on appropriate number of class intervals (k) based on the guidelines provided below :

No. of observations	Recommended No. of Class Intervals
50-100	7
101-200	8
201-500	9
500-100	10
Over 1000	11-20

30

DESCRIPTIVE STATISTICS

These guidelines are not rigid and can be modified as and when found necessary. Generally, the number of classes should not be too large; otherwise the purpose of classification, viz, summarization of data will not be served. Moreover, by taking a large number of classes, one will introduce an irregular pattern in the frequencies which may be completely absent in the actual distribution.

The number of classes should not be too small either, for this also may obscure the true nature of the distribution, i.e, some useful information may be lost due to too much condensation. Further, if the number of classes is too small, each observation within any class is equal to the midpoint of that class will make the computed value of measures of central tendency and dispersion very unreliable.

Lastly, the classes should preferably be of equal width. Otherwise, the class frequencies will not be directly comparable and the computations of statistical measures will be laborious.

31

DESCRIPTIVE STATISTICS

4. Calculate the approximate length of class interval as equal to

$$C = \frac{R+W}{K}; W \text{ be the smallest unit of measurement.}$$

5. Specify the class boundaries. The lower limit (boundary) of the first class is given by $X_{\min} - 0.5W$. The subsequent ones are obtained by adding the class width C.

6. Read the observations in order, and for each observation, put a tally mark against the particular class containing the observation.

7. Count the total number of tally marks in each class. This is called the frequency of that class.

32

DESCRIPTIVE STATISTICS

Response Time (in minutes) to 100 customer complaints

93.7									
100.4									
109.2									
101.2									
105.3									
143.9									
63.9									
79.3									
111.5									
76.3									

DESCRIPTIVE STATISTICS

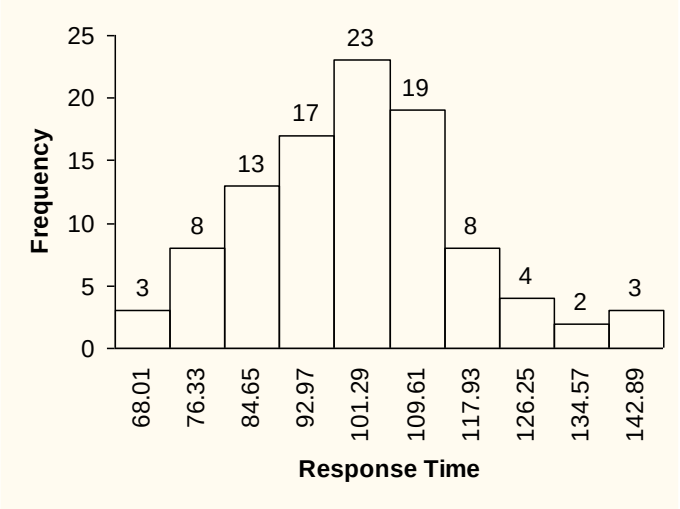
Frequency Distribution of Response Time

Class	Midpoint of class (x_i)	Frequency (f_i)	% Cumulative Frequency (F_i)
63.85 – 72.17	68.01	3	3
72.17 – 80.49	76.33	8	11
80.49 – 88.81	84.65	13	24
88.81 – 97.13	92.97	17	41
97.13 – 105.45	101.29	23	64
105.45 – 113.77	109.61	19	83
113.77 – 122.09	117.93	8	91
122.09 – 130.41	126.25	4	95
130.41 – 138.73	134.57	2	97
138.73 – 147.05	142.89	3	100

$n = \sum f_i = 100$

DESCRIPTIVE STATISTICS

Histogram of Response Time



DESCRIPTIVE STATISTICS

Specimen of Histogram

Figure shows 12 typical histogram.

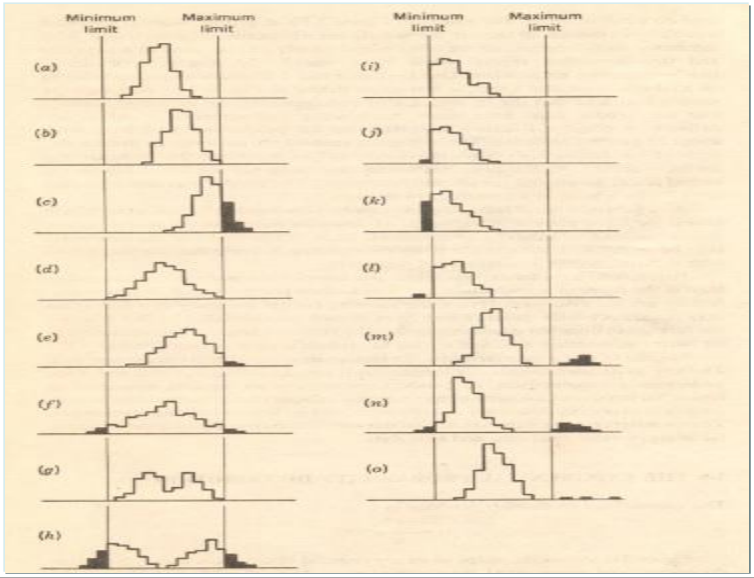
1. Does the process have the ability to meet the specification limits?
2. What action, if any, is appropriate on the process?

These questions can be answered by analyzing.

1. The centering of the histogram. This defines the aim of the process.
2. The width of the histogram. This defines the variability about the aim.
3. The shape of the histogram. When a normal or bell-shaped curve is expected, then any significant deviation or other aberration is usually caused by a manufacturing (or other) condition that may be the root of the quality problem. For example, histogram with two or more peaks may reveal that several “populations” have been mixed together.

Histogram illustrate how variables data provide much more information than do attributes data.

DESCRIPTIVE STATISTICS



37

DESCRIPTIVE STATISTICS

For example, Figure b,d,g, and i warn of potential trouble even though all units in the sample are within specification limits. With attributes measurement, all the units would simply be classified as acceptable and the inspection report would have stated “50 inspected, 0 defective” – therefore no problem. One customer had a dramatic experience based on a lot which yielded a sample histogram similar to figure "i". Although the sample indicated that the lot met quality requirements, the customer realized that the vendor must have made much scrap and screened it out before delivery. A rough calculation indicated that full production must have been about 25 percent defective. The histogram enabled the customer to deduce this without ever having been inside the vendor’s plant. Note how the “ product tells on the process”. As the customer would eventually pay for this scrap (in the selling price), he wanted the situation corrected. The vendor was contacted and advice was offered in a constructive manner.

As a general rule, at least 50 measurements are needed for the histogram to reveal the basic pattern of variation. Histogram based on too few measurements can lead to incorrect conclusions, because the shape of the histogram may be incomplete without the observer realizing it.

38

DESCRIPTIVE STATISTICS

Histogram have limitations. Since the samples are taken at random rather than in the order of manufacture, the time to time process trends during manufacture are not disclosed. Hence the seeming central tendency of a histogram may be illusory – the process may have drifted substantially. In like manner, the histogram does not disclose whether the vendor's process was operating at its best, i.e., whether it was in a state of statistical control.

In spite of these shortcomings, the histogram is an effective analytical tool. The key to its usefulness is its simplicity. It speaks a language that everyone understands – comparison of product measurements against specification limits. To draw useful conclusions from this comparison requires little experience in interpreting frequency distribution, and no formal training in statistics. The experience soon expands, to include applications in development, manufacturing, vendor relations, and field data.

DESCRIPTIVE STATISTICS

Exercise 1: The data shown below are chemical process yield on successive days. Construct a histogram for these data.

94.1	87.3	94.1	92.4	84.6	85.4
93.2	84.1	92.1	90.6	83.6	86.6
90.6	90.1	96.4	89.1	85.4	91.7
91.4	95.2	88.2	88.8	89.7	87.5
88.2	86.1	86.4	86.4	87.6	84.2
86.1	94.3	85	85.1	85.1	85.1
95.1	93.2	84.9	84	89.6	90.5
90	86.7	87.3	93.7	90	95.6
92.4	83	89.6	87.7	90.1	88.3
87.3	95.3	90.3	90.6	94.3	84.1

DESCRIPTIVE STATISTICS

Exercise 2: The time to failure in hours of an electronic component subjected to an accelerated life test is given below. Construct histogram of these data.

127	124	121	118
125	123	136	131
131	120	140	125
124	119	137	133
129	128	125	141
121	133	124	125
142	137	128	140
151	124	129	132
160	142	130	129
125	123	122	126

41

DESCRIPTIVE STATISTICS

Measuring the Shape

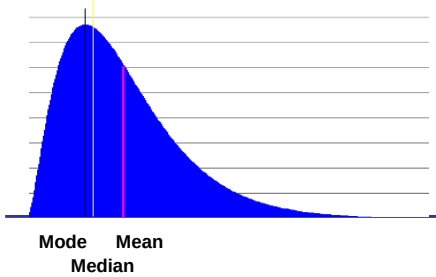
If data pattern is Symmetric

- It's a data set in which spread of the data set around its mean is identical
- For such a data set - mean = mode = median

If data pattern is Asymmetric

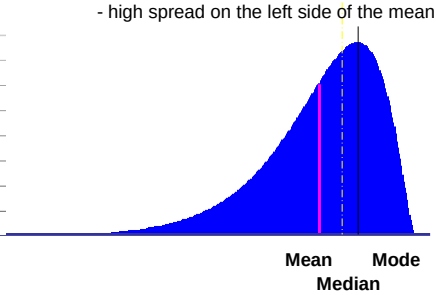
–Positive / Right skewed

high spread on the right side of the mean



Negative / Left skewed

- high spread on the left side of the mean



DESCRIPTIVE STATISTICS

Continuous Data Characteristics

Location / Central Tendency

- Measure of the center point of any data set

Spread / Dispersion

- Measure of the spread of any data set around its center

Shape

- Measure of symmetry of any data set around its center

43

DESCRIPTIVE STATISTICS

Measures of location

- Mean
- Median
- Mode
- Quartiles

44

DESCRIPTIVE STATISTICS

Mean

- Mean is the arithmetic average of all data points in a data set

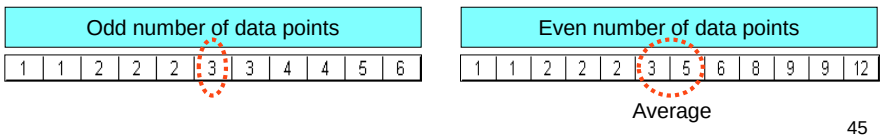
$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots\ldots\ldots + X_n}{n}$$ Where n = number of data points

Mode

- Mode is the most frequently occurring data point in a data set

Median

- Median is the middle data point of a data set arranged in an ascending / descending order



DESCRIPTIVE STATISTICS

Spread Measures

Range

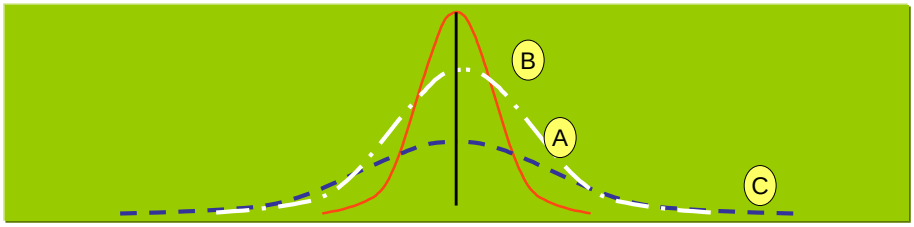
- Range is the difference between the maximum & minimum data value Variance / Standard Deviation
- Variance and Standard Deviation tell us how individual data points are spread around mean

$$\text{Variance} = s^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots\ldots + (X_n - \bar{X})^2}{(n-1)}$$

$$\text{Standard Deviation} = s = \sqrt{s^2}$$

DESCRIPTIVE STATISTICS

Importance of Spread



Mean of Curve 'A' is more representative of its data set as compared to Curves 'B' & 'C'

If this graph represents say, delivery time of an item to different customers, then individual customer's experience about delivery time would be different.

47

DESCRIPTIVE STATISTICS

Inter Quartile Range

The lower quartile Q_1 is the value such that one-fourth of the observations fall below it and three fourths fall above it.

The middle quartile is the median

The third quartile Q_3 is the value such that three -fourths of the observations fall below it and one-fourth above it.

The Inter Quartile Range IQR is the difference between the third quartile and the first quartile.

Thus IQR= Q_3-Q_1

48

DESCRIPTIVE STATISTICS

Box Plot

Box Plots are simple means of providing a useful picture of how the data are distributed. To draw Box Plot

- Determine Q_1, Q_3 and IQR
- A line is drawn at the median to divide the box
- Two lines, known as Whiskers are drawn outward from the box.

One line extends the top edge of the box at Q_3 to either maximum data value or Q_3+1.5 (IQR). Another line from the bottom edge of the box at Q_1 extends downward to a value that is either the minimum data value or Q_1 – 1.5 IQR whichever is greater.

- The end points of the whiskers are known as upper and lower adjacent values
- Values that fall outside the adjacent values are candidates for consideration as outliers. They are plotted as asterisks (*).

DESCRIPTIVE STATISTICS

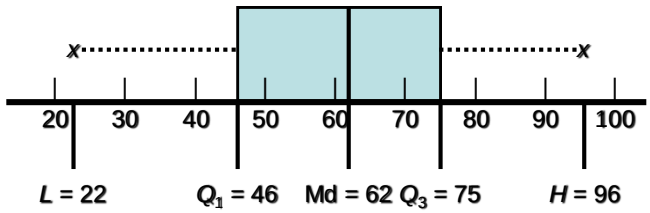
93.7
100.4
109.2
101.2
105.3
143.9
63.9
79.3
111.5
76.3
98.8
116.1
91.1
116.1
81.8
96.3
87.2
89.8
88.3
87.9
100.5
82.9
89.7
80.2
93.3

DESCRIPTIVE STATISTICS

Box Plots

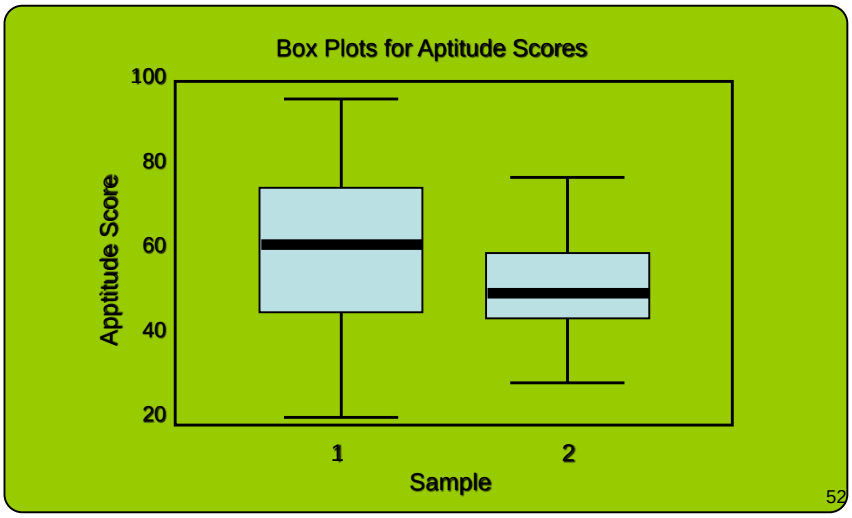
For a given data set:

L = 22	Q_3 = 75
Q_1 = 46	IQR = 75 - 46 = 29
Q_2 = Md = 62	H = 96



51

DESCRIPTIVE STATISTICS



52

DESCRIPTIVE STATISTICS

Exercise 1: The data of 30 customers on credit card usage in INR1000, are given.

1. Summarize and interpret the credit card usage?
2. Plot histogram?

NORMAL DISTRIBUTION

NORMAL DISTRIBUTION

Introduction to Normal Distribution

Developed by *Karl Gauss*

Most prominently used distribution in statistics

Applicability to many situations where given the population knowledge, we need to predict the sample behavior

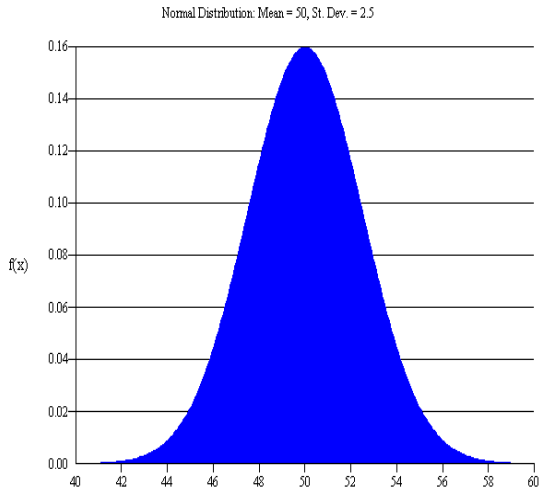
Many natural phenomena follows Normal Distribution

- Human characteristics such as weights, heights & IQ's

Also Physical process outputs such as yield, follow Normal Distribution

NORMAL DISTRIBUTION

Normal Distribution Properties



Normal Distribution

The normal curve is bell-shaped and has a single peak at the center of the distribution

It is symmetric about mean

The arithmetic mean, median, and mode of the distribution are equal and located at the peak

Thus half the area under the curve is above the mean and half is below it

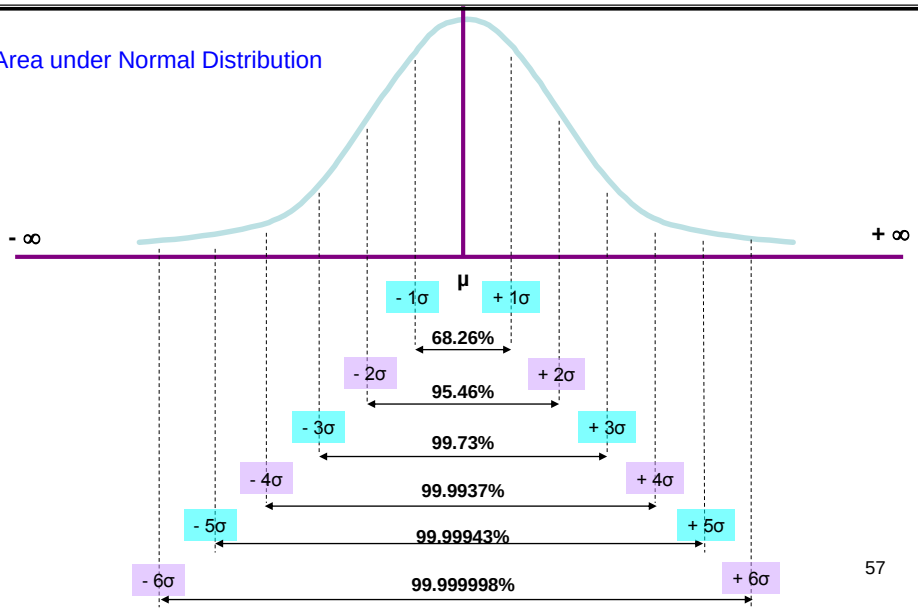
Area under the curve is ~1

It is asymptotic

Indian Statistical Institute

NORMAL DISTRIBUTION

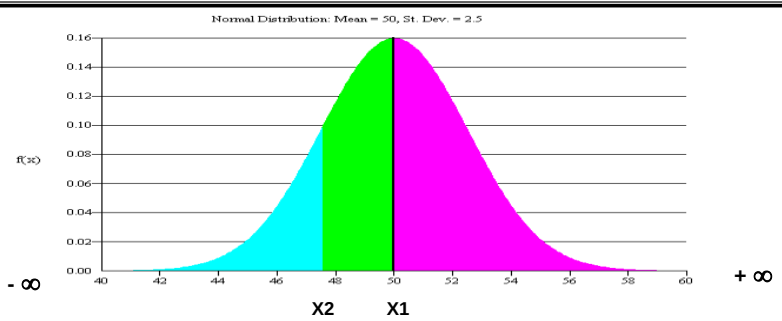
Area under Normal Distribution



57

Indian Statistical Institute

NORMAL DISTRIBUTION

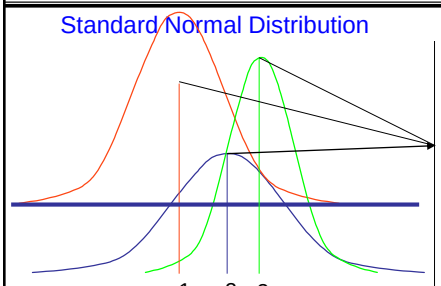


- Area covered within X1 & X2 represents the probability that a data point lies between X1 & X2
- If mean, X1 = 50 & X2 = 47.5, then the area between them = ½ of 68.26% = 34.13%
- In other words, if the distribution (say Productivity of a Programmer in LOC/day) follows normal distribution, then probability, that the productivity on a particular day in future would lie between 47.5 and 50 is 34.13%

58

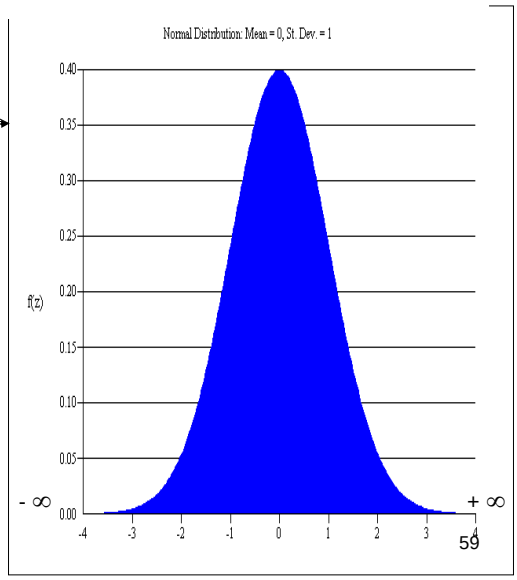
NORMAL DISTRIBUTION

Standard Normal Distribution



Instead of dealing with a family of normal distributions with varying means & standard deviations, a standard normal curve standardizes all the distributions with a single curve that has a mean of 0 & standard deviation of 1

It's illustrated as N ~ (0,1), i.e. mean = 0 & standard deviation = 1



59

NORMAL DISTRIBUTION

Concept of Z Value

To standardize different measurement units, the Z variable is used

$$Z = \frac{X - \mu}{\sigma}$$

Where X = Value of the data point we are concerned with

μ = Mean of the data points

σ = Standard Deviation of the data points

Z = Number of standard deviations between X & the mean (μ)

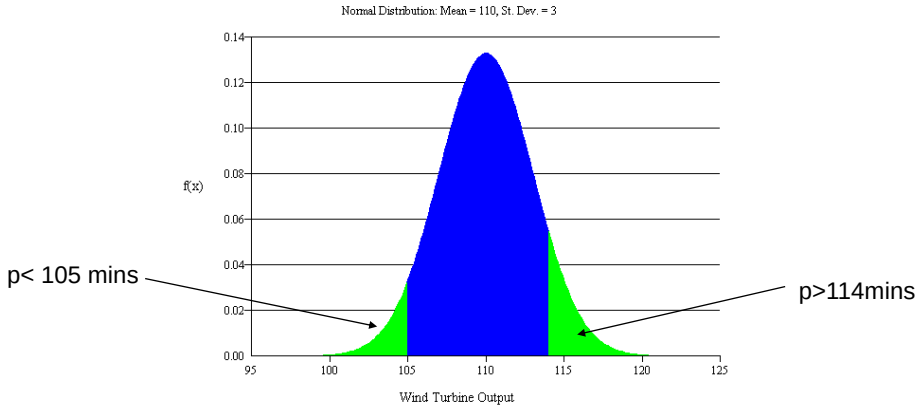
Z value is unique for each probability within the normal distribution

It helps in finding probabilities of data points anywhere within the distribution

60

NORMAL DISTRIBUTION

In a certain organization, the response time to customer complaints is normally distributed with a mean of 110 minutes with a standard deviation of 3 minutes. What is the probability that for any complaint the response time will be more than 114 mins? Less than 105 mins?



Introduction
to
Python

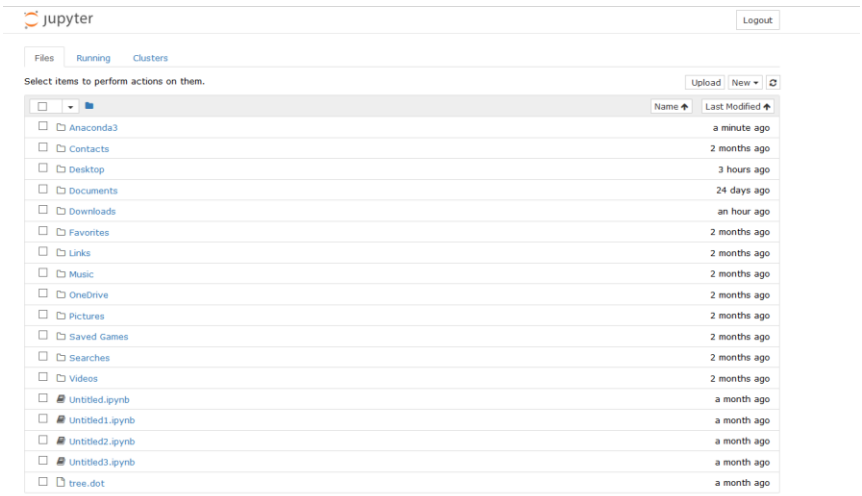
PYTHON INSTALLATION

1. Download Anaconda from http://jupyter.readthedocs.io/en/latest/install.html
2. Run the set up (exe) file and follow instructions
3. Check Jupyter notebook is installed

63

PYTHON INSTALLATION

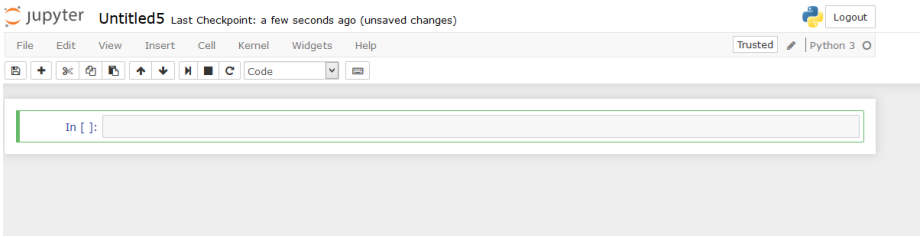
3. Open Jupyter Notebook



64

PYTHON INSTALLATION

3. Open Jupyter Notebook



65

DESCRIPTIVE STATISTICS *using* Python

66

DESCRIPTIVE STATISTICS

Exercise 1: The monthly credit card expenses of an individual in 1000 rupees is given in the file Credit_Card_Expenses.csv.

a. Read the dataset to Python
b. Compute mean, median minimum, maximum, range, variance, standard deviation, skewness, kurtosis and quantiles of Credit Card Expenses
c. Compute default summary of Credit Card Expenses
d. Draw Histogram of Credit Card Expenses

DESCRIPTIVE STATISTICS

Reading a csv file : Source code

```
import pandas as mypd
mydata = mypd.read_csv("E:/ISI/PM-01//Data/Credit_Card_Expenses.csv")
mydata
```

To read a particular column or variable of data set to a ne variable

Example: Read CC_Expenses to CC

```
cc = mydata.CC_Expenses
cc
```

DESCRIPTIVE STATISTICS

Operators - Arithmetic

Operator	Description
+	addition
-	subtraction
*	multiplication
/	division
**	exponentiation
%	modulus (x mod y) 5%2 is 1

69

DESCRIPTIVE STATISTICS

Operators - Logical

Operator	Description
<	less than
<=	less than or equal to
>	greater than
>=	greater than or equal to
==	exactly equal to
! =	not equal to

70

DESCRIPTIVE STATISTICS

Descriptive Statistics

Computation of descriptive statistics for variable CC

Function	Code	Value
Mean	cc.mean()	59.2
Median	cc.median()	59
Mode	cc.mode()	59
Standard deviation	cc.std()	3.105
Variance	cc.var()	9.642
Minimum	cc.min()	53
Maximum	cc.max()	65
Percentile	cc.quantile(0.9)	63
Skewness	cc.skew()	-0.09
Kurtosis	cc.kurt()	-0.436

71

DESCRIPTIVE STATISTICS

Descriptive Statistics

Statistics	Code
Summary	cc.describe()

Statistics	Value
Count	20
Mean	59.2
Standard Deviation	3.1052
Minimum	53
Q1	57
Median	59
Q3	61
Maximum	65

72

DESCRIPTIVE STATISTICS

Descriptive Statistics

Arithmetic functions for variable CC

Function	Code	Value
Count	cc.count()	20
Sum	cc.sum()	1148
Product	cc.prod()	6.21447E+18

Function	Code	Value
Square root	Import math as mymath mymath.sqrt(49)	7
Sum of Squares	sum(cc**2)	70276

73

DESCRIPTIVE STATISTICS

Graphs:

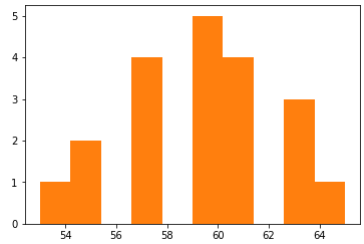
Graph	Code
Histogram	import matplotlib.pyplot as myplot myplot.hist(cc) myplot.show()
Box Plot	myplot.boxplot(cc) myplot.show()

74

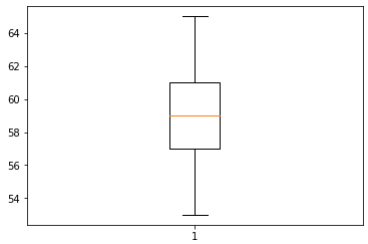
DESCRIPTIVE STATISTICS

Graphs:

Histogram



Box plot



75

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

a. Import the file to Python

b. Compute descriptive summary of variable Credit Card Usage

c. Check whether the average usage varies with sex?

d. Check whether the average credit card usage vary with those who do shopping with credit card and those who don't do shopping?

e. Check whether the average credit card usage vary with those who do banking with credit card and those who don't do banking?

f. Compute the aggregate average of usage with sex & shopping?

g. Compute the aggregate average of usage with all three factors?

76

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Reading dataset to variable: mydata

```
import pandas as mypd
mydata = mypd.read_csv("E:/ISI\PM-01/Data/CC_Expenses_Exercise.csv")
mydata
```

Reading the variable

```
cc = mydata.Credit_Card_usage
gender = mydata.Sex
shopping = mydata.Shopping
banking = mydata.Banking
```

77

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Computing descriptive statistics for variable : CC

```
cc.describe()
```

Count	Mean	SD	Minimum	25%	50%	75%	Maximum
30	66	42.9595	20	30	55	90	150

78

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Computing average credit card usage for different sex

```
cc.groupby(gender).mean()
```

Group	Sex	Average Credit Card Usage
1	Male	93.33333
2	Female	38.66667

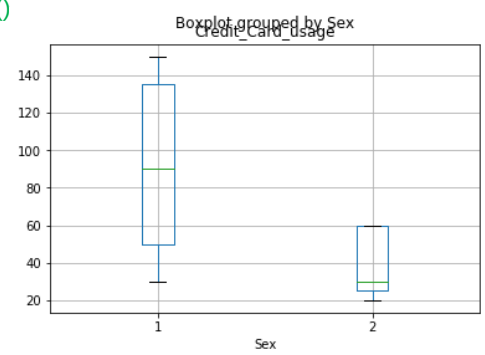
79

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Box plot of Credit Card usage by sex

```
import matplotlib.pyplot as myplot
mydata.boxplot(column = 'Credit_Card_usage', by = 'Sex')
myplot.show()
```



80

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Computing aggregate average of credit card usage for different sex and shopping

```
cc.groupby([gender, banking]).mean()
```

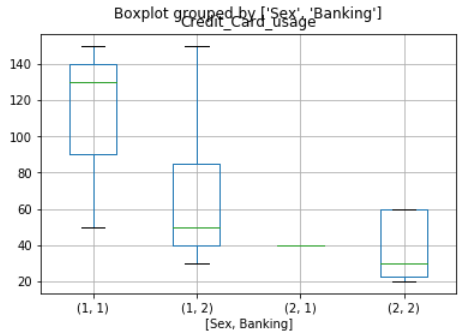
Sex	Banking	Average Credit Card Usage
Male	Yes	115.00000
Male	No	68.57143
Female	Yes	40.00000
Female	No	38.57143

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Box plot of Credit Card usage by sex sex and banking

```
mydata.boxplot(column = 'Credit_Card_usage', by =[ 'Sex', 'Banking'])
myplot.show()
```



DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Computing aggregate average of credit card usage by 3 factors

```
cc.groupby([gender, shopping, banking]).mean()
```

Sex	Shopping	Banking	Average Credit Card Usage
Male	Yes	Yes	130.00000
Male	Yes	No	62.00000
Male	No	Yes	70.00000
Male	No	No	85.00000
Female	Yes	Yes	40.00000
Female	Yes	No	48.00000
Female	No	No	33.33333

83

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Computing aggregate summary of credit card usage by 3 factors

```
cc.groupby([gender, shopping, banking]).describe()
```

Sex	Shopping	Banking	Count	Mean	SD	Min	25%	50%	75%	Max
Male	Yes	Yes	6	130	20.97618	90	130	135	140	150
Male	Yes	No	5	62	49.699	30	40	40	50	150
Male	No	Yes	2	70	28.284	50	60	70	80	90
Male	No	No	2	85	7.071	80	82.5	85	87.5	90
Female	Yes	Yes	1	40	-	40	40	40	40	40
Female	Yes	No	5	48	16.432	30	30	60	60	60
Female	No	No	9	33.3333	16.583	20	20	30	40	60

84

TEST *of* HYPOTHESIS

TEST OF HYPOTHESIS

Introduction:

In many situations, it is required to accept or reject a statement or claim about some parameter

Example:

1. The average cycle time is less than 24 hours
2. The % rejection is only 1%

The statement is called the hypothesis

The procedure for decision making about the hypothesis is called hypothesis testing

Advantages

1. Handles uncertainty in decision making
2. Minimizes subjectivity in decision making
3. Helps to validate assumptions or verify conclusions

TEST OF HYPOTHESIS

Some of the commonly used hypothesis tests:

- Checking mean equal to a specified value ($mu = mu_0$)
- Two means are equal or not ($mu_1 = mu_2$)
- Two variances are equal or not ($sigma_1^2 = sigma_2^2$)

- Proportion equal to a specified value ($P = P_0$)
- Two Proportions are equal or not ($P_1 = P_2$)

TEST OF HYPOTHESIS

Null Hypothesis:

A statement about the status quo

One of no difference or no effect

Denoted by H0

Alternative Hypothesis:

One in which some difference or effect is expected

Denoted by H1

TEST OF HYPOTHESIS

Types of errors in hypothesis testing

The decision procedure may lead to either of the two wrong conclusions

Type I Error

Rejecting the null hypothesis H0 when it is true

Type II Error

Failing to reject the null hypothesis H0 when it is false

Alpha (Significance level) = Probability of making type I error

Beta = Probability of making type II error

Power = 1 – Beta : Probability of correctly rejecting a false null hypothesis

TEST OF HYPOTHESIS

Hypothesis Testing: General Procedure

1. Formulate the null hypothesis H0 and the alternative hypothesis H1
2. Gather evidence (data collection)
3. Based on evidence take a decision to accept or reject H0

TEST OF HYPOTHESIS

Methodology demo: To Test Mean = Specified Value (mu = mu_0)

Suppose we want to test whether mean of a process characteristic is 5 based on the following sample data from the process

4	4	5	5	6
5	4.5	6.5	6	5.5

Calculate the mean of the sample, xbar = 5.15

Compare xbar with specified value 5

or xbar - specified value = xbar - 5 with 0

If xbar - 5 is close to 0

then conclude mean = 5

else mean $\neq$ 5

TEST OF HYPOTHESIS

Methodology demo : To Test Mean = Specified Value (mu = mu_0)

Consider another set of sample data. Check whether mean of the process characteristic is 500

400	400	500	500	600
500	450	650	600	550

Mean of the sample, xbar = 515

xbar - 500 = 515 - 500 = 15

Can we conclude mean $\neq$ 500?

Conclusion:

Difficult to say mean = specified value by looking at xbar - specified value alone

TEST OF HYPOTHESIS

Methodology demo: To Test Mean = Specified Value (mu = mu_0)

Test statistic is calculated by dividing (xbar - specified value) by a function of standard deviation

To test Mean = Specified value

Test Statistic t_0 = (xbar - Specified value) / (SD / $\sqrt{n}$)

If test statistic is close to 0, conclude that Mean = Specified value

To check whether test statistic is close to 0, find out p value from the sampling distribution of test statistic

TEST OF HYPOTHESIS

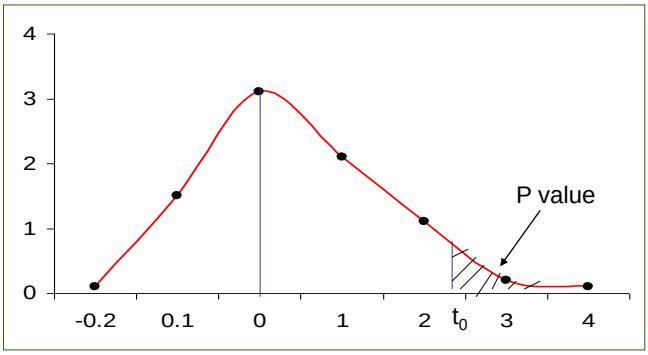
Methodology demo: To Test Mean = Specified Value

P value

The probability that such evidence or result will occur when H0 is true

Based on the reference distribution of test statistic

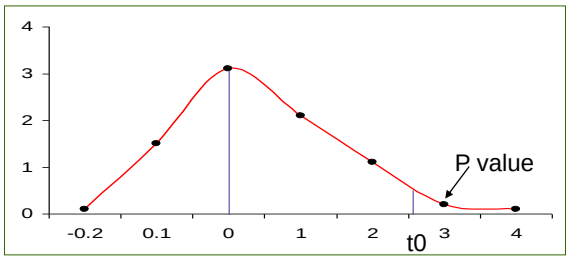
The tail area beyond the value of test statistic in reference distribution



TEST OF HYPOTHESIS

Methodology demo : To Test Mean = Specified Value

P value



If test statistic t_0 is close to 0 then p will be high

If test statistic t_0 is not close to 0 then p will be small

If p is small , $p < 0.05$ (with alpha = 0.05), conclude that $t \neq 0$, then

Mean ≠ Specified Value, H0 rejected

TEST OF HYPOTHESIS

To Test Mean = Specified Value (mu = mu_0)

Example: Suppose we want to test whether mean of the process characteristic is 5 based on the following sample data

4	4	5	5	6
5	4.5	6.5	6	5.5

H0: Mean = 5

H1: Mean ≠ 5

Calculate xbar = 5.15

SD = 0.8515

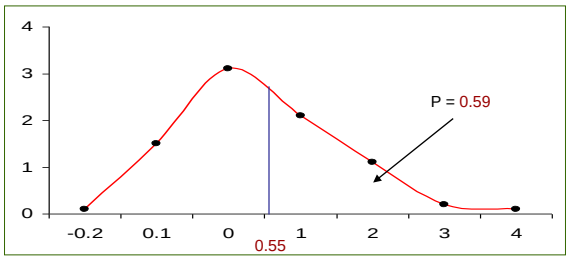
n = 10

Test statistic t_0 = (xbar - 5)/(SD / √n) = (5.15 - 5) / (0.8515 / √10) = 0.5571

TEST OF HYPOTHESIS

Example: To Test Mean = Specified Value (mu = mu_0)

$t_0 = 0.5571$



P ≥ 0.05, hence Mean = Specified value = 5.

H0: Mean = 5 is not rejected

TEST OF HYPOTHESIS

Hypothesis Testing: Steps

1. Formulate the null hypothesis H0 and the alternative hypothesis H1
2. Select an appropriate statistical test and the corresponding test statistic
3. Choose level of significance alpha (generally taken as 0.05)
4. Collect data and calculate the value of test statistic
5. Determine the probability associated with the test statistic under the null hypothesis using sampling distribution of the test statistic
6. Compare the probability associated with the test statistic with level of significance specified

TEST OF HYPOTHESIS

One sample t test

Exercise 1 : A company claims that on an average it takes only 40 hours to process any purchase order. Based on the data given below, can you validate the claim? The data is given in PO_Processing.csv

TEST OF HYPOTHESIS

One sample t test

Exercise 1 : A company claims that on an average it takes only 40 hours to process any purchase order. Based on the data given below, can you validate the claim? The data is given in PO_Processing.csv

Reading data to mydata

```
import pandas as mypd
from scipy import stats
mydata = mypd.read_csv("E:/ISI/PM-01/Data/PO_Processing.csv")
mydata
PT = mydata.Processing_Time
PT
```

Performing one sample t test

```
stats.ttest_1samp(PT, 40)
```

TEST OF HYPOTHESIS

One sample t test

Exercise 1 : A company claims that on an average it takes only 40 hours to process any purchase order. Based on the data given below, can you validate the claim? The data is given in PO_Processing.csv

Statistics	Value
t	3.7031
P value	0.00035

TEST OF HYPOTHESIS

One sample t test

Exercise 2 : A computer manufacturing company claims that on an average it will respond to any complaint logged by the customer from anywhere in the world in 24 hours. Based on the data, validate the claim? The data is given in Compaint_Response_Time.csv

Response Time	
24	26
31	27
29	24
26	23
28	27
26	28
29	27
29	23
27	27
31	23
25	25
29	27
29	26
25	28
26	27

TEST OF HYPOTHESIS

To Test Two Means are Equal:

Null hypothesis H0: $Mean_1 = Mean_2$ ($mu_1 = mu_2$)

Alternative hypothesis H1: $Mean_1 \neq Mean_2$ ($mu_1 \neq mu_2$)

or

H1: $Mean_1 > Mean_2$ ($mu_1 > mu_2$)

or

H1: $Mean_1 < Mean_2$ ($mu_1 < mu_2$)

TEST OF HYPOTHESIS

To Test Two Means are Equal: Methodology

Calculate both sample means xbar1 & xbar2

Calculate SD1 & SD2

Compare xbar1 with xbar2

Or xbar1 - xbar2 with 0

Calculate test statistic t_0 by dividing (xbar1 – xbar2) by a function of SD1 & SD2

t_0 = (xbar1 – xbar2) / (Sp $\sqrt{}$ ((1/n1)+(1/n2)))

Calculate p value from t distribution

If $p \geq 0.05$ then H0: $Mean_1 = Mean_2$ is not rejected

TEST OF HYPOTHESIS

Two sample t test

Exercise 1: A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. Based on the data given below, check whether the promotional activity resulted in increasing the sales. The outlets where promotional activity introduced are denoted by 1 and others by 2? The data is given in Sales_Promotion.csv

Outlet	Sales	Outlet	Sales
1	1217	2	1731
1	1416	2	1420
1	1381	2	1065
1	1413	2	1612
1	1800	2	1361
1	1724	2	1259
1	1310	2	1470
1	1616	2	622
1	1941	2	1711
1	1792	2	2315
1	1453	2	1180
1	1780	2	1515

TEST OF HYPOTHESIS

Two sample t test

Exercise 1: A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. Based on the data given below, check whether the promotional activity resulted in increasing the sales. The outlets where promotional activity introduced are denoted by 1 and others by 2?

Reading data to mydata

```
import pandas as mypd
from scipy import stats
mydata = mypd.read_csv("E:/ISI/PM-01/Data/Sales_Promotion.csv")
mydata
```

Reading the variables

```
sales_1 = mydata.Sales_Out1
sales_2 = mydata.Sales_Out2
```

TEST OF HYPOTHESIS

Two sample t test

Exercise 1: A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. Based on the data given below, check whether the promotional activity resulted in increasing the sales. The outlets where promotional activity introduced are denoted by 1 and others by 2?

2 sample t Test
stats.ttest_ind(sales_1, sales_2)

Statistics	Value
t	0.9625
p value	0.3463

TEST OF HYPOTHESIS

Two sample t test

Exercise 2: A bpo company have developed a new method for better utilization of its resources. 10 observations on utilization from both methods are given below: Check whether the mean utilization for both methods are same or not? Data is given in Utilization.csv.

Method	Utilization	Method	Utilization
Old	89.5	New	89.5
Old	90	New	91.5
Old	91	New	91
Old	91.5	New	89
Old	92.5	New	91.5
Old	91	New	92
Old	89	New	92
Old	89.5	New	90.5
Old	91	New	90
Old	92	New	91

TEST OF HYPOTHESIS

Exercise 3: The data of 30 customers on credit card usage in INR1000, gender (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in table below.

1. Check whether the average credit card usage is same for both gender?
2. Check whether the average credit card usage is same for those who do shopping with credit card and those who don't do shopping?
3. Check whether the average credit card usage is same for those who do banking with credit card and those who don't do banking?

TEST OF HYPOTHESIS

Paired t test:

A special case of two sample t test

When observations on two groups are collected in pairs

Each pair of observation is taken under homogeneous conditions

Procedure

Compute d: difference in paired observations

Let difference in means be $\mu_D = \mu_1 - \mu_2$

Null hypothesis H0: $\mu_D = 0$

Alternative hypothesis H1: $\mu_D \neq 0$ or $\mu_D > 0$ or $\mu_D < 0$

Test statistics t0 = $\frac{\bar{d}}{s_d / \sqrt{n}}$

Reject H0 if p – value < 0.05

TEST OF HYPOTHESIS

Paired t test: Exercise 1

The manager of a fleet of automobiles is testing two brands of radial tires. He assigns one tire of each brand at random to the two rear wheels of eight cars and runs the cars until the tire wear out. Is both brands have equal mean life? The data in kilometers is given in tires.csv

Brand 1	Brand 2
36925	34318
45300	42280
36240	35500
32100	31950
37210	38015
48360	47800
38200	37810
33500	33215

111

Indian Statistical Institute

TEST OF HYPOTHESIS

Paired t test: Exercise 1

The manager of a fleet of automobiles is testing two brands of radial tires. He assigns one tire of each brand at random to the two rear wheels of eight cars and runs the cars until the tire wear out. Is both brands have equal mean life? The data in kilometers is given in tires.csv

Reading the file and variables

```
mydata = mypd.read_csv("E:/ISI/PM-01/Course_Material/Data/Tires.csv")
mydata
b1 = mydata.       Brand1
b2 = mydata.Brand2
```

Paired t test

```
stats.ttest_rel(b1,b2)
```

Statistics	Value
t	1.9039
P value	0.09863

112

TEST OF HYPOTHESIS

Paired t test: Exercise 2

Ten individuals have participated in a diet – modification program to stimulate weight loss. Their weights (in kg) both before and after participation in the program is given in Diet.csv. One an average is the program successful?

Subject	Before	After
1	88	85
2	97	88
3	112	100
4	91	86
5	85	79
6	95	89
7	98	90
8	112	100
9	133	126
10	141	129

NORMALITY TEST

NORMALITY TEST

Normality test

A methodology to check whether the characteristic under study is normally distributed or not

Two Methods

1. Quantile – Quantile (Q- Q) plot
2. Shapiro – Wilk test

NORMALITY TEST

Normality test - Quantile – Quantile (Q- Q) plot

- Plots the ranked samples from the given distribution against a similar number of ranked quantiles taken from a normal distribution
- If the sample is normally distributed then the line will be straight in the plot

NORMALITY TEST

Normality test – Shapiro – Wilk test

H0: Deviation from bell shape (normality) = 0

H1 : Deviation from bell shape ≠ 0

If p value ≥ 0.05 (5%), then H0 is not rejected, distribution is normal

NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is the processing time normally distributed?

NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is processing time normally distributed?

Reading the data and variable

```
import pandas as mypd
from scipy import stats
import matplotlib.pyplot as myplot
mydata = mypd.read_csv("E:/ISI/PM-01/Data/PO_Processing.csv")
PT = mydata.Processing_Time
```

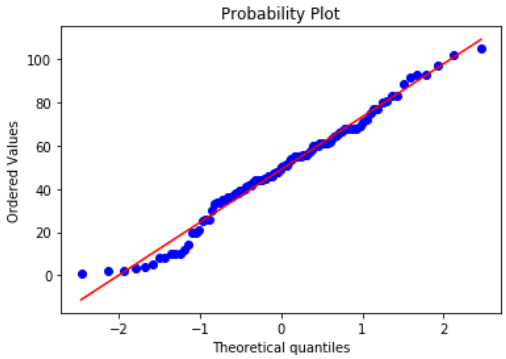

NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is processing time normally distributed?

Normality Check using Normal Q – Q plot

```
stats.probplot(PT, plot = myplot)
myplot.show()
```



NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is processing time normally distributed?

Normality test
stats.mstats.normaltest(PT)

Statistics	Value
W	0.33965
p value	0.84381

NORMALITY TEST

Normality test

Exercise 2 : The time taken to respond to customer complaints is is given in Compaint_Response_Time.csv. Check whether the complaint response time follows normal distribution?

Response Time	
24	26
31	27
29	24
26	23
28	27
26	28
29	27
29	23
27	27
31	23
25	25
29	27
29	26
25	28
26	27

NORMALITY TEST

Normality test

Exercise 3 : The impurity level (in ppm) is routinely measured in an intermediate chemical process. The data is given in Impurity.csv. Check whether the impurity follows normal distribution?

ANALYSIS *of* VARIANCE

ANALYSIS OF VARIANCE

ANOVA

Analysis of Variance is a test of means for two or more populations

Partitions the total variability in the variable under study to different components

H0 = $Mean_1$ = $Mean_2$ = - - - =$Mean_k$

Reject H0 if p – value < 0.05

Example:

To study location of shelf on sales revenue

ANALYSIS OF VARIANCE

One Way Anova : Example

An electronics and home appliance chain suspect the location of shelves where television sets are kept will influence the sales revenue. The data on sales revenue in lakhs from the television sets when they are kept at different locations inside the store are given in sales revenue data file. The location is denoted as 1:front, 2: middle & 3: rear. Verify the doubt? The data is given in Sales_Revenue_Anova.csv.

ANALYSIS OF VARIANCE

One Way Anova : Example

Factor: Location(A)

Levels : front, middle, rear

Response: Sales revenue

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

Level 1 Sum(A_1):

Sum of all response values when location is at level 1 (front)

= 1.55 + 2.36 + 1.84 + 1.72

= 7.47

nA_1: Number of response values with location is at level 1 (front)

= 4

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

Level 1 Average:

Sum of all response values when location is at level 1 / number of response values with location is at level 1

$= A_1 / nA_1 = 7.47 / 4 = 1.87$

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

	Level 1 (front)	Level 2 (middle)	Level 3 (rear)
Sum	A_1: 7.47	A_2: 30.31	A_3: 15.55
Number	nA_1: 4	nA_2: 8	nA_3: 6
Average	1.87	3.79	2.59

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 2: Calculate the grand total (T)

T = Sum of all the response values

= 1.55 + 2.36 + - - - + 2.72 + 2.07 = 53.33

Step 3: Calculate the total number of response values (N)

N = 18

Step 4: Calculate the Correction Factor (CF)

CF = (Grand Total)^2 / Number of Response values

= T^2 / N = $(537.33)^2$ / 18 = 158.0049

131

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 5: Calculate the Total Sum of Squares (TSS)

TSS = Sum of square of all the response values - CF

= $1.55^2 + 2.36^2$ + - - - + $2.72^2 + 2.07^2$ – 158.0049

= 15.2182

132

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 6: Calculate the between (factor) sum of square

$SS_A = A_1^2 / nA_1 + A_2^2 / nA_2 + A_3^2 / nA_3 - CF$

$= 7.47^2 / 4 + 30.31^2 / 8 + 15.55^2 / 4 – 158.0049$

$= 11.0827$

Step 7: Calculate the within (error) sum of square

SS_e = Total sum of square – between sum of square

$= TSS - SS_A = 15.2182 – 11.0827 = 4.1354$

133

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 8: Calculate degrees of freedom (df)

Total df = Total Number of response values - 1

= 18 - 1 = 17

Between df

= Number of levels of the factor - 1

= 3 - 1 = 2

Within df = Total df – Between df

= 17 - 2 = 15

134

ANALYSIS OF VARIANCE

One Way Anova : Example

Anova Table:

Source	df	SS	MS	F	F Crit	P value
Between	2	11.08272	5.541358	20.09949	3.68	0.0000
Within	15	4.135446	0.275696			
Total	17	15.21816				

MS = SS / df

$F = MS_{Between} / MS_{Within}$

F Crit =finv (probability, between df, within df) , probability = 0.05

P value = fdist (F, between df, within df)

ANALYSIS OF VARIANCE

One Way Anova : Python Code

Reading data and variables to R

```
import pandas as mypd
from scipy import stats
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
mydata = mypd.read_csv("E:\ISI\PM-01\Data/Sales_Revenue_Anova.csv")
sales = mydata.Sales_Revenue
location = mydata.Location
```

Computing ANOVA table

```
mymodel = ols('sales ~ C(location)', mydata).fit()
anova_table = anova_lm(mymodel)
anova_table
```

ANALYSIS OF VARIANCE

One Way Anova :

	df	SS	MS	F	p-value
Location	2	11.08272	5.541358	20.09949	5.7E-05
Residual	15	4.135446	0.275696		

ANALYSIS OF VARIANCE

One Way Anova : Decision Rule

If p value < 0.05, then

The factor has significant effect on the process output or response.

Meaning:

When the factor is changed from 1 level to another level, there will be significant change in the response.

ANALYSIS OF VARIANCE

One Way Anova : Example Result

For factor Location, p = 0.000 < 0.05

Conclusion:

Location has significant effect on sales revenue

Meaning:

The sales revenue is not same for different locations like front, middle & rear

ANALYSIS OF VARIANCE

One Way Anova : Example Result

The expected sales revenue for different location under study is equal to level averages.

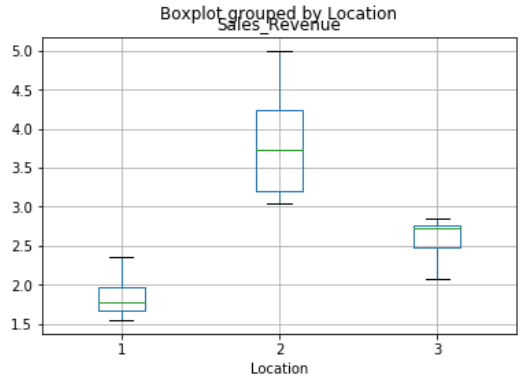
Location	Expected Sales Revenue
Front	1.8675
Middle	3.78875
Rear	2.591667

```
sales.groupby(location).mean()
```

ANALYSIS OF VARIANCE

One Way Anova : Example Result

```
import matplotlib.pyplot as myplot
mydata.boxplot(column ='Sales_Revenue', by = 'Location')
myplot.show()
```



141

ANALYSIS OF VARIANCE

Anova logic:

Two Types of Variations:

1. Variation within the level of a factor
2. Variation between the levels of factor

142

ANALYSIS OF VARIANCE

Anova logic :

Variation between the level of a factor:

The effect of Factor.

Variation within the levels of a factor:

The inherent variation in the process or Process Error.

	Location		
	Front	Middle	Rear
Sales Revenue	1.34 1.89 1.35 2.07 2.41 3.06	3.20 2.81 4.52 4.40 4.75 5.19 3.42 9.80	2.30 1.91 1.40 1.48

143

ANALYSIS OF VARIANCE

Anova logic :

If the variation between the levels of a factor is significantly higher than the inherent variation

then the factor has significant effect on response

To check whether a factor is significant:

Compare variation between levels with variation within levels

144

Indian Statistical Institute

ANALYSIS OF VARIANCE

Anova logic :

Measure of variation between levels: MS of the factor ($MS_{between}$)

Measure of variation within levels: MS Error (MS_{within})

To check whether a factor is significant:

Compare MS of between with MS within

i.e. Calculate $F = MS_{between} / MS_{within}$

If F is very high, then the factor is significant.

145

Indian Statistical Institute

ANALYSIS OF VARIANCE

Exercise 1: An insurance company wants to check whether the waiting time of customer at their single window operation across 4 cities is same or not. The data is given in Insurance_waiting_time.csv?

Exercise 2: An two wheeler manufacturing company wants to study the effect of four engine tuning techniques on the mileage. The data collected is given in Mileage.csv file. Test whether the tuning techniques impacts the mileage?

146

CROSS TABULATION

CROSS TABULATION

- An approach to summarize and identify the relation between two or more variables or parameters
- Describes two variables simultaneously
- Expressed as two way table
- Variables need to be categorical or grouped

Input or Process Variable	Output Variable				
	Very Good	Good	Average	Below Average	Poor
0 – 3					
3 - 6					
6 - 12					

CROSS TABULATION

Example: A branded apparel manufacturing company has collected the data from 50 customers on usage, gender, awareness of brand and preference of the brand. Usage has been coded as 1, 2 ,and 3 representing light, medium and heavy usage. The gender has been coded as 1 for female and 2 for male users. The attitude and preference are measured on a 7 point scale (1: unfavorable to 7 : very favorable). The data is given in apparel_data.csv file .

1. Does male and female differ in their usage?
2. Does male and female differ in their awareness of the brand?
3. Does male and female differ in their preference?
4. Does higher the awareness means higher preference?

149

CROSS TABULATION

b. Constructing cross tabulation of Gender vs. Usage

```
import pandas as mypd
mydata = mypd.read_csv("E:/ISI/PM-01/Data/Apparel_Data.csv")
usage = mydata.Usage
gender = mydata.Gender
mytable = mypd.crosstab(gender, usage)
mytable
```

Gender	Usage		
	Light	Medium	Heavy
Female	15	6	5
Male	6	6	12

150

CHI SQUARE TEST

CHI SQUARE TEST

Objective:

To test whether two variables are related or not

To check whether a metric is depends on another metric

Usage:

When both the variables (x & y) need to be categorical (grouped)

H0: Relation between x & y = 0 or x and y are independent

H1: Relation between x & y $\neq$ 0 or x and y are not independent

If p value < 0.05, then H0 is rejected

CHI SQUARE TEST

Exercise:

A project is undertaken to improve the CSat score of transaction processing. Based on brainstorming, the project team suspects that lack of experience is a cause of low CSat score.

The following data was collected. Analyze the data and verify whether CSat score dependents on experience

Experience (Months)	CSat Score				
	VD	D	N	S	VS
0 – 3	50	40	30	10	10
3- 6	5	30	50	35	7
6 - 9	6	7	30	40	50

Note: Table gives the count of CSat score of very dissatisfied to very satisfied for agents belonging to three different experience groups

153

CHI SQUARE TEST

Exercise:

Step 1: Calculate the row and column sum

Experience (Months)	CSat Score					
	VD	D	N	S	VS	Row Sum
0 – 3	50	40	30	10	10	140
3 - 6	5	30	50	35	7	127
6 - 9	6	7	30	40	50	133
Col Sum	61	77	110	85	67	400

154

CHI SQUARE TEST

Exercise:

Step 2: Calculate expected count for each cell

Expected count of CSat score VD for group 0 – 3 months experience

= Expected count of cell (1,1) = (Row 1 sum x Column 1 sum) / Total

= (140 x 61) / 400 = 21.4

Table of expected count (the count expected if variables are not related)

Experience (Months)	CSat Score					
	VD	D	N	S	VS	Row Sum
0 – 3	21.4	27	38.5	29.8	23.5	140
3 - 6	19.4	24.4	34.9	27	21.3	127
6 - 9	20.3	25.6	36.6	28.3	22.3	133
Col Sum	61	77	110	85	67	400

155

CHI SQUARE TEST

Exercise:

Step 3: Take difference between observed count and expected count

For cell (1,1)

observed Count = 50

expected Count = 21.4

difference = 28.7

Table of observed count – expected count

Experience (Months)	CSat Score				
	VD	D	N	S	VS
0 – 3	28.7	13.1	-8.5	-20	-13
3 - 6	-14.4	5.55	15.1	8.01	-14
6 - 9	-14.3	-19	-6.6	11.7	27.7

156

CHI SQUARE TEST

Exercise:

Step 4: Calculate (observed - expected)^2 / expected for each cell

Table of (observed - expected)^2 / expected

Experience (Months)	CStat Score				
	VD	D	N	S	VS
0 – 3	38.45	6.32	1.88	13.11	7.71
3 - 6	10.66	1.26	6.51	2.38	9.58
6- 9	10.06	13.52	1.18	4.87	34.50

CHI SQUARE TEST

Exercise:

Step 5: Calculate Chi Square = Sum of all ((observed - expected)^2 / expected)

Chi Square calculated = 38.45 + 6.32 + - - - + 34.5

Chi Square Calculated χ^2= 161.98

If variables are not related then χ^2 will be close to 0

Step 6: Calculate p value

P value = chidist(chi Sq, df)

= chidist(161.98,8)

= 0.00

Conclusion:

Since p value 0.00 < 0.05, Csat score depends on experience or the variables are related

CHI SQUARE TEST

Example: A branded apparel manufacturing company has collected the data from 50 customers on usage, gender, awareness of brand and preference of the brand. Usage has been coded as 1, 2 ,and 3 representing light, medium and heavy usage. The gender has been coded as 1 for female and 2 for male users. The attitude and preference are measured on a 7 point scale (1: unfavorable to 7 : very favorable). The data is given in apparel_data.csv file .

1. Estimate the relation between gender and usage?
2. Estimate the relation between gender and awareness of the brand?
3. Estimate the relation between gender and preference?
4. Does higher the awareness means higher preference?

CHI SQUARE TEST

b. Constructing cross tabulation of Gender vs. Usage

```
import pandas as mypd
mydata = mypd.read_csv("E:/ISI/PM-01/Data/Apparel_Data.csv")
usage = mydata.Usage
gender = mydata.Gender
mytable = mypd.crosstab(gender, usage)
mytable
```

Gender	Usage		
	Light	Medium	Heavy
Male	15	6	5
Female	6	6	12

CHI SQUARE TEST

c. Chi Square test of independence - Gender vs. Usage

```
from scipy import stats
stats.chi2_contingency(mytable)
```

Statistics	Value
Chi Square	6.6702
P value	0.03561

CORRELATION
&
REGRESSION

CORRELATION & REGRESSION

Correlation:

Correlation analysis is a technique to identify the relationship between two variables.

Type and degree of relationship between two variables.

CORRELATION & REGRESSION

Correlation: Usage

Explore the relationship between the output characteristic and input or process variable.

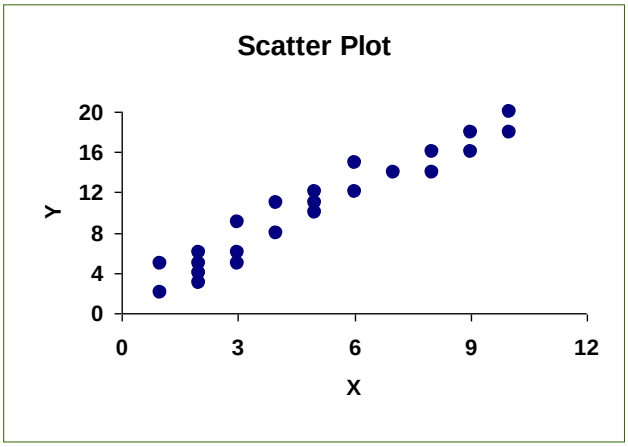
Output variable : y : Dependent variable

Input / Process variable : x : Independent variable

Indian Statistical Institute

CORRELATION & REGRESSION

Positive Correlation: y increases as x increases & vice versa

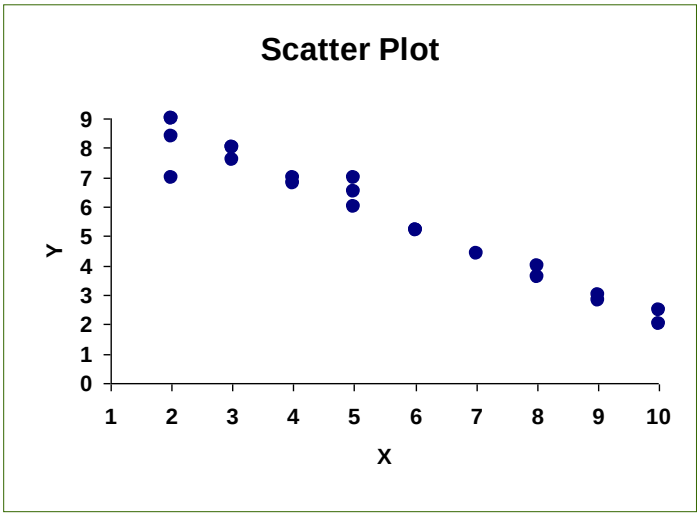


165

Indian Statistical Institute

CORRELATION & REGRESSION

Negative Correlation: y decreases as x increases & vice versa

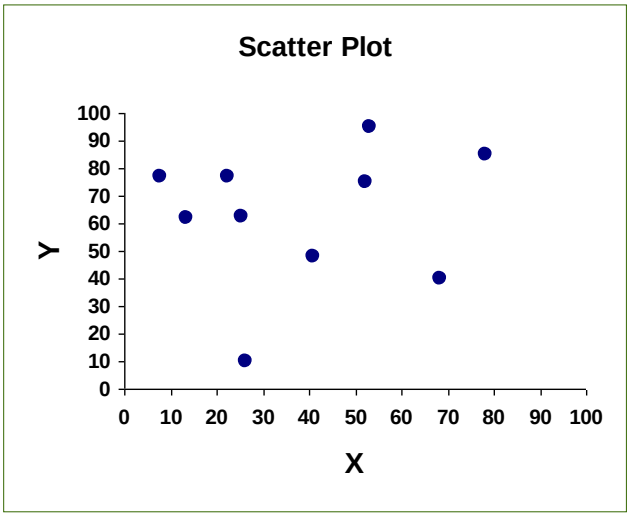


166

Indian Statistical Institute

CORRELATION & REGRESSION

No Correlation: Random Distribution of points

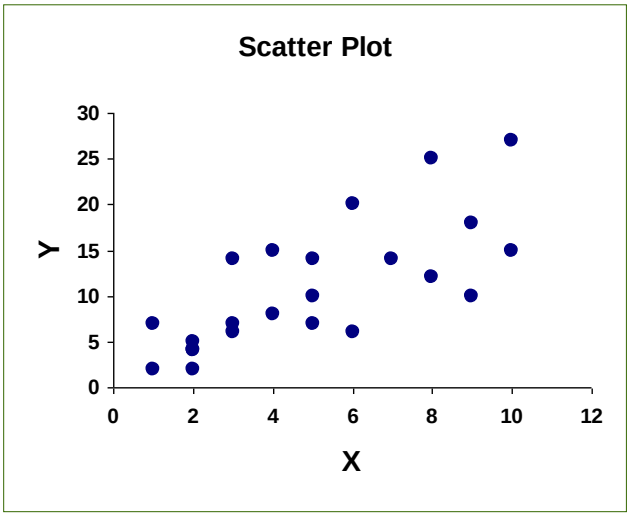


167

Indian Statistical Institute

CORRELATION & REGRESSION

Is there any correlation ?



168

Indian Statistical Institute

CORRELATION & REGRESSION

Measure of Correlation: Coefficient of Correlation

Symbol : r

Range : -1 to 1

Sign : Type of correlation

Value : Degree of correlation

Examples:

r = 0.6 , 60 % positive correlation

r = -0.82, 82% negative correlation

r = 0, No correlation

169

Indian Statistical Institute

CORRELATION & REGRESSION

Coefficient of Correlation Computation: Positive Correlation

Collect data on x and y: When x is low, y is also low & vice versa

x	y
2	5
3	7
1	3
5	11
6	12
7	15

170

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Calculate Mean of x & y values

SL No.	x	y
1	2	5
2	3	7
3	1	3
4	5	11
5	6	12
6	7	15
Mean	4	8.83

171

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Take x – Mean x and y – Mean y

SL No.	x – Mean x	y – Mean y
1	-2	-3.83
2	-1	-1.83
3	-3	-5.83
4	1	2.17
5	2	3.17
6	3	6.17

Conclusion:

Low values will become negative & high values will become positive

172

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Generally when x values are negative, y values are also negative & vice versa

SL No.	x – Mean x	y – Mean y
1	-2	-3.83
2	-1	-1.83
3	-3	-5.83
4	1	2.17
5	2	3.17
6	3	6.17

173

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Then

Product of x & y values will be generally positive

SL No.	x – Mean x	y – Mean y	Product
1	-2	-3.83	7.66
2	-1	-1.83	1.83
3	-3	-5.83	17.49
4	1	2.17	2.17
5	2	3.17	6.34
6	3	6.17	18.51
		Sum = Sxy	54

174

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Sum of Product of x & y values (Sxy) will be positive

SL No.	x – Mean x	y – Mean y	Product
1	-2	-3.83	7.66
2	-1	-1.83	1.83
3	-3	-5.83	17.49
4	1	2.17	2.17
5	2	3.17	6.34
6	3	6.17	18.51
		Sum = Sxy	54

175

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Collect data on x and y: When x is low then y will be high & vice versa

x	y
2	12
3	11
1	15
5	7
6	5
7	3

176

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Calculate Mean of x & y values

SL No.	x	y
1	2	12
2	3	11
3	1	15
4	5	7
5	6	5
6	7	3
Mean	4	8.83

177

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Take x – Mean x and y – Mean y

SL No.	x – Mean x	y – Mean y
1	-2	3.67
2	-1	2.67
3	-3	6.67
4	1	-1.33
5	2	-3.33
6	3	-5.33

Conclusion:

Low values will become negative & high values will become positive

178

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Generally when x values are negative, y values are positive & vice versa

SL No.	x – Mean x	y – Mean y
1	-2	3.67
2	-1	2.67
3	-3	6.67
4	1	-1.33
5	2	-3.33
6	3	-5.33

179

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Then

Product of x & y values will be generally negative

SL No.	x – Mean x	y – Mean y	Product
1	-2	3.67	-7.34
2	-1	2.67	-2.67
3	-3	6.67	-20.01
4	1	-1.33	-1.33
5	2	-3.33	-6.66
6	3	-5.33	-15.99
		Sum = Sxy	- 54

180

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Sum of Product of x & y values Sxy will be negative

SL No.	x – Mean x	y – Mean y	Product
1	-2	3.67	-7.34
2	-1	2.67	-2.67
3	-3	6.67	-20.01
4	1	-1.33	-1.33
5	2	-3.33	-6.66
6	3	-5.33	-15.99
		Sum = Sxy	- 54

181

CORRELATION & REGRESSION

Coefficient of Correlation Computation :

In Short

If correlation is positive

Sxy will be positive

If correlation is negative

Sxy will be negative

182

CORRELATION & REGRESSION

Coefficient of Correlation Computation :

Sxy is divided by $\sqrt{}$ (Sxx.Syy)

Sxy = Σ(x-Mean x)(y-Mean y)

Sxx = Σ(x-Mean x)^2

Syy = Σ(y-Mean y)^2

Correlation Coefficient r = Sxy / $\sqrt{}$ (Sxx.Syy)

183

CORRELATION & REGRESSION

Coefficient of Correlation Computation :

SL No.	x – Mean x	y – Mean y	Product	(x – Mean x)^2	(y – Mean y)^2
1	-2	3.67	-7.34	4	14.6689
2	-1	2.67	-2.67	1	3.3489
3	-3	6.67	-20.01	9	33.9889
4	1	-1.33	-1.33	1	4.7089
5	2	-3.33	-6.66	4	10.0489
6	3	-5.33	-15.99	9	38.0689
Sum			Sxy: -54	Sxx: 28	Syy:104.83

r = Sxy / $\sqrt{}$Sxx.Syy = -54 / $\sqrt{}$(28 x 104.83) = -0.9967

184

Indian Statistical Institute

CORRELATION & REGRESSION

Exercise: The data on vapor pressure of water at various temperatures are given in Correlation.csv file.

1. Construct the scatter plot and interpret?
2. Compute the correlation coefficient?

185

Indian Statistical Institute

CORRELATION & REGRESSION

Exercise: The data on vapor pressure of water at various temperatures are given in Correlation.csv file.

1. Reading the data and variables

```
import pandas as mypd
import numpy as mynp
import matplotlib.pyplot as myplot
mydata = mypd.read_csv("E:/ISI/PM-01/Data/Correlation.csv")
temp = mydata.Temperature
pressure = mydata.Vapor_Pressure
```

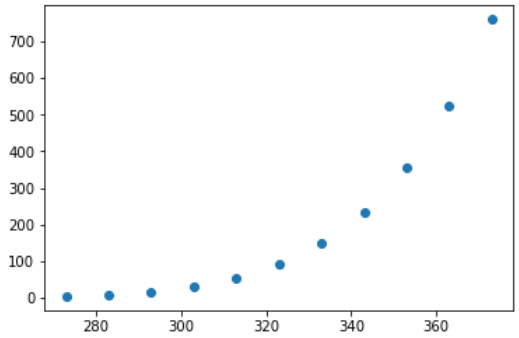
186

CORRELATION & REGRESSION

Exercise: The data on vapor pressure of water at various temperatures are given in Correlation.csv file.

2. Constructing Scatter plot

```
myplot.scatter(temp, pressure)
myplot.show()
```



187

CORRELATION & REGRESSION

Exercise: The data on vapor pressure of water at various temperatures are given in Correlation.csv file.

Computing correlation coefficient

```
mynp.corrcoef(temp, pressure)
```

Statistics	Value
r	0.893

188

MULTIPLE REGRESSION ANALYSIS

CORRELATION & REGRESSION

Regression

Correlation helps

To check whether two variables are related

If related

Identify the type & degree of relationship

Indian Statistical Institute

CORRELATION & REGRESSION

Regression

Regression helps

- To identify the exact form of the relationship
- To model output in terms of input or process variables

Examples:

Expected (Yield) = 5 + 3 x Time - 2 x Temperature

191

Indian Statistical Institute

CORRELATION & REGRESSION

Simple Linear Regression Illustration

Output variable is modeled in terms of only one variable

x	y
2	7
1	4
5	16
4	13
3	10
6	19

Regression Model

$y = 1 + 3x$

192

CORRELATION & REGRESSION

Simple Linear Regression

General Form:

$$y = a + bx + \varepsilon$$

where

a: intercept (the value of y when x is equal to 0)

b: slope (indicates the amount of change in y with every unit change in x)

CORRELATION & REGRESSION

Simple Linear Regression: Parameter Estimation

Model: $y = a + bx + \varepsilon$

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

$$\hat{b} = S_{xy} / S_{xx}$$

Test for Significance (Testing b = 0 or not) of relation between x & y

H0: $b = 0$

H1: $b \neq 0$

Test Statistic $t_0 = (\hat{b} - 0)/se(\hat{b})$

If p value < 0.05, then H0 is rejected & y can be modeled with x

CORRELATION & REGRESSION

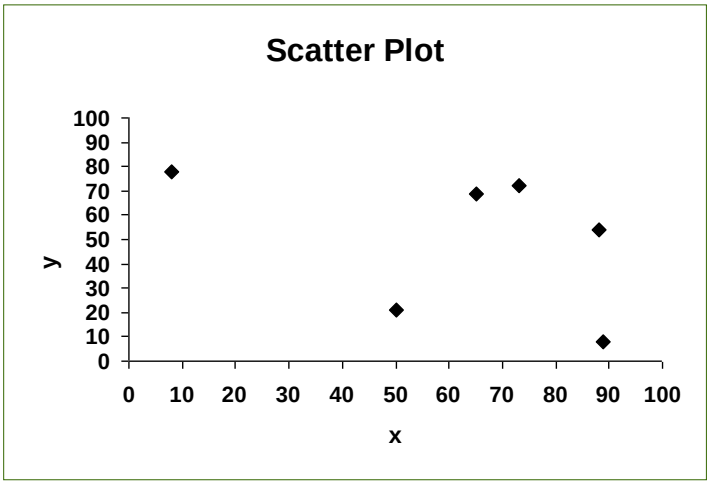
Regression illustration: Issues

x	y
65	69
8	78
89	8
88	21
50	24
73	72

195

CORRELATION & REGRESSION

Regression Model $y = 76.32 - 0.42x + \varepsilon$



196

CORRELATION & REGRESSION

Regression: Issues

For any set of data,

a & b can be calculated

Regression model $y = a + bx + \varepsilon$ can be build

But all the models may not be useful

CORRELATION & REGRESSION

Coefficient of Regression: Measure of degree of Relationship

Symbol : R^2

$R^2 = SS_R / Syy = b.Sxy / Syy$

$SS_R = \Sigma(y_{predicted} - \text{Mean } y)^2$

$Syy = \Sigma(y_{actual} - \text{Mean } y)^2$

R^2 : amount variation in y explained by x

Range of R^2 : 0 to 1

If $R^2 \geq 0.6$, the model is reasonably good

CORRELATION & REGRESSION

Coefficient of Regression: Testing the significance of Regression

Regression ANOVA

Model	SS	df	MS	F	p value
Regression	SS_R				
Residual	Syy – SS_R				
Total	Syy				

If p value < 0.05, then the regression model is significant

REGRESSION ANALYSIS

Multiple Linear Regression

To model output variable y in terms of two or more variables.

General Form:

$$y = a + b_1x_1 + b_2x_2 + - - - + b_kx_k + \varepsilon$$

Two variable case:

$$y = a + b_1x_1 + b_2x_2 + \varepsilon$$

Where

a: intercept (the predicted value of y when all x's are zero)

b_j: slope (the amount change in y for unit change in x_j keeping all other x's constant, j = 1,2,---,k)

REGRESSION ANALYSIS

Exercise : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time?

Step 1: Read data

```
import pandas as mypd
from scipy import stats
import matplotlib.pyplot as myplot
from pandas.tools.plotting import scatter_matrix
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

mydata = mypd.read_csv("E:/ISI/PM-01/Data/Mult_reg_Yield.csv")
time = mydata.Time
temp = mydata.Temperature
output = mydata.Yield
```

201

REGRESSION ANALYSIS

Exercise : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time?

Step 1: Correlation Analysis

```
scatter_matrix(mydata)
myplot.show()
```

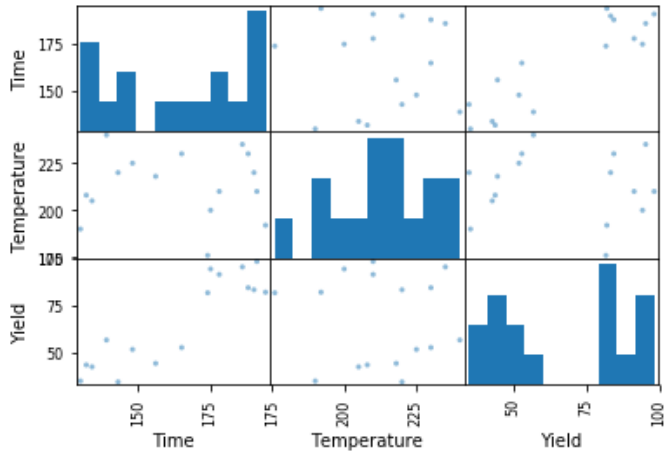
Correlation between xs & y should be high

Correlation between xs should be low

202

REGRESSION ANALYSIS

Exercise : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time?



REGRESSION ANALYSIS

Step 2: Regression Output

```
mymodel = ols("output ~ time + temp", mydata).fit()
mymodel.summary()
```

Statistics	Value	Criteria
R-squared:	0.806	≥ 0.6
Adj. R-squared:	0.777	≥ 0.6
F-statistic:	27.07	
Prob (F-statistic):	2.32e-05	< 0.05
Log-Likelihood:	-59.703	
AIC:	125.4	
BIC:	127.7	

REGRESSION ANALYSIS

Step 2: Regression Output

```
anova_table = anova_lm(mymodel)
anova_table
```

	df	SS	MS	F	p-value
Time	1	6777.81	6777.81	53.98722	0.000006
Temp	1	19.25253	19.25253	0.153352	0.701696
Residual	13	1632.081	125.5447		

Criteria: p value < 0.05

REGRESSION ANALYSIS

Step 2: Regression Output

Regression ANOVA

Model	SS	df	MS	F	p value
Regression	6797.063	2	3398.531	27.07	0.0000
Residual	1632.08138	13	125.5447		
Total	8429.14438	15			

Criteria: P value < 0.05

Indian Statistical Institute

REGRESSION ANALYSIS

Step 2: Regression Output – Identify the model

	Coefficients	Std error	t	p-value	[0.025	0.975]
Intercept	-67.8844	40.587	-1.673	0.118	-155.57	19.797
Time	0.9061	0.123	7.344	0.000	0.64	1.173
Temp	-0.0642	0.164	-0.392	0.702	-0.418	0.29

Interpretation: Only time is related to yield or output as p value < 0.05

207

Indian Statistical Institute

REGRESSION ANALYSIS

Step 2: Regression Output – Identify the model

	Coefficients	Std error	t	p-value	[0.025	0.975]
Intercept	- 81.6205	19.791	-4.124	0.001	-124.067	-39.174
Time	0.9065	0.120	7.580	0.000	0.650	1.163

Model Yield= 0.9065 x Time - 81.621

Statistics	Value	Criteria
R-squared:	0.804	≥ 0.6
Adj. R-squared:	0.79	≥ 0.6
F-statistic:	57.46	
Prob (F-statistic):	2.55e-06	< 0.05

208

REGRESSION ANALYSIS

Step 3: Residual Analysis

```
pred = mymodel.predict()
res = output - pred
```

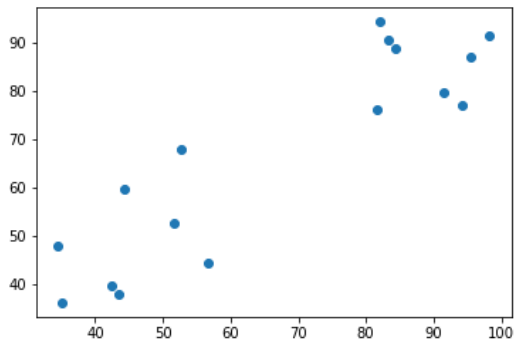
SL No	Actual	Predicted	Residuals
1	35	36.22	-1.22
2	81.7	76.10	5.60
3	42.5	39.84	2.66
4	98.3	91.51	6.79
5	52.7	67.94	-15.24
6	82	94.23	-12.23
7	34.5	48.00	-13.50
8	95.4	86.98	8.42
9	56.7	44.38	12.32
10	84.4	88.79	-4.39
11	94.3	77.01	17.29
12	44.3	59.79	-15.49
13	83.3	90.61	-7.31
14	91.4	79.73	11.67
15	43.5	38.03	5.47
16	51.7	52.53	-0.83

209

REGRESSION ANALYSIS

Step 3: Residual Analysis – Actual Vs Fitted

```
myplot.scatter(output, pred)
myplot.show()
```



Note: There need to be strong positive correlation between actual and fitted response

210

REGRESSION ANALYSIS

Step 3: Residual Analysis: Normality test

```
stats.mstats.normaltest(res)
```

Normality Test: Yield data	
W	p value
1.8945	0.3878

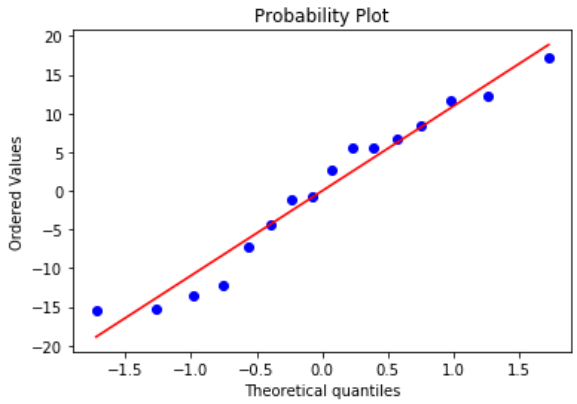
```
res_sq = res**2
mse = res_sq.mean()

import math as mymath
rmse = mymath.sqrt(mse)
rmse
```

Statistic	Value
MSE	103.21
RMSE	10.159

REGRESSION ANALYSIS

7: Residual Analysis: Normality test

```
stats.probplot(res, plot = myplot)
myplot.show
```

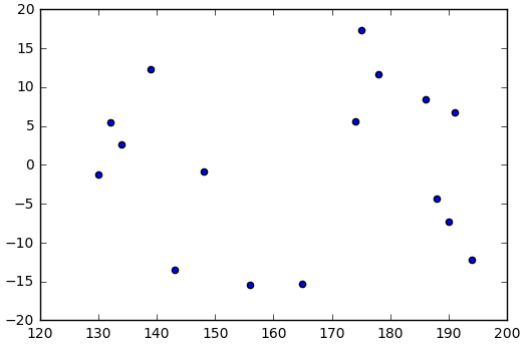


REGRESSION ANALYSIS

7: Model adequacy check

Residuals Vs Independent variables

```
myplot.scatter(time, res)
myplot.show()
```



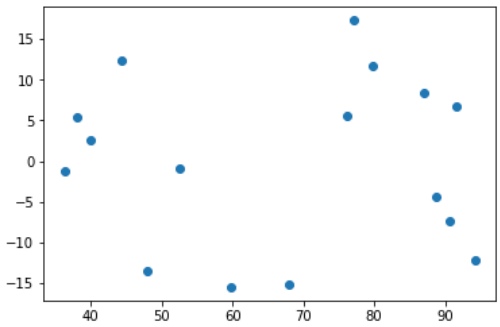
Note: There should not be any pattern or trend, the points should be distributed randomly

REGRESSION ANALYSIS

7: Model adequacy check

Residuals Vs Fitted

```
myplot.scatter(pred, res)
myplot.show()
```



Note: There should not be any pattern or trend, the points should be distributed randomly

REGRESSION ANALYSIS

Regression with dummy variables

When x's are not numeric but nominal

Each nominal or categorical variable is converted into dummy variables

Dummy variables takes values 0 or 1

Number of dummy variable for one x variable is equal to number of distinct values of that variable - 1

Example: A study was conducted to measure the effect of gender and income on attitude towards vocation. Data was collected from 30 respondents and is given in Travel_dummy_reg file. Attitude towards vocation is measured on a 9 point scale. Gender is coded as male = 1 and female = 2. Income is coded as low=1, medium = 2 and high = 3. Develop a model for attitude towards vocation in terms of gender and Income?

REGRESSION ANALYSIS

Regression with dummy variables

Variable		Dummy
Gender	Code	gender_Code
Male	1	0
Female	2	1

Variable		Dummy	
Income	Code	Income1	Income 2
Low	1	0	0
Medium	2	1	0
High	3	0	1

REGRESSION ANALYSIS

Regression with dummy variables

Read the fie and variables

```
import pandas as mypd
from scipy import stats
import matplotlib.pyplot as myplot
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

mydata = mypd.read_csv("E:/ISI/PM-01/Data/Travel_dummy_Reg.csv")
gender = mydata.Gender
income = mydata.Income
attitude = mydata.Attitude
```

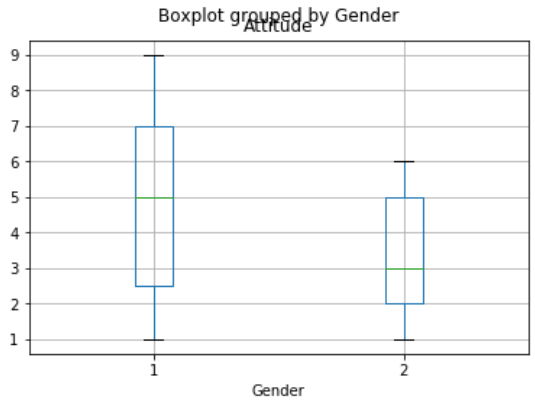

REGRESSION ANALYSIS

Regression with dummy variables

Checking relation between x and y

Attitude Vs gender

```
mydata.boxplot(column = 'Attitude', by = 'Gender')
myplot.show()
```



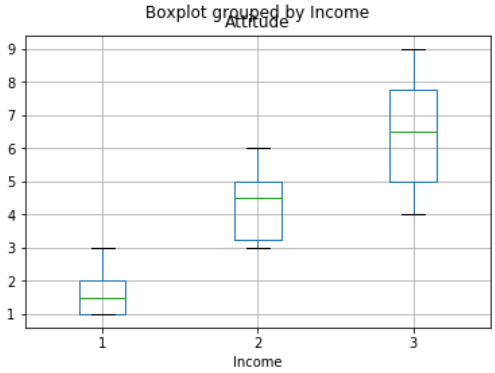
REGRESSION ANALYSIS

Regression with dummy variables

Checking relation between x and y

Attitude Vs income

```
mydata.boxplot(column = 'Attitude', by = 'Income')
myplot.show()
```



REGRESSION ANALYSIS

Regression with dummy variables – Output

```
mymodel = ols('attitude ~ C(gender) + C(income)', mydata).fit()
mymodel.summary()
```

R^2	0.86
Adjusted R^2	0.844
F Statistics	53.37
p value	0.0000

	Coef	Std err	t	p-value	[0.025	0.975]
Intercept	2.4	0.336	7.145	0.00	1.71	3.09
C(gender)[T.2]	-1.6	0.336	-4.763	0.00	-2.29	-0.91
C(income)[T.2]	2.8	0.411	6.806	0.00	1.954	3.646
C(income)[T.3]	4.8	0.411	11.668	0.00	3.954	5.646

REGRESSION ANALYSIS

Regression with dummy variables – Anova Table

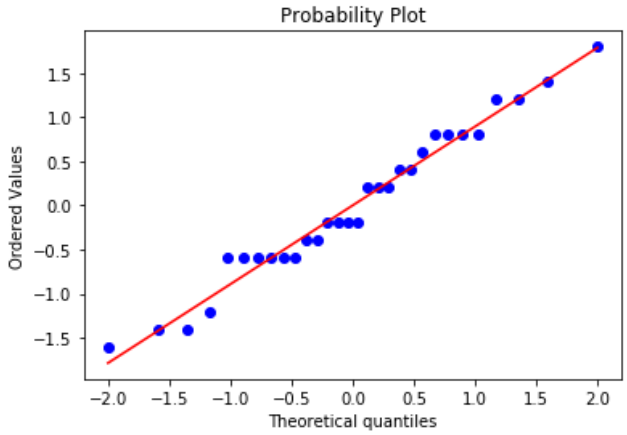
```
anova_table = anova_lm(mymodel)
anova_table
```

	df	SS	MS	F	P-value
C(gender)	1	19.2	19.2	22.69091	0.00006
C(income)	2	116.26667	58.13333	68.70303	0.00000
Residual	26	22	0.846154		

REGRESSION ANALYSIS

Regression with dummy variables – Normality test of Residuals

```
stats.probplot(res, plot = myplot)
myplot.show()
```



REGRESSION ANALYSIS

Regression with dummy variables – Normality test of Residuals

stats.mstats.normaltest(res)

Statistics	Value
w	0.5211
p-value	0.7706

BINARY LOGISTIC REGRESSION

BINARY LOGISTIC REGRESSION

Used to develop models when the output or response variable y is binary

The output variable will be binary, coded as either success or failure

Models probability of success p which lies between 0 and 1

Linear model is not appropriate

$$p = \frac{e^{a+b_1x_1+b_2x_2+---+b_kx_k}}{1+e^{a+b_1x_1+b_2x_2+---+b_kx_k}}$$

p: probability of success
x_i's : independent variables
a, b_1, b_2, ---: coefficients to be estimated

If estimate of $p \geq 0.5$, then classified as success, otherwise as failure

BINARY LOGISTIC REGRESSION

Example: Develop a model to predict the non payment of overdrafts by customers of a multinational banking institution. The data collected is given in Logistic_Reg.csv file. The factors and response considered are given below.

SL No	Factor
1	Individual expected level of activity score
2	Transaction speed score
3	Peer comparison score in terms of transaction volume

Response	Values
Outcome	0: Not Paid and 1: Paid

BINARY LOGISTIC REGRESSION

Example: Develop a model to predict the non payment of overdrafts by customers of a multinational banking institution. The data collected is given in Logistic_Reg.csv file.

Reading the file and variables

```
import pandas as mypd
from sklearn.linear_model import LogisticRegression
mydata = mypd.read_csv("E:/ISI/PM-01/Data/Logistic_Reg.csv")
import statsmodels.api as mysm

x = mydata[["Ind_Exp_Act_Score", "Tran_Speed_Score", "Peer_Comb_Score"]]
y = mydata.Outcome
x["Intercept"]=1
```


BINARY LOGISTIC REGRESSION

Example: Develop a model to predict the non payment of overdrafts by customers of a multinational banking institution. The data collected is given in Logistic_Reg.csv file.

Developing the model

```
mymodel = mysm.Logit(y,x)
myresult = mymodel.fit()
myresult.summary()
```

Logistic Regression Results

Statistic	Value	Statistic	Value
Response	Outcome	No. of values	980
Model	Logit	Df Residuals	976
Method	MLE	Df Model	3

BINARY LOGISTIC REGRESSION

Example: Develop a model to predict the non payment of overdrafts by customers of a multinational banking institution. The data collected is given in Logistic_Reg.csv file.

Logistic Regression Results

Statistic	Value	Criteria
Pseudo R^2	0.893	≥ 0.6
Log-Likelihood:	-63.416	
LL-Null:	-577.85	
LLR p-value:	0.00	< 0.05

BINARY LOGISTIC REGRESSION

Example: Develop a model to predict the non payment of overdrafts by customers of a multinational banking institution. The data collected is given in Logistic_Reg.csv file.

Logistic Regression Results

	Code	Coef	Std err	z	p-value	95 % CI
Ind_Exp_Act_Score	x_1	2.7957	0.355	7.867	0.00	2.099 3.492
Tran_Speed_Score	x_2	2.7532	0.343	8.032	0.00	2.081 3.425
Peer_Comb_Score	x_3	3.5153	0.434	8.095	0.00	2.664 4.366
Intercept		-35.5062	4.406	-8.058	0.00	-71.012

Criteria: p-value < 0.05

BINARY LOGISTIC REGRESSION

Example: Develop a model to predict the non payment of overdrafts by customers of a multinational banking institution. The data collected is given in Logistic_Reg.csv file.

Logistic Regression Results

	Code	Coef	Std err	z	p-value	95 % CI
Ind_Exp_Act_Score	x_1	2.7957	0.355	7.867	0.00	2.099 3.492
Tran_Speed_Score	x_2	2.7532	0.343	8.032	0.00	2.081 3.425
Peer_Comb_Score	x_3	3.5153	0.434	8.095	0.00	2.664 4.366
Intercept		-35.5062	4.406	-8.058	0.00	-71.012

The Model

$$y = \frac{e^{-35.5062+2.7957x_1+2.7532x_2+3.5153x_3}}{1+e^{-35.5062+2.7957x_1+2.7532x_2+3.5153x_3}}$$

231

Indian Statistical Institute

BINARY LOGISTIC REGRESSION

Example: Develop a model to predict the non payment of overdrafts by customers of a multinational banking institution. The data collected is given in Logistic_Reg.csv file.

Exporting the Predicted values

```
pred = myresult.predict(x)
myoutput = mypd.DataFrame(pred)
myoutput.to_csv("E:\ISI\PM-01/output.csv")
```

Actual	Predicted	
	0	1
0	257	14
1	14	695

Statistics	Computation	Value
Accuracy %	(257 + 695) / (257 + 695 + 14 + 14)	97.14
Misclassification Error %	100 – Accuracy %	2.85

232

Accuracy of ≥ 80 % is good

CLASSIFICATION and REGRESSION TREE

CLASSIFICATION AND REGRESSION TREE

Objective

To develop a predictive model to classify dependant or response metric (y) in terms of independent or exploratory variablesxs).

When to Use

xs : Continuous or discrete

y : Discrete or continuous

CLASSIFICATION AND REGRESSION TREE

Classification Tree

When response y is discrete

Method = “DecisionTreeClassifier”

Regression Tree

When response y is numeric

Method = “DecisionTreeRegressor”

CLASSIFICATION AND REGRESSION TREE

Example:

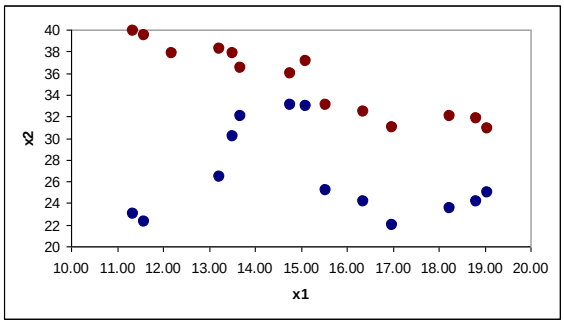
Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)

x1	x2	Y	x1	x2	Y
11.35	23	Blue	11.85	39.9	Red
11.59	22.3	Blue	12.09	39.5	Red
12.19	24.5	Blue	12.69	37.8	Red
13.23	26.4	Blue	13.73	38.2	Red
13.51	30.2	Blue	14.01	37.8	Red
13.68	32	Blue	14.18	36.5	Red
14.78	33.1	Blue	15.28	36	Red
15.11	33	Blue	15.61	37.1	Red
15.55	25.2	Blue	16.05	33.1	Red
16.37	24.1	Blue	16.87	32.4	Red
16.99	22	Blue	17.49	31	Red
18.23	23.5	Blue	18.73	32	Red
18.83	24.1	Blue	19.33	31.8	Red
19.06	25	Blue	19.56	30.9	Red

CLASSIFICATION AND REGRESSION TREE

Example:

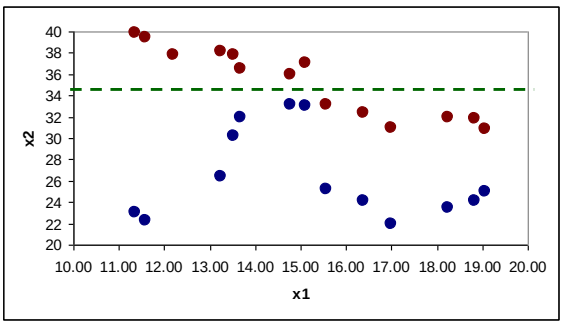
Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)

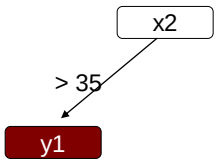


CLASSIFICATION AND REGRESSION TREE

Example:

Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)

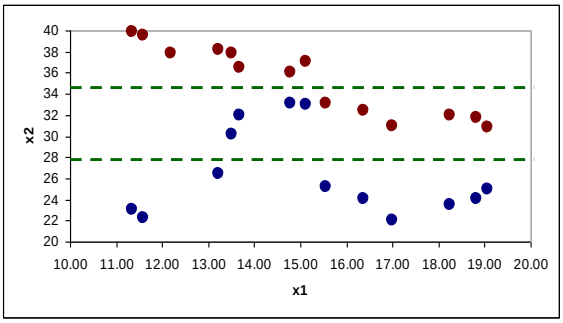


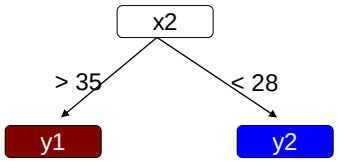


CLASSIFICATION AND REGRESSION TREE

Example:

Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)



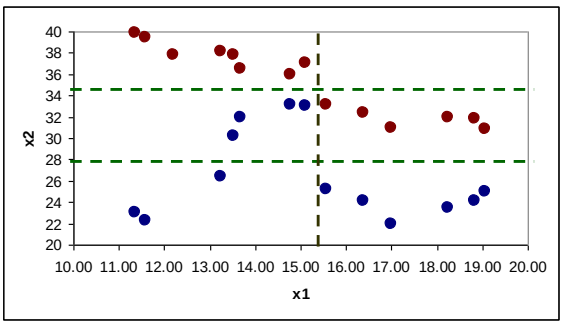


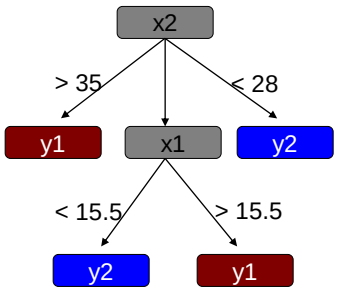
239

CLASSIFICATION AND REGRESSION TREE

Example:

Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)





240

CLASSIFICATION AND REGRESSION TREE

Example: Rules

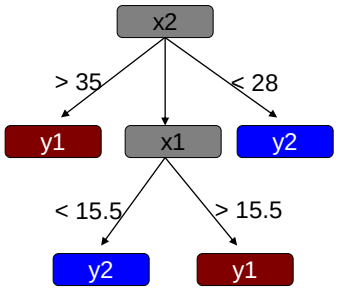
Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)

If x2 > 35 then y = y1

If x2 < 28, then y = y2

If 28 > x2 > 35 & x1 > 15.5, then y = y1

If 28 > x2 > 35 & x1 < 15.5, then y = y2



241

CLASSIFICATION AND REGRESSION TREE

Challenges

How to represent the entire information in the dataset using minimum number of rules?

How to develop the smallest tree?

Solution

Select the variable with maximum information (highest relation with y) for first split

242

CLASSIFICATION AND REGRESSION TREE

Example: A marketing company wants to optimize their mailing campaign by sending the brochure mail only to those customers who responded to previous mail campaigns. The profile of customers are given below. Can you develop a rule to identify the profile of customers who are likely to respond (Mail_Respond.csv)?

Profile Variable	Values
District	0:Urban, 1: Suburban & 2: Rural
House Type	0:Detached, 1: Semi Detached & 2: Terrace
Income	0:Low & 1: High
Previous Customer	0:No & 1:Yes

Output Variable	Value
Outcome	0:No & 1:Yes

243

CLASSIFICATION AND REGRESSION TREE

Example: A marketing company wants to optimize their mailing campaign by sending the brochure mail only to those customers who responded to previous mail campaigns. The profile of customers are given in mail_respond.csv? Can you develop a rule to identify the profile of customers who are likely to respond?

Number of variables = 4

SL No	Variable Name	Number of values
1	District	3
2	House Type	3
3	Income	2
4	Previous Customer	2

Total Combination of Customer Profiles = 3 x 3 x 2 x 2 = 36

244

CLASSIFICATION AND REGRESSION TREE

Read file and variables

```
import pandas as mypd
from sklearn import tree
mydata = mypd.read_csv("E:/ISI/PM-01/Data/Mail_Respond.csv")
x = mydata[["District", "House_Type", "Income", "Previous_Customer"]]
y = mydata.Outcome
```


CLASSIFICATION AND REGRESSION TREE

Develop the model

```
mymodel = tree.DecisionTreeClassifier(min_samples_split = 10)
mymodel.fit(x,y)
mymodel.score(x,y)
```

Statistics	Value (%)
Accuracy	100
Misclassification Error	0.00

CLASSIFICATION AND REGRESSION TREE

Model Accuracy measures

```
pred = mymodel.predict(x)
mytable = mypd.crosstab(y, pred)
mytable
```

Actual Vs predicted: %

Actual	Predicted	
	No	Yes
No	34	0
Yes	0	66

Accuracy = 34 + 66 = 100%

247

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep (0: No & 1: Yes) using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

Variables	Values
Age	Numeric
Sex	0:Male & 1: Female
Region	0: Inner City, 1: Rural, 2: Suburban & 3: Town
Income	Numeric
Married	0: No, 1: Yes
Children	Numeric
Car	0: No, 1: Yes
Saving Account	0: No, 1: Yes
Current Account	0: No, 1: Yes
Mortgage	0: No, 1: Yes

248

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

Reading data

```
import pandas as mypd
from sklearn import tree
from sklearn.cross_validation import train_test_split

mydata = mypd.read_csv("E:/ISI/PM-01/Data/bank-data.csv")
x = mydata.values[:, 0:9]
y = mydata.values[:, 10]
```


CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

Split data into training and test data

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2,
random_state = 100)
```

Develop model using training data

```
mymodel = tree.DecisionTreeClassifier(min_samples_split=50)
mymodel.fit(x_train, y_train)
mymodel.score(x_train, y_train)
```

Statistics	Value (%)
Accuracy	83.3
Misclassification Error	16.7

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

```
pred = mymodel.predict(x_train)
mytable = mypd.crosstab(y_train, pred)
mytable
```

Actual vs Predicted

Actual	Predicted	
	No	Yes
No	232	30
Yes	50	168

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

Validating the Model using test data

```
pred_test = mymodel.predict(x_test)
mytesttable = mypd.crosstab(y_test, pred_test)
mytesttable
```

Actual Vs predicted: %

Actual	Predicted	
	No	Yes
No	58	6
Yes	15	41

Accuracy = (58 + 41)/(58 + 6 + 15 +41) = 82.5 %

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

Data	Accuracy	Misclassification Error
Training	83.33	16.67
Test	82.5	17.5

RANDOM FOREST
and
BAGGING

RANDOM FOREST

Improves predictive accuracy

Generates large number of bootstrapped trees

Classifies a new case using each tree in the new forest of trees

Final predicted outcome by combining the results across all of the trees

Regression tree – average

Classification tree – majority vote

255

RANDOM FOREST

- Uses trees as building blocks to construct more powerful prediction models
- Decision trees suffer from high variance

  If we split the data into two parts and construct two different trees for each half of the data, the trees can be quite different
- In contrast, a proceedure with low varaince will yield similar results if applied repeatedly to distinct datasets
- Bagging is a general purpose procedure for reducing the variance of a statistiocal learning method

256

Indian Statistical Institute

RANDOM FOREST

Procedure

- Take many training sets from the population
- Build seperate prediction models using each training set
- Average the resulting predictions
- Averaging of a set of observatins reduce variance
- Different training datasets are taken using bootstrap sampling
- Generally bootstraped sample consists of two third of the observations and the model is tested on the remainng one third ofthe out of the bag observations

For discrete response – will take the majority vote instead of average

Major difference between bagging and Random Forest

Bagging generally uses all the p predictors while random forest uses $\sqrt{p}$ predictors

257

Indian Statistical Institute

RANDOM FOREST

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Python Code

Call libraries and import data

```
import pandas as mypd
from sklearn.ensemble import RandomForestRegressor
from sklearn.cross_validation import train_test_split
import math as mymath

mydata = mypd.read_csv("E:/ISI/PM-01/Data/Boston_Housing_Data.csv“)
x = mydata.values[:, 0:12]
y = mydata.values[:,13]
```

258

RANDOM FOREST

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Python Code

Split data into training and test

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 100)
```

Develop the model using training data - Bagging

```
mymodel = RandomForestRegressor(n_estimators = 500, min_sample_split = 40,
                max_features = None)
mymodel.fit(x_train,y_train)
```

n_estimators : Number of trees

max_features = None, include all (p) explanatory variable (x's)

max_features = 'auto', include subset ($\sqrt{p}$) explanatory variable (x's)

RANDOM FOREST

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Python Code

```
mymodel.score(x_train, y_train)
pred = mymodel.predict(x_train)
res = y_train – pred
res_sq = res**2
res_ss = res_sq.sum()
total_ss = y_train.var()*404
r_sq = 1 - res_ss/total_ss
mse = res_sq.mean()
rmse = mymath.sqrt(mse)
```

RANDOM FOREST

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Statistics	Value
MSE	3.733
RMSE	1.932
R^2	95.41

Indian Statistical Institute

RANDOM FOREST

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Python Code

Validate the model using test data

```
pred_test = mymodel.predict(x_test)
res_test = y_test- pred_test
res_test_sq = res_test**2
res_test_ss = res_test_sq.sum()
total_test_ss = t_test.var()*101
r_test_sq = 1 - res_test_ss/total_test_ss
mse = res_test_sq.mean()
rmse = mymath.sqrt(mse)
```

Indian Statistical Institute

RANDOM FOREST

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Statistics	Training	Test
MSE	3.733	18.007
RMSE	1.932	4.243
R^2	95.41	81.17

263

Indian Statistical Institute

RANDOM FOREST

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Developing model with random forest

```
mymodel = RandomForestRegressor(n_estimators = 500, min_samples_split =
          40, max_features= 'auto']
```

Developing model with CART

```
mymodel = tree.DecisiontreeRegressor(min_samples_split=40)
```

Statistics	Bagging		Random Forest		Regression Tree	
	Training	Test	Training	Test	Training	Test
MSE	3.733	18.007	4.449	20.169	13.287	28.879
RMSE	1.932	4.243	2.109	4.491	3.645	5.373
R^2	95.41	81.17	94.52	78.91	83.65	69.81

264

RANDOM FOREST

Exercise 1

Develop a model to predict whether a customer will take personal equity plan or not using bank-data .csv. Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

RANDOM FOREST

Exercise 2

Develop a model to predict the plant variety using Iris data. Validate the model using Iris_test data?

NAÏVE BAYES CLASSIFIER

NAÏVE BAYES CLASSIFIER

- Used to predict the probability that the value of the output variable will fall in an interval for a given set of values of input or predictor variables
- Assigns each observation to the most likely class, given its predictor values
- Uses the conditional probability of *P(y / x)* for making prediction

Methodology

Assign a test observation with predictor vector x_0 to the class *j* for which

$$P(y = j / x = x_0)$$

is the largest

NAÏVE BAYES CLASSIFIER

Example : The data on code review duration and defect density obtained for 10 code reviews are given below. Predict the defect density for review duration = Low using Naïve Bayes classifier?

SL No	Review Duration	Defect Density
1	Low	High
2	Low	Medium
3	Low	Low
4	Low	Medium
5	Low	Medium
6	Low	High
7	High	Low
8	High	High
9	High	Low
10	High	High
11	High	Low
12	High	Low

Predict *y* given *x* = Low

Indian Statistical Institute

NAÏVE BAYES CLASSIFIER

Example : The data on code review duration and defect density obtained for 10 code reviews are given below. Predict the defect density for review duration = Low using Naïve Bayes classifier?

SL No	Review Duration	Defect Density
1	Low	High
2	Low	Medium
3	Low	Low
4	Low	Medium
5	Low	Medium
6	Low	High
7	High	Low
8	High	High
9	High	Low
10	High	High
11	High	Low
12	High	Low

P(y = Low / x = Low)
= Number of cases when both x & y are Low / Number of cases x is Low = 1/6 = 0.17

P(y = Medium / x = Low)
= Number of cases both x is Low and y is Medium / Number of cases x is Low = 3/6 = 0.50

P(y = High / x = Low)
= Number of cases both x is Low and y is Medium / Number of cases x is Low = 2/6 = 0.33

NAÏVE BAYES CLASSIFIER

Example : The data on code review duration and defect density obtained for 10 code reviews are given below. Predict the defect density for review duration = Low using Naïve Bayes classifier?

SL No	Review Duration	Defect Density
1	Low	High
2	Low	Medium
3	Low	Low
4	Low	Medium
5	Low	Medium
6	Low	High
7	High	Low
8	High	High
9	High	Low
10	High	High
11	High	Low
12	High	Low

Maximum of

P(y = Low / x = Low), P(y = Medium / x = Low) and P(y = High / x = Low)

max (0.17, 0.50, 0. 33) = 0.50

= P(y = Medium / x = Low)

Predicted value of y for x = Low is y = Medium

NAÏVE BAYES CLASSIFIER

Used to develop models when the output or response variable y is categorical

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

NAÏVE BAYES CLASSIFIER

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Read Data

```
import pandas as mypd
from sklearn.naive_bayes import GaussianNB

mydata = mypd.read_csv("E:/ISI/PM-01/Data/Iris_data.csv")
x = mydata.values[:, 0:4]
y = mydata.values[:, 4]
```

Develop Model

```
mymodel = GaussianNB()
mymodel.fit(x, y)
pred = mymodel.predict(x)
mytable = mypd.crosstab(y, pred)
mytable
```


NAÏVE BAYES CLASSIFIER

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Actual	predicted		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	37	0	0
Iris-versicolor	0	32	3
Iris-virginica	0	2	40

Statistics	Value
Accuracy	95.61
Misclassification Error	4.39

NAÏVE BAYES CLASSIFIER

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Validating the model on test data

```
mytestdata = mypd.read_csv("E:/ISI/PM-01/Data/Iris_test.csv")
test_x =mytestdata.values[:,0:4]
test_y = mytestdata.values[:,4]

pred_test = mymodel.predict(test_x)
mytesttable = mypd.crosstab(test_y, pred_test)
mytesttable
```


NAÏVE BAYES CLASSIFIER

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Validation Results

Actual	Predicted		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	13	0	0
Iris-versicolor	0	14	1
Iris-virginica	0	1	7

Statistics	Value
Accuracy	94.44
Misclassification Error	5.56

NAÏVE BAYES CLASSIFIER

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Results

Statistics	Training	Test
Accuracy	95.61	94.44
Misclassification Error	4.39	5.56

K - NEAREST NEIGHBORS

K - NEAREST NEIGHBORS

A non parametric approach for developing models

No assumptions are made about the shape of the decision boundary or underlying distribution

Performs better than regression when the relationship between x's and y is non linear

Will not provide information about which x's are important

K - NEAREST NEIGHBORS

Methodology

Let (x_i, y_i) be the training dataset consists of large number of n records

Let x_0 be the test observation set for which the value of y_0 need to be predicted

Step 1: Identify k records from (x_i, y_i) with x_i values are close to x_0

Step 2: Compute the predicted y_0 from the k y_i values

if y is continuous then y_0 = average of k y_i's

else y_0 = maximum occurring value of y_i in k y_i's

K - NEAREST NEIGHBORS

Example 1

A develop a methodology to predict the value of y in terms of x1 and x2 based on the data given below. Use k – nearest neighbors approach with k = 3. using the methodology predict the value of y for x1 = 15.2 and x2 = 33.1

Training Data set			
Record No.	x1	x2	Y
1	11.35	23	Blue
2	11.59	22.3	Blue
3	12.19	24.5	Blue
4	13.23	26.4	Blue
5	13.51	30.2	Blue
6	13.68	32	Blue
7	14.78	33.1	Blue
8	15.11	33	Blue
9	15.55	25.2	Blue
10	11.85	39.9	Red
11	12.09	39.5	Red
12	12.69	37.8	Red
13	13.73	38.2	Red
14	14.01	37.8	Red
15	14.18	36.5	Red
16	15.28	36	Red
17	15.61	37.1	Red
18	16.05	33.1	Red

Test data		
x1	x2	y
15.20	33.1	?

281

Indian Statistical Institute

K - NEAREST NEIGHBORS

Example 1

A develop a methodology to predict the value of y in terms of x1 and x2 based on the data given below. Use k – nearest neighbors approach with k = 3. using the methodology predict the value of y for x1 = 15.2 and x2 = 33.1

Step 1: Compute the distance (Euclidean) of each record in training data from test data

Record No.	x1	x2	Y	Ecludean distance
1	11.35	23	Blue	10.81
2	11.59	22.3	Blue	11.39
3	12.19	24.5	Blue	9.11
4	13.23	26.4	Blue	6.98
5	13.51	30.2	Blue	3.36
6	13.68	32	Blue	1.88
7	14.78	33.1	Blue	0.42
8	15.11	33	Blue	0.13
9	15.55	25.2	Blue	7.91
10	11.85	39.9	Red	7.58
11	12.09	39.5	Red	7.12
12	12.69	37.8	Red	5.33
13	13.73	38.2	Red	5.31
14	14.01	37.8	Red	4.85
15	14.18	36.5	Red	3.55
16	15.28	36	Red	2.90
17	15.61	37.1	Red	4.02
18	16.05	33.1	Red	0.85

282

K - NEAREST NEIGHBORS

Example 1

A develop a methodology to predict the value of y in terms of x1 and x2 based on the data given below. Use k – nearest neighbors approach with k = 3. using the methodology predict the value of y for x1 = 15.2 and x2 = 33.1

Step 2: identify k = 3 records closest (with minimum distance) to test data

Record No.	x1	x2	Y	Ecludean distance
1	11.35	23	Blue	10.81
2	11.59	22.3	Blue	11.39
3	12.19	24.5	Blue	9.11
4	13.23	26.4	Blue	6.98
5	13.51	30.2	Blue	3.36
6	13.68	32	Blue	1.88
7	14.78	33.1	Blue	0.42
8	15.11	33	Blue	0.13
9	15.55	25.2	Blue	7.91
10	11.85	39.9	Red	7.58
11	12.09	39.5	Red	7.12
12	12.69	37.8	Red	5.33
13	13.73	38.2	Red	5.31
14	14.01	37.8	Red	4.85
15	14.18	36.5	Red	3.55
16	15.28	36	Red	2.90
17	15.61	37.1	Red	4.02
18	16.05	33.1	Red	0.85

283

Indian Statistical Institute

K - NEAREST NEIGHBORS

Example 1

A develop a methodology to predict the value of y in terms of x1 and x2 based on the data given below. Use k – nearest neighbors approach with k = 3. using the methodology predict the value of y for x1 = 15.2 and x2 = 33.1

Step 3: Count different y values in k = 3 records. The predicted value is the mode

Record No.	x1	x2	Y	Euclidean distance
7	14.78	33.1	Blue	0.42
8	15.11	33	Blue	0.13
18	16.05	33.1	Red	0.85

y	Number of Occurrences
Blue	2
Red	1
Mode	Blue

x1	x2	Predicted y
15.20	33.1	Blue

284

K - NEAREST NEIGHBORS

Example 2 : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time? Use k – nearest neighbors approach with k = 2. Predict the yield for the following temperature & time?

Variable	Value
Time	185
Temperature	225
Yield	?

285

Indian Statistical Institute

K - NEAREST NEIGHBORS

Example 2 : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time? Use k – nearest neighbors approach with k = 2. Predict the yield for the following temperature & time?

Record Number	Time	Temperature	%Yield	Euclidean Distance
1	130	190	35	65.192
2	174	176	81.7	50.220
3	134	205	42.5	54.781
4	191	210	98.3	16.155
5	165	230	52.7	20.616
6	194	192	82	34.205
7	143	220	34.5	42.297
8	186	235	95.4	10.050
9	139	240	56.7	48.384
10	188	230	84.4	5.831
11	175	200	94.3	26.926
12	156	218	44.3	29.833
13	190	220	83.3	7.071
14	178	210	91.4	16.553
15	132	208	43.5	55.660
16	148	225	51.7	37.000

Variable	Value
Time	185
Temperature	225
Yield	=(84.4 + 83.3)/2 = 83.85

286

K - NEAREST NEIGHBORS

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data? Use k = 5

Call libraries and import data

```
import pandas as mypd
from sklearn.neighbors import KNeighborsClassifier

mydata = mypd.read_csv("E:/ISI/PM-01/Data/Iris_data.csv")
x = mydata.values[:, 0:4]
y = mydata.values[:, 4]
```

Develop Model

```
mymodel = KNeighborsClassifier(n_neighbors = 5)
mymodel.fit(x, y)
mymodel.score(x, y)
pred = mymodel.predict(x)
mytable = mypd.crosstab(y, pred)
mytable
```


K - NEAREST NEIGHBORS

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data? Use k = 5

Actual vs Predicted

Actual	Predicted		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	37	0	0
Iris-versicolor	0	34	1
Iris-virginica	0	1	41

Statistics	Value
Accuracy	98.24
Misclassification Error	1.76

K - NEAREST NEIGHBORS

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data? Use k = 5

Validating the model on test data

```
mytestdata = mypd.read_csv("E:/ISI/PM-01/Data/Iris_test.csv")
test_x = mytestdata.values[:, 0:4]
test_y = mytestdata.values[:, 4]
pred_test = mymodel.predict(test_x)
mytesttable = mypd.crosstab(test_y, pred_test)
mytesttable
```


K - NEAREST NEIGHBORS

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data? Use k = 5

Validating the model on test data

Actual	Predicted		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	13	0	0
Iris-versicolor	0	13	2
Iris-virginica	0	0	8

Statistics	Value
Accuracy	94.44
Misclassification Error	5.56

K - NEAREST NEIGHBORS

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data? Use k = 5

Result

Statistics	Training	Test
Accuracy	98.24	94.44
Misclassification Error	1.76	5.56

SUPPORT VECTOR MACHINE

SUPPORT VECTOR MACHINE

Hyperplane

In two dimensions, a hyperplane is a one dimension subspace namely a line

In three dimensions, a hyperplane is a flat two dimension subspace namely a plane

In a *p* dimensional space, a hyperplane is a flat affine subspace of *p-1* dimension

Mathematical Equation

In *2* dimension $\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$

Any point x = (x_1, x_2) satisfying the above equation will be in the hyperplane

In *p* dimension

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + --- + \beta_p x_p = 0$$

Any point x = (x_1, x_2, - - - , x_p) satisfying the above equation will be in the hyperplane

293

SUPPORT VECTOR MACHINE

Hyperplane

Suppose for a x = (x_1, x_2, - - - , x_p)

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + --- + \beta_p x_p > 0$$

Then the x = (x_1, x_2, - - - , x_p) lies in one side of the hyperplane

Suppose for a x = (x_1, x_2, - - - , x_p)

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + --- + \beta_p x_p < 0$$

Then the x = (x_1, x_2, - - - , x_p) lies on the other side of the hyperplane

Hence

Hyperplane is dividing *p* dimensional space into *2* halves

We can easily determine which side of the hyperplane a point lies by evaluating the hyperplane

294

SUPPORT VECTOR MACHINE

Procedure

Suppose it is possible to construct a hyperplane that separate the training observations perfectly into two classes according to their class labels (say y = 1 0r y = -1).

Then a separating hyperplane has the property that

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + --- + \beta_p x_p > 0$ If y = 1 and

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + --- + \beta_p x_p < 0$ If y = -1 and

If a separating hyperplane exists, it can be used to construct a natural classifier

A test observation is assigned to a class depending on which side of the hyperplane it is located

295

Indian Statistical Institute

SUPPORT VECTOR MACHINE

Maximal Marginal Classifier

If the data can be perfectly separated using a hyperplane, then there exists many such hyperplanes

Then the best separating hyperplane (maximal marginal hyperplane) is the one which is furthest from the training observations

Margin: The minimal distance from hyperplane to an observation

Maximal marginal classifier is the separating hyperplane with maximum margin.

296

SUPPORT VECTOR MACHINE

Support Vector Machine

In many cases no separating hyperplane exists
So no maximum marginal classifier exists

The generalisation of maximum marginal hyperplane to no separable cases is Support Vector Machine

In SVM, a hyperplane is chosen to separate most of the observations into the two classes but may misclassify a few observations

C: The number of misclassified observations. Optimum C can be obtained through cross validation.

297

SUPPORT VECTOR MACHINE

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Support Vector Machine. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

298

SUPPORT VECTOR MACHINE

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Support Vector Machine. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Call libraries and import data

```
import pandas as mypd
from sklearn import svm

mydata = mypd.read_csv("E:/ISI/PM-01/Data/Iris_data.csv")
mydata x = mydata.values[:, 0:4]
y = mydata.values[:, 4]
```

Develop Model

```
mymodel = svm.SVC()
mymodel.fit(x, y)
mymodel.score(x, y)
pred = mymodel.predict(x)
mytable = mypd.crosstab(y, pred)
mytable
```


SUPPORT VECTOR MACHINE

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Support Vector Machine. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Actual vs Predicted

Actual	Predicted		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	37	0	0
Iris-versicolor	0	33	2
Iris-virginica	0	0	42

Statistics	Value
Accuracy	98.24
Misclassification Error	1.76

SUPPORT VECTOR MACHINE

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Support Vector Machine. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Validating the model on test data

```
mytestdata = mypd.read_csv("E:/ISI/PM-01/Data/Iris_test.csv")
test_x = mytestdata.values[:, 0:4]
test_y = mytestdata.values[:, 4]
pred_test = mymodel.predict(test_x)
mytesttable = mypd.crosstab(test_y, pred_test)
mytesttable
```


SUPPORT VECTOR MACHINE

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Support Vector Machine. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Validating the model on test data

Actual	Predicted		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	13	0	0
Iris-versicolor	0	14	1
Iris-virginica	0	0	8

Statistics	Value
Accuracy	97.22
Misclassification Error	2.78

SUPPORT VECTOR MACHINE

Example: Develop a model to predict the iris plant class (1: Iris-setosa, 2: Iris-versicolor & 3: Iris-virginica) based on sepal length, sepal width, petal length and petal width using Support Vector Machine. The data is given in Iris_data.csv file. Validate the model with Iris_test.csv data?

Result

Statistics	Training	Test
Accuracy	98.24	97.22
Misclassification Error	1.76	2.78

303

ARTIFICIAL NEURAL NETWORKS

304

ARTIFICAL NEURAL NETWORKS

Introduction

One of the most fascinating machine learning modeling technique

Generally uses back propagation algorithm

Relatively complex (due to deep learning with many hidden layers)

Structure is inspired by brain functioning

Generally computationally expensive

ARTIFICAL NEURAL NETWORKS

Instructions

1. Normalize the data – Use Min – Max transformation (optional)

   Normalized data = Data – Minimum / (Maximum – Minimum)

2. Number of hidden layers required = 1 for vast number of application

3. Number of neurons required = 2/3 of the number of predictor variables or input layers

Remark: The optimum number of layers and neurons are the ones which would minimize mean square error or misclassification error which can be obtained by testing again and again

ARTIFICAL NEURAL NETWORKS

Example: Develop a model to predict the non payment of overdrafts by customers of a multinational banking institution. The data collected is given in Logistic_Reg.csv file. The factors and response considered are given below. Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

SL No	Factor
1	Individual expected level of activity score
2	Transaction speed score
3	Peer comparison score in terms of transaction volume

Response	Values
Outcome	0: Not Paid and 1: Paid

ARTIFICAL NEURAL NETWORKS

Example

Importing packages

```
import pandas as mypd
from sklearn.cross_validation import train_test_split
from sklearn.neural_network import MLPClassifier
```

Reading the data

```
mydata = mypd.read_csv("E:/ISI/PM03/Course_Material/Data/Logistic_Reg.csv")
x = mydata.values[:, 0:3]
y = mydata.Outcome
```

Splitting the data into training and test

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 100)
```

ARTIFICAL NEURAL NETWORKS

Example

Develop the model

```
mymodel =MLPClassifier(solver = 'lbfgs', alpha = 1e-5, hidden_layer_sizes = (2), random_state = 100)
mymodel.fit(x_train, y_train)
```

Note:

Classification problem: Use MLPCLassifier

Value estimation: Use MLPRegressor

Solver:
'lbfgs' : Uses quasi-Newton method optimization algorithm.
'sgd' :Uses stochastic gradient descent optimization algorithm.
'adam' :Uses stochastic gradient-based optimizer

ARTIFICAL NEURAL NETWORKS

Example: Interpretation

hidden_layer_sizes : a vector representing hidden layers and hidden neurons in each layer

hidden_layer_sizes = (l) : one hidden layers with *l* hidden neurons

ARTIFICAL NEURAL NETWORKS

Output

mymodel.score(x_train, y_train)

Statistics	Value
% Accuracy	96.81
% Error	3.19

mymodel.predict_proba(x_train)

ARTIFICAL NEURAL NETWORKS

Output: Validation

predtest = mymodel.predict(x_test)

mytable = mypd.crosstab(y_test, predtest)

mytable

Actual Vs Predicted

		Predicted	
		0	1
Actual	0	54	4
	1	0	138

ARTIFICAL NEURAL NETWORKS

Output: Validation

Actual Vs Predicted (%)

		Predicted	
		0	1
Actual	0	27.55	2.04
	1	0.00	70.41

Statistics	Training	Test
% Accuracy	96.81	97.96
% Error	3.19	2.04

ARTIFICAL NEURAL NETWORKS

Output

```
> mse = mean(res^2)
> rmse = sqrt(mse)
> residual_ss = sum(res^2)
> total_ss = var(myzdata$Conversion)*15
> r_sq = 1 – residual_ss / total_ss
```

Statistics	Value
Mean Square Error	0.0009994
Root Mean Square Error	0.0316128
R Square	0.9905

ARTIFICAL NEURAL NETWORKS

Prediction for new data set

```
> test <- read_csv("E:/Infosys/output.csv“)
> output = compute(mymodel, test)
> output$net.result
```

Temperature	Time	Kappa_Number	Conversion	Predicted Conversion
1	0.0058	0.1243	0.9577	0.9882
1	0.0058	0.2090	0.9915	0.9813
1	0.0000	0.3220	1.0000	0.9782
1	0.0173	0.4633	0.9437	0.9269
1	0.0231	0.6610	0.9155	0.8871

315

ARTIFICAL NEURAL NETWORKS

Exercise 1

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

316

ARTIFICAL NEURAL NETWORKS

Exercise 1

Python Code – Import the packages

```
import pandas as mypd
from sklearn.cross_validation import train_test_split
from sklearn.neural_network import MLPRegressor
```

Import the data

```
mydata = mypd.read_csv("E:/ISI/PM- 03/Course_Material/Data/ Boston_Housing_Data.csv")
x = mydata.values[:, 0:12]
y = mydata.values[:,13]
```

Split data into training and test

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state
= 100)
```


ARTIFICAL NEURAL NETWORKS

Exercise 1

Develop the model

```
mymodel = MLPRegressor(solver = 'lbfgs', alpha = 0.001, hidden_layer_sizes =
          (6), random_state= 100)
mymodel.fit(x_train, y_train)
mymodel.score(x_train,y_train)
```

Statistic	Value
R^2	66.76

ARTIFICAL NEURAL NETWORKS

Validation: Test data

```
pred = mymodel.predict(x_test)
res = y_test – pred
res_sq = res**2
res_ss = sum(res_sq)
total_ss = y_test.var()*100

rsq = 1 - res_ss/total_ss
rsq
```

Statistic	Training	Test
R^2	66.76	63.43

Foundation Course

on

Predictive Modeling

using

Python

Thank You