BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Overview of the course
& Basic Probability & Statistics (CS -1)

(Session 1: 20th /21st May,2023)

Overview of the course

- M 1 : Basic Probability & Statistics
- M 2 : Conditional Probability & Bayes' theorem
- M 3 : Probability Distributions
- M 4 : Hypothesis Testing
- M 5 : Prediction & Forecasting
- M 6 : Prediction & Forecasting Gaussian Mixture model & Expectation Maximization

TEXT BOOKS

T1 : Statistics for Data Scientists, An introduction to probability ,statistics and Data Analysis, Maurits Kaptein et al,Springer 2022

T2 : Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning

T3 : Introduction to Time Series and Forecasting, Second Edition, Peter J Brockwell, Richard A Davis, Springer.

Evaluation Components

No	Name	Type	Weight
EC-1(a)	Quizzes – 1 ,2 & 3 (Best two will be considered)	Online	10%
EC-1(b)	Assignments - 2	Online	20%
EC-2	Mid-Semester Test	Closed Book	30%
EC-3	Comprehensive Exam	Open Book	40%

Module 1: (Basic Probability & Statistics)

Contact Session	List of Topic Title	Reference
CS - 1	Measures of Central Tendency & Measures of Variability, Data – Symmetric & Asymmetric, outlier detection, 5 point summary, Introduction to probability	T1 & T2

“Statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write”

H G Wells

Statistics

Statistics may be defined as science that is employed to

- Collect the data
- Present and organize the data in a systematic manner
- Analyse the data
- Infer about the data
- Take decision from the data.

In other words, Statistics can also be defined as numerical data with a view to analyse it.

Types of Variable
Qualitative (Categorical): express a qualitative attribute such as hair color, eye color, religion.
Quantitative(Numerical): measured in terms of numbers such as height, weight, number of people.
Nominal: no ordering is possible such as hair color, eye color, religion.
Ordinal: ordering is possible such as health, which can take values such as poor, reasonable, good, or excellent.
Discrete: countable and have a finite number of possibilities such as number of people
Continuous: not countable and have an infinite number of possibilities such as height
INTERVAL: ratio of values of variable do not have any meaning and it does not have an inherently defined zero value such as temperature
RATIO: ratio of values of variable have meaning and it have an inherently defined zero value such as length

Measures of Central Tendency

- Measure of central tendency provides a very convenient way of describing a set of scores with a single number that describes the **PERFORMANCE** of the group.

- Also defined as a single value that is used to describe the "**center**" of the data.

- Three commonly used measures of central tendency:
  1. Mean
  2. Median
  3. Mode

Mean

- Also referred as the "**arithmetic average**"
- The most commonly used measure of the center of data
- Numbers that describe what is average or typical of the distribution
- Computation of Sample Mean:

$$\bar{Y} = \frac{\sum Y}{N}$$ = Σ Y/ N = (Y1 + Y2 + Y3 + … Yn) / N where

"Y bar" equals the sum of all the scores, Y, divided by the number of scores, N.

- Computation of the Mean for grouped Data

$$\bar{Y} = \frac{\sum f\,Y}{N}$$ Where f Y = a score multiplied by its frequency

Mean: Grouped Scores

Hours Spent Watching TV	Frequency (f)	fY	Percentage	C%
1	104	104	31.3	31.3
2	130	260	39.2	70.5
3	98	294	29.5	100.0
Total	332	658	100.0	

$$\bar{Y} = \frac{\sum fY}{N} = \frac{658}{332} = 1.98$$

Data of Children watching TV in Bengaluru

Mean

Properties

- It measures stability. Mean is the most stable among other measures of central tendency because every score contributes to the value of the mean.
- It may easily affected by the extreme scores.
- The sum of each score's distance from the mean is zero.
- It can be applied to interval level of measurement
- It may not be an actual score in the distribution
- It is very easy to compute.

Mean

When to Use the Mean

- Sampling stability is desired.
- Other measures are to be computed such as standard deviation, coefficient of variation and skewness

The Mode

- The category or score with the largest frequency (or percentage) in the distribution.
- The mode can be calculated for variables with levels of measurement that are: nominal, ordinal, or interval-ratio.

Example:

- Number of Votes for Candidates for Lok Sabha MP. The mode, in this case, gives you the "central" response of the voters: the most popular candidate.

  - Candidate A – 11,769 votes
  - Candidate B – 39,443 votes
  - Candidate C – 78,331 votes

The Mode:
"Candidate C"

Mode

Properties

- It can be used when the data are qualitative as well as quantitative.
- It may not be unique.
- It is not affected by extreme values.

When to Use the Mode

- When the “typical” value is desired.
- When the data set is measured on a nominal scale

The Median

- The score that divides the distribution into two equal parts, so that half the cases are above it and half below it.

- The median is the middle score, or average of middle scores in a distribution.
  - Fifty percent (50%) lies below the median value and 50% lies above the median value.
  - It is also known as the middle score or the 50th percentile.

Measures of central tendency

$$\frac{x_1 + x_2 + x_3 + \cdots x_n}{n}$$

μ

$n-1$ $\bar{x}$

- The mean

Drawback?

- the median

10, 15, 20, 25, 26

10, 15, 20, 25, 28, 32

Average of these

- the mode

2, 5, 5, 2, 3, 2, 2, 2

5, 2, 5, 5, 2, 3, 2, 2, 2, 5, 5

Data : Symmetrical and Asymmetrical

Negative Skew

Mean < Median

symmetric

Mean = Median

Positive Skew

Mean > Median

Shape of the distribution of data

- Symmetrical : Mean is equal to median
- Skewed
  - Negatively : mean < median
  - Positively : mean > median
- Bimodal : has two distinct modes
- Multi-modal : has more than 2 distinct modes

Distribution Shape

Sl. No.	X_1	X_2
1	2	1
2	8	15
3	5	5
4	3	5
5	7	6
6	8	3
7	5	5
8	2	2
9	5	3
Total	45	45

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

Statistical measures	Group 1
Mean	5
Median	5
Mode	5

Sl. No.	X_2
1	1
2	15
3	5
4	5
5	6
6	3
7	5
8	2
9	3
Total	45

Statistical measures	Group 2
Mean	5
Median	5
Mode	5

Sl. No.	X_1	X_2
1	2	1
2	8	15
3	5	5
4	3	5
5	7	6
6	8	3
7	5	5
8	2	2
9	5	3
Total	45	45

Statistical measures	Group 1 & 2
Mean	5
Median	5
Mode	5

Do we need any other measure?

Answer: Yes

Measures of variability

Three Measures of Variability:

- The Range
- The Variance
- The Standard Deviations

Measure of Variability

Variability can be defined several ways:

- A quantitative distance measure based on the differences between scores
- Describes distance of the spread of scores or distance of a score from the mean

Purposes of Measure of Variability:

- Describe the distribution
- Measure how well an individual score represents the distribution

The Three Measures

Three Measures of Variability:

- The Range
- The Variance
- The Standard Deviations

The Ranges

- The distance covered by the scores in a distribution – From smallest value to highest value
- For continuous data, real limits are used

Range = URL for X_{max} - LRL for X_{min}

- Based on two scores, not all the data – An imprecise, unreliable measure of variability

Example: For a set of scores: 7, 2, 7, 6, 5, 6, 2

Range = Highest Score minus Lowest score = 7 - 2 = 5

The Standard Deviation

- Most common and most important measure of variability is the standard deviation
  - A measure of the standard, or average, distance from the mean
  - Describes whether the scores are clustered closely around the mean or are widely scattered
- Calculation differs for population and samples
- Variance is a necessary *companion concept* to standard deviation but *not the same* concept

The Standard Deviation

Exercise : Find out the deviations of all the data points with the mean….and then find the ‘mean deviation’.

The Standard Deviation

- Mean deviations will always be ‘zero’ !
  (because Mean is a balance point)

Then, how do you find ‘Standard Deviation’ ?

Need a new strategy

The Standard Deviation

New Strategy :

a) First square each deviation score
b) Then sum the Squared Deviations (SS)
c) Average the squared deviations

- Mean Squared Deviation is known as **"Variance"**
- Variability is now measured in squared units

$$Standard\ Deviation = \sqrt{Variance}$$

The Variance

Variance equals mean (average) squared deviation (distance) of the scores from the mean

$$\text{Variance} = \frac{\text{sum of squared deviations}}{\text{number of scores}}$$

where $$SS = \sum (X - \mu)^2$$

The Population Variance

- Population variance equals mean (average) squared deviation (distance) of the scores from the population mean

- Variance is the average of squared deviations, so we identify population variance with a lowercase Greek letter sigma squared: σ 2

- Standard deviation is the square root of the variance, so we identify it with a lowercase Greek letter sigma: σ

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

Statistical measures	Group 1
Mean	5
Median	5
Mode	5

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{45}{5} = 5$$

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

$$\mathbf{S} = \sqrt{\frac{44}{8}} = 2.345$$

Sl. No.	X_2
1	1
2	15
3	5
4	5
5	6
6	3
7	5
8	2
9	3
Total	45

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{45}{5} = 5$$

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

$$S = \sqrt{\frac{134}{8}} = 4.093$$

Standard Deviation and Variance for a Sample

- Goal of inferential statistics:
  - Draw general conclusions about population
  -
  - Based on limited information from a sample

- Samples differ from the population
  - Samples have less variability

  - Computing the Variance and Standard Deviation in the same way as for a population would give a biased estimate of the population values

Sample Standard Deviation and Variance

- Sum of Squares (SS) is computed as before
- Formula for Variance has n-1 rather than N in the denominator
- Notation uses s instead of σ

Population of Adult Heights

$$\text{variance of sample} = s^2 = \frac{SS}{n-1}$$

$$\text{standard deviation of sample} = s = \sqrt{\frac{SS}{n-1}}$$

Degrees of Freedom

- Population variance
  - Mean is known
  - Deviations are computed from a known mean

- Sample variance as estimate of population
  - Population mean is unknown
  - Using sample mean restricts variability

- Degrees of freedom
  - Number of scores in sample that are independent and free to vary
  - Degrees of freedom (df) = n – 1

Descriptive Statistics

- A standard deviation describes scores in terms of distance from the mean
- Describe an entire distribution with just two numbers (M and s)
- Reference to both allows reconstruction of the measurement scale from just these two numbers
- Means and standard deviations together provide extremely useful descriptive statistics for characterizing distributions

Five point summary of Data

The five number summary of data includes 5 items:

- **Minimum**.
- **Q1** (the first quartile, or the 25% mark).
- **Median**.
- **Q3** (the third quartile, or the 75% mark).
- **Maximum**.

Interquartile range (IQR)

- It is measure of Variation
- Also Known as Midspread : Spread in the Middle 50%
- Difference Between Third & First Quartiles:
- Not Affected by Extreme Values

Interquartile Range = IQR = $Q_3 - Q_1$

Data in Ordered Array: 11 12 13 16 16 17 17 18 21

Position of $Q_1 = \frac{1 \cdot (9 + 1)}{4} = 2.50$, $Q_1 = 12.5$

Position of $Q_3 = \frac{3 \cdot (9 + 1)}{4} = 7.50$, $Q_3 = 17.5$

Interquartile Range = IQR= $Q_3 - Q_1 = 17.5 - 12.5 = 5$

Box-and-Whisker plot

Box-and-Whisker plot

Box-and-Whisker plot

Potential outliers

- The lower limit and upper limit of a data set are given by:

  Lower limit = Q_1 - 1.5 x IQR

  Upper limit = Q_3 + 1.5 x IQR

- Data points that lie below the lower limit or above the upper limit are **potential outliers.**

HW problem :

For the data set below:

82	45	64	80	82	74	79	80	80	78	80	80	48	73	80	79	81	70	78	73

(a.) Obtain and interpret the quartiles.
(b.) Determine and interpret the interquartile range.
(c.) Find and interpret the five-number(point) summary.
(d.) Identify potential outliers, if any.
(e.) Construct and interpret a boxplot.

HW problem :

Human measurements provide a rich area of application for statistical methods. The article "A Longitudinal Study of the Development of Elementary School Children's Private Speech" (*Merrill-Palmer Q.*, 1990: 443–463) reported on a study of children talking to themselves (private speech). It was thought that private speech would be related to IQ, because IQ is supposed to measure mental maturity, and it was known that private speech decreases as students progress through the primary grades. The study included 33 students whose first-grade IQ scores are given here:

82	96	99	102	103	103	106	107	108	108	108	108
109	110	110	111	113	113	113	113	115	115	118	118
119	121	122	122	127	132	136	140	146			

Describe the data and comment on any interesting features.

Introduction to probability

Random Experiment :

❖ The term "**random experiment**" is used to describe any action whose outcome is not known in advance. Here are some examples of experiments dealing with statistical data:

- Tossing a coin
- Counting how many times a certain word or a combination of words appears in the text of the “King Lear” or in a text of Confucius.
- Counting occurrences of a certain combination of amino acids in a protein database.
- Pulling a card from the deck.

the experiment.

Sample spaces and events

- If the experiment is to throw a die and record the outcome, then sample space is S = { 1,2,3,4,5,6}

Event : An event is a subset of sample space of the random experiment.

Definition of probability

Classical approach :

CLASSICAL PROBABILITY

$$\text{Probability of an event} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Empirical approach :

Empirical or **relative frequency** is the second type of objective probability. It is based on the number of times an event occurs as a proportion of a known number of trials.

EMPIRICAL PROBABILITY The probability of an event happening is the fraction of the time similar events happened in the past.

In terms of a formula:

$$\text{Empirical probability} = \frac{\text{Number of times the event occurs}}{\text{Total number of observations}}$$

The empirical approach to probability is based on what is called the law of large numbers. The key to establishing probabilities empirically is that more observations will provide a more accurate estimate of the probability.

LAW OF LARGE NUMBERS Over a large number of trials, the empirical probability of an event will approach its true probability.

Axiomatic approach :

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:

If S is the sample space and E is any event in a random experiment,

(1) $P(S) = 1$

(2) $0 \leq P(E) \leq 1$

(3) For two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Thank You

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

IMP Note to Self

innovate achieve lead

BITS Pilani

Session No 2

Axioms of Probability, Probability basics, mutually exclusive and independent events,

(Session 2: 27th/28th MAY 2023)

Contact Session 2

Contact Session 2: Module 1(Module 1:Basic Probability & Statisitcs)

Contact Session	List of Topic Title	Reference
CS - 2	Axioms of Probability,Mutually exclusive and independent events,Problem solving to understand basic probability concepts	T1 & T2
HW	Problems on probability	T1 & T2
Lab		

Agenda

- Experiments, assignment of probabilities
- Events and their probability
- Some basic relationships of probability
- Basic problem solving

Introduction to probability

Random Experiment :

The term "**random experiment**" is used to describe any action whose
outcome is not known in advance. Here are some examples of experiments dealing with statistical data:

- Tossing a coin
- Counting how many times a certain word or a combination of words appears in the text of the “King Lear” or in a text of Confucius.
- Counting occurrences of a certain combination of amino acids in a protein database.
- Pulling a card from the deck.

Sample spaces and events

<u>Sample space</u> :

- The sample space of a random experiment is a set S that includes all possible outcomes of the experiment.

If the experiment is to throw a die and record the outcome, then,

sample space is S = { 1,2,3,4,5,6}

Event : An event is a subset of sample space of the random experiment.

Definition of probability

Classical approach :

CLASSICAL PROBABILITY

$$\text{Probability of an event} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Empirical approach :

Empirical or **relative frequency** is the second type of objective probability. It is based on the number of times an event occurs as a proportion of a known number of trials.

EMPIRICAL PROBABILITY The probability of an event happening is the fraction of the time similar events happened in the past.

In terms of a formula:

$$\text{Empirical probability} = \frac{\text{Number of times the event occurs}}{\text{Total number of observations}}$$

The empirical approach to probability is based on what is called the law of large numbers. The key to establishing probabilities empirically is that more observations will provide a more accurate estimate of the probability.

LAW OF LARGE NUMBERS Over a large number of trials, the empirical probability of an event will approach its true probability.

Axiomatic approach :

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:

If S is the sample space and E is any event in a random experiment,

(1) $P(S) = 1$

(2) $0 \leq P(E) \leq 1$

(3) For two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

RECALL: Random Experiment

Term "**random experiment**" is used to describe any action whose outcome is not known in advance. Here are some examples of experiments dealing with statistical data:

- Tossing a coin
- Counting how many times a certain word or a combination of words appears in the text of the “King Lear” or in a text of Confucius
- counting occurrences of a certain combination of amino acids in a protein database.
- pulling a card from the deck

Sample spaces, sample sets and events

The ***sample space*** of a random experiment is a set S that includes all possible outcomes of the experiment.

For example, if the experiment is to throw a die and record the outcome, the sample space is S = { 1,2,3,4,5,6}

- Discrete sample spaces.
- Continuous sample spaces

Discrete Random Variables

- A discrete random variable is one which may take on only a countable number of distinct values such as 0,

  1, 2, 3, 4,....

- • Discrete random variables are usually (but not necessarily) counts.

Examples:

- number of children in a family
- the Friday night attendance at a cinema
- the number of patients a doctor sees in one day
- the number of defective light bulbs in a box of ten
- the number of “heads” flipped in 3 trials

Continuous Random Variable

- A continuous random variable is one which takes an infinite number of possible values.

- Examples:
  - ✓ height
  - ✓ weight
  - ✓ the amount of sugar in an orange
  - ✓ the time required to run a mile.

Event

An **event** is a subset of the sample space of a random experiment.

An event is a set of outcomes of the experiment. This includes the *null* (empty) set of outcomes and the set of *all* outcomes. Each time the experiment is run, a given event *A* either *occurs*, if the outcome of the experiment is an element of *A*, or *does not occur*, if the outcome of the experiment is not an element of *A*.

Complement of an Event

- The <u>complement</u> of event *A* is defined to be the event consisting of all sample points that are not in *A*.
- The complement of *A* is denoted by A^c.
- The <u>Venn diagram</u> below illustrates the concept of a complement.

Union of Two Events

- The <u>union</u> of events *A* and *B* is the event containing all sample points that are in *A* or *B* or both.
- The union is denoted by *A* U *B*
- The union of *A* and *B* is illustrated below.

Intersection of Two Events

- The <u>intersection</u> of events *A* and *B* is the set of all sample points that are in both *A* and *B*.
- The intersection of *A* and *B* is the area of overlap in the illustration below.

Mutually Exclusive Events

- Two events are said to be <u>mutually exclusive</u> if the events have no sample points in common. That is, two events are mutually exclusive if, when one event occurs, the other cannot occur.

Axioms of Probability

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:

If S is the sample space and E is any event in a random experiment,

(1) $P(S) = 1$

(2) $0 \leq P(E) \leq 1$

(3) For two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Probability as a Numerical Measure of the Likelihood of Occurrence

Increasing Likelihood of Occurrence

Probability: 0 .5 1

The occurrence of the event is just as likely as it is unlikely.

THE ADDITION RULE

- The probability that event *A* or *B* will occur is given by

  $$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

- If events *A* and *B* are **mutually exclusive**, then the rule can be simplified to

  $$P(A \cup B) = P(A) + P(B).$$

Probability and Venn Diagram

$$P(A) = 1 - P(A^c)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Independent & Dependent

Events are either

- *Independent* (the occurrence of one event has no effect on the probability of occurrence of the other) or
- *Dependent* (the occurrence of one event gives information about the occurrence of the other)

Example

An experiment has the four possible mutually exclusive outcomes A, B, C, D. Check whether the following assignments of probability are permissible:

(a) P(A)= 0.38, P(B) = 0.16, P(C) =0.11, P(D) = 0.35 Permissible
(b) P(A)= 0.31, P(B) = 0.27, P(C) =0.28, P(D) = 0.16 NOT
(c) P(A)= 0.32, P(B) = 0.27, P(C) = -0.06, P(D) = 0.47 NOT
(d) P(A)= 1/2, P(B) = 1/4, P(C) = 1/8, P(D) = 1/16 NOT
(e) P(A)= 5/8, P(B) = 1/6, P(C) = 1/3, P(D) = 2/9 NOT

EXAMPLE: 1

If two dice are thrown , what is the probability that the sum is

a) Greater than 8

b) Less than 6

c) Neither 7 nor 11

Solution:

$$P(X = 2) = P\{(1, 1)\} = \frac{1}{36}$$

$$P(X = 3) = P\{(1, 2), (2, 1)\} = \frac{2}{36}$$

$$P(X = 4) = P\{(1, 3), (2, 2), (3, 1)\} = \frac{3}{36}$$

$$P(X = 5) = P\{(1, 4), (2, 3), (3, 2), (4, 1)\} = \frac{4}{36}$$

$$P(X = 6) = P\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} = \frac{5}{36}$$

$$P(X = 7) = P\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\} = \frac{6}{36}$$

$$P(X = 8) = P\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\} = \frac{5}{36}$$

$$P(X = 9) = P\{(3, 6), (4, 5), (5, 4), (6, 3)\} = \frac{4}{36}$$

$$P(X = 10) = P\{(4, 6)\ (5, 5), (6, 4)\} = \frac{3}{36}$$

$$P(X = 11) = P\{(5, 6), (6, 5)\} = \frac{2}{36}$$

$$P(X = 12) = P\{(6, 6)\} = \frac{1}{36}$$

If two dice are thrown , what is the probability that the sum is

a) Greater than 8

b) Less than 6

c) Neither 7 nor 11

a) $P(\text{sum} > 8)$

$= P(9) + P(10) + P(11) + P(12)$

$= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{10}{36}$

b) $P(\text{sum} < 6)$

$= P(5) + P(4) + P(3) + P(2)$

$= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{10}{36}$

If two dice are thrown , what is the probability that the sum is

a) Greater than 8

b) Less than 6

c) Neither 7 nor 11

$$P(\text{neither } 7 \text{ nor } 11)$$

$$= 1 - [P(7) + P(11)]$$

$$= 1 - \left[\frac{6}{36} + \frac{2}{36}\right]$$

$$= 28/36 = 7/9$$

Example:2

The probability that a student passes in statistics examination is 2/3 and the probability that he /she will not pass in mathematics examination is 5/9. The probability that he/she will pass in at least one of the examination is 4/5. Find the probability that he /she will pass in both the examinations

Solution:

The probability of passing in statistics P(S)=2/3

The probability of passing in Mathematics P(M)=1-5/9 =4/9

Probability of Passing at least one of these examination

= P(S∪M) = 4/5

$$\boldsymbol{P(S \cup M) = P(S) + P(M) - P(S \cap M).}$$

$$\frac{4}{5} = \frac{2}{3} + \frac{4}{9} - P(S \cap M).$$

$$P(S \cap M) = \frac{2}{3} + \frac{4}{9} - \frac{4}{5} = \frac{14}{45}.$$

The probability of passing both examinations $P(S \cap M) = \frac{14}{45}$

Example: 3

Suppose that 75% of all investors invest in traditional annuities and 45% of them invest in the stock market. If 85% invest in the stock market and/or traditional annuities, what percentage invest in both?

Solution:

Let A be the event that a randomly selected investor invests in traditional annuities.

Let B be the event that he or she invests in the stock market.

Then $P(A) = 0.75$, $P(B) = 0.45$, and $P(A \cup B) = 0.85$

Since,

$$P(AB) = P(A) + P(B) - P(A \cup B)$$
$$= 0.75 + 0.45 - 0.85$$
$$= 0.35.$$

Example: 4

Suppose the manufacturer's specifications for the length of a certain type of computer cable are 2000 ± 10 millimeters. In this industry, it is known that small cable is just as likely to be defective (not meeting specifications) as large cable. That is, the probability of randomly producing a cable with length exceeding 2010 millimeters is equal to the probability of producing a cable with length smaller than 1990 millimeters. The probability that the production procedure meets specifications is known to be 0.99.

(a) What is the probability that a cable selected randomly is too large?

(b) What is the probability that a randomly selected cable is larger than 1990 millimeters?

Solution:

***Solution*:** Let M be the event that a cable meets specifications. Let S and L be the events that the cable is too small and too large, respectively. Then

(a) $P(M) = 0.99$ and $P(S) = P(L) = (1 - 0.99)/2 = 0.005$.

(b) Denoting by X the length of a randomly selected cable, we have

$$P(1990 \leq X \leq 2010) = P(M) = 0.99.$$

Since $P(X \geq 2010) = P(L) = 0.005$,

$$P(X \geq 1990) = P(M) + P(L) = 0.995.$$

This also can be solved by using Theorem 2.9:

$$P(X \geq 1990) + P(X < 1990) = 1.$$

Thus, $P(X \geq 1990) = 1 - P(S) = 1 - 0.005 = 0.995$.

Exercise problems:

Q 1: A Survey conducted by a bank revealed that 40% of the accounts are savings accounts and 35% of the accounts are current accounts and the balance are loan accounts.

- What is the probability that an account taken at random is a loan account ?
- What is the probability that an account taken at random is **NOT** savings account ?
- What is the probability that an account taken at random is **NOT** a current account
- What is the probability that an account taken at random is a current account or a loan account?

Exercise problems

Q 2. A speaks truth in 80% cases and B speaks in 60% cases. What percentage of cases are they likely to contradict each other in stating the same fact.

Q 3. In a certain residential suburb, 60% of all households get Internet service from the local cable company, 80% get television service from that company, and 50% get both services from that company. If a household is randomly selected,

i) What is the probability that it gets at least one of these two services from the company, and

ii) What is the probability that it gets exactly one of these services from the company?

Exercise problems

Q 4: The next generation of miniaturised wireless capsules with active locomotion will require two miniature electric motors to manoeuvre each capsule. Suppose 10 motors have been fabricated but that, in spite of test performed on the individual motors 2 will not operate satisfactorily when placed into capsule, to fabricate a new capsule,2 motors will be randomly selected(that is, each pair of motors has the same chance of being selected) find the probability that

a) Both motors will operate satisfactorily in the capsule.

a) One motor will operate satisfactorily and other will not.

Exercise problems

Q 5. Suppose that 55% of all adults regularly consume coffee, 45% regularly consume carbonated soda, and 70% regularly consume at least one of these two products.

i) What is the probability that a randomly selected adult regularly consumes both coffee and soda?

ii) What is the probability that a randomly selected adult doesn't regularly consume at least one of these two products?

Exercise problems

Q 6. Suppose a student is selected at random from 80 students where 30 are taking mathematics, 20 are taking chemistry and 10 are taking both. Find the probability ‘p’ that the student is taking Mathematics or chemistry?.

Q7: If A and B are events with $P(A \cup B) = 7/8$, $P(A \cap B) = ¼$ and $P(A') = 5/8$, find $P(A)$, $P(B)$ and $P(A \cap B')$**).**

Thanks

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani
Pilani Campus

Session 3: Conditional Probability, Independent events and, Total Probability

(Session 3: 3rd/4th June 2023)

IMP Note to Self

RECAPTIULATION:

- Introduction to Probability
- Review of Set Theory
- Counting Principles
- Definition of Probability
- Axioms of Probability
- Addition Rule of Probability

Contact Session 3

Contact Session 3: Module 2(Conditional Probability & Bayes theorem)

Contact Session	List of Topic Title	Reference
CS - 3	Introduction to conditional probability,indepents events, Total probability	T1 & T2
HW	Problems on conditional probability	T1 & T2
Lab		

CONTENTS:

- Conditional Probability
- Independent Events
- Total Probability

CONDITIONAL PROBABILITY

The probabilities assigned to various events depend on what is known about the experimental situation when the assignment is made. Subsequent to the initial assignment, partial information relevant to the outcome of the experiment may become available. Such information may cause us to revise some of our probability assignments

CONDITIONAL PROBABILITY

We examine how the information “an event B has occurred” affects the probability assigned to A.

For example, A might refer to an individual having a particular disease in the presence of certain symptoms. If a blood test is performed on the individual and the result is negative , then the probability of having the disease will change (it should decrease, but not usually to zero, since blood tests are not infallible). We will use the notation to represent the conditional probability of A given that the event B has occurred. B is the “conditioning event.”

CONDITIONAL PROBABILITY

Let A and B be two events in sample space. The **conditional probability** that event A occurs given that event B has occurred and it is denoted by

$$P\left(\frac{A}{B}\right) = \frac{P(A \text{ and } B)}{P(B)} \quad \text{OR} \quad P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

It can also written as $P(A \cap B) = P(B)\, P(A/B) \qquad P(B) \neq 0$

$= P(A)\, P(B/A) \qquad P(A) \neq 0$

Let A,B and C be three events in a sample space S,

then $P(A \cap B \cap C) = P(A)\, P(B/A)\, P(C/A \cap B)$

and it is called Multiplication Rule

CONDITIONAL PROBABILITY

Multiplication Rule

In general, $A_1, A_2, \ldots\ldots A_n$ are events in S, then

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\, P(A_2/A_1) P(A_3/A_1 \cap A_2) \ldots\ldots P(A_n/A_1 \cap A_2 \cap \ldots\ldots \cap A_{n-1})$$

Examples:1

In a certain college, 25% of the students failed Maths, 15% of the students failed chemistry and 10% of the students failed both maths and chemistry. A student is selected at random, find

a. If he failed chemistry, what is the probability that he failed Maths?

$$P(M) = 0.25, \quad P(C) = 0.15, \quad P(M \cap C) = 0.1$$

$$P\left(\frac{M}{C}\right) = \frac{P(M \cap C)}{P(C)} = \frac{0.1}{0.15} = 0.6667$$

b. If the failed maths, what is the probability he failed chemistry?

$$P\left(\frac{C}{M}\right) = \frac{P(C \cap M)}{P(M)} = \frac{0.1}{0.25} = 0.4$$

c. What is the probability that the student failed in Maths or chemistry?

Examples:2

In a housing colony, 70% of the houses are well planned and 60% of the houses are well planned and well built. Find the probability that an arbitrarily chosen house in this colony is well built given that it is well planned.

Solution:

Let A be the event that the house is well planned B be the event that the house is well built therefore

$$P(A) = 0.70, \qquad P(A \cap B) = 0.60$$

$$P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)} = \frac{0.60}{0.70} = 0.8571$$

Examples: 3

The probabilities of a regularly scheduled flight departs on time is 0.83, arrives on time is 0.82 & it departs and arrives on time is 0.78. Find the probability that a plane

(i) arrives on time given that it departed on time,

(ii) departed on time given that it has arrived on time and

(iii) find $\boldsymbol{P}\left(\frac{\boldsymbol{A}}{\overline{\boldsymbol{D}}}\right)$

Ans:

Let D and A be the events that the flight departs and arrives on time respectively. Then, $\boldsymbol{P\ (D) = 0.83, P(A) = 0.82}$ and $\boldsymbol{P\ (D \cap A) = 0.78}$

(i) Probability that the plane arrives on time given that it departed on time is

$$\boldsymbol{P\left(\frac{A}{D}\right) = \frac{P(D \cap A)}{P(D)} = \frac{0.78}{0.83} = 0.9398}$$

(ii) Probability that the plane departed on time given that it has arrived on time is $\boldsymbol{P\left(\frac{D}{A}\right) = \frac{P(D \cap A)}{P(A)} = \frac{0.78}{0.82} = 0.9512}$

(iii) $\boldsymbol{P\ \left(\frac{A}{\bar{D}}\right) = \frac{P(A \cap \bar{D})}{P(\bar{D})} = \frac{0.82-0.78}{1-0.83} = 0.24}$

This is the probability that the flight arrives on time given that it did not depart on time

Example: 4

In a housing colony, 70% of the houses are well planned and 60% of the houses are well planned and well built. Find the probability that an arbitrarily chosen house in this colony is well built given that it is well planned.

Solution: Let A be the event that the house is well planned

B be the event that the house is well built

therefore $P(A) = 0.70, \quad P(A \cap B) = 0.60$

$$P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)} = \frac{0.60}{0.70} = 0.8571$$

BITS Pilani, Pilani Campus

Example: 5

A news magazine publishes three columns entitled "Art" (A), "Books" (B), and "Cinema" (C). Reading habits of a randomly selected reader with respect to these columns are

Read regularly	A	B	C	$A \cap B$	$A \cap C$	$B \cap C$	$A \cap B \cap C$
Probability	.14	.23	.37	.08	.09	.13	.05

Find

A). P(A/B)

B). P(A/BUC)

C). P(A/ reads at least one)

Solution

We thus have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.08}{.23} = .348$$

$$P(A|B \cup C) = \frac{P(A \cap (B \cup C))}{P(B \cup C)} = \frac{.04 + .05 + .03}{.47} = \frac{.12}{.47} = .255$$

$$P(A|\text{reads at least one}) = P(A|A \cup B \cup C) = \frac{P(A \cap (A \cup B \cup C))}{P(A \cup B \cup C)}$$

$$= \frac{P(A)}{P(A \cup B \cup C)} = \frac{.14}{.49} = .286$$

and

$$P(A \cup B|C) = \frac{P((A \cup B) \cap C)}{P(C)} = \frac{.04 + .05 + .08}{.37} = .459$$

INDEPENDENT EVENTS

We can deduce an important result from the conditional probability:

If B has no effect on A, then, $\boldsymbol{P\left(\frac{A}{B}\right) = P(A)}$ Also $\boldsymbol{P\left(\frac{B}{A}\right) = P(B)}$

and we say the events are independent.

i.e., The probability of A does not depend on B.

so, $$\boldsymbol{P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}}$$

becomes, $$\boldsymbol{P(A) = \frac{P(A \cap B)}{P(B)}}$$

or $$\boxed{P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}) \times P(\mathrm{B})}$$

Example: 1

A problem in statistics is given to 3 students A,B,C. Their chances of solving it are 1/2,1/3/,1/4. Find the probability that the problem is solved.

Solution: Problem can be solved by either A or B or C

Therefore we have to use

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

Using complement of the event i.e the problem is not solved . Therefore P(problem solved is)=1-P(not solved)

$$= 1 - P(\overline{A} \cap \overline{B} \cap \overline{C})$$

$$= 1 - P(\overline{A})\, P(\overline{B})\, P(\overline{C})$$

$$= 1 - \left[1 - \frac{1}{2}\right]\left[1 - \frac{1}{3}\right]\left[1 - \frac{1}{4}\right]$$

$$= \frac{1}{4}$$

Note: If A, B,C are independent then $\overline{A}, \overline{B}, \overline{C}$ are also independent.

Example: 2

A box contains 20 fuses of which 5 are defective. If two fuses are choosen at random one after the other. What is probability that both the fuses are defective if

(i) the first fuse is replaced,

(ii) (ii) the first fuse is not replaced.

Solution: Let A be the event that the first fuse is defective and
B be the event that the second fuse is defective

(i) When the first fuse is replaced, the events are independent hence

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}) \times P(\mathrm{B}) = \frac{5C_1}{20C_1} \times \frac{5C_1}{20C_1} = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

(ii) When first fuse is not replaced, the events are not independent then

$$P(B \cap A) = \mathrm{P}(\mathrm{A}) \times P\left(\frac{B}{A}\right) = \frac{5C_1}{20C_1} \times \frac{4C_1}{19C_1} = \frac{1}{19}$$

Miscellaneous examples

1. A box A contains 2 white balls and 4 black balls. Another box B contains 5 white balls and 7 black balls. A ball is transferred from box A to box B, than a ball is drawn from box B. What is the probability that it is white.

2. A person X can hit a target three times in five shots, a second person Y twice in five shots and a third person Z thrice in four shots. They all fire once simultaneously. Find the probability that (i) atleast two shots hit, (ii) exactly two shots hit.

Continued

3. If A and B are two events with P(A)= ½, P(B)= 1/3 and $P(A \cap B) = 1/4$
Find

$$P\left(\frac{A}{B}\right), P\left(\frac{B}{A}\right), P\left(\frac{\overline{A}}{\overline{B}}\right), P\left(\frac{\overline{B}}{\overline{A}}\right), P\left(\frac{A}{\overline{B}}\right)$$

4.Three students X,Y,Z write an examination. Their chances of passing are $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{4}$ respectively. Find the probability that

(i) all of them pass,

(ii) at least one of them passes,

(iii) at least two of them pass.

Continued

5. The probability that Ram will hit a target is 0.56 and the probability that shyam will hit the target is 0.45. The probability that both will hit the target is 0.25. find the probability that

(i) at least one will hit the target

(ii) exactly one will hit the target.

6. If A and B are two events such that

$$P(A \cap B) = 0.05, P(B) = 0.7, P(A \cap \overline{B}) = 0.02$$

Find (i) $P(A \cup B)$, (ii) $P(A \cup \overline{B})$

Total Probability

The Law of Total Probability

Let $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B,

$$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k)$$
$$= \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

$S = A_1 \cup A_2 \cup \ldots \cup A_n$ and $A_1 \cap A_2 \cap \ldots \cap A_n = \emptyset$

$$\therefore \boldsymbol{B} = \boldsymbol{B} \cap \boldsymbol{S} = \boldsymbol{B} \cap \{\boldsymbol{A}_1 \cup \boldsymbol{A}_2 \cup \boldsymbol{A}_3, \ . \quad . \quad . \cup \boldsymbol{A}_n\}$$

Proof:

We have $S = \{A_1 \cup A_2 \cup A_3 \cup \ . \ \ . \ \ . \cup A_n\}$ and $A \subset S$

$$\therefore B = B \cap S = B \cap \{A_1 \cup A_2 \cup A_3, \ . \ \ . \ \ . \cup A_n\}$$

Using distributive law in the R.H.S, we get

Since $B \cap A_i$ (i = 1 to n) are mutually exclusive, we have by applying addition rule of probability,

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots\cdots + P(B \cap A_n)$$

$$i.e., \ P(B) = \sum_{i=1}^{i=n} P(B \cap A_i)$$

Using multiplication rule on each term on R.H.S, namely

$$P(B \cap A_i) = P(A_i) \cdot P(B \mid A_i) \qquad \text{(A)}$$

$$P(B) = \sum_{i=1}^{i=n} P(A_i) P(B \mid A_i)$$

(B)-Total theorem on Probability

The Theorem of Total Probability
(tree diagram)

Examples :1

A certain firm has plants A, B, C producing, respectively 35 %, 15% and 50% of the total output. The probabilities of a non – defective product are, respectively, 0.75, 0.95 and 0.85.A Customer receives a bad product, what is the Chance that product came from the plant C?

Tree Diagram

Solution

Let X: " Customer receives a defective product".

Clearly, P(X)= $P(A)P\left[\frac{X}{A}\right]+P(B)P\left[\frac{X}{B}\right]+P(C)P\left[\frac{X}{C}\right]$

=0.17

Therefore, the chance that product is manufactured by the plant C is

$$P(C \mid X)=\frac{P(C \cap X)}{P(X)} = \frac{0.5 \cdot 0.15}{0.17} = 0.4412$$

Example 2

The chances that an academician, a business man and a politician becoming Vice Chancellor of an university are 0.5, 0.3 and 0.2 respectively. The probability that research work will be promoted in the university by these 3 gentlemen are respectively are 0.8, 0.6 and 0.4. It is found Research work has been promoted by the university. What is the chance that an academician has become the VC?

Tree Diagram

Solution

Let **X: " Research work is promoted"**

Clearly, $P(X) = 0.5 \times 0.8 + 0.3 \times 0.6 + 0.2 \times 0.4 = 0.66$

Now to find P[" An Academician is VC" / "Research work is promoted i.e. event X"] $= \frac{0.5 \times 0.8}{0.66} = 0.6061$

Practice problems

Example 1.

An office has 4 secretaries handling respectively 20%, 60%,15% and 5% of the files of all government reports. The probability that they misfile such reports are respectively 0.05, 0.1, 0.1 and 0.05. Find the probability that the misfiled report can be blamed on the first secretary.

Example 2.

In a class 70% are boys and 30% are girls. 5% of boys and 3% of girls are irregular to the classes. What is the probability of a student selected at random is irregular to the classes and what is the probability that the irregular student is a girl?

Continued

Example 3.

Three machines A, B and C produce respectively 60%, 30% and 10% of the total number of items of a factory. The percentage of defective outputs of these machines are 2%, 3% and 4%. An item is selected at random and is found to be defective.

(i) Find the probability that the item was produced by machine C?

(ii) What is the probability that the item was produced by machine C or B?

Example: 4

A card is randomly drawn from an incomplete deck of cards from which the ace of diamonds is missing.

1. What is the probability that the card is “clubs”?
2. What is the probability that the card is a “queen”?
3. Are the events “clubs” and “queen” independent

Example: 5

In a group of children from primary school there are 18 girls and 15 boys. Of the girls, 9 have had measles. Of the boys, 6 have had measles.

1. What is the probability that a randomly chosen child from this group has had measles?

2. If we randomly choose one person from the group of 18 girls, what is the probability that this girl has had measles?

3. Are the events “boy” and “measles” in this example independent?

IMP Note to Self

Thank you

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani
Pilani Campus

Session 4: Bayes theorem and Naïve Bayes theorem

(Session 4: 10th/11th June 2023)

IMP Note to Self

RECAPTIULATION:

- Introduction to Probability
- Review of Set Theory
- Counting Principles
- Definition of Probability,
- Axioms of Probability,
- Addition Rule of Probability
- Conditional Probability,
- Independent events and,
- Total Probability

Contact Session 4

Contact Session	List of Topic Title	Reference
CS - 4	Bayes theorem(with proof),Introduction to Naïve Bayes concept.	T1 & T2
HW	Problems on Bayes theorem	T1 & T2
Lab	Bayes theorem & Naïve Bayes Concept	Lab 2

Agenda

- Bayes Theorem
- Introduction to Naïve Bayes concept

Text Books

No	Author(s), Title, Edition, Publishing House
T1	Statistics for Data Scientsists, An introduction to probability, statistics and Data Analysis, Maurits Kaptein et al, Springer 2022
T2	Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning

BAYES THEOREM:

Let $P = \{A_1,\ A_2,\ A_3,\ \ldots\ , A_k\}$ be a set of exhaustive and mutually exclusive events of a sample space S with $P(A_i) \neq 0$ For each i. If B is any other event associated with A_i with $P(B) \neq 0$, then

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{P(B)}$$

$$= \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{\sum_{i=1}^{n} P(A_i)P\left(\frac{B}{A_i}\right)}$$

Geometrical Representation of Bayes theorem

$S = A_1 \cup A_2 \cup \ldots \cup A_n$ and $A_1 \cap A_2 \cap \ldots \cap A_n = \emptyset$

$$\therefore B = B \cap S = B \cap \{A_1 \cup A_2 \cup A_3, \ . \quad . \quad . \cup A_n\}$$

Proof:

The conditional Probability of A_i for any i given B is given as

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$$

But we know that

$$P(A_i \cap B) = P(A_i) \cdot P(B \mid A_i) \;\&\; P(B) = \sum_{i=1}^{i=n} P(A_i) P(B \mid A_i)$$

$$P(A_i \mid B) = \frac{P(A_i) \cdot P(B \mid A_i)}{\sum_{i=1}^{i=n} P(A_i) P(B \mid A_i)}$$

Hence the theorem proved

Example: 1

Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time, whereas an individual without the disease will show a positive test result only 2% of the time. If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

Solution:

To use Bayes' theorem, let A_1 = individual has the disease, A_2 = individual does not have the disease, and B = positive test result. Then $P(A_1) = .001$, $P(A_2) = .999$, $P(B|A_1) = .99$, and $P(B|A_2) = .02$. The tree diagram for this problem is in Figure

Next to each branch corresponding to a positive test result, the multiplication rule yields the recorded probabilities. Therefore, $P(B) = .00099 + .01998 = .02097$, from which we have

$$P(A_1 | B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.00099}{.02097} = .047$$

Example: 2

A certain firm has plants A, B, C producing, respectively 35 %, 15% and 50% of the total output. The probabilities of a non – defective product are, respectively, 0.75, 0.95 and 0.85.A Customer receives a bad product, what is the Chance that product came from the plant C?

Tree Diagram

Solution

Let X: " Customer receives a defective product".

Clearly, P(X)= $P(A)P\left[\frac{X}{A}\right]+P(B)P\left[\frac{X}{B}\right]+P(C)P\left[\frac{X}{C}\right]$

=0.17

Therefore, the chance that product is manufactured by the plant C is

$$P(C\mid X)=\frac{P(C\cap X)}{P(X)} = \frac{0.5\cdot 0.15}{0.17} = 0.4412$$

Example: 2

The chances that an academician, a business man and a politician becoming Vice Chancellor of an university are 0.5, 0.3 and 0.2 respectively. The probability that research work will be promoted in the university by these 3 gentlemen are respectively are 0.8, 0.6 and 0.4. It is found Research work has been promoted by the university. What is the chance that an academician has become the VC?

Tree Diagram

Let **X: " Research work is promoted"**

Clearly, $P(X) = 0.5 \times 0.8 + 0.3 \times 0.6 + 0.2 \times 0.4 = 0.66$

Now to find P[" An Academician is VC" / "Research work is promoted i.e. event X"] $= \frac{0.5 \times 0.8}{0.66} = 0.6061$

Example: A manufacturer of tablets receives its LED screens from three different suppliers, 60% from supplier $B1$, 30% from supplier $B2$, and 10% from supplier $B3$. In other words, the probabilities that any one LED screens received by the plant comes from these three suppliers are 0.60, 0.30, and 0.10. Also suppose that 95% of the LED screens from $B1$, 80% of those from $B2$, and 65% of those from $B3$ perform according to specifications.

1) What is the probability that any one LED screen received by the plant will perform according to specifications?

2) Determine the probability that a particular LED screen, which is known to perform according to specifications, came from supplier $B3$.

If A denotes the event that a LED screen received by the plant performs according to specifications, and $B1$, $B2$, and $B3$ are the events that it comes from the respective suppliers

$$A = A \cap [B1 \cup B2 \cup B3]$$
$$= (A \cap B1) \cup (A \cap B2) \cup (A \cap B3)$$
$$= P(A \cap B1) + P(A \cap B2) + P(A \cap B3)$$
$$P(A) = P(B1) \cdot P(A \mid B1) + P(B2) \cdot P(A \mid B2) + P(B3) \cdot P(A \mid B3)$$
$$P(A) = (0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65) = 0.875$$

$$P(B3 \mid A) = \frac{(0.10)(0.65)}{(0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65)} = 0.074$$

Example: 3

Example: Two firms V and W consider bidding on a road-building job, which may or may not be awarded depending on the amounts of the bids. Firm V submits a bid and the probability is ¾ that it will get the job provided firm W does not bid. The probability is ¾ that W will bid, and if it does, the probability that V will get the job is only1/3. (a) what is the probability that V will get the job? (b) If V gets the job, what is the probability that W did not bid?

Solution:

Answer: Given $P(V/W^1)=3/4$, $P(W) = 1/3$, $P(V/W)=1/3$ $P(W^1) = 2/3$

(a) $V = (V \cap W) \cup (V \cap W^1) \quad => P(V) = P(V \cap W) + P(V \cap W^1)$

$$P(V) = P(V/W)\,P(W) + P(V/W^1)\,P(W^1) = 11/18$$

(b) $P(W^1/V) = P(V/W^1)\,P(W^1) / P(V) = 9/11$

Suggested problems:

Example . An office has 4 secretaries handling respectively 20%, 60%,15% and 5% of the files of all government reports. The probability that they misfile such reports are respectively 0.05, 0.1, 0.1 and 0.05. Find the probability that the misfiled report can be blamed on the first secretary.

Example . In a class 70% are boys and 30% are girls. 5% of boys and 3% of girls are irregular to the classes. What is the probability of a student selected at random is irregular to the classes and what is the probability that the irregular student is a girl?

Example:

Three machines A, B and C produce respectively 60%, 30% and 10% of the total number of items of a factory. The percentage of defective outputs of these machines are 2%, 3% and 4%. An item is selected at random and is found to be defective. (i) Find the probability that the item was produced by machine C? (ii) What is the probability that the item was produced by machine C or B?

Bayesian Learning

- Naive Bayes is a set of simple and efficient machine learning algorithms for solving a variety of classification and regression problems.
- Naive Bayes assumes conditional independence where Bayes theorem does not. This means the relationship between all input features are independent.
- Bayesian learning algorithms that calculate explicit probabilities for hypotheses, such as the naive Bayes classifier, are among the most practical approaches to certain types of learning problems
- For example: Problem of learning to classify text documents such as electronic news articles.
- For such learning tasks, the naive Bayes classifier is among the most effective algorithms known

Features of Bayesian learning

- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.
- Prior knowledge is provided by asserting
  - prior probability for each candidate hypothesis, and
  - probability distribution over observed data for each possible hypothesis.
- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.

Bayes Theorem

- $P(h)$ = prior probability of hypothesis h, before seeing the training data
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Machine Learning

- Generative models
  - Build model to estimate the posterior probability P(Y|X) by estimating
  - likelihood of data given target (hypothesis) P(X|Y)
  - Prior probabilities over target P(Y)
  - In general, for a specific class Y=ck,

$$P(Y = c_k|X) = \frac{P(X|Y = c_k) * P(Y=c_k)}{P(X)}$$

Choosing Hypotheses

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Generally want the most probable hypothesis given the training data
- *Maximum a posteriori* hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

- If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

Brute Force MAP Hypothesis

- For each hypothesis *h* in *H*, calculate the posterior probability

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

MAP Hypothesis

- Using Bayes theorem, we compute the MAP hypothesis for all probable hypothesis (or all unique class labels)
- Identify the best hypothesis describing the data as

$$h_{MAP} = argmax_{h \in H} P(h|D)$$

$$= argmax_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

$$= argmax_{h \in H} P(D|h)P(h)$$

H: set of all hypothesis

P(D) is independent of h and is same for all hypothesis, therefore dropped

Maximum Likelihood Estimation

- When no prior information is available, all hypothesis are equally likely i.e p(hi) = p(hj)

- This is also true for a balanced class problem where all the classes are equally likely

- This is known as uniform prior
- MAP hypothesis further simplified to
- h_{ML} = argmax P(D/h) (where h belongs to H)

Conditional independence

- **Definition**: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k)\ P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z_k)$$

$$P(X|Y, Z) = P(X|Z)$$

Example:

$$P(\text{Thunder}|\text{Rain}, \text{Lightning}) = P(\text{Thunder}|\text{Lightning})$$

Applying conditional independence

Naïve Bayes assumes X_i are conditionally independent given Y

e.g., $P(X_1|X_2,Y) = P(X_1|Y)$

$$P(X_1,X_2|Y) = P(X_1|X_2,Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

General form: $P(X_1,\cdots,X_n|Y) = \prod_{j=1}^{n} P(X_j|Y)$

How many parameters to describe $P(X_1,\cdots,X_n|Y)$? $P(\mathrm{Y})$?

Without conditional independence assumption?

With conditional independence assumption?

Naïve Bayes Independence assumption

Assumption:

$$P(X_1, \cdots, X_n | Y) = \prod_{j=1}^{n} P(X_j | Y)$$

i.e., X_i and X_j are <u>conditionally independent</u>
given Y for $i \neq j$

Naïve Bayes classifier

- **Bayes rule:**

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) P(X_1, \cdots, X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1, \cdots, X_n | Y = y_j)}$$

- **Assume conditional independence among X_i's:**

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) \Pi_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \Pi_i P(X_i | Y = y_j)}$$

- **Pick the most probable (MAP) Y**

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Prior Probability

MLE

NAÏVE BAYES CLASSIFIER

- Assume independence among attributes X_i when class is given:
  - $P(X_1, X_2, \ldots, X_d \mid Y_j) = P(X_1 \mid Y_j)\, P(X_2 \mid Y_j) \ldots P(X_d \mid Y_j)$
  - Now we can estimate $P(X_i \mid Y_j)$ for all X_i and Y_j combinations from the training data
  - New point is classified to Y_j if $P(Y_j)\, \Pi\, P(X_i \mid Y_j)$ is maximal.

Example 1:

If the weather is sunny, then the player will play or not?

i.e. Play/ Sunny = Yes or No

Note if we know P(Yes/Sunny) and P(No/Sunny) then we can answer the question asked

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Steps to Apply Bayes Theorem

Step 1- View or collect "raw" data.

Step 2 - Convert long data to a frequency table

	Play		
	no	yes	Row Total
weather			
Sunny	2	3	5
Overcast	0	4	4
Rainy	3	2	5
Column Total	5	9	14

Step 3 - Row and column sums to get probabilities

Weather probabilities

sunny = 5/14 , rainy = 5/14

Overcast = 4/14

Play probabilities

no = 5/14

yes = 9/14

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Steps to Apply Bayes Theorem

	Play			
	no	yes	Row Total	
weather				
Sunny	2	3	5	P(Sunny)= 5/14
Overcast	0	4	4	P(Overcast) = 4/14
Rainy	3	2	5	P(Rainy)=5/14
Column Total	5	9	14	
	P(no)=5/14	P(yes)=9/14		

Step 4 - Apply probabilities from frequency table to Bayes theorem

$$P(yes \mid sunny) = \frac{P(sunny \mid yes)\, P(yes)}{P(sunny)}$$

Weather	No	Yes	
Rainy	3	2	$5/14 = 0.36$
Sunny	2	3	$5/14 = 0.36$
overcast	0	4	$4/14 = 0.29$
Total	5	9	
	$5/14 = 0.36$	$9/14 = 0.64$	

Now $P(\text{Yes} \mid \text{sunny}) = \dfrac{P(\text{sunny} \mid \text{Yes})\, P(\text{Yes})}{P(\text{sunny})}$

$$= \frac{(3/9)(9/14)}{5/14} = \underline{0.60}$$

$$P(\text{No} \mid \text{sunny}) = \frac{(2/5)(5/14)}{5/14} = \underline{0.40}$$

Example 2:

If the features of today = (Outlook is Sunny, Temp is Hot, Humidity is Normal, Windy is False), then the player will play or not?

S. No	Outlook	Temp	Humidity	Windy	Play Tennis
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

today = (Sunny, Hot, Normal, False)

$X \quad X_1 \quad X_2 \quad X_3 \quad X_4$

yes
no

$$P(yes \mid X) = \frac{P(X \mid yes)\, P(yes)}{P(X)}$$

$$\cancel{P(no \mid X)}$$

$$= \frac{P(X_1, X_2, X_3, X_4 \mid yes)\, yes}{P(X_1, X_2, X_3, X_4)} = \frac{P(X_1 \mid Y)\, P(X_2 \mid Y)\, P(X_3 \mid Y)\, P(X_4 \mid Y)\, P(Y)}{P(X_1)\, P(X_2)\, P(X_3)\, P(X_4)}$$

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

Example 3:

Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex
Yes	No	No	No	Male
Yes	Yes	Yes	Yes	Female
No	No	No	No	Male
Yes	Yes	Yes	Yes	Male
Yes	No	Yes	No	Female
No	No	No	No	Female
Yes	Yes	Yes	Yes	Male
No	No	No	No	Male
Yes	No	No	No	Male
Yes	Yes	Yes	No	Female

New Instance: Magazine Promotion = Yes , Watch Promotion = Yes, Life Insurance Promotion = No, Credit Card Insurance = No then Sex = ?

D = { magazine promotion, watch promotion, Life Insurance promotion, credit card insurance }

h_i = male or Female

P (male / Yes, yes, NO, NO) = ?

P (Female / yes, Yes, NO, NO) = ?

	Magazine Promotion		Watch Promotion		L.I Promotion		Credit Card Promotion	
	Male	Female	Male	Female	Male	F	M	F
YES	4	3	2	2	2	3	2	1
NO	2	1	4	2	4	1	4	3
Ratios (YES)	$4/6$	$3/4$	$2/6$	$2/4$	$2/6$	$3/4$	$2/6$	$1/4$
NO	$2/6$	$1/4$	$4/6$	$2/4$	$4/6$	$1/4$	$4/6$	$3/4$

$$P(\text{male} \mid E)$$

$$= \frac{P(E \mid \text{male})\, P(\text{male})}{P(E)}$$

$$= \frac{\left(\frac{4}{6} \cdot \frac{2}{6} \cdot \frac{4}{6} \cdot \frac{4}{6}\right)\left(\frac{3}{5}\right)}{P(E)}$$

$$= \frac{0.0593}{P(E)}$$

YES

YES

NO

NO

$\frac{6}{10} \cdot \frac{3}{5}$

$$P(\text{Female} / E)$$

$$= \frac{P(E/\text{Female})\, P(\text{Female})}{P(E)}$$

$$= \frac{\left(\frac{3}{4}\right)\left(\frac{2}{4}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) \cdot \frac{2}{5}}{P(E)}$$

$$= \frac{\left(\frac{9}{128}\right)\left(\frac{2}{5}\right)}{P(E)} = \frac{0.0281}{P(E)}$$

$0.0593 > 0.0281$

ie male ✓

Example 4:

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Issues with Naïve Bayes Classifier

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- P(Yes) = 3/10
  P(No) = 7/10

- P(Yes | Married) = 0 x 3/10 / P(Married)
  P(No | Married) = 4/7 x 7/10 / P(Married)

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

$P(X \mid No) = 2/6 \times 0 \times 0.0083 = 0$

$P(X \mid Yes) = 0 \times 1/3 \times 1.2 \times 10^{-9} = 0$

Naïve Bayes will not be able to classify X as Yes or No!

Naïve Bayes for Text Classification

- Naïve Bayes is commonly used for text classification
- For a document with k terms $d = (t_1, \ldots, t_k)$

Fraction of documents in c

$$P(c|d) = P(c)P(d|c) = P(c)\prod_{t_i \in d} P(t_i|c)$$

- $P(t_i|c)$= Fraction of terms from all documents in c that are t_i.

Number of times t_i appears in some document in c

$$P(t_i|c) = \frac{N_{ic} + 1}{N_c + T}$$

Laplace Smoothing

Number of unique words (vocabulary size)

Total number of terms in all documents in c

- Easy to implement and works relatively well
- Limitation: Hard to incorporate additional features (beyond words).
  - E.g., number of adjectives used.

A Simple Example:

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Which tag does the sentence *A very close game* belong to? i.e. P(sports| *A very close game*)

Feature Engineering: Bag of words i.e use word frequencies without considering order

Using Bayes Theorem:

P(sports| *A very close game*)

$$= \frac{P(\textit{A very close game} \mid \text{sports})\, P(\text{sports})}{P(\textit{A very close game})}$$

We assume that every word in a sentence is **independent** of the other ones

"close" doesn't appear in sentences of sports tag, So P(close | sports) = 0, which makes product 0

A Simple Example

Draw | Naive Bayes

Text classification

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Which tag does the sentence "*A very close game*" belong to? i.e. P(sports| *A very close game*)

Feature Engineering: Bag of words i.e use word frequencies without considering order

Using Bayes Theorem:

P(sports| *A very close game*)

$$= \frac{P(\textit{A very close game} \mid \text{sports})\, P(\text{sports})}{P(\textit{A very close game})}$$

We assume that every word in a sentence is **independent** of the other ones

P(A very close game) = P(A) P(very) P(close) P(game)

P(A very closed game | sports) = P(a | sports) P(very | sports) P(close | sports) P(game | sports)

"close" doesn't appear in sentences of sports tag, So P(close | sports) = 0, which makes product 0

$$\frac{P(\text{sports} \mid \text{A very close game})}{P(\text{not sports} \mid \text{A very close game})} = \frac{P(\text{A very close game} \mid \text{sports}) \cdot P(\text{sports})}{P(\text{A very close game})}$$

Laplace smoothing

- Laplace smoothing: we add 1 or in general constant k to every count so it's never zero.
- To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1
- In our case, the possible words are ['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match'].

Apply Laplace Smoothing

Word	P(word | Sports)	P(word | Not Sports)
a	2+1 / 11+14	1+1 / 9+14
very	1+1 / 11+14	0+1 / 9+14
close	0+1 / 11+14	1+1 / 9+14
game	2+1 / 11+14	0+1 / 9+14

$$P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports) \times P(Sports)$$
$$= 2.76 \times 10^{-5}$$
$$= 0.0000276$$

$$P(a|Not\,Sports) \times P(very|Not\,Sports) \times P(close|Not\,Sports) \times P(game|Not\,Sports) \times P(Not\,Sports)$$
$$= 0.572 \times 10^{-5}$$
$$= 0.00000572$$

Example :

Doc No	Text
1	I LOVED THE MOVIE
2	I HATED THE MOVIE
3	A GREAT MOVIE ,GOOD MOVIE
4	POOR ACTING
5	GREAT ACTING , A GOOD MOVIE
NEW	I HATED THE POOR ACTING

Example :

Doc No	Text	
1	I LOVED THE MOVIE	_POSITIVE
2	I HATED THE MOVIE	_ NEGATIVE
3	A GREAT MOVIE ,GOOD MOVIE	_POSITIVE
4	POOR ACTING	_NEGATIVE
5	GREAT ACTING , A GOOD MOVIE	_POSITIVE
NEW	I HATED THE POOR ACTING	_????

$P(c / x)$

$P(+ / \text{I hated the poor acting}) =$

$P(- / \text{I hated the poor acting}) =$

Based on these probabilities, we can decide the class which the new text belongs

$P(+ \mid I \text{ hated the acting})$

ie $P(c_1 \mid x) = \dfrac{P(x \mid c_1)\, P(c_1)}{\cancel{P(x)}}$

$= P(I \mid +)\, P(\text{hated} \mid +)\, P(\text{the} \mid +)\, P(\text{acting} \mid +)\, P(+)$

$\dfrac{P(I, +)}{P(+)}$ $\qquad$ $\dfrac{P(\text{hated}, +)}{P(+)}$

words	positive	negative
I	1	1
loved	1	0
the	1	1
movie	4	1
hated	0	1
a	2	0
great	2	0
Poor	0	1
acting	1	1
good	2	0

$$P(I \mid +) = \frac{1+1}{14+10}$$

$$P(I \mid -) = \frac{1+1}{6+10}$$

"I hated the poor acting"

word	positive	negative
I	$\frac{1+1}{14+10} = 0.0833$	$\frac{1+1}{6+10} = 0.125$
hated	$\frac{0+1}{14+10} = 0.0417$	$\frac{1+1}{6+10} = 0.125$
the	$\frac{1+1}{14+10} = 0.0833$	$\frac{1+1}{6+10} = 0.125$
poor	$\frac{0+1}{14+10} = 0.0417$	$\frac{1+1}{6+10} = 0.125$
acting	$\frac{1+1}{14+10} = 0.0833$	$\frac{1+1}{6+10} = 0.125$

x : I hate the Poor acting

$P(+/x)$

$= (\quad)(\quad)(\quad)(\quad)(\quad) \times P(+)$

$\downarrow$

$3/5$

$= 6.03 \times 10^{-7}$

$P(-/x)$

$= (\quad)(\quad)(\quad)(\quad)(\quad) P(-1)$:

$\downarrow 2/5$

$= 1.22 \times 10^{-5}$

$\therefore$ Negative class

Example:

Suppose we got the new message with the words '**Dear Friend**', Decide whether this new message is a normal or spam message?

i.e. Normal/ Dear, Friend = Yes or No

Note if we know P (Normal/ Dear, Friend) and P(Spam/ Dear, Friend) then we can answer the question asked

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem

Step 1- View or collect "raw" data.

Step 2 - Convert long data to a frequency table

	Play normal	spam	Row Total
word			
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

Step 3 - Row and column sums to get probabilities

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem

Play word	normal	spam	Row Total
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

As P (N/ D, F) > P (S/ D, F) , we can decide that Dear Friend is Normal message.

Step 4 - Apply probabilities from frequency table to Bayes theorem

$$P(N/D,F) = \frac{P(D,F/N).P(N)}{P(D,F)} = \frac{P(D/N).P(F/N).P(N)}{P(D).P(F)} = \frac{\left(\frac{8}{17}\right).\left(\frac{5}{17}\right).\left(\frac{17}{24}\right)}{\left(\frac{10}{24}\right).\left(\frac{6}{24}\right)} = \frac{0.098}{0.104} = 0.9423$$

$$P(S/D,F) = \frac{P(D,F/S).P(S)}{P(D,F)} = \frac{P(D/S).P(F/S).P(S)}{P(D).P(F)} = \frac{\left(\frac{2}{7}\right).\left(\frac{1}{7}\right).\left(\frac{7}{24}\right)}{\left(\frac{10}{24}\right).\left(\frac{6}{24}\right)} = \frac{0.012}{0.104} = 0.1153$$

Example continued:

Suppose we got the new message contains the word **'Lunch Money Money Money Money'** , Decide whether this new message is a normal or spam message?

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem

Play word	normal	spam	Row Total
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

We can observe that we have to classify any message with Lunch as Normal message, no matter how many times we see the word Money and that's the problem.

To work around this problem add 1 count to the frequency table to each word(Laplace smoothing)

Step 4 - Apply probabilities from frequency table

$$P(N).\ P(L/N).\ P(M/N)^4 = \left(\frac{17}{24}\right).\left(\frac{3}{17}\right).\left(\frac{1}{17}\right)^4 = 0.0000015$$

$$P(S).\ P(L/S).\ P(M/S)^4 = \left(\frac{7}{24}\right).\left(\frac{0}{7}\right).\left(\frac{4}{7}\right)^4 = 0$$

Steps to Apply Bayes Theorem

Play

word	normal	spam	Row Total
Dear	8+1	2+1	12
Friend	5+1	1+1	8
Lunch	3+1	0+1	5
Money	1+1	4+1	7
Column Total	21	11	32

Step 4 - Apply probabilities from frequency table

$$P(N).\ P(L/N).\ P(M/N)^4 = \left(\frac{21}{32}\right).\left(\frac{4}{21}\right).\left(\frac{2}{21}\right)^4 = 0.00001$$

$$P(S).\ P(L/S).\ P(M/S)^4 = \left(\frac{11}{32}\right).\left(\frac{1}{11}\right).\left(\frac{5}{11}\right)^4 = 0.00133$$

As $P(S/L, M^4) > P(N/L, M^4)$, we can decide that **Lunch Money Money Money Money is Spam** message.

Naïve Bayes Classifier Applications

Categorizing News

Email Spam Detection

Face Recognition

Sentiment Analysis

Naive Bayes Classifier

- ✓ Along with decision trees, neural networks, one of the most practical learning methods.
- ✓ When to use
  - ✓ Moderate or large training set available
  - ✓ Attributes that describe instances are conditionally independent given classification
- ✓ Successful applications:
  - ✓ Diagnosis
  - ✓ Classifying text documents

Learning to Classify Text

- Why?
  - Learn which news articles are of interest
  - Learn to classify web pages by topic

- Naive Bayes is among most effective algorithms
- What attributes shall we use to represent text documents??

Baseline: Bag of Words Approach

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	
...	
oil	
...	
Zaire	

Practical Issues of Bayesian learning

- Require initial knowledge of many probabilities
  - Often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
- Significant computational cost required to determine the Bayes optimal hypothesis in the general case (linear in the number of candidate hypotheses)

HW: Exercise

Consider the car theft problem with attributes Color, Type, Origin, and the target, Stolen can be either Yes or No.

we need to classify whether the car is stolen, given the features of the car.

Given the Red color Domestic SUV car Find the probability of whether the car is stolen?

Color	Type	Origin	Stolen?
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

HW: Exercise

If the weather is Snowy, then the player will play or not?

weather	Player play
Sunny	yes
Rainy	no
Cloudy	yes
Sunny	no
Sunny	yes
snowy	no
Rainy	yes
Cloudy	no
Cloudy	yes
Sunny	yes
snowy	no
Cloudy	yes
Rainy	no
snowy	no
snowy	yes

Thanks

BITS Pilani
Pilani Campus

M.Tech. (AIML)

Session-5 (Random Variables)

Team ISM

Session-5 Agenda

Random variables –

Discrete & continuous Expectation of a random variable, mean and variance of a random variable –

Single random random variable &

Joint distributions

Contact Session 5: Module 3: Probability Distributions

Contact Session	List of Topic Title	Reference
CS - 5	Random variables - Discrete & continuous Expectaion of a random variable,mean and variance of a random variable – Sinlge random random variable & Joint distributions	T1 & T2
HW	Problems on random variables	T1 & T2
Lab	Probability Distributions & Sampling	Lab 3

Random Variables

- A **random variable** is a variable that assumes numerical values associated with the random outcome of an experiment, where one (and only one) numerical value is assigned to each sample point.
- In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.

Random Variables

- A random variable can be classified as being either discrete or continuous depending on the numerical values it assumes.
- A <u>discrete random variable</u> may assume either finite or countably infinite number of values
- A <u>continuous random variable</u> may assume any numerical value in an interval or collection of intervals.
- Continuous random variables are generated in experiments where things are "measured' as opposed to "counted".
- Experimental outcomes based on measurement of time, distance, weight, volume etc. generate continuous RV.

Types of random Variables

- A **discrete random variable** can assume a countable number of values.
  - Number of steps to the top of the Eiffel Tower*

- A **continuous random variable** can assume any value along a given interval of a number line.
  - The time a tourist stays at the top once s/he gets there

Two Types of Random Variables

- **Discrete random variables**
  - Number of sales
  - Number of calls
  - Shares of stock
  - People in line
  - Mistakes per page

- **Continuous random variables**
  - Length
  - Depth
  - Volume
  - Time
  - Weight

Discrete Probability Distributions

- The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable.
- The probability distribution is defined by a probability function, denoted by $f(x)$, which provides the probability for each value of the random variable.

The required conditions for a discrete probability function are:

$$f(x) \geq 0$$

$$\Sigma f(x) = 1$$

- We can describe a discrete probability distribution with a table, graph, or equation.
- Advantage: once the probability distribution is known, it is relatively easy to determine the probability of a variety of events that may be of interest to the decision maker.

Probability Distributions for Discrete Random Variables

- Say a random variable x follows this pattern: $p(x) = (.3)(.7)^{x-1}$ for $x > 0$.
  - This table gives the probabilities (rounded to two digits) for x between 1 and 10.

x	$P(x)$
1	.30
2	.21
3	.15
4	.11
5	.07
6	.05
7	.04
8	.02
9	.02
10	.01

Expected Value and Variance

- The <u>expected value</u>, or mean, of a random variable is a measure of its central location.
  - The mean or Expected value of a discrete random variable:

$$E(x) = \mu = \Sigma x f(x)$$

- The <u>variance</u> summarizes the variability in the values of a random variable.
  - Variance of a discrete random variable:

$$\mathrm{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x)$$

- The <u>standard deviation</u>, σ, is defined as the positive square root of the variance.

Rules of Expected Value

- Multiplying RV by a constant a, $E(aX) = a.E(X)$
- Adding a constant b, $E(X+b) = E(X) + b$
- Therefore, $E(aX + b) = ?$

Variability of Discrete Random Variables

- The **variance** of a **discrete random variable** *x* is

$$\sigma^2 = E[(x-\mu)^2] = \sum (x-\mu)^2 p(x).$$

- The **standard deviation** of a **discrete random variable** *x* is

$$\sqrt{\sigma^2} = \sqrt{E[(x-\mu)^2]} = \sqrt{\sum (x-\mu)^2 p(x)}.$$

Rules of variability

- Multiplying RV by a constant a, $V(aX) = a^2 . V(X)$
- Adding a constant b, $V(X+b) = V(X)$
- $$\sigma_{aX} = |a| \cdot \sigma_X, \quad \sigma_{X+b} = \sigma_X$$

Example

At a shooting range, a shooter is able to hit a target in either 1, 2 or 3 shots. Let x be a random variable indicating the number of shots fired to hit the target. The following probability function was proposed.

$$f(x) = x/6$$

Is this probability function valid?

Identify the r.v to be discrete or continuous?

EXPERIMENT	Random Variable (x)
Audit 50 tax returns	Number of returns that contains error
Operate a restaurant for one day	Number of customers
Observe an employee's work	No. of productive hours in an 8-hour workday

Discrete Uniform Probability Distribution

- The discrete uniform **probability** distribution is the simplest example of a discrete probability distribution given by a formula.
  - The discrete uniform probability function is

$$f(x) = 1/n$$

where:

n = the number of values the random variable may assume

Note that the values of the random variable are equally likely.

Random Variables

Discrete

$X = \{1, 2, 3, 4\}$

→ No of students

→ No of bits transmitted

Continuous

$X \in (1, 3)$

→ pressure

→ Temperature

→ Time

→ Voltage

Random Variables

Discrete

$P(x)$

continuous

$f(x)$

<u>Validation</u> :-

1) $0 \leq P(x) \leq 1$

2) $\sum P(x) = 1$

1) $0 \leq f(x) \leq 1$

2) $\int f(x)\,dx = 1$

probability distribution function

probability density function

Example:

x	1	2	4	8	16
$P(x)$	0.05	0.10	0.35	0.40	0.10

compute the following

a) $E(X)$

b) $V(X)$ directly from the definition

c) standard deviation of X

d) $V(X)$ using the shortcut formula

Example

Let x be a random variable with PDF given by

$$f(x) = \begin{cases} cx^2, & -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

a) Find constant 'c'.

b) Find $E(x)$ and $V(x)$

c) Find $P(x \geq 1/2)$

Continuous Probability Distributions

- A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.
- It is not possible to talk about the probability of the random variable assuming a particular value.
- Instead, we talk about the probability of the random variable assuming a value within a given interval.
- The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the probability density function between x_1 and x_2.

Continuous Random Variables

A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.

The probability of the random variable assuming a value within some given interval from x1 to x2 is defined to be the area under the graph of the **probability density function** between x1 and x2

Example:

- Height of students in a class
- Amount of ice tea in a glass
- Change in temperature throughout a day
- Price of a car in next year

Continuous Random Variables

Probability Density Function

For a continuous random variable X, a **probability density function** is a function such that

(1) $f(x) \geq 0$

(2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

(3) $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx =$ area under $f(x)$ from a to b for any a and b (4.1)

Continuous Random Variables

Cumulative Distribution Function

The **cumulative distribution function** of a continuous random variable X is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\,du$$

for $-\infty < x < \infty$.

Probability Density Function from the Cumulative Distribution Function

Given $F(x)$,

$$f(x) = \frac{dF(x)}{dx}$$

as long as the derivative exists.

Continuous Random Variables

Mean and Variance

Suppose that X is a continuous random variable with probability density function $f(x)$. The **mean** or **expected value** of X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

The **variance** of X, denoted as $V(X)$ or σ^2, is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

The **standard deviation** of X is $\sigma = \sqrt{\sigma^2}$.

Integration Formulas

$$\int kf(u)du = k\int f(u)du$$

$$\int u^n du = \frac{u^{n+1}}{n+1}$$

$$\int e^u du = e^u$$

$$\int \sin u \, du = -\cos u$$

$$\int \cos u \, du = \sin u$$

$$\int [f(u) \pm g(u)]du = \int f(u)du \pm \int g(u)du$$

$$\int u dv = uv - \int v du$$

Continuous Random Variables

EXAMPLE 1 **Calculating probabilities from the probability density function**

If a random variable has the probability density

$$f(x) = \begin{cases} 2e^{-2x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

find the probabilities that it will take on a value

(a) between 1 and 3;

(b) greater than 0.5.

Solution Evaluating the necessary integrals, we get

(a)
$$\int_1^3 2e^{-2x}\,dx = e^{-2} - e^{-6} = 0.133$$

(b)
$$\int_{0.5}^{\infty} 2e^{-2x}\,dx = e^{-1} = 0.368$$

With reference to the preceding example, find the distribution function and use it to determine the probability that the random variable will take on a value less than or equal to 1.

Performing the necessary integrations, we get

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \displaystyle\int_0^x 2e^{-2t}\,dt = 1 - e^{-2x} & \text{for } x > 0 \end{cases}$$

and substitution of $x = 1$ yields

$$F(1) = 1 - e^{-2} = 0.865$$

Determining the mean and variance using the probability density function

With reference to Example 1, find the mean and the variance of the given probability density.

Performing the necessary integrations, using integrations by parts, we get

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx = \int_{0}^{\infty} x \cdot 2e^{-2x}\, dx = \frac{1}{2}$$

Alternatively, the expectation of x is $E(X) = 0.5$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \int_{0}^{\infty} \left(x - \frac{1}{2}\right)^2 \cdot 2e^{-2x}\, dx = \frac{1}{4}$$

A probability density function assigns probability one to $(-\infty, \infty)$

Find k so that the following can serve as the probability density of a random variable:

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ kxe^{-4x^2} & \text{for } x > 0 \end{cases}$$

Solution

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} kxe^{-4x^2}\,dx = \int_{0}^{\infty} \frac{k}{8} \cdot e^{-u}\,du = \frac{k}{8} = 1$$

so that $k = 8$.

Continuous Random Variables

5.4 If the probability density of a random variable is given by

$$f(x) = \begin{cases} x & \text{for } 0 < x < 1 \\ 2 - x & \text{for } 1 \leq x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

find the probabilities that a random variable having this probability density will take on a value

(a) between 0.2 and 0.8; (b) between 0.6 and 1.2.

5.14 Find μ and σ^2 for the probability density of Exer cise 5.4.

Continuous Random Variables

$$f(x) = \begin{cases} x & \text{for } 0 < x < 1 \\ 2 - x & \text{for } 1 \leq x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

Find μ and σ^2 for the probability density

Continuous Random Variables

5.6 Given the probability density $f(x) = \dfrac{k}{1+x^2}$ for $-\infty < x < \infty$, find k.

Continuous Random Variables

5.10 The length of satisfactory service (years) provided by a certain model of laptop computer is a random variable having the probability density

$$f(x) = \begin{cases} \dfrac{1}{4.5}\, e^{-x/4.5} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

Find the probabilities that one of these laptops will pro vide satisfactory service for

(a) at most 2.5 years; (b) anywhere from 4 to 6 years; (c) at least 6.75 years.

Continuous Random Variables

$$f(x) = \begin{cases} \dfrac{1}{4.5} e^{-x/4.5} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

(a) at most 2.5 years;

(b) anywhere from 4 to 6 years;

(c) at least 6.75 years.

The cumulative distribution function of the random variable Y is

$$F_Y(y) = \begin{cases} 0 & y < 0, \\ y/4 & 0 \le y \le 4, \\ 1 & y > 4. \end{cases}$$

Sketch the CDF of Y and calculate the following probabilities:

(1) $P[Y \le -1]$ *(2)* $P[Y \le 1]$

(3) $P[2 < Y \le 3]$ *(4)* $P[Y > 1.5]$

The CDF of Y is

From the CDF $F_Y(y)$, we can calculate the probabilities:

(1) $P[Y \le -1] = F_Y(-1) = 0$

(2) $P[Y \le 1] = F_Y(1) = 1/4$

(3) $P[2 < Y \le 3] = F_Y(3) - F_Y(2) = 3/4 - 2/4 = 1/4$

(4) $P[Y > 1.5] = 1 - P[Y \le 1.5] = 1 - F_Y(1.5) = 1 - (1.5)/4 = 5/8$

The probability density function of the random variable Y is

$$f_Y(y) = \begin{cases} 3y^2/2 & -1 \le y \le 1, \\ 0 & \textit{otherwise.} \end{cases}$$

Sketch the PDF and find the following:

(1) the expected value $E[Y]$

(2) the second moment $E[Y^2]$

(3) the variance $\mathrm{Var}[Y]$

(4) the standard deviation σ_Y

Recall - Continuous Random Variables

Properties:

$$f(x) \geq 0$$

$$\int_a^b f(x)\,dx = 1$$

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

$$f(x) = P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

$$f(c) = P(X = c) = P(c \leq X \leq c) = \int_c^c f(x)\,dx = 0$$

$$E(x) = \int_a^b x f(x)\,dx$$

$$E(x^2) = \int_a^b x^2 f(x)\,dx$$

$$V(X) = E(X^2) - ((E(X))^2$$

Exercise - Continuous Random Variables

5.7 If the distribution function of a random variable is given by

$$F(x) = \begin{cases} 1 - \dfrac{4}{x^2} & \text{for } x > 2 \\ 0 & \text{for } x \le 2 \end{cases}$$

find the probabilities that this random variable will take on a value

(a) less than 3; (b) between 4 and 5.

5.9 Let the phase error in a tracking device have probability density

$$f(x) = \begin{cases} \cos x & 0 < x < \pi/2 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability that the phase error is

(a) between 0 and $\pi/4$; (b) greater than $\pi/3$.

Find μ and σ for the distribution of the phase error

4.3.1 The random variable X has probability density function

$$f_X(x) = \begin{cases} cx & 0 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

Use the PDF to find

(a) the constant c,

(b) $P[0 \le X \le 1]$,

(c) $P[-1/2 \le X \le 1/2]$,

(d) the CDF $F_X(x)$.

4.4.4 The probability density function of random variable Y is

$$f_Y(y) = \begin{cases} y/2 & 0 \le y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

What are $E[Y]$ and $Var[Y]$?

4.4.5 The cumulative distribution function of the random variable Y is

$$F_Y(y) = \begin{cases} 0 & y < -1, \\ (y+1)/2 & -1 \le y \le 1, \\ 1 & y > 1. \end{cases}$$

What are $E[Y]$ and $Var[Y]$?

5.1 Joint Probability distribution

Introduction:

- Let $X = \{x_1, x_2, \dots, x_m\}$ and $Y = \{y_1, y_2, \dots, y_n\}$ be two discrete random variables. Then $P(x, y) = J_{ij}$ is called joint probability function of X and Y if it satisfies the conditions:

$$(i)\ J_{ij} \geq 0 \quad (ii) \sum_{i=1}^{m} \sum_{j=1}^{n} J_{ij} = 1$$

- Set of values of this joint probability function J_{ij} is called joint probability distribution of X and Y.

	y_1	y_2	$\dots$	y_n	Sum
x_1	J_{11}	J_{12}	$\dots$	J_{1n}	$f(x_1)$
x_2	J_{21}	J_{22}	$\dots$	J_{2n}	$f(x_2)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_m	J_{m1}	J_{m2}	$\dots$	J_{mn}	$f(x_m)$
Sum	$g(y_1)$	$g(y_2)$	$\dots$	$g(y_n)$	$Total = 1$

- So far we have been talking about the probability of a single variable, or a variable conditional on another.
- We often want to determine the joint probability of two variables, such as X and Y. Suppose we are able to determine the following information for education (X) and age (Y) for all Indian citizens based on the census.

Age (Y): / Education (X)		Age : 25-35 30	Age: 35-55 45	Age: 55-85 70
None	0	.01	.02	.05
Primary	1	.03	.06	.10
Secondary	2	.18	.21	.15
College	3	.07	.08	.04

Each cell is the relative frequency (f/N).

We can define the joint probability distribution as:
$p(x, y) = Pr(X = x \text{ and } Y = y)$

Example: what is the probability of getting a 30 year old college graduate?
$p(x,y) = Pr(X=3 \text{ and } Y=30) = .07$

We can see that: $p(x) = \Sigma_y\, p(x,y)$
$p(x=1) = .03 + .06 + .10 = .19$

Education (X)	Age (Y):	Age : 25-35 30	Age: 35-55 45	Age: 55-85 70
None	0	.01	.02	.05
Primary	1	.03	.06	.10
Secondary	2	.18	.21	.15
College	3	.07	.08	.04

Marginal Probability

- We call this the **marginal probability** because it is calculated by summing across rows or columns and is thus reported in the margins of the table.

We can calculate this for our entire table.

Age (Y): / Education (X)	30	45	70	p(x)
None: 0	.01	.02	.05	**.08**
Primary: 1	.03	.06	.10	**.19**
Secondary: 2	.18	.21	.15	**.54**
College: 3	.07	.08	.04	**.19**
p(y)	**.29**	**.37**	**.34**	**1**

If X and Y are discrete random variables, the joint probability distribution of X and Y is a description of the set of points (x,y) in the range of (X,Y) along with the probability of each point.

The joint probability distribution of two discrete random variables is sometimes referred to as the **bivariate probability distribution** or **bivariate distribution.**

Thus we can describe the joint probability distribution of two discrete random variables is through a **joint probability mass function**

$$f(x,y)=P(X=x,Y=y)$$

Joint Probability Mass Function

The function $f(x,y)$ is a **joint probability distribution** or **probability mass function** of the discrete random variables X and Y if

1. $f(x,y) \geq 0$ for all (x,y),
2. $\sum_x \sum_y f(x,y) = 1$,
3. $P(X = x, Y = y) = f(x,y)$.

For any region A in the xy plane, $P[(X,Y) \in A] = \sum\sum_A f(x,y)$.

Joint Density Function

When X and Y are continuous random variables, the **joint density function** $f(x, y)$ is a surface lying above the xy plane, and $P[(X, Y) \in A]$, where A is any region in the xy plane, is equal to the volume of the right cylinder bounded by the base A and the surface.

The function $f(x, y)$ is a **joint density function** of the continuous random variables X and Y if

1. $f(x, y) \geq 0$, for all (x, y),
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$,
3. $P[(X, Y) \in A] = \int\int_A f(x, y)\, dx\, dy$, for any region A in the xy plane.

Marginal Distributions

The **marginal distributions** of X alone and of Y alone are

$$g(x) = \sum_y f(x,y) \quad \text{and} \quad h(y) = \sum_x f(x,y)$$

for the discrete case, and

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\, dy \quad \text{and} \quad h(y) = \int_{-\infty}^{\infty} f(x,y)\, dx$$

for the continuous case.

Consider the joint distribution of X and Y.
Compute the following probabilities:
(i) $P(X = 1, Y = 2)$ (ii) $P(X \geq 1, Y \geq 2)$
(iii) $P(X \leq 1, Y \leq 2)$ (iv) $P(X + Y \geq 2)$ (v) $P(X \geq 1, Y \leq 2)$.

X \ Y	0	1	2	3
0	0	1/8	1/4	1/8
1	1/8	1/4	1/8	0

Solution:

(i) $X = \{0, 1\}, Y = \{0, 1, 2, 3, 4\}$

$P(X = 1, Y = 2) = P(1, 2) = \frac{1}{8}$

(ii) If $X \geq 1, X = \{1\}$. If $Y \geq 2, Y = \{2, 3\}$

$P(X \geq 1, Y \geq 2) = P(1, 2) + P(1, 3) = \frac{1}{8} + 0 = \frac{1}{8}$

(iii) If $X \leq 1, X = \{0, 1\}$. If $Y \leq 2, Y = \{0, 1, 2\}$

$P(X \leq 1, Y \leq 2) = P(0, 0) + P(0, 1) + P(0, 2) + P(1, 0) + P(1, 1) + P(1, 2)$

$$= 0 + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}$$

Cont.

(iv) If $X + Y \geq 2$ then

$X + Y = 0 + 2 \; or \; 0 + 3 \; or \; 1 + 1 \; \; or \; 1 + 2 \; or \; 1 + 3$

$$P(X + Y \geq 2) = P(0,2) + P(0,3) + P(1,1) + P(1,2) + P(1,3)$$

$$= \frac{1}{4} + \frac{1}{8} + \frac{1}{4} + \frac{1}{8} + 0 = \frac{3}{4}$$

(v) If $X \geq 1, X = \{1\}$. If $Y \leq 2, Y = \{0,1,2\}$

$$P(X \geq 1, Y \leq 2) = P(1,0) + P(1,1) + P(1,2)$$

$$= \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{1}{2}$$

Problem:

- Two ballpoint pens are selected at random form a box that contains blue pens, 2 red pens and 3 green pens. If X is the number of blue pens selected and Y is the number of red pens selected, find the joint probability function f(x,y)
- **Solution:**

The possible pairs of values (x,y) are (0,0),(0,1),(1,0),(1,1),(0,2),(2,0)
The joint probability distribution can be represented by the formula

$$f(x,y) = \frac{\binom{3}{x}\binom{2}{y}\binom{3}{2-x-y}}{\binom{8}{2}},$$

for $x = 0, 1, 2$; $y = 0, 1, 2$; and $0 \leq x + y \leq 2$.

Joint distribution

$f(x,y)$		X			Rows Total
		0	1	2	
y	0	3/28	9/28	3/28	15/28
	1	3/14	3/14	0	3/7
	2	1/28	0	0	1/28
Columns Total		5/14	15/28	3/28	1

3. Find the joint distribution of X and Y which are the independent random variables with the following respective distributions.

x_i	1	2
$f(x_i)$	0.7	0.3

y_j	-2	5	8
$g(y_j)$	0.3	0.5	0.2

Solution:

Since X and Y are independent random variables,

$$J_{ij} = f(x_i)g(y_j)$$

Therefore,

$x \backslash y$	-2	5	8	$f(x)$
1	0.21	0.35	0.14	0.7
2	0.09	0.15	0.06	0.3
$g(y)$	0.3	0.5	0.2	$Total = 1$

6. The joint probability distribution of two discrete random variables X and Y is given by $f(x,y) = k(2x+y)$ for $0 \le x \le 2$, $0 \le y \le 3$. (i) Find the value of k. (ii) The marginal distribution of X and Y (iii) Show that X and Y are dependent.

Q.9 A candy company distributed boxes of chocolates with a mixture of creams, toffees, and nuts coated in both light and dark chocolate. For a randomly selected box, let X and Y, respectively, be the proportions of the light and dark chocolates that are creams and suppose that the joint density function is

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \le x \le 1, 0 \le y \le 1 \\ 0, & \text{elsewhere} \end{cases}$$

a) Verify whether

b) Find $P[(X,Y) \in A]$, where A is the region $\{(x,y) \mid 0<x<\frac{1}{2}, \frac{1}{4}<y<\frac{1}{2}\}$

c) Find g(x) and h(y) for the joint density function.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1$$

Example 9 – Solution

a) $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dxdy = \int_{0}^{1}\int_{0}^{1} \frac{2}{5}(2x + 3y)\,dxdy$$

$$= \int_{0}^{1} \left.\frac{2x^2}{5} + \frac{6xy}{5}\right|_{x=0}^{x=1} dy$$

$$= \int_{0}^{1} \left(\frac{2}{5} + \frac{6y}{5}\right) dy = \left.\frac{2y}{5} + \frac{3y^2}{5}\right|_{0}^{1}$$

$$= \frac{2}{5} + \frac{3}{5} = 1$$

Example – Solution

b)

$$P\left[(X,Y)\in A\right] = P\left(0 < X < \tfrac{1}{2}, \tfrac{1}{4} < Y < \tfrac{1}{2}\right)$$

$$= \int_{\frac{1}{4}}^{\frac{1}{2}}\int_{0}^{\frac{1}{2}} \frac{2}{5}(2x+3y)\,dx\,dy$$

$$= \int_{\frac{1}{4}}^{\frac{1}{2}} \left.\frac{2x^2}{5} + \frac{6xy}{5}\right|_{x=0}^{x=\frac{1}{2}} dy$$

$$= \int_{\frac{1}{4}}^{\frac{1}{2}} \left(\frac{1}{10} + \frac{3y}{5}\right) dy = \left.\frac{y}{10} + \frac{3y^2}{10}\right|_{\frac{1}{4}}^{\frac{1}{2}}$$

$$= \frac{1}{10}\left[\left(\frac{1}{2}+\frac{3}{4}\right) - \left(\frac{1}{4}+\frac{3}{16}\right)\right] = \frac{13}{160}$$

Example – Solution

By definition,

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\,dy = \int_{0}^{1} \frac{2}{5}(2x + 3y)\,dy = \left.\frac{4xy}{5} + \frac{6y^2}{10}\right|_{y=0}^{y=1} = \frac{4x+3}{5}$$

For $0 \le x \le 1$, and $g(x)=0$ elsewhere.

Similarly, $$h(y) = \int_{-\infty}^{\infty} f(x, y)\,dx = \int_{0}^{1} \frac{2}{5}(2x + 3y)\,dx = \frac{4(1+3y)}{5}$$

For $0 \le y \le 1$, and $h(y)=0$ elsewhere.

Problem:

Find 'K' if the joint probability density function of a bivariate random variable (X,Y) is given by

$$f(x,y)=\begin{cases} K(1-x)(1-y) & if\ 0<x<4 \\ 0 & otherwise \end{cases}$$

7. The joint probability distribution of X and Y is given by $f(x,y) = c(x^2 + y^2)$ for $x = -1, 0, 1, 3.$ *and* $y = -1, 2, 3.$ (i) Find the value of c. (ii) $P(x = 0, y \leq 2)$ (iii) $P(x \leq 1, y > 2)$ (iv) $P(x \geq 2 - y)$

Solution:

By data, $X = \{-1, 0, 1, 3\}$ and $Y = \{-1, 2, 3\}$

$f(x,y) = c(x^2 + y^2)$

The joint probability distribution of X and Y:

X \ Y	-1	2	3	$f(X)$
-1	2c	$5c$	$10c$	$17c$
0	c	$4c$	$9c$	$14c$
1	$2c$	$5c$	$10c$	$17c$
3	$10c$	$13c$	$18c$	$41c$
g(Y)	$15c$	$27c$	$47c$	89c

(i) **Find c:** $1 = \Sigma f(x,y) = 89c$

$c = \frac{1}{89}$

(ii) $x = 0, y = \{-1, 2\}$

$P(x = 0, y \leq 2)$

$= P(0,-1) + P(0,2)$

$= c + 4c = 5c$

$= 5/89$

(iii) $x = \{-1, 0, 1\},\ y = \{3\}$

$P(x \leq 1, y > 2)$

$= P(-1,3) + P(0,3) + P(1,3)$

$= 10c + 9c + 10c$

$= 29c = 29/89$

Cont.

By data, $X = \{-1, 0, 1, 3\}$ and $Y = \{-1, 2, 3\}$

(iv) $P(x \geq 2 - y) = P(x + y \geq 2)$

$$= P(-1, 3) + P(0, 2) + P(0, 3) + P(1, 2) + P(1, 3) + P(3, -1) + P(3, 2) + P(3, 3)$$

$$= 10c + 4c + 9c + 5c + 10c + 10c + 13c + 18c$$

$$= 79c = 79/89$$

Thank You !

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session No 6
Probability Distributions

(24th/25th June 2023)

Session : 6 Agenda

Distributions:-

- Bernoulli
- Binomial
- Poisson
- Normal distributions.
- Introduction to t-distribution
- F-Distribution and
- Chi Square distributions

Bernoulli Distribution

Definition

A random variable 'X' is said to have Bernouli distribution if its probability mass function is given by

$$p(x) = \begin{cases} p^{x}\, q^{1-x}, & x = 0,\ 1 \\ 0, & \text{elsewhere} \end{cases}$$

Mean & Variance

Binomial Distribution

Definition

A random variable 'X' is said to have Binomial distribution if its probability mass function is given by

$$p(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0, 1, 2, 3, \ldots, n \\ 0, & \text{elsewhere} \end{cases}$$

Binomial Distribution

- Binomial distribution is a discrete probability distribution.
- Binomial distribution will be applied under the following experimental conditions

1) The number of trails (n) is finite

2) The trails are independent of each other

3) The probability of success p is constant for each trail.

4) Each trail results in two mutually exclusive events known as success and failure.

Mean & variance

$$P(x) = {}^nC_x\, p^x q^{n-x}, \quad x = 0, 1, 2, \ldots n$$

$$\text{mean} = \mu = E(x) = \sum x\, P(x)$$

$$= \sum x\, {}^nC_x\, p^x q^{n-x} = \sum x \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

$$= \sum \frac{n!}{(x-1)!\,(n-x)!} p^x q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!\,[(n-1)-(x-1)]!} p^{x-1} q^{(n-1)-(x-1)}$$

$${}^{(n-1)}C_{x-1}$$

$$= np\,(q+p)^{n-1}$$

$$(q+p)^n = {}^nC_0\, q^n \cdot p^0 + {}^nC_1\, q^{n-1} \cdot p + \ldots + {}^nC_n\, q^0 p^n$$

$$= np$$

Variance $\sigma^2 = E(x^2) - [E(x)]^2$

$$E(x^2) = \sum x^2 p(x) = \sum x^2 \, {}^nC_x p^x q^{n-x}$$

$$= \sum [x(x-1) + x] \, {}^nC_x p^x q^{n-x}$$

$\sum x p(x) \rightarrow$ mean $= np$

$$= \sum x(x-1) \cdot {}^nC_x p^x q^{n-x} + \sum x \, {}^nC_x p^x q^{n-x}$$

$$= \sum x(x-1) \frac{n!}{x!\,(n-x)!} p^x q^{n-x} + np$$

$$= \sum \frac{n(n-1)(n-2)!}{(x-2)!\,[(n-2)-(x-2)]!} p^2 p^{(x-2)} \cdot q^{(n-2)-(x-2)} + np$$

$$= n(n-1)p^2 \sum {}^{(n-2)}C_{x-2} p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2 + np$$

Variance $= \sigma^2 = n(n-1)p^2 + np - (np)^2$

$$= \cancel{n^2p^2} - np^2 + np - \cancel{n^2p^2}$$

$$= np(1-p) = npq$$

Binomial Distribution

4.11 Which conditions for the binomial distribution, if any, fail to hold in the following situations?

(a) The number of persons having a cold at a family reunion attended by 30 persons.

Ans: Getting cold is not independent. So it is not binomial

(b) Among 8 projectors in the department office, 2 do not work properly but are not marked defective. Two are selected and the number that do not work properly will be recorded

Ans: Probability of success is not same in each trial. So this is not binomial

Binomial Distribution Thinking Challenge

You're a telemarketer selling service contracts for Macy's. You've sold 20 in your last 100 calls (***p* = .20**). If you call **12** people tonight, what's the probability of

A. No sales?

B. Exactly 2 sales?

C. At most 2 sales?

D. At least 2 sales?

Binomial Distribution Solution*

***n* = 12, *p* = .20**

A. $p(0)$ = **.0687**

B. $p(2)$ = **.2835**

C. $p(\text{at most } 2) = p(0) + p(1) + p(2)$
$= .0687 + .2062 + .2835$
= **.5584**

D. $p(\text{at least } 2) = p(2) + p(3) \ldots + p(12)$
$= 1 - [p(0) + p(1)]$
$= 1 - .0687 - .2062$
= **.7251**

Understandings

Phrase	***Math Symbol***
"at least"	$\geq$
"more than" or "greater than"	$>$
"fewer than" or "less than"	$<$
"no more than"	$\leq$
"exactly"	$=$

Binomial Distribution

If a coin is tossed 6 times, what is the probability of getting 2 or fewer heads?

$P(x \le 2) = \sum p(x) = P(0) + P(1) + P(2)$

$$P(X = 0) = \binom{6}{0}(0.5)^0(0.5)^6 = \frac{6!}{6!\,0!}(0.5)^6 = 0.015625$$

$$P(X = 1) = \binom{6}{1}(0.5)^1(0.5)^5 = \frac{6!}{5!\,1!}(0.5)^6 = 0.09375$$

$$P(X = 2) = \binom{6}{2}(0.5)^1(0.5)^4 = \frac{6!}{4!\,2!}(0.5)^6 = 0.078125$$

$P(x \le 2) = \sum p(x) = 0.015625 + 0.09375 + 0.078125 = 0.1875$

Example

The probability that a bomb dropped from a plane will strike the target is 1/5. if six such bombs are dropped.

Find the probability that

1) exactly two bombs hit the target

2) At least two will hit the target

Example

An unbiased dice is thrown 5 times and occurrence of 1 or 6 considered as success find,

1) Probability of exactly one success
2) Probability of at least 4 success
3) Probability of at least one success
4) Mean and variance

Example

Seven dice are thrown 729 times. How many times do you expect at least four dice to show three or five?

Example

Find the binomial distribution if the mean is 4 and variance is 3

Additional problem

In a large number of parts manufactured by a machine , the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples , how many would be expected to contain atleast 3 defective parts.

Sol: given n=20;

$$\text{mean} = \mu = 2, \text{ but } \mu = np \Rightarrow 2 = 20p \therefore p = 0.1$$

$$\therefore P(x) = 20c_x (0.1)^x (0.9)^{20-x}$$

Therefore the probability that there are atleast 3 defective parts in a sample of 20 is

$$P(X \geq 3) = 1 - P(X < 3) = 1 - [P(0) + P(1) + P(2)]$$

$$\therefore P(X \geq 3) = 0.323$$

Hence out of 1000 samples we get 0.323*1000=323 samples

Discrete Probability Distribution

Discrete Probability Distributions

- **Binomial**
- **Hypergeometric**
- **Poisson**

Binomial Distribution Recall

A **binomial experiment** is a probability experiment that satisfies the following conditions.

1. The experiment is repeated for a fixed number of trials, where each trial is **independent** of other trials.
2. There are only two possible outcomes of interest for each trial. The outcomes can be classified as a success (S) or as a failure (F).
3. The probability of a success $P(S)$ is the same for each trial.
4. The random variable x counts the number of successful trials.

Poisson Distribution

Definition

A random variable 'X' is said to have Poisson distribution if its probability mass function is given by

$$p(x) = \begin{cases} e^{-\lambda} \dfrac{\lambda^x}{x!}, & x = 0, 1, 2, \ldots \\ 0, \text{ elsewhere} & \end{cases}$$

Poisson distribution is the discrete probability distribution of a discrete random variable X, which may not have upper bound. It is defined for non negative values of x as follows

Cont…

- Poisson distribution is suitable for rare events for which the probability of occurrence 'p' is very small and the number of trials 'n' is very large.

- Also binomial distribution can be approximated by Poisson distribution when $n \to \infty$ and $p \to 0$ such that $\lambda = np$ = constant

Example

If the probability of a bad reaction from a certain injection is 0.001.

- Determine the chance that out of 2000 individuals more than two will get a bad reaction.

Solution

Poisson Distribution

Police records show that number of accident victims died in road traffic accidents is 0.1%. What is the probability that among 500 randomly selected accident victims

(i) none have died?

(ii) at least 3 have died

(iii) between 2 and 6 have died

Example:

Example 6: A hospital switch board receives an average of 4 emergency calls in a 10 minute interval. What is the probability that

(i) there are at most 2 emergency calls in 10 minute interval

(ii) there are exactly 3 emergency calls in a 10 minute interval

Ans: Given Mean = 4calls. We have

$$P(x) = \frac{(4)^{x} e^{-4}}{x!}$$

(i) P(at most 2 calls) = P(x ≤ 2) = P(0)+P(1)+P(2)=0.2381

(ii) P(Exactly 3 calls)= P(x = 3) = 0.1954

Example:

Example 2: The mean number of power outages in the city of Bangalore follows a Poisson variate and is 4 per year. Find the probability that in a given year,

a) there are exactly 3 outages, b) there are more than 3 outages.

Poisson Mean and Variance

Mean:

$$\mu = \lambda$$

For a Poisson random variable, the variance and mean are the same!

Variance and Standard Deviation

$$\sigma^2 = \lambda$$

$$\sigma = \sqrt{\lambda}$$

where λ= expected number of occurrences in a given experiment.

Example

4.55 In a factory, 8% of all machines break down at least once a year. Use the Poisson approximation to the binomial distribution to determine the probabilities that among 25 machines (randomly chosen in the factory):

(a) 5 will break down at least once a year;

(b) at least 4 will break down once a year;

(c) anywhere from 3 to 8, inclusive, will break down at least once a year.

Suggested Problems

Given that a switch board of a consultant's office receives on the average 0.6 calls per minute, find the probabilities that

(a) in a given minute there will be at least 1 call;

(b) in a 4-minute interval there will be at least 3 calls.

Solution:

a) here

$$\lambda = 0.6 \text{ and } t = 1 \therefore \quad p(x, \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$

b)

$$p(x, \lambda t) = \frac{e^{-0.6}(0.6)^x}{x!}$$

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(0)$$

$$= 1 - e^{-0.6} = 0.4512$$

$$\lambda = 0.6 \text{ and } t = 4 \therefore \quad p(x, \lambda t) = \frac{e^{-2.4}(2.4)^x}{x!}$$

$$P(X \geq 3) = 1 - P(X < 3) = 1 - [P(0) + P(1) + P(2)]$$

$$= 0.4303$$

Exercise:

The number of customers arriving at a cafeteria at an average rate of 0.3 per minute.

(a) Find the probability that exactly 2 customers arrive in a 10-minute span.

(b) Find the probability that 2 or more customers arrive in a 10-minute span.

(c) Find the probability that exactly one customer arrives in a 5-minute span and one customer arrives in the next 5-minute span.

Exercise:

The average number of trucks arriving on any one day at a truck depot in a certain city is known to be 12. what is the probability that on a given day fewer than nine trucks will arrive at this depot?

Exercise:

A certain kind of sheet metal has on the average, five defects per 10 square feet. If we assume a Poisson distribution, what is the probability that a 15-square-foot sheet of the metal will have at least six defects?

Exercise:

The number of flaws in a fiber optic cable follows a poisson process with an average of 0.6 per 100 feet.

a) Find the probability of exactly 2 flaws in 200 foot cable.

b) Find the probability of exactly 1 flaw in first 100 feet and exactly 1 flaw in the second 100 feet.

Exercise:

Suppose that on an average 1 out of 10000 houses catch fire in a year in a district. If there are 2000 houses in the district, find the probability that exactly 5 house will catch fire during that year.

RECALL: Continuous Random Variables

A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.

The probability of the random variable assuming a value within some given interval from x1 to x2 is defined to be the area under the graph of the **probability density function** between x1 and x2

Example:

- Height of students in a class
- Amount of ice tea in a glass
- Change in temperature throughout a day
- Price of a car in next year

Continuous Probability Distributions:

Normal Distribution:

Normal distribution is a general distribution. Any information such as heights, weights, lengths, widths, sizes, profits, prices, wages, time etc. containing sufficiently large data with mean and variance follows normal distribution. The distribution of errors of repeated measurements follows normal distribution.

Normal Distribution

A continuous random variable X which assumes all possible values in the entire real space, i.e., $-\infty<x<\infty$ is said to follow normal distribution with two parameters such as mean μ and variance σ^2 if it has the probability density function given by

$$f(x;\ \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

$$-\infty < \mu < \infty, \sigma > 0$$

Normal Distribution

Change in Mean determines the shift in the distribution
Change in the deviation determines the spread of the data points

$$f(x;\ \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ = mean
σ = standard deviation
π = 3.14159
e = 2.71828

This is a bell shaped curve with different centers and spreads depending on μ and σ

Normal Distribution

Properties:

1. Normal curve is bell shaped and symmetric about the mean
2. Mean = Mode = Median
3. Total area under normal curve is equal to 1
4. Normal curve approaches but never touches the x axis as it extends farther and farther away from the mean

Normal Distribution

$E(X)=\mu =$ $$\int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

$Var(X)=\sigma^2 =$ $$\int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx) - \mu^2$$

Standard Deviation(X)=σ

No matter what μ and σ are, the area between μ-σ and μ+σ is about 68%; the area between μ-2σ and μ+2σ is about 95%; and the area between μ-3σ and μ+3σ is about 99.7%. Almost all values fall within 3 standard deviations.

68-95-99.7 Rule for Normal Distributions

68% of the AUC(Area under curve) within ±1σ of μ

95% of the AUC within ±2σ of μ

99.7% of the AUC within ±3σ of μ

Standard Normal Distribution

The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1.

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt = P(Z \le z)$$

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

All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z i.e.

$$P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\ dZ$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Problem

If a random variable has the standard normal distribution, find the probability that it will take on a value

a)less than 1.75 = $P(Z < 1.75) = F(1.75) = 0.9599$

b)less than -1.25 = $P(Z < -1.25) = F(-1.25) = 1 - F(1.25) = 0.1056$

c)greater than 2.06 = $P(Z > 2.06) = 1 - F(2.06) = 0.0197$

d)greater than -1.82 = $P(Z > -1.82) = 1 - F(-1.82) = 1 - 0.0344 = 0.9656$

$P(Z > -1.82) = 1 - F(-1.82) = 1 - (1 - F(1.82) = F(1.82) = 0.9656$

Probability and Normal Distributions

Example:
The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score less than 90.

$$z = \frac{x - \mu}{\sigma} = \frac{90 - 78}{8}$$
$$= 1.5$$

The probability that a student receives a test score less than 90 is 0.9332.

$P(x < 90) = P(z < 1.5) = 0.9332$

Probability and Normal Distributions

Example:
The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score greater than 85.

$$z = \frac{x - \mu}{\sigma} = \frac{85 - 78}{8}$$
$$= 0.875 \approx 0.88$$

The probability that a student receives a test score greater than 85 is 0.1894.

$P(x > 85) = P(z > 0.88) = 1 - P(z < 0.88) = 1 - 0.8106 = 0.1894$

Probability and Normal Distributions

Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score between 60 and 80.

$$z_1 = \frac{x - \mu}{\sigma} = \frac{60 - 78}{8} = -2.25$$

$$z_2 = \frac{x - \mu}{\sigma} = \frac{80 - 78}{8} = 0.25$$

The probability that a student receives a test score between 60 and 80 is 0.5865.

$$P(60 < x < 80) = P(-2.25 < z < 0.25) = F(0.25) - F(-2.25) = 0.5987 - 0.0122 = 0.5865$$

Problems

If $Z \sim N(0,1)$, to find constants a and b such that $P(Z \le a) = 0.9147$ and $P(Z \ge b) = 0.0526$

$a = 1.37$ and $b = 1.62$

If $X \sim N(25,36)$, find a constant c such that $P(|X - 25| \le c) = 0.9544$.

Solution :

$P(|X - 25| \le c) = 0.9544$

$\Rightarrow P(\frac{-c}{6} \le \frac{X-25}{6} \le \frac{c}{6}) = 0.9544 \quad \Rightarrow \Phi(\frac{c}{6}) - [1 - \Phi(\frac{c}{6})] = 0.9544$

$\Rightarrow \Phi(\frac{c}{6}) = 0.9772 \quad \Rightarrow \frac{c}{6} = 2 \Rightarrow c = 12$

Problem

The time for oil to percolate to all parts of an engine can be treated as a random variable having a normal distribution with mean 20 seconds. Find its standard deviation if the probability is 0.25 that will take a value greater than 31.5 seconds.

Given : Mean μ = 20 seconds, P(X > 31.5) = 0.25

σ = ?

$$z = \frac{x - \mu}{\sigma} = \frac{31.5 - 20}{\sigma}$$

$$\sigma = \frac{11.5}{z} = \frac{11.5}{0.675} = 17.04$$

Ex.

Monthly Salary X in a big organisation is normally distributed with mean Rs 3000 and standard deviation of Rs.250 what should be the minimum salary of a worker in this organisation so that the probability that he belongs to top 5%workers?

Problem

Butterfly-style valves used in heating and ventilating industries have a high flow coefficient. Flow coefficient can be modeled by a normal distribution with mean 496 C_v and standard deviation 25C_v. Find the probability that a valve will have a flow coefficient of

a) atleast 450C_v

$P(X > 450) = P[(450 – 496) /25] = P(Z > -1.84) = 1 – F(-1.84) = 1- 00329 = .9671$

b. between 445.5 and 522C_v

$P(445.5 < X < 522) = P(-2.02 < Z < 1.04) = F(1.04) – F(-2.02)= 0.8508 – 0.0217 = 0.8291$

Normal Approximation to Binomial Distribution:

If n tends to infinity (preferably n>30) and neither p nor 1-p is so small, that is p and 1-p both are large preferably np>15 and n(1-p)>15 then binomial probabilities can be approximated by normal probabilities using continuity correction and taking

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1-p)}}$$

Where X is binomial random variable approximated as normal random variable.

Introduction to t-distribution

The pdf of the t distribution with ν degrees of freedom:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}}$$

for $-\infty < t < \infty$

Introduction to t-distribution (Sampling distribution of mean when σ unknown)

The t-distribution, also known as the **Student's t-distribution**, is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails.

It is used for estimating population parameters for **small sample sizes when**

t - Distribution When σ is unknown and $n < 30$.

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \quad with\ (n-1)\ d.o.f.$$

Where sample S.D. is calculated by formula $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Introduction to Chi- Square(χ^2) distribution

- Chi-Square distribution pdf is given by

$$f(x) = \begin{cases} \dfrac{1}{2^{\vartheta/2}\Gamma(\frac{\vartheta}{2})} x^{\frac{\vartheta}{2}-1} e^{-\frac{x}{2}} & for\ x > 0 \\ 0 & otherwise \end{cases}$$

where ϑ is the parameter of distribution, also known as degrees of freedom.

- Introduction to Chi- Square(χ^2) distribution

Introduction to Chi- Square(χ^2) distribution

- Chi-Square distribution curve is not symmetrical, and hence not a normal curve.
- Chi-Square varies from 0 to ∞(Curve lies entirely in first quadrant).
- It depends only on degrees of freedom ϑ.
- It is very important in estimation and hypothesis testing.
- It is used in sampling distributions of the sample variance, analysis of variance.
- It is used as a measure of goodness of fit.

Introduction to F-distribution (Sampling distribution of the ratio of two sample variances)

- The pdf for a random variable $X \sim F(n_1, n_2)$ is given by

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} \quad , x > 0$$

- F- distribution curve lies entirely in first quadrant.
- The F- curve depends not only on the two parameters ϑ_1 and ϑ_2 but also on the order in which they are stated.

Introduction to F-distribution
(Sampling distribution of the ratio of two sample variances)

- The F-distribution was **developed by Fisher to study the behavior of two variances from random samples taken from two independent normal populations.**
- In applied problems we may be interested in knowing whether the population variances are equal or not, based on the response of the random samples.
- If S_1^2 and S_2^2 are variances of independent random sample of size n_1 and n_2 from a normal populations with variances σ_1^2 and σ_2^2, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

which follows F –distribution with $\vartheta_1 = n_1-1$ and $\vartheta_2=n_2-1$ d.o.f.

Thanks

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session No 7

Testing of Hypothesis

($1^{st}/2^{nd}$ July, 2023)

Contact Session 7: Module 4: Hypothesis Testing

Contact Session	List of Topic Title	Reference
CS - 7	Sampling – random sampling and Stratified sampling, Sampling distribution – Central Limit theorem, Estimation– Interval Estimation, Confidence level	T1 & T2

Sampling: Terminology

POPULATION (Big data or Huge or Massive data)

A population is the set of all elements of interest in a particular study. The process of conducting a survey to collect data for the entire population is called a census.

SAMPLE (Small data)

A sample is a subset of the population. The process of conducting a survey to collect data for a sample is called a sample survey.

As one of its major contributions, statistics uses data from a sample to make estimates and test hypotheses about the characteristics of a population in the point of view of cost constraint. This part of statistics is called inferential statistics.

Terminology

Parameter

The various statistical characteristics or constants of a population such as Population mean, Population SD, Population Proportion etc., are called parameters.

Statistic

The various statistical functions of a sample such as sample mean, sample sd, sample proportion etc., are called statistics.

A statistic is a function of sample observations and is used to estimate its corresponding parameter. It is also known as estimator and its calculated value is called estimate.

A statistic is a variable and becomes a random variable if the sample is random sample.

Sampling

- Sampling is a method or process of selecting samples from populations.
- Data are gathered from samples and conclusions are drawn about the population as a part of the inferential statistics process
- A sample provides a reasonable means for gathering useful decision-making information that might be otherwise unattainable and unaffordable.

Reasons for Sampling

Taking a sample instead of conducting a census offers several advantages

1. The sample can save money.

2. The sample can save time.

3. For given resources, the sample can broaden the scope of the study.

4. If accessing the population is impossible, the sample is the only option.

Types of Sampling

Sampling are of two types:

- Random Sampling
- Non-Random Sampling

Random Sampling	Non –Random Sampling
Simple Random Sampling	Judgemental Sampling
Stratified Sampling	Quota Sampling
Systematic Sampling	Convenience Sampling
Clustered Sampling	Snowball Sampling

Random Versus Non random Sampling

- In **random(probability)**sampling every unit of the population has the same probability of being selected into the sample.

- In **nonrandom(non-probability)**sampling not every unit of the population has the same probability of being selected into the sample.

Simple Random Sampling

In a simple random sample, every member of the population has an equal chance of being selected. The sampling frame should include the whole population.

To conduct this type of sampling, we can use tools like random number generators or other techniques that are based entirely on chance.

Example:
Suppose we want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

Stratified Random Sampling

- In this, the population is divided into non overlapping **subpopulations** called **strata**.

- The researcher then extracts a random sample from each of the subpopulations.

- The main reason for using stratified random sampling is that it has the potential for reducing sampling error.

- With stratified random sampling, the potential to match the sample closely to the population is greater than it is with simple random sampling because portions of the total sample are taken from different population subgroups.

Example:

- A survey about timekeeping might divide the population by time zone, then take 100 random samples per zone.

- A study on tax reform might stratify a population according to income, then take random samples from each stratum.

Non-probability sampling

Non-probability sampling is a sampling method that uses non-random criteria like the availability, geographical proximity, or expert knowledge of the individuals you want to research in order to answer a research question.

Non-probability sampling is used when the population parameters are either unknown or not possible to individually identify.

Example:

A visitors to a website that doesn't require users to create an account could form part of a non-probability sample.

Sampling Error

- **Sampling error** occurs *when the sample is not representative of the population.*

- When random sampling techniques are used to select elements for the sample, sampling error occurs by chance.

Population of Wages of employees of an organization

1861	2495	1000	2497	1865	791	2090	2637	1327	1678
1680	2858	795	2495	2496	2501	1160	1480	1860	2490
2090	2840	2490	2640	659	827	2646	2638	2643	868
1327	1866	1861	2486	2865	3011	2494	1489	1865	2855
2840	2499	2093	2660	1165	2600	2085	2640	2998	1861
2956	2495	2865	1865	3000	3019	1670	2858	2642	1680
3038	3000	1313	596	656	3240	590	2501	2485	3015
2092	1679	3024	2497	2825	2630	2070	2900	1861	2636
2495	2637	2497	1159	2640	3050	870	2896	2500	2638
926	2860	1481	875	2482	1860	2086	934	3200	2490

Select different samples of varied sizes

Sample 1

12

3000	2486	820	1678	2070	2638	2490	1865	1000	2090	596	3200

Sample 2

11

2840	2858	3000	2490	2998	3050	2070	2896	3200	2490	3280

Sample 3

2858	3240	2497	2865	656	2093	934	1861	868	795

Sample 4

2086	1000	2497	596	656	875	2085	934	1313

Sample 5

820	1313	3000	2640	596	2640	2600	2495	934	2500

Select different samples of varied sizes

Sample 6

2840 2499 1327 1861 2495 3024 3038 2497

Sample 7

2858 2490 868 1670 1480 2643 1480 1680 2085 2490

Sample 8

2495 2858 1861 2092 2499 3000 2660 1000 1679 926 2660

Sample 9

795 791 3200 2085 2638 2497 2486 1159 2640

Sample 10

3019 3240 3200 3050 3000 3015 2900 2896 2998

Compute sample mean of these samples

Sample No.	Sample size	Mean	SD
1	12	1994.42	843.23
2	11	2830.18	349.94
3	10	1866.70	988.57
4	9	1338.00	704.36
5	10	1953.80	920.44
6	8	2447.63	590.64
7	10	1974.40	638.05
8	11	2157.27	715.10
9	9	2032.33	891.53
10	9	3035.33	117.40
Overall	100	2162.24	732.26

Sampling Variability

- The term "sampling variability" refers to the fact that the statistical information from a sample (called a *statistic*) will vary as the random sampling is repeated.

- **Sampling variability** will **decrease** as the **sample size increases.**

- the samples must be randomly chosen, must be of the same size (not smaller than 30), and the more samples that are used, the more reliable the information gathered will be.

Do you consider these sample means and sample SDs as variable?

If yes, should we not describe the distribution of these variables?

The distribution of the sample estimates is called sampling distribution

For example the distribution of sample means is called Sampling distribution of mean

Definition

- The probability distribution of a statistic (sample estimate) is called sampling distribution.
- The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection.

Sampling Distribution Of $\bar{x}$

- The sample mean is one of the more common statistics used in the inferential process.

- The **distribution** of the values of the sample mean ($\bar{x}$) in repeated **samples** is called the **sampling distribution of $\bar{x}$**

- One way to examine the distribution possibilities is to take a population with a particular distribution, randomly select samples of a given size, compute the sample means, and attempt to determine how the means are distributed.

Example

- Suppose a small finite population consists of only N = 8 numbers:
- 54 55 59 63 64 68 69 70
- Using an Excel-produced histogram, we can see the shape of the distribution of this population of data.

- Suppose we take all possible samples of size n = 2 from this population with replacement.

Example

The result is the following pairs of data.

(54,54)	(55,54)	(59,54)	(63,54)
(54,55)	(55,55)	(59,55)	(63,55)
(54,59)	(55,59)	(59,59)	(63,59)
(54,63)	(55,63)	(59,63)	(63,63)
(54,64)	(55,64)	(59,64)	(63,64)
(54,68)	(55,68)	(59,68)	(63,68)
(54,69)	(55,69)	(59,69)	(63,69)
(54,70)	(55,70)	(59,70)	(63,70)
(64,54)	(68,54)	(69,54)	(70,54)
(64,55)	(68,55)	(69,55)	(70,55)
(64,59)	(68,59)	(69,59)	(70,59)
(64,63)	(68,63)	(69,63)	(70,63)
(64,64)	(68,64)	(69,64)	(70,64)
(64,68)	(68,68)	(69,68)	(70,68)
(64,69)	(68,69)	(69,69)	(70,69)
(64,70)	(68,70)	(69,70)	(70,70)

The means of each of these samples follow.

54	54.5	56.5	58.5	59	61	61.5	62
54.5	55	57	59	59.5	61.5	62	62.5
56.5	57	59	61	61.5	63.5	64	64.5
58.5	59	61	63	63.5	65.5	66	66.5
59	59.5	61.5	63.5	64	66	66.5	67
60	61.5	63.5	65.5	66	68	68.5	69
61.5	62	64	66	66.5	68.5	69	69.5
62	62.5	64.5	66.5	67	69	69.5	70

Example

- Again using an Excel-produced histogram, we can see the shape of the distribution of these sample means.

Conclusions

- Notice that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population.
- The sample means appear to "pile up" toward the middle of the distribution and "tail off" toward the extremes.
- As sample sizes become much larger, the sample mean distributions begin to approach a **normal distribution** and the variation among the means decreases.

Sample Means from 90 Samples Ranging in Size from n = 2 to n = 30 from a Uniformly Distributed Population with a = 10 and b = 30

Shapes of the Distributions of Sample Means

Central Limit Theorem

- If samples of size *n* are drawn randomly from a population that has a mean of μ and a standard deviation of σ, the sample means, $\bar{x}$, are approximately normally distributed for sufficiently large sample sizes (*n* >= 30) regardless of the shape of the population distribution.

- If the population is normally distributed, the sample means are normally distributed for any size sample.

- From mathematical expectation

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Z score for sample means

- The central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed populations.
- Thus, **sample means** can be **analyzed** by using ***z* scores**
- The formula to determine z scores for individual values from a normal distribution:

$$z = \frac{x - \mu}{\sigma}$$

- If sample means are normally distributed, the z score formula applied to sample means would be

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

- The standard deviation of the statistic of interest is $\sigma_{\bar{x}}$, sometimes referred to as the **standard error of the mean.**

Example

Suppose the mean expenditure per customer at a tire store is \$85.00, with a standard deviation of \$9.00.

If a random sample of 40 customers is taken, what is the probability that the sample average expenditure per customer for this sample will be \$87.00 or more?

Solution

Because the sample size is greater than 30, the central limit theorem

Can be used, and the sample means are normally distributed.

μ= \$85 σ= \$9

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\$87.00 - \$85.00}{\frac{\$9.00}{\sqrt{40}}} = \frac{\$2.00}{\$1.42} = 1.41$$

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

Example

Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers.

What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?

Solution

For this problem, μ= 448, σ= 21, and n = 49. The problem is to determine

P(441<= $\bar{x}$ <=446).

The following

$$z = \frac{441 - 448}{\frac{21}{\sqrt{49}}} = \frac{-7}{3} = -2.33$$

$$z = \frac{446 - 448}{\frac{21}{\sqrt{49}}} = \frac{-2}{3} = -0.67$$

$$P(-2.33 < Z < -0.67)$$

$$= P(0.67 < Z < 2.33)$$

$$= F(2.33) - F(0.67)$$

$$= 0.9901 - 0.7485$$

Sampling from a Finite Population

- The earlier example was based on the assumption that the population was infinitely or extremely large.

- In cases of a finite population, *a statistical adjustment can be made to the z formula for sample means*. The adjustment is called the **finite correction factor**

$$\sqrt{\frac{N-n}{N-1}}.$$

- Following is the z formula for sample means when samples are drawn from finite populations.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

Rules for finite population

- As the size of the finite population becomes larger in relation to sample size, the finite correction factor approaches 1.

- In theory, whenever researchers are working with a finite population, they can use the finite correction factor.

- A rough rule of thumb for many researchers is that, if the sample size is **less** than **5%** of the finite population size or ***n/N* < 0.05**, the finite correction factor does **not** significantly modify the solution.

Example

A production company's 350 hourly employees average 37.6 years of age, with a standard deviation of 8.3 years.

If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

Solution

- The population mean is 37.6, with a population standard deviation of 8.3.
- The sample size is 45, but it is being drawn from a finite population of 350; that is, n = 45 and N = 350.
- The sample mean under consideration is 40
- Using the z formula with the finite correction factor gives

$$z = \frac{40 - 37.6}{\frac{8.3}{\sqrt{45}}\sqrt{\frac{350 - 45}{350 - 1}}} = \frac{2.4}{1.157} = 2.07$$

...solution

- This z value yields a probability of .480
- Therefore, the probability of getting a
- sample average age of less than
- 40 years is **.4808 + .5000 = .9808.**

0.98077
≃ 0.9808

Sampling Distribution Of Sample Proportion

- If research results in ***countable*** items such as how many people in a sample have a flexible work schedule, the sample proportion is often the statistic of choice.

SAMPLE PROPORTION

$$\hat{p} = \frac{x}{n}$$

where

x = number of items in a sample that have the characteristic
n = number of items in the sample

Example

- In a sample of 100 factory workers, 30 workers might belong to a union.

- The value of sample proportion for this characteristic, union membership, is

$$30/100 = 0.30 = 30\%$$

How does a researcher use the sample proportion in analysis?

- The central limit theorem applies to sample proportions in that the normal distribution approximates the shape of the distribution of sample proportions

- If $n*p > 15$ and $n*q > 15$(p is the population proportion and $q = 1 - p$).

- The mean of sample proportions for all samples of size n randomly drawn from a population is p (the population proportion) and the standard deviation of sample proportions is $\sqrt{\frac{p \cdot q}{n}}$

- sometimes referred to as the **standard error of the proportion**

Z Formula For Sample Proportions

For $n*p > 1$ 5 and $n*q > 15$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

where

$\hat{p}$ = sample proportion
n = sample size
p = population proportion
$q = 1 - p$

Example

Suppose 60% of the electrical contractors in a region use a particular brand of wire. What is the probability of taking a random sample of size 120 from these electrical contractors and finding that .50 or less use that brand of wire?

Solution

$p = .60 \quad \hat{p} = .50 \quad n = 120$

The z formula yields

$$z = \frac{.50 - .60}{\sqrt{\frac{(.60)(.40)}{120}}} = \frac{-.10}{.0447} = -2.24$$

= -2.24 is .4875.

For $z < -2.24$ (the tail of the distribution), the answer is .5000 - .4875 = **.0125**.

Example

If 10% of a population of parts is defective, what is the probability of randomly selecting 80 parts and finding that 12 or more parts are defective?

Solution

Here, $p = .10$, $\hat{p} = 12/80 = .15$, and $n = 80$. Entering these values in the z formula yields

$$z = \frac{.15 - .10}{\sqrt{\frac{(.10)(.90)}{80}}} = \frac{.05}{.0335} = 1.49$$

- The probability of .4319 for a z value of 1.49, which is the area between the sample proportion, .15, and the population proportion, .10. The answer to the question is

$$P(\hat{p} \geq .15) = .5000 - .4319 = .0681.$$

Forms Of Statistical Inference

- Three forms of statistical inference

  - Point estimation
  - Interval estimation
  - Hypothesis testing

Point Estimate

- A **point estimate** is a statistic taken from a sample that is used to estimate a population parameter.
- A point estimate is only as good as the representativeness of its sample.
- If other random samples are taken from the population, the point estimates derived from those samples are likely to vary.

Interval Estimate

- Because of variation in sample statistics, estimating a population parameter with an interval estimate is often preferable to using a point estimate.

- An interval estimate (**confidence interval**) is a range of values within which the analyst can declare, with some confidence, the population parameter lies.

Confidence Interval to Estimate μ

100(1 − α)% CONFIDENCE INTERVAL TO ESTIMATE μ: σ KNOWN (8.1)

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where

α = the area under the normal curve outside the confidence interval area

$\alpha/2$ = the area in one end (tail) of the distribution outside the confidence interval

Confidence Intervals

Standard normal variable value $Z_{\alpha/2}$ (Table Value)	Level of significance $\propto$		
	1%=0.01	5%=0.05	10%=0.1
$\propto/2$	0.01/2 =0.005	0.05/2=0.025	0.1/2=0.05
$F(Z_{\alpha/2})=1-\propto/2$ $P(Z<=Z_{\alpha/2})=1-\propto/2$	P(Z<=2.58)= 0.995 then $Z_{\alpha/2}$= 2.58	P(Z<=1.96)= 0.975 then $Z_{\alpha/2}$= 1.96	P(Z<=1.645)= 0.95 then $Z_{\alpha/2}$= 1.645

Example

In the cellular telephone company, problem of estimating the population mean number of minutes called per residential user per month, from the sample of 85 bills it was determined that the sample mean is 510 minutes.

Suppose past history and similar studies indicate that the population standard deviation is 46 minutes.

Determine a 95% confidence interval.

Solution

The business researcher can now complete the cellular telephone problem. To determine a 95% confidence interval for $\bar{x} = 510$, $\sigma = 46$, $n = 85$, and $z = 1.96$, the researcher estimates the average call length by including the value of z in formula 8.1.

$$510 - 1.96\frac{46}{\sqrt{85}} \leq \mu \leq 510 + 1.96\frac{46}{\sqrt{85}}$$

$$510 - 9.78 \leq \mu \leq 510 + 9.78$$

$$500.22 \leq \mu \leq 519.78$$

…solution

- The confidence interval is constructed from the point estimate, which in this problem is 510 minutes, and the error of this estimate, which is 9.78 minutes.

- The resulting confidence interval is 500.22 <= μ <= 519.78.

- The cellular telephone company researcher is 95%, confident that the average length of a call for the population is between 500.22 and 519.78 minutes.

Example

A study is conducted in a company that employs 800 engineers. A random sample of 50 engineers reveals that the average sample age is 34.3 years. Historically, the population standard deviation of the age of the company's engineers is approximately 8 years.

Construct a 98% confidence interval to estimate the average age of all the engineers in this company.

Solution

- This problem has a finite population. The sample size, 50, is greater than 5% of the population, so the finite correction factor may be helpful.

- In this case N = 800, n = 50, $\bar{x}$= 34.3 and σ= 8

- The z value for a 98% confidence interval is 2.33

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$34.30 - 2.33\frac{8}{\sqrt{50}}\sqrt{\frac{750}{799}} \le \mu \le 34.30 + 2.33\frac{8}{\sqrt{50}}\sqrt{\frac{750}{799}}$$

$$34.30 - 2.55 \le \mu \le 34.30 + 2.55$$

$$31.75 \le \mu \le 36.85$$

Estimating The Population Proportion

- Methods similar to those used earlier can be used to estimate the population proportion.
- The central limit theorem for sample proportions led to the following formula

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

- where q = 1 - p. Recall that this formula can be applied only when n*p and n*q are **greater** than **5**.

- for confidence interval purposes only and for large sample sizes— is substituted for p in the denominator, yielding

$$z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}}$$

Confidence Interval To Estimate P

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$$

where

$\hat{p}$ = sample proportion
$\hat{q} = 1 - \hat{p}$
p = population proportion
n = sample size

In this formula, $\hat{p}$ is the point estimate and $\pm z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$ is the error of the estimation.

Example

A study of 87 randomly selected companies with a telemarketing operation revealed that 39% of the sampled companies used telemarketing to assist them in order processing.

Using this information, how could a researcher estimate the *population* proportion of telemarketing companies that use their telemarketing operation to assist them in order processing?

Solution

The sample proportion, $\hat{p} = .39$, is the *point estimate* of the population proportion, p. For $n = 87$ and $\hat{p} = .39$, a 95% confidence interval can be computed to determine the interval estimation of p. The z value for 95% confidence is 1.96. The value of $\hat{q} = 1 - \hat{p} = 1 - .39 = .61$. The confidence interval estimate is

$$.39 - 1.96\sqrt{\frac{(.39)(.61)}{87}} \le p \le .39 + 1.96\sqrt{\frac{(.39)(.61)}{87}}$$

$$.39 - .10 \le p \le .39 + .10$$

$$.29 \le p \le .49$$

Example

Coopers & Lybrand surveyed 210 chief executives of fast-growing small companies. Only 51% of these executives had a management succession plan in place. A spokesperson for Cooper & Lybrand said that many companies do not worry about management succession unless it is an immediate problem. However, the unexpected exit of a corporate leader can disrupt and unfocus a company for long enough to cause it to lose its momentum.

Use the data given to compute a 92% confidence interval to estimate the proportions

Solution

The point estimate is the sample proportion given to be .51. It is estimated that .51, or 51% of all fast-growing small companies have a management succession plan. Realizing that the point estimate might change with another sample selection, we calculate a confidence interval.

The value of n is 210; $\hat{p}$ is .51, and $\hat{q} = 1 - \hat{p} = .49$. Because the level of confidence is 92%, the value of $z_{.04} = 1.75$. The confidence interval is computed as

$$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$.51 - 1.75\sqrt{\frac{(.51)(.49)}{210}} \le p \le .51 + 1.75\sqrt{\frac{(.51)(.49)}{210}}$$

$$.51 - .06 \le p \le .51 + .06$$

$$.45 \le p \le .57$$

Home Work Problems

Question :

Car mufflers are constructed by nearly automatic machine. One manufacturer finds that, for any type of car muffler, the time for a person to set up and complete a production run has a normal distribution with mean 1.82 hours and standard deviation 1.20.

What is the probability that the sample mean of the next 40 runs will be from 1.65 to 2.04 hours ?

Question :

Engine bearings depend on a film of oil to keep shaft and bearing surfaces separated. Insufficient lubrication causes bearings to be overloaded. The insufficient lubrication can be modeled as a random variable having a mean 0.6520 ml and standard deviation 0.0125 ml.

The sample mean of insufficient lubrication will be obtained frm a random sample of 60 bearings.

What is the probability that sample mean $\bar{x}$ will be between 0.600 ml and 0.640 ml ?

Question :

A random sample size of $n = 100$ is taken from a population with $\sigma = 5.1$.

Given that the sample mean is $\bar{x} = 2.16$,

construct a 95% confidence interval for the population mean μ.

Question :

With reference to the data in section 2.1 (of R1) , we have

n = 50 , $\overline{x}$ = 305.58 nm, and s2 = 1366.86(hence, s=36.97 nm),

Construct a 99% confidence interval for the population mean of all nanopillars.

*

245	333	296	304	276	336	289	234	253	292
366	323	309	284	310	338	297	314	305	330
266	391	315	305	290	300	292	311	272	312
315	355	346	337	303	265	278	276	373	271
308	276	364	390	298	290	308	221	274	343

Exercise

A survey was taken of U.S. companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India?

A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years.

Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of U.S. companies trading with firms in India.

Solution

Here, n= 44, $\bar{x}$= 10.455 and σ= 7.7. To determine the value of $z_{\alpha/2}$, divide the 90% confidence in half, or take .5000 - α/2= .5000 - .0500= 0.45 where α= 10%.

Z table yields a z value of 1.645 for the area of .45

The confidence interval is

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

$$10.455 - 1.645\frac{7.7}{\sqrt{44}} \le \mu \le 10.455 + 1.645\frac{7.7}{\sqrt{44}}$$

$$10.455 - 1.910 \le \mu \le 10.455 + 1.910$$

$$8.545 \le \mu \le 12.365$$

Exercise

A clothing company produces men's jeans. The jeans are made and sold with either a regular cut or a boot cut. In an effort to estimate the proportion of their men's jeans market in Oklahoma City that prefers boot-cut jeans, the analyst takes a random sample of 212 jeans sales from the company's two Oklahoma City retail outlets. Only 34 of the sales were for boot-cut jeans.

Construct a 90% confidence interval to estimate the proportion of the population in Oklahoma City who prefer boot-cut jeans.

Solution

The sample size is 212, and the number preferring boot-cut jeans is 34. The sample proportion is $\hat{p} = 34/212 = .16$. A point estimate for boot-cut jeans in the population is .16, or 16%. The z value for a 90% level of confidence is 1.645, and the value of $\hat{q} = 1 - \hat{p} = 1 - .16 = .84$. The confidence interval estimate is

$$.16 - 1.645\sqrt{\frac{(.16)(.84)}{212}} \le p \le .16 + 1.645\sqrt{\frac{(.16)(.84)}{212}}$$

$$.16 - .04 \le p \le .16 + .04$$

$$.12 \le p \le .20$$

Thank You

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session No: 9
Testing of Hypothesis
(5th /6th Aug ,2023)

Need for testing of hypothesis

Often the decisions are made based on samples estimates to generalize on population parameter (as described in sampling and estimation).

In this process, there may be a difference between the estimate and the parameter which needs to be examined.

$$|\text{Estimate-Parameter}| = \begin{cases} 0 \\ Small \\ Large \end{cases}$$

Need for testing of hypothesis

Case(i):

If the difference is zero, it is called unbiased

Case(ii):

If the difference is small, it may due to chance or sampling error (improper sampling technique used leads to sampling error)

Case(iii):

If the difference is large, it may a real one or due to sampling error (improper sampling technique used leads to sampling error)

Hence, there is a need to test what type of difference is between estimate and parameter.

Hypothesis

A statement which is yet to be proved/ established or a statement on the parameter(s) of the Probability distribution to be tested

Null Hypothesis

Hypothesis of no difference or neutral or may be due to Sampling variation

Alternative Hypothesis

Hypothesis of difference which is yet to be proved/ established

Types of test

A statistical rule which decides whether to accept or reject the null hypothesis on the basis of data

Parametric tests

Based on the assumption of some probability distribution

Non-parametric tests

Not based on any assumption of probability distribution

Parametric tests

It is assumed that the data do follow some probability. distribution which is characterized by any parameters.

Large Sample Test → $n \geq 30$

Standard Normal Test → Z-Test

Small Sample Test → $n < 30$

Student's t-test → Unpaired t-Test, Paired t-Test

Analysis of Variance → ANOVA, Rm ANOVA

Non - Parametric tests

It is assumed that the data do not follow any probability distribution which is not characterized by any parameters.

Distribution - free tests

- **Chi - Square Test**
- **Mann – Whitney U test**
- **Kruskal - WallisTest**
- **Fisher's exact probabilities**
- **Wilcoxon Signed Rank Test**
- **Friedman'sTest**

Steps involved in Testing of Hypothesis

State null and alternative hypotheses

↓

Specify the level of significance 'α'

↓

Define the probability distribution the data follows

↓

Compute the test statistic based defined population

↓

Define the rejection criteria/critical regional

↓

Conclusion

Errors in Hypothesis Testing

In hypothesis testing, there are two types of errors.

Type I error: A type I error occurs when we incorrectly reject H_0 (i.e., we reject the null hypothesis, when H_0 is true).

Type II error: A type II error occurs when we incorrectly fail to reject H_0 (i.e., we accept H_0 when it is not true).

	Observation	
Decision	H_0 is true	H_0 is false
H_0 is accepted	Decision is correct	Type II error
H_0 is rejected	Type I error	Decision is correct

Errors in decision making

Any example based on data		
Statistical Example		
Decision	Null Hypothesis (H_0)	
	True	False
Accept		
Reject		

Any example based on data		
Statistical Example		
Decision	Null Hypothesis (H_0)	
	True	False
Accept		Type – II Error
Reject	Type – I Error	

Errors in decision making

Any example based on data		
Statistical Example		
Decision	Null Hypothesis (H_0)	
	True	False
Accept		β - error
Reject	α- Error	

Any example based on data		
Statistical Example		
Decision	Null Hypothesis (H_0)	
	True	False
Accept	Confidence Level (1-α)	
Reject		Power (1-β)

Decision on α –error and β - error

P – value: The strength of the evidence against the null hypothesis that the true difference in the population is zero

P-value

If the P-value is less than 1% (< 0.01),

→ Overwhelming evidence that supports the alternative hypothesis

If the P-value is between 5% and 10%,

→ Weak evidence that supports the alternative hypothesis

If the P-value is between 1% and 5%,

→ Strong evidence that supports the alternative hypothesis

If the P-value exceeds 10%,

→ No evidence that supports the alternative hypothesis.

Univariate analysis

Dependent Variable

	Nominal	Ordinal	Interval/ Ratio
Statistical measure	Proportion	Median	Mean and Variance
Large sample tests ($n \geq 30$)	Z - test	Sign test	Z-test (If normality assumption holds and σ^2 known) t-test (If normality assumption holds and σ^2 unknown)
Small sample tests ($n < 30$)	Binomial test	Sign test	t-test (If normality assumption holds and σ^2 unknown)

Multivariate analysis

Dependent Variable

			Nominal	Ordinal	Interval/ Ratio
Statistical measure			Proportion	Median	Mean and Variance
Large sample tests ($n \geq 30$)	Independent variable	2 groups	Z - test	Mann-Whitney U-test	Z-test (If normality assumption holds and σ^2 known) t-test (If normality assumption holds and σ^2 unknown)
		> 2 groups	Chi-square test	Kruskal-Wallis test	ANOVA (If normality assumption holds)
Small sample tests ($n < 30$)	Independent variable	2 groups	Chi-square test	Mann-Whitney U-test	Z-test (If normality assumption holds and σ^2 known) t-test (If normality assumption holds and σ^2 unknown)
		> 2 groups	Chi-square test	Kruskal-Wallis test	ANOVA (If normality assumption holds)

Parametric tests

Z-test → This is a test based on Standard Normal Distribution

Used for testing the

- Mean of a single population (μ)
- Difference between means of two populations ($\mu_1 - \mu_2$)
- Proportion of a single population (P)
- Difference between proportions of two populations ($P_1 - P_2$)

Assumptions on Z-test

- Samples are drawn from normal distribution
- The population variances should be known

Z-test

- Two groups should be independent
- Subjects should be allocated randomly to both groups
- **The sample size should be more than 30 (i.e., $n \geq 30$)**

Testing mean of a single population

1. State null and alternative hypothesis
2. Specify the level of significance 'α'
3. Standard Normal Distribution
4. Compute the test statistic
5. Define the critical region/ rejection criteria
6. Conclusion

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu < \mu_0$$
$$\text{or } H_1 : \mu > \mu_0$$
$$\text{or } H_1 : \mu \neq \mu_0$$

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \cong N(0,1)$$

Rejection criteria

Summary of One- and Two-Tail Tests

One-Tail Test (left tail)	One-Tail Test (right tail)	Two-Tail Test (Either left or right tail)
$H_0: \mu = \mu_0$ vs $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ vs $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$

Example:1

- It is claimed that sports-car owners drive on the average 17000 kms per year. A consumer firm believes that the average milage is probably higher. To check, the consumer firm obtained information from randomly selected 40 sports-car owners that resulted in a sample mean of 17352 kms with a population standard deviation of 1348 kms. At what can be concluded about this claim at
- 5% level of significance (Critical value is 1.645)
- 1% level of significance (Critical value is 2.331)

H_0 ⇨ The average milage of sports-car as claimed and the sample average milage may be same

$$H_0 : \mu = \mu_0 = 17000$$

H_1 ⇨ The average milage of sports-car as claimed may be higher than the sample average milage

$$H_1 : \mu > \mu_0 = 17000$$

(a) At 5% level of significance with critical value 1.645

$$Z = \frac{17352 - 17000}{\frac{1348}{\sqrt{40}}} = 1.650$$

Critical value for

$\alpha = 0.05$ is 1.645

Since $Z = 1.650 > 1.645$,

Reject H_0 and Accept H_1

Hypothesis to test

$H_0: \mu = 17000$

vs

$H_1: \mu > 17000$

(b) At 1% level of significance with critical value 2.331

$$Z = \frac{17352 - 17000}{\frac{1348}{\sqrt{40}}} = 1.650$$

Hypothesis to test

H_0:μ=17000

vs

H_1:μ>17000

Critical value for

α = 0.01 is 2.331

Since Z =1.624 < 2.330,

Accept H_0 Reject H_1

Example:2

The manager of a courier service believes that packets delivered at the beginning of the month are heavier than those delivered at the end of month. As an experiment, he weighed a random sample of 20 packets at the beginning of the month and found that the mean weight was 5.25 kg. A randomly selected 10 packets at the end of the month had a mean weight of 4.96 kg. It was observed from the past experience that the population variances are 1.20 kg and 1.15 kg. At 5% level of significance, can it be concluded that the packets delivered at the beginning of the month weigh more? Also find P-value and 95% confidence interval for the difference between the means.

H_0 ⇨ The mean weight of packets delivered at the early in the month and at the end of month may be same

$$H_0 : \mu_1 = \mu_2$$

H_1 ⇨ The mean weight of packets delivered at the end of the month may be higher than at the early of month

$$H_1 : \mu_1 > \mu_2$$

At 5% (0.05) level of significance with critical value 1.645

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{5.25 - 4.96}{\sqrt{\frac{(1.20)^2}{20} + \frac{(1.15)^2}{10}}} = 0.642$$

$$P(Z < 0.642) = 0.7389$$

Hypothesis to test

$H_0: \mu_1 - \mu_2 = 0$

vs

$H_1: \mu_1 - \mu_2 > 0$

???

95% CI for μ is [- 454, 1.033]

Critical value for α = 0.05 is 1.645. Since Z = 0.642 < 1.645, Accept H_0 Reject H_1

Z-test for proportion of a single population

Z-test → Proportion of a single population (P)

Assumptions

- Assume that the samples are drawn from normal distribution
- The sample size should be more than or equal to 30
- Subjects should be selected randomly

1. State null and alternative hypothesis

$$H_0 : P = P_0 \text{ vs } H_1 : P < P_0$$
$$\text{or } H_1 : P > P_0$$
$$\text{or } H_1 : P \neq P_0$$

2. Specify the level of significance 'α'

3. Standard Normal Distribution

4. Compute the test statistic

$$Z = \frac{p - P_0}{\sqrt{\frac{pq}{n}}} \cong N(0,1)$$

5. Define the critical region/ rejection criteria

6. Conclusion

Note: Rejection criteria same as in one sample test or P-value

Example:

A researcher claims that Republican Party will win in the next Senate elections, especially in Florida State. Statistical data reported that 23% voted for Republican Party in the last election. To test the claim a researcher surveyed 80 people and found 22 said they voted for Republican Party in the last election. Is there enough evidence at $\alpha=0.05$ to support this claim?

- $P_0=0.23$
- $n=80$
- $\hat{p}=22/80=0.275$

Then, define the Null and Alternative hypothesis

- Null Hypothesis: $p= 0.23$
- Alternative Hypothesis: $p\neq0.23$

Then, calculate the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

$$z = \frac{0.275 - 0.23}{\sqrt{0.23 * 0.77/80}}$$

= 0.045 / 0.047 = 0.957

Since the calculated value is between -1.96 and 1.96 and it is not in the critical region, hence failed to reject the null hypothesis

Z-test for difference between proportions

1. State null and alternative hypothesis

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 < P_2$$
$$\text{or } H_1 : P_1 > P_2$$
$$\text{or } H_1 : P_1 \neq P_2$$

2. Specify the level of significance 'α'
3. Standard Normal Distribution
4. Compute the test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \cong N(0,1)$$

5. Define the critical region/ rejection criteria
6. Conclusion

Note: Rejection criteria same as in one sample test or P-value

Testing of Hypothesis ➡ Z-test for proportions of two popln.

Based on sample size standard error for proportion may be calculated:

If the sample sizes are equal then SE (P_1-P_2) is calculated by

$$SE(p_1 - p_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Based on sample size standard error for proportion may be calculated:

If the sample sizes are equal then SE (P_1-P_2) is calculated by

$$SE(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, p = \frac{x_1 + x_2}{n_1 + n_2} \text{ and } q = 1 - p$$

Testing of Hypothesis ➡ Z-test for proportions of two popln.

A cigarette manufacturing company claims that its brand A cigarettes outsells its brand B cigarettes by 8%. If it is found that 42 out of a random sample of 200 smokers prefer brand A and 18 out of 100 smokers prefer brand B, test at 5% level of significance, whether 8% difference a valid claim. Also construct 95% CI for (P_1-P_2) and find P-value.

Testing of Hypothesis ➡ Z-test for proportions of two popln.

H_0 ➡ **There may be 8% difference in the sale of two brands of cigarettes may be a valid claim**

$$H_0 : P_1 - P_2 = 0.08$$

H_1 ➡ **There may be 8% difference in the sale of two brands of cigarettes may not be a valid claim**

$$H_1 : P_1 - P_2 \neq 0.08$$

Finding Confidence Interval for difference in population proportion ($P_1 - P_2$)

The100 (1-α)% confidence interval for difference between two population proportions ($P_1 - P_2$)

$$(p_1 - p_2) \pm Z_{\alpha/2}\, SE\,(p_1 - p_2)$$

Estimation ➡ Interval estimate

95% confidence interval for difference between two population proportions ($P_1 - P_2$)

$$(p_1 - p_2) \pm Z_{\alpha/2}\ SE\ (p_1 - p_2) = (0.21 - 0.18) \pm 1.96 * 0.049$$
$$= (-0.066,\ 0.126)$$

Testing of Hypothesis ➡ Z-test for proportions of two popln.

At 5% level of significance with critical value ±1.96

$$Z = \frac{(0.21 - 0.18) - 0.08}{\sqrt{0.2 * 0.8\left(\frac{1}{200} + \frac{1}{100}\right)}} = -1.02$$

Hypothesis to test

$H_0: P_1 - P_2 = 0.08$

vs

$H_1: P_1 - P_2 \neq 0.08$

$P(Z < -2.074) = 0.1539$

Critical value for $\alpha = 0.05$ is -1.96

Since Z = - 1.02 > -1.96, Don't reject H_0 and reject H_1

$Z_{cal} = -1.02$

$Z_{0.025} = -1.96$

95% CI for P is (-0.066, 0.126)

Testing of Hypothesis ➡ Z-test for proportions of two popln.

Problem on Z-test on difference in proportions

Use α = 1%.

Bangalore 2019

387/1500 BSNL telephones connections were disconnected

Delhi 2019

310/1200 BSNL telephones connections were disconnected

- **Is there a significant difference between the two proportion of telephones disconnection by BSNL?**

Construct 99% confidence interval for difference in proportions

Testing of Hypothesis ➡ Student's t-test

t-test ➡ Testing mean of single population (μ)

- Samples are drawn from normal population
- The population variance should be unknown
- The sample size should be less than 30 (i.e., n < 30)

t-test

- Sample should be allocated randomly
- However even if sample size more than 30 (i.e., n > 30) and population variance unknown, t-test should be continue to apply, because of central limit theorem it approaches normal.

t-test → Degrees of freedom (df): No. of independent observations

Suppose

a+b = 20. If we assign a=9 then b=11 or vice-versa. ∴ df=(2-1)=1

a+b+c = 20. If we assign a=9 and b=6 then c=5. ∴ df=(3-1)=2

In general, if there are n observations df = n-1

1. State null and alternative hypothesis

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu < \mu_0$$
$$\text{or } H_1 : \mu > \mu_0$$
$$\text{or } H_1 : \mu \neq \mu_0$$

2. Specify the level of significance 'α'

3. Student's t-distribution

4. Compute the test statistic

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \cong t_{(\alpha, n-1)}$$

5. Define the critical region/ rejection criteria

6. Conclusion

Note: Rejection criteria may be based on critical value or P-value

5 Define the critical region/ rejection criteria

(i) Reject H_0, if computed value of t is less than the critical value, ie., $P(t < -t_\alpha)$, otherwise do not reject H_0

(ii) Reject H_0, if computed value of t is greater than the critical value, ie., $P(t > t_\alpha)$, otherwise do not reject H_0

By combining both (i) and (ii), Reject H_0, if computed value of $|t|$ is greater than the critical value, ie., $P(|t| > t_\alpha)$, otherwise do not reject H_0. Besides α, the df is also important.

Conclusion

5 Define the critical region/ rejection criteria

(iii) Reject H_0, if computed value of t is less than or greater than the critical value, ie., $P(t < -t_{\alpha/2})$ or $P(t > t_{\alpha/2})$, otherwise do not reject H_0

Alternatively, reject H_0, if computed value of $|t|$ is greater than the critical value, ie., $P(|t| > t_{\alpha/2})$, otherwise do not reject H_0. Besides α, the degrees of freedom is also important.

Conclusion

Example

It is claimed that sports-car owners drive on the average 18580 kms per year. A consumer firm believes that the average milage is probably higher. To check, the consumer firm obtained information from randomly selected 10 sports-car owners that resulted in a sample mean of 17352 kms with a sample standard deviation of 2012 kms. What can be concluded about this claim at

- 5% level of significance
- 1% level of significance

H_0 ⇨ The average milage of sports-car as claimed and the sample average milage may be same

$$H_0 : \mu = \mu_0 = 18580$$

H_1 ⇨ The average milage of sports-car as claimed may be higher than the sample average milage

$$H_1 : \mu > \mu_0 = 18580$$

(a) At 5% level of significance with critical value 1.645

$$|t| = \frac{|17352 - 18580|}{\frac{2012}{\sqrt{10}}} = 1.929$$

95% CI for μ is

$$\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = [16184.91, 18519.09]$$

$$P - value = 0.0428$$

Hypothesis to test

H_0:μ=18580 vs H_1:μ>18580

Critical value for α = 0.05 is 1.833 for 9 degree of freedom

Since | t | =1.929 > 1.833, Reject H_0 and Accept H_1

Testing the difference between means

1. State null and alternative hypothesis

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 < \mu_2$$
$$\text{or } H_1 : \mu_1 > \mu_2$$
$$\text{or } H_1 : \mu_1 \neq \mu_2$$

2. Specify the level of significance 'α'

3. Standard Normal Distribution

4. Compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \cong t_{(\alpha, n_1+n_2-2)}$$

Note: If sample sizes are unequal compute pooled SE

5. Define the critical region/ rejection criteria

6. Conclusion

Note: Rejection criteria may be based on critical value or P-value

t-test → Difference between means of two populations ($\mu_1 - \mu_2$)

- Samples are drawn from normal populations
- The population variances should be unknown
- The sample size should be less than 30 (i.e., $n < 30$)

t-test

- The population variances should be equal
- Two groups shouldbe independent
- Subjects should be allocated randomly to both groups
- However even if sample size more than 30 (i.e., $n > 30$) and population variances unknown, t-test should be continue to apply, because of central limit theorem it approaches normal.

Example

Random samples of 15 and 10 were selected from two thermocouples. The sample means were 315, 303 and sample standard deviations were 3.8, 4.9 respectively.

- Construct 95% CI for difference in means
- Test whether there is any significant difference in the means of two thermocouples at 5% level of significance
- Find the P-value

H_0 ⇨ The mean of two thermocouples may be same

$$H_0 : \mu_1 = \mu_2$$

H_1 ⇨ The mean of two thermocouples may be different

$$H_1 : \mu_1 \neq \mu_2$$

At 5% (0.05) level of significance with critical value

$$|t| = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{315 - 303}{\sqrt{\frac{(3.8)^2}{15} + \frac{(4.9)^2}{10}}} = 3.571$$

Hypothesis to test

$H_0: \mu_1 - \mu_2 = 0$

vs

$H_1: \mu_1 - \mu_2 > 0$

???

95% CI for μ is [6.24, 17.76] not includes 0

95% CI for μ is [6.24, 17.76]

Critical value for $\alpha = 0.05$ is 1.714. Since $|t| = 3.571 > 1.714$, Reject H_0 & Accept H_1

PROBLEM:

The manager of a courier service believes that packets delivered at the beginning of the month are heavier than those delivered at the end of month. As an experiment, he weighed a random sample of 15 packets at the beginning of the month and found that the mean weight was 5.25 kg. A randomly selected 10 packets at the end of the month had a mean weight of 4.56 kg. It was observed from the past experience that the sample variances are 1.20 kg and 1.15 kg.

- At 5% level of significance, can it be concluded that the packets delivered at the beginning of the month weigh more?
- Also find P-value and 95% confidence interval for the difference between themeans.

H_0

The mean weight of packets delivered at the early in the month and at the end of month may be same

$$H_0 : \mu_1 = \mu_2$$

H_1

The mean weight of packets delivered at the early in the month may be higher than at the end of month

$$H_1 : \mu_1 > \mu_2$$

Estimation ➡ Confidence interval for (μ_1-μ_2) based t-test

Finding Confidence Interval for difference between two population means (μ_1 -μ_2)

The100 (1-α)% confidence interval for difference between two means

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\ SE\ (\bar{x}_1 - \bar{x}_2)$$

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\, SE(\bar{x}_1 - \bar{x}_2) \le (\mu_1 - \mu_2) \le (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\, SE(\bar{x}_1 - \bar{x}_2)$$

95% Confidence Interval for difference between two population means (μ_1 - μ_2)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\, SE\,(\bar{x}_1 - \bar{x}_2) = (5.25 - 4.26) \pm 2.069 * 0.443$$

$$= (0.073, 1.907)$$

At 5% (0.05) level of significance with critical value 1.714

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{5.25 - 4.26}{0.443} = 2.233$$

$$0.025 \le P \le 0.01$$

($\mu_1 - \mu_2$) =0 not included in

95% CI for $\mu_1 - \mu_2$ is (0.073, 1.907)

Hypothesis to test

$H_0: \mu_1 - \mu_2 = 0$

vs

$H_1: \mu_1 - \mu_2 > 0$

???

t_{cal}= 2.233

Z_α= 1.714

Critical value for α = 0.05 is 1.714. Since t = 2.233 > 1.714, Reject H_0, Don't reject H_1

Student's paired t-test

1. State null and alternative hypothesis
2. Specify the level of significance 'α'
3. Student's t-distribution
4. Compute the test statistic
5. Define the critical region/ rejection criteria
6. Conclusion

$$H_0 : \mu_d = 0 \text{ vs } H_1 : \mu_d < 0$$
$$\text{or } H_1 : \mu_d > 0$$
$$\text{or } H_1 : \mu_d \neq 0$$

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \cong t_{(\alpha, n-1)}$$

But μ_d under H_0 will be 0

Note: Rejection criteria may be based on critical value or P-value

Student's paired t-test

- The HRD manager wishes to see if there has been any change in the ability of trainees after a specific training programme.
- The trainees take a aptitude test

Before and after training programme.

Subjects	Before (x)	After (y)
1	75	70
2	70	77
3	46	57
4	68	60
5	68	79
6	43	64
7	55	55
8	68	77
9	77	76

Subjects	Before (x)	After (y)	d = y-x	(d-mean)^2
1	75	70	-5	100
2	70	77	7	4
3	46	57	11	36
4	68	60	-8	169
5	68	79	11	36
6	43	64	21	256
7	55	55	0	25
8	68	77	9	16
9	77	76	-1	36
Total			45	678

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{45}{9} = 5$$

$$S_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - d)^2}{n-1}}$$

$$S_d = \sqrt{\frac{678}{8}} = 9.21$$

At 5% (0.05) level of significance with critical value is 3.31

$$|t| = \frac{\bar{d} - \mu_d}{\frac{S_d}{\sqrt{n}}} = \frac{5-0}{\frac{9.21}{\sqrt{9}}} = 3.07$$

Hypothesis to test

$H_0: \mu_d = 0$

vs

$H_1: \mu_d > 0$

???

95% CI for μ is [-2.52, 12.52]

95% CI for μ is [[-2.52, 12.52] not includes 0

Critical value for $\alpha = 0.05$ is 1.895. Since $|t| = 1.63 < 2.31$, Accept H_0 & Reject H_1

Exercise

Diet-modification Program

Ten individuals have participated

Subject	1	2	3	4	5	6	7	8	9	10
Weight Before	195	213	247	201	187	210	215	246	294	310
Weight After	187	195	221	190	175	197	199	221	278	285

Is there sufficient evidence to support claim that this program is effective in reducing weight?

Use $\alpha = 0.05$.
Construct 95% confidence interval for mean difference.

Test 1

Test 2

Is there sufficient evidence to conclude that both tests give the same mean impurity level

Specimen	1	2	3	4	5	6	7	8
Test 1	1.2	1.3	1.5	1.4	1.7	1.8	1.4	1.3
Test 2	1.4	1.7	1.5	1.3	2.0	2.1	1.7	1.6

Using $\alpha = 0.01$

Construct 99% confidence interval for mean difference

Problem-1

The target thickness for silicon wafers used in a certain type of integrated circuit is 245mm. A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18mm and a sample standard deviation of 3.60mm.

Does this data suggest that true average wafer thickness is something other than the target value?

Problem 2

In a survey of buying habits. 400 women shopper's are chosen at random in super Market A located in a certain city. Their average weekly food expenditure is Rs.250 with a SD of Rs.40. For 400 women shopper's are chosen at random in super Market B in another city. The average weekly food expenditure is Rs.220 with a SD of Rs.55.

Test at 1% LOS whether the average weekly food expenditure of the two populations of shoppers are equal.

Problem 3

- Two types of new cars produced in INDIA are tested for petrol mileage. One sample consisting of 42 cars averaged 15 kmpl while the other sample consisting of 80 cars averaged 11.5 kmpl with population variance as ${\sigma_1}^2 = 2$ and ${\sigma_2}^2 = 1.5$ respectively.

- Test whether there is any significance difference in the petrol consumption of these two types of cars.

Problem 4

Studying the flow of traffic at two busy intersections between 4pm and 6pm to determine the possible need for turn signals. It was found that on 40 week days there were on the average 247.3 cars approaching the first intersection from the south which made left turn, while on 30 week days there were on the average 254.1 cars approaching the second intersection from the south made left turns. The corresponding sample SD are 15.2 and 12.Test the significance between the difference of two means at 5% level.

Problem 5

- Is there any systematic tendency for part-time college faculty to hold their students to different standards than do full-time faculty? The article "Are There Instructional Differences Between Full-Time and Part-Time Faculty?" (College Teaching, 2009: 23–26) reported that for a sample of 125 courses taught by full-time faculty, the mean course GPA was 2.7186 and the standard deviation was .63342, whereas for a sample of 88 courses taught by part-timers, the mean and standard deviation were 2.8639 and .49241, respectively.

- Does it appear that true average course GPA for part-time faculty differs from that for faculty teaching full-time? Test the appropriate hypotheses at significance level 1%.

Problem 6

It is thought that the front cover and the nature of the first question on mail surveys influence the response rate. The article "The Impact of Cover Design and First Questions on Response Rates for a Mail Survey of Skydivers" (*Leisure Sciences*, 1991: 67–76) tested this theory by experimenting with different cover designs. One cover was plain; the other used a picture of a skydiver. The researchers speculated that the return rate would be lower for the plain cover.

Cover	Number Sent	Number Returned
Plain	207	104
Skydiver	213	109

Does this data support the researchers' hypothesis?

Problem 7

A spectrophotometer used for measuring CO concentration [ppm (parts per million) by volume] is checked for accuracy by taking readings on a manufactured gas (called span gas) in which the CO concentration is very precisely controlled at 70 ppm. If the readings suggest that the spectrophotometer is not working properly, it will have to be recalibrated.

Assume that if it is properly calibrated, measured concentration for span gas samples is normally distributed.

On the

basis of the six readings—85, 77, 82, 68, 72, and 69—is recalibration necessary?

Chi-square Statistic

- Although the technique is still rather widely presented as a mechanism for constructing confidence intervals to estimate a population variance, you should proceed with extreme caution and **apply** the technique **only** in cases **where** the **population** is known to be **normally distributed**. We can say that this technique lacks robustness.

χ^2 FORMULA FOR SINGLE VARIANCE (8.5)

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$$df = n - 1$$

Three Chi-Square Distributions

Chi-square test

Chi-square Test

- **Independence**
- **Goodness-of-fit**

- Should be applied ONLY for Frequencies
- Not for percentages, ratios, mean etc.

Testing of Hypothesis ➡ Chi-square test: Independence

Based on attributes used to test

(a) INDEPENCE of two different categorical variables

or

(b) GOODNESS OF FIT

Caution:

Should be applied ONLY for FREQUENCIES not for percentages, ratios, mean etc

Hypothesis for testing independence

The hypothesis to be tested for independence will be

H_0: The two categorical variables may be independent (may not be associated)

H_1: The two categorical variables may not be independent (may be associated)

Procedure for testing independence

To check the independence (no association) between the two categorical variables, the statistical test used is Chi-square test given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, k = r \text{ x } c \text{ \# of cells}$$

The test-statistic follows Chi-square distribution with (r-1)(c-1) degrees of freedom. r = # of rows, c = # of columns

Expected frequencies

$$E_{ij} = \frac{r_i c_j}{n},$$

$$\text{for } i = 1, 2, \ldots, m; \; j = 1, 2, \ldots, n$$

Chi-square is calculated by

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2_{[(r-1)(c-1)]}$$

where k = r x c is the total number of cells in the r x c contingency table, r = total no. of rows and c is total no. of columns.

Testing of Hypothesis ➡ Chi-square test: Independence

Assumptions of Chi-square test

If the expected cell frequencies is < 5

→ Yate's correction should be applied for continuity

In a 2 x 2 contingency table, if one or more of the cell has the expected cell frequencies is < 5,

→ Fisher's exact probabilities should be computed

For the use of Chi-square test

→ The sample size should not be less than 20.

The Fisher's exact Probability

$$P = \frac{1}{n!}\frac{r_1!}{a!}\frac{r_2!}{b!}\frac{c_1!}{c!}\frac{c_2!}{d!}$$

For an r x c table, if the expected frequencies in any cells are < 5, merge the rows are columns meaningfully

Testing of Hypothesis ➡ Chi-square test: Independence

A study to find the association between smoking and ca. lung has revealed the following data? Find is there any association exists between smoking and ca. lung?

Smoking	Carcinoma of lung		Total
	Present	Absent	
Smokers	69	2431	2500
Non-smokers	24	1476	1500
Total	93	3907	4000

Calculation of expected frequencies

$$E_1 = \frac{c_1 r_1}{n} = \frac{93*2500}{4000} = 58.125$$

$$E_2 = \frac{c_2 r_1}{n} = \frac{3907*2500}{4000} = 2441.875$$

$$E_3 = \frac{c_1 r_2}{n} = \frac{93*1500}{4000} = 34.875$$

$$E_4 = \frac{c_2 r_1}{n} = \frac{3907*1500}{4000} = 1465.125$$

Testing of Hypothesis ➡ Chi-square test: Independence

Calculation of Chi-square statistic - $\chi 2$

Sl No	Observed frequencies (O_i)	Expected frequencies (E_i)	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	69	58.125	10.875	118.266	2.035
2	2431	2441.875	-10.875	118.266	0.048
3	24	34.875	-10.875	118.266	3.391
4	1476	1465.125	10.875	118.266	0.081
Total	4000	4000			$\chi^2 = 5.555$

Testing of Hypothesis ➡ Chi-square test: Independence

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21

Calculation of P – value

$$\frac{6.63 - 3.84}{5.56 - 3.84} = \frac{0.05 - 0.01}{P - 0.01}$$

$$P = 0.01 + \frac{(0.05 - 0.01) * (5.56 - 3.84)}{(6.63 - 3.84)} = \mathbf{0.035}$$

Interpretation

H_0: Smoking habit and Cancer of lung may be independent (may not be associated)

H_1: Smoking habit and Cancer of lung may not be independent (may be associated)

$\chi^2 = 5.555$

df = 1, Critical value at $\alpha = 0.05$ is 3.841, P = 0.035

Inference: There may be an association between smoking and Cancer of lung

By replacing O_1, O_2, O_3, and O_4, by a, b, c, and d the 2 x 2 contingency table can also be written as

Categorical variable 1	Categorical variable 2		Total
	Response 1	Response 2	
Response 1	**a**	**b**	r_1
Response 2	**c**	**d**	r_2
Total	c_1	c_2	n

Alternate formula for calculation of chi-square statistic

When the expected frequencies in all the cells are more than 5, alternatively Chi-square statistic can be calculated using the formula for 2 x 2 table only by

$$\chi^2 = n\frac{(ad - bc)^2}{r_1 r_2 c_1 c_2}$$

Where a, b, c, d are cell frequencies; r_1 and r_2 are row totals; c_1 and c_2 are column totals

Testing of Hypothesis ➡ Chi-square test: Independence

To assess the length of hospital stay and the type of insurance, data were taken on 70 individuals

Type of Insurance	Length of Hospital Stay (days)		Total
	≤10	>10	
Type 1	42	3	45
Type 2	18	7	25
Total	60	10	70

Examine whether Chi-square test can be applied to this data to test the independence between type of insurance and length of hospital stay?

Calculation of expected frequencies

$$E_1 = \frac{c_1 r_1}{n} = \frac{60*45}{70} = 38.75$$

$$E_2 = \frac{c_2 r_1}{n} = \frac{10*45}{70} = 6.43$$

$$E_3 = \frac{c_1 r_2}{n} = \frac{60*25}{70} = 21.43$$

$$E_4 = \frac{c_2 r_1}{n} = \frac{10*25}{70} = 3.57$$

Testing of Hypothesis ➡ Chi-square test: Independence

Calculation of Chi-square statistic - $\chi 2$

Sl No	Observed frequencies (O_i)	Expected frequencies (E_i)	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	42	38.57	-	-	-
2	3	6.43	-	-	-
3	18	21.43	-	-	-
4	7	3.57	-	-	-
Total	70	70			???

Since the expected frequency in the 4 is less than 5 Chi-square cannot be applied and hence the Fisher's exact probabilities has to be calculated.

Testing of Hypothesis ➡ Chi-square test: Independence

H_0: Type of insurance plan and length of hospital stay may be independent

H_1: Type of insurance plan and length of hospital stay may be associated

Computed Probability

$$P = \frac{1}{70!}\frac{45!}{42!}\frac{25!}{3!}\frac{60!}{18!}\frac{10!}{7!}$$

One tailed: P=0.0201 Two tailed: P=0.0282

Conclusion: H_0 may be rejected and hence the type of insurance plan and length at may not be independent (may be associated)

Testing of Hypothesis ➡ Chi-square test: Independence

To assess the length of hospital stay and the type of insurance, data were taken on 70 individuals

Type of Insurance	Length of Hospital Stay (days)		Total
	≤10	>10	
Type 1	42	3	45
Type 2	13	12	25
Total	55	15	70

Examine whether Chi-square test can be applied to this data to test the independence between type of insurance and length of hospital stay?

Calculation of expected frequencies

$$E_1 = \frac{r_1 c_1}{n} = \frac{45 * 55}{70} = 35.36$$

$$E_2 = \frac{r_1 c_2}{n} = \frac{45 * 15}{70} = 9.64$$

$$E_3 = \frac{r_2 c_1}{n} = \frac{25 * 55}{70} = 19.64$$

$$E_4 = \frac{r_2 c_2}{n} = \frac{25 * 15}{70} = 5.36$$

$$\chi^2 = \frac{(42-35.36)^2}{35.36} + \frac{(3-9.64)^2}{9.64} + \frac{(13-19.64)^2}{19.64} + \frac{(12-5.36)^2}{5.36} = 16.307$$

Testing of Hypothesis ➡ Chi-square test: Independence

- H_0: Duration of hospital stay and type of insurance plan may be independent (not associated)
- H_1: Duration of hospital stay and type of insurance plan may not be independent (Associated)
- $\chi^2 = 16.307$
- df=1
- $P < 0.001$
- Inference: Reject H_0, which shows duration of hospital stay and type of insurance may be associated

Testing of Hypothesis ➡ Chi-square test: Independence

Hypertension	Non smokers	Moderate smokers	Heavy smokers	Total
A	O_1	O_2	O_3	r_1
B	O_4	O_5	O_6	r_2
Total	c_1	c_2	c_3	n

The expected frequencies and Chi-square statistic are computed by

$$E_1 = \frac{r_1 c_1}{n}, \quad E_2 = \frac{r_1 c_2}{n}, \quad E_3 = \frac{r_1 c_3}{n}, \quad E_4 = \frac{r_2 c_1}{n},$$

$$E_5 = \frac{r_2 c_2}{n}, \quad E_6 = \frac{r_2 c_3}{n}, \quad \text{and} \quad \chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

Testing of Hypothesis ➡ Chi-square test: Independence

Under the null hypothesis, the observed frequencies and the calculated expected frequencies will be as follows:

Hypertension	Non smokers	Moderate smokers	Heavy smokers	Total
A	O_1 E_1	O_2 E_2	O_3 E_3	r_1
B	O_4 E_4	O_5 E_5	O_6 E_6	r_2
Total	c_1	c_2	c_3	n

Testing of Hypothesis ➡ Chi-square test: Independence

Three pension plans

Independent of job classification

Use **α = 0.05**

The opinion of a random sample of 500 employees are shown below

Job Classification	Pension Plan			Total
	1	2	3	
Salaried workers	166	86	68	320
Hourly workers	84	64	32	180
Total	250	150	100	500

$$E_1 = \frac{r_1c_1}{n} = \frac{320 \times 250}{500} = 106.24$$

$$E_2 = \frac{r_1c_2}{n} = \frac{320 \times 150}{500} = 96.00$$

$$E_3 = \frac{r_1c_3}{n} = \frac{320 \times 100}{500} = 64.00$$

$$E_4 = \frac{r_2c_1}{n} = \frac{180 \times 250}{500} = 90.00$$

$$E_5 = \frac{r_2c_2}{n} = \frac{180 \times 150}{500} = 54.60$$

$$E_6 = \frac{r_2c_3}{n} = \frac{180 \times 100}{500} = 36.00$$

Testing of Hypothesis ➡ Chi-square test: Independence

Sl No	(O_i)	(E_i)	(O_i-E_i)	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
1	166	106.24	59.76	3571.26	33.62
2	86	96.00	-10.00	100.00	1.04
3	68	64.00	4.00	16.00	0.25
4	84	90.00	-6.00	36.00	0.40
5	64	54.60	9.40	88.36	1.62
6	32	36.00	-4.00	16.00	0.44
Total	180	180	Chi-square value		37.37

Testing of Hypothesis ➡ Chi-square test: Independence

- H_0: Job satisfaction and pension plan may be independently distributed (not associated)
- H_1: Job satisfaction and pension plan may not be independently distributed (Associated)
- $\chi^2 = 37.37$
- df=2
- $P<0.001$
- Inference: Reject H_0, which shows Job satisfaction and pension plan are associated

Problem

A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in second class and 20 got a first class.

Do these figures commensurate with the general examination result which is in ratio of 4:3:2:1 for the various categories respectively.

H_0: The observed results commensurate with the general examination results.

H_1: The observed results not commensurate with the general examination results.

Expected frequencies are in the ratio of 4:3:2:1

Total frequency = 500

If we divide total frequency 500 in the ratio 4:3:2:1, We get the Expected frequencies

As $200\left(500\ X\ \frac{4}{10}\right)$, $150\left(500\ X\ \frac{3}{10}\right)$,$100\left(500\ X\ \frac{2}{10}\right)$,$50\left(500\ X\ \frac{1}{10}\right)$

S. NO	(Oi)	(Ei)	(Oi-Ei)	$(O-E)^2$	$(O-E)^2$/Ei
1	220	200	20	400	2
2	170	150	20	400	2.667
3	90	100	10	100	1
4	20	50	30	900	18

The Chi-square value is

❖$\chi^2 = \sum_1^4 \frac{(O-E)^2}{E}$ = 2 + 2.667 + 1 +18

= 23.667

Dof is 4-1 = 3 and Level of significance $\propto$ = 0.05

χ^2 tab value is 7.81

χ^2 calculated = 23.667 >χ^2 tab = 7.81 at $\propto$ = 0.05 for 3 dof

❖Decision : we reject the Null hypothesis H_0 at α = 5 % LOS.

i.e, we accept H_1.

i.e, The observed results not commensurate with the general examination results

Hypothesis Testing

(Problems for Discussion)

EXAMPLE

The mean lifetime of a sample of 100 items of a product by a company is 1,560 hours with a S.D of 90 hours.

Test the hypothesis that the mean lifetime of the product is 1,580 hours.

EXAMPLE:

The mean life time of a sample of 400 fluoresent light bulbs produced by a company is found to be 1600 hours with a S.D of 150 hours.

Test the hypothesis that the mean life time of the bulbs produced is higher than 1570 hours at $\alpha = 0.01$.

Example:- Discussion

It is claimed that the mean of the population is 67 gms at 5% level of significance. Mean obtained from a random sample of size 100 is 64 with SD 3. Validate the claim.

$H_0 : \mu = 67$

$H_1 : \mu \neq 67$

$\alpha = 5\%$

population

$\mu = 64$

$n = 100$
$\bar{x} = 64$
$s = 3$

sample

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{64 - 67}{3/\sqrt{100}} = -10$$

So H_0 is rejected

Example

There is an assumption that there is no significant difference between boys and girls wrt intelligence. Tests are conducted on two groups and the following are the observations

	Mean.	SD.	Size
Girls.	75.	8.	60
Boys.	73.	10.	100

Validate this at 5% LoS

Example

An automobile tyre manufacturer claims that the average life of a particular grade of tyre is more than 20,000 km. A random sample of 16 tyres is having mean 22,000 km with a standard deviation of 5000 km.

Validate the claim of the manufacturer at 5% LoS.

Example

As a manager of finance you are assigned the task of choosing the best product in terms of life of the product.

Product A: Mean 1456 hours with SD 423, size 10

Product B: Mean 1280 hours with SD 398, size 17

Use 5% level of significance.

Example :-

From the data available, it is observed that 400 out of 850 customers purchased the groceries online.

Can we say that most of the customers are moving towards online shopping even for groceries?

Example:-

It is found 290 errors in the randomly selected 400 lines of code from Team A, and 160 errors in 300 lines of code from Team B.

Can we assume that team B's performance is superior to that of team A: ?

Example:

Following is the record of number of accidents took place during the various days of the week.

Monday	Tuesday	Wednesday	Thursday	Fri day	Sat day	Sun day
184	148	145	153	150	154	116

Can we conclude that accident is independent of the day in a week?

Example:

	with cancer	without cancer	total
Smokers	400	300	700
Non-smokers	300	500	800
total	700	800	1500

Can we conclude that smoking causes cancer?

Example

- Random samples of 15 and 10 were selected from two thermocouples. The sample means were 315, 303 and sample standard deviations were 3.8, 4.9 respectively
- Test whether there is any significant difference in the means of two thermocouples at 5% level of significance

Thanks

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session 10

(Maximum Likelihood Estimation & ANOVA)

12th and 13th Aug 2023

Maximum Likelihood Estimation (MLE)

- Method of Maximum Likelihood Estimation is the best and most popular one among all methods to obtain an almost good or best estimator for a population parameter.
- It is a method of obtaining an estimator which most (maximum) likely estimates the true value of the parameter i.e., finding an estimator that can give most likely nearer value for the unknown true value of parameter.
- The corresponding estimator is called maximum likelihood estimator (MLE).

Maximum Likelihood Estimation (MLE)

Suppose we have a random sample $x_1, x_2, .., x_n$ whose assumed probability distribution depends on some unknown parameter θ.

Ex:

1) For Binomial unknown parameters are n, p.

2) For Poisson unknown parameter is λ.

Our goal is to find good estimator of θ(population parameter) using sample and which can be done with the help of MLE.

Maximum Likelihood function

- Let $x_1, x_2, .., x_n$ be i.i.d. random variables drawn from some probability distribution that depends on some unknown parameter θ.
- The goal of MLE to maximize likelihood function

$$L(\theta) = f(x_1, x_2, \ldots, x_n \mid \theta)$$
$$= f(x_1|\theta) f(x_2|\theta) \ldots f(x_n|\theta)$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

Maximum Likelihood Estimation (MLE)

- The maximum likelihood estimate (MLE) of θ is that value of θ that maximizes likelihood(θ).

It is defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i / \theta)$$

For maximization, we have

$$\frac{dL}{d\theta} = 0 \; ; \quad \frac{d^2L}{d\theta^2} < 0$$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}, \frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Solution: Here the distribution is the binomial distribution with $n = 100$.

$$P\left(H = 61 \mid p = \frac{1}{3}\right) = \binom{100}{61}\left(\frac{1}{3}\right)^{61}\left(\frac{2}{3}\right)^{39} \approx 9.6 \times 10^{-9}$$

$$P\left(H = 61 \mid p = \frac{1}{2}\right) = \binom{100}{61}\left(\frac{1}{2}\right)^{61}\left(\frac{1}{2}\right)^{39} = 0.007$$

$$P\left(H = 61 \mid p = \frac{2}{3}\right) = \binom{100}{61}\left(\frac{2}{3}\right)^{61}\left(\frac{1}{3}\right)^{39} = 0.040$$

p.m.f.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$0 \le p \le 1$$
$$x = 0, 1, 2, \ldots, n;$$

Since $P\left(H = 61 \mid p = \frac{2}{3}\right)$ is maximum and hence MLE is $p = \frac{2}{3}$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Solution: Since the distribution follow is Binomial distribution, with parameter p. Here $n = 100$. The likelihood function (MLE) is

$$P(H = 61|p) = \binom{100}{61} p^{61}(1-p)^{39}$$

For maximization

$$\frac{d}{dp} P(H = 61|p) = 0$$

$$\Rightarrow \binom{100}{61}\left[61p^{60}(1-p)^{39} - 39p^{61}(1-p)^{38}\right] = 0$$

$$\Rightarrow p^{60}(1-p)^{38}(61 - 100p) = 0$$

$$\Rightarrow p = 0, \frac{61}{100}, 1$$

Thus, the likelihoods are

$$P(H = 61|p = 0) = 0$$

$$P\left(H = 61\middle|p = \frac{61}{100}\right) = \binom{100}{61}\left(\frac{61}{100}\right)^{61}\left(\frac{39}{100}\right)$$

$$P(H = 61|p = 1) = 0$$

Since $P\left(H = 61\middle|p = \frac{61}{100}\right)$ is maximum and hence $p = \frac{61}{100}$ is the MLE.

Maximum Likelihood for a Binomial distribution

- Suppose we wish to find the maximum likelihood estimate (MLE) of θ for a Binomial distribution,

$$p_k(k,\theta) = nC_k \theta^k (1-\theta)^{n-k}$$

$$\log p_k(k,\theta) = \log(nC_k) + k\log(\theta) + (n-k)\log((1-\theta)$$

$$\frac{\partial \log p_k(k,\theta)}{\partial \theta} = 0 \Rightarrow 0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

$$k - k\theta = n\theta - k\theta \Rightarrow \theta = \frac{k}{n}$$

Example

If 48 of 60 transceivers passed inspection. (a) Obtain the maximum likelihood estimate of the probability that a transceiver will pass inspection. (b) Obtain the maximum likelihood estimate that the next two transceivers tested will pass inspection.

Solution:

a) n=60, k=48, the MLE P(to pass test) = $\frac{48}{60}$=0.8

b) For next 2 consecutive test, the MLE to pass test = (0.8)(0.8)

Maximum Likelihood Estimation

Example:

Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X=0]=(1-p), P[X=1]=p. Find the maximum likelihood estimate of p.

Soln :

Maximum Likelihood Estimation

Example:

Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X=0]=(1-p), P[X=1]=p. Find the maximum likelihood estimate of p.

Soln :

L(p)=P[X=0] P[X=1] P[X=0] P[X=0] P[X=1] P[X=0]

=(1-p) p (1-p) (1-p) p (1-p)

$=(1-p)^3p^2$.

Log L(p)=log[$(1-p)^3p^2$.]=log[$(1-p)^3$]+log[($p^{2)}$.]=3log(1-p)+2logp

$$\frac{\partial LogL(p)}{\partial p} = 0$$

$$\frac{-3}{1-p} + \frac{2}{p} = 0 \Rightarrow \frac{-3p+2-2p}{p(1-p)} = 0 \Rightarrow p = 2/5$$

That is , there is 1/3 chance to observe this sample if we believe the population to be Binomial distributed .

MLE for Poisson Distribution Parameter

Let $X_1, X_2, \ldots, X_n \in R$ be a random sample from a Poisson distribution

The p.d.f. of a Poisson Distribution is :

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \Bigg| \text{; where x =0,1,2,...}$$

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-\lambda n} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i} \Bigg|$$

The log-likelihood is:

$$lnL(\lambda) = -\lambda n + \sum_{i=1}^{n} xi.\, ln(\lambda) - ln(\prod_{i=1}^{n} xi) \Bigg|$$

Setting its derivative with respect to λ to zero, we have:

$$\frac{d}{d\lambda} lnL(\lambda) = -n + \sum_{i=1}^{n} xi. \frac{1}{\lambda} = 0$$

giving,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X}$$

which is the maximum likelihood estimate

Example:

In one area along the interstate, the number of dropped wireless phone connections per call follows a Poisson distribution. From four calls, the number of dropped connections is 2, 0, 3, 1

(a) Find the maximum likelihood estimate of λ.

(b) Obtain the maximum likelihood estimate that the next two calls will be completed without any accidental drops.

Solution: a) $\bar{x} = \lambda = \frac{2+0+3+1}{4} = 1.5$

b) For next one call, P(x=0) = $\frac{e^{-1.5}(1.5)^0}{0!} = e^{-1.5}$

For next two calls = P(X=0)P(X=0)= $(e^{-1.5})(e^{-1.5})$= e^{-3}

EXAMPLE **Maximum likelihood estimator of function of λ**

The number of defective hard drives produced daily by a production line can be modeled as a Poisson distribution. The counts for ten days are

$$7 \quad 3 \quad 1 \quad 2 \quad 4 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2$$

Obtain the maximum likelihood estimate of the probability of 0 or 1 defectives on one day.

Solution From the maximum likelihood estimate of λ is $\widehat{\lambda} = \bar{x} = 26/10 = 2.6$. Consequently, by the invariance property, the maximum likelihood estimate of

$$P(X = 0 \quad \text{or} \quad 1) = e^{-\lambda} + \frac{\lambda e^{-\lambda}}{1!}$$

is

$$e^{-\widehat{\lambda}} + \frac{\widehat{\lambda} e^{-\widehat{\lambda}}}{1!} = e^{-2.6} + \frac{2.6 \cdot e^{-2.6}}{1!} = 0.267$$

There will 1 or fewer defectives on just over one-quarter of the days. ■

The method of maximum likelihood applies to more than one parameter.

Testing of Hypothesis ⇨ Analysis of Variance (ANOVA)

Types of ANOVA

- One-way
- Two-way
- Three-way (LSD)
- Repeated measures of ANOVA (Rm ANOVA)

Note: In this unit we discuss only One-way ANOVA and Two ANOVA in detail

Testing of Hypothesis ➡ Why Analysis of Variance (ANOVA)

Student's t-test cannot be applied

⬇

No. of groups are more than two (say k) and are independent

⬇

If t-test is applied, the type-I error will increase

ANOVA Used to test equality of more than two population means against not equal

Testing of Hypothesis ➡ One-way Analysis of Variance

ANOVA ➡ Testing equality of k group means against not equal

ANOVA

- Samples are drawn from normal population
- The population variances should be equal
- The sample size should be less than 30 (i.e., n < 30)
- Groups should be independent
- Subjects should be allocated randomly to both groups
- However even if sample size more than 30 (i.e., n > 30) ANOVA should be continue to apply, because of central limit theorem it approaches normal.

One-way ANOVA

ANOVA is a statistical technique used to determine whether differences exist among three or more population means.

In one-way ANOVA the effect of one factor on the mean is tested. It is based on independent random samples drawn from k – different levels of a factor, also called treatments.

The following notations are used in one-way ANOVA. The data can be represented in the following tabular structure.

Data representation for one-way ANOVA

Treatments					Total
Treatment 1	x_{11}	x_{12}	…	x_{1n_1}	$x_{1.}$
Treatment 2	x_{21}	x_{22}	…	x_{2n_2}	$x_{2.}$
⋮	⋮	⋮	…	⋮	⋮
Treatment k	x_{k1}	x_{k2}	…	x_{kn_k}	$x_{k.}$

x_{ij} - the j th sample value from the ith treatment, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$

k - number of treatments compared.

$x_{i.}$ - the sample total of ith treatment.

n_i - the number of observations in the ith treatment.

$$\sum_{i=1}^{k} n_i = n$$

Test Procedure

Let the observations x_{ij}, $j = 1, 2, \ldots, n_i$ for treatment i, be assumed to come from $N(\mu_i, \sigma^2)$ population, $i = 1, 2, \ldots, k$ where σ^2 is unknown.

Step 1 : Framing Hypotheses

Null Hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

That is, there is no significant difference among the population means of k treatments.

Alternative Hypothesis

H_1: $\mu_i \neq \mu_j$ for atleast one pair (i,j); $i, j = 1,2,\ldots,k$; $i \neq j$

That is, at least one pair of means differ significantly.

Step 2 : Data

Data is presented in the tabular form as described in the previous section

Step 3 : Level of significance : α

Step 4 : Test Statistic

$F = MST/MSE$ which follows $F_{(k-1,\, n-k)}$, under H_0

To evaluate the test statistic we compute the following:

(i) Correction factor: $C.F = \dfrac{G^2}{n}$ where $G = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$

(ii) Total Sum of Squares: $TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - C.F$

(iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^{k} \dfrac{x_{i.}^2}{n_i} - C.F,$

where $x_{i.} = \sum_{j=1}^{n_i} x_{ij},\ i = 1, 2, \ldots, k$

(iv) Sum of Squares due to Error: $SSE = TSS - SST$

Mean Sum of Squares

Mean Sum of Squares due to treatment: $MST = SST / k - 1$

Mean Sum of Squares due to Error: $MSE = SSE / n - k$

Step 5 : Calculation of Test statistic

ANOVA Table (one-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Treatments	SST	k-1	$MST = \frac{SST}{k-1}$	$F_0 = \frac{MST}{MSE}$
Error	SSE	n-k	$MSE = \frac{SSE}{n-k}$	
Total	TSS	n-1		

Step 6 : Critical value

$f_e = f_{(k-1,\, n-k),\alpha}$.

Step 7 : Decision

If $F_0 < f_{(k-1,\, n-k),\alpha}$ then reject H_0.

Example:1

Three different techniques namely medication, exercises and special diet are randomly assigned to (individuals diagnosed with high blood pressure) lower the blood pressure. After four weeks the reduction in each person's blood pressure is recorded. Test at 5% level, whether there is significant difference in mean reduction of blood pressure among the three techniques.

Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	12	8	4

Solution:

Step 1 : Hypotheses

Null Hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3$

That is, there is no significant difference among the three groups on the average reduction in blood pressure.

Alternative Hypothesis: $H_1: \mu_i \neq \mu_j$ for atleast one pair (i, j); $i, j = 1, 2, 3$; $i \neq j$.

That is, there is significant difference in the average reduction in blood pressure in atleast one pair of treatments.

Step 2 : Data

Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	12	8	4

Step 3 : Level of significance $\alpha = 0.05$

Step 4 : Test statistic

$F_0 = MST / MSE$

Step 5 : Calculation of Test statistic

						Total	Square
Medication	10	12	9	15	13	59	3481
Exercise	6	8	3	0	2	19	361
Diet	5	9	12	8	4	38	1444
						G = 116	5286

Individual squares

Medication	100	144	81	225	169
Exercise	36	64	9	0	4
Diet	25	81	144	64	16

$$\sum\sum x_{ij}^2 = 1162$$

1. Correction Factor: $CF = \frac{G^2}{n} = \frac{(116)^2}{15} = \frac{13456}{15} = 897.06$

2. Total Sum of Squares: $TSS = = \sum\sum x_{ij}^2 - C.F$

$$= 1162 - 897.06 = 264.94$$

3. Sum of Squares between Treatments: $SST = \frac{\sum x_i^2}{n_i} - C.F$

$$= \frac{5286}{5} - 897.06$$
$$= 1057.2 - 897.06$$
$$= 160.14$$

4. Sum of Squares due to Error: $SSE = TSS - SST$

$$= 264.94 - 160.14 = 104.8$$

ANOVA Table (one-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Between treatments	160.14	$3 - 1 = 2$	80.07	$F_0 = \frac{80.07}{8.73} = 9.17$
Error	104.8	12	8.73	
Total	264.94	$n - 1 = 15 - 1 = 14$		

Step 6 : Critical value

$f(2, 12),0.05 = 3.8853.$

Step 7 : Decision

As $F_0 = 9.17 > f_{(2,\ 12),0.05} = 3.8853$, the null hypothesis is rejected. Hence, we conclude that there exists significant difference in the reduction of the average blood pressure in atleast one pair of techniques.

Example :2

Three composition instructors recorded the number of spelling errors which their students made on a research paper. At 1% level of significance test whether there is significant difference in the average number of errors in the three classes of students.

Instructor 1	2	3	5	0	8		
Instructor 2	4	6	8	4	9	0	2
Instructor 3	5	2	3	2	3	3	

Solution:

Step 1 : Hypotheses

Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$

That is there is no significant difference among the mean number of errors in the three classes of students.

Alternative Hypothesis

$H_1 : \mu_i \neq \mu_j$ for at one pair (i, j); $i,j = 1,2,3$; $i \neq j$.

That is, atleast one pair of groups differ significantly on the mean number of errors.

Step 2 : Data

Instructor 1	2	3	5	0	8		
Instructor 2	4	6	8	4	9	0	2
Instructor 3	5	2	3	2	3	3	

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

$F_0 = MST / MSE$

Step 5 : Calculation of Test statistic

								Total	Square
Instructor 1	2	3	5	0	8			18	324
Instructor 2	4	6	8	4	9	0	2	33	1089
Instructor 3	5	2	3	2	3	3		18	324
								69	

Individual squares

Instructor 1	4	9	25	0	64		
Instructor 2	16	36	64	16	81	0	4
Instructor 3	25	4	9	4	9	9	

$$\sum\sum x_{ij}^2 = 379$$

Correction Factor: $$CF = \frac{G^2}{n} = \frac{(69)^2}{18} = \frac{4761}{18} = 264.5$$

Total Sum of Squares: $$TSS = \sum\sum x_{ij}^2 - C.F$$
$$= 379 - 264.5 = 114.5$$

Sum of Squares between Treatments: $SST = \dfrac{\sum x_i^2}{n_i} - C.F$

$$= \left(\frac{324}{5} + \frac{1089}{7} + \frac{324}{6}\right) - 264.5$$

$$= (64.8 + 155.6 + 54) - 264.5$$

$$= (274.4) - 264.5$$

$$= 9.9$$

Sum of Squares due to Error: $SSE = TSS - SST$

$$= 114.5 - 9.9$$

$$= 104.6$$

ANOVA Table

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Between treatments	9.9	$3 - 1 = 2$	$\frac{9.9}{2} = 4.95$	$F_0 = \frac{4.95}{6.97} = 0.710$
Error	104.6	15	$\frac{104.6}{15} = 6.97$	
Total		$n - 1 = 18 - 1 = 17$		

Step 6 : Critical value

The critical value = $f_{(15,\,2),0.05} = 3.6823$.

Step 7 : Decision

As $F_0 = 0.710 < f_{(15,\,2),0.05} = 3.6823$, null hypothesis is not rejected. There is no enough evidence to reject the null hypothesis and hence we conclude that the mean number of errors made by these three classes of students are not equal. p-value equals **0.508553**

TWO-WAY CLASSIFICATION

In two-way ANOVA a study variable is compared over three or more groups, controlling for another variable. The grouping is taken as one factor and the control is taken as another factor. The grouping factor is usually known as Treatment. The control factor is usually called Block. The accuracy of the test in two -way ANOVA is considerably higher than that of the one-way ANOVA, as the additional factor, block is used to reduce the error variance.

In two-way ANOVA, the data can be represented in the following tabular form.

Blocks

Blocks

Groups or Treatments	1	2	3	…	m	$x_{i.}$
1	x_{11}	x_{12}	$x_{13.}$	…	x_{1m}	$x_{1.}$
2	x_{21}	x_{22}	$x_{2.}$	…	x_{2m}	$x_{2.}$
3	x_{31}	x_{32}	$x_{3.}$	…	x_{3m}	$x_{3.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	x_{k1}	x_{k2}	x_{k3}	…	x_{km}	$x_{k.}$
$x_{.j}$	$x_{.1}$	$x_{.2}$	$x_{.3}$	…	$x_{.m}$	G

1. Test Procedure

Steps involved in two-way ANOVA are:

Step 1 : In two-way ANOVA we have two pairs of hypotheses, one for treatments and one for the blocks.

Framing Hypotheses

Null Hypotheses

H_{01}: There is no significant difference among the population means of different groups (Treatments)

H_{02}: There is no significant difference among the population means of different Blocks

Alternative Hypotheses

H_{11}: Atleast one pair of treatment means differs significantly

H_{12}: Atleast one pair of block means differs significantly

Step 2 : **Data** is presented in a rectangular table form as described in the previous section.

Step 3 : **Level of significance** α.

Step 4 : **Test Statistic**

F_{0t} (treatments) = MST / MSE

F_{0b} (block) = MSB / MSE

To find the test statistic we have to find the following intermediate values.

i) Correction Factor: $C.F = \frac{G^2}{n}$ where $G = \sum_{j=1}^{m}\sum_{i=1}^{k} x_{ij}$

ii) Total Sum of Squares: $TSS = \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}^2 - C.F$

iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^{k} \frac{x_{i.}^2}{m} - C.F$

iv) Sum of squares between blocks: $SSB = \sum_{j=1}^{m} \frac{x_{.j}^2}{k} - C.F$

v) Sum of Squares due to Error: $SSE = TSS\text{-}SST\text{-}SSB$

vi) Degrees of freedom

Degrees of freedom (d.f.)	d.f.
Total Sum of Squares	$n-1$
Treatment Sum of Squares	$k-1$
Block Sum of Squares	$m-1$
Error of Sum Squares	$(m-1)(k-1)$

vii) Mean Sum of Squares

Mean sum of Squares due to Treatments: $MST = \dfrac{SST}{k-1}$

Mean sum of Squares due to Blocks: $MSB = \dfrac{SSB}{m-1}$

Mean sum of Squares due to Error: $MSE = \dfrac{SSE}{(k-1)(m-1)}$

Step 5 : Calculation of the Test Statistic

ANOVA Table (two-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Treatments	SST	k-1	MST	$F_{0t} = \frac{MST}{MSE}$
Blocks	SSB	m-1	MSB	$F_{0b} = \frac{MSB}{MSE}$
Error	SSE	$(k-1)(m-1)$	MSE	
Total	TSS	n-1		

Step 6 : Critical values

Critical value for treatments = $f_{(k-1,(m-1)(k-1)),\alpha}$

Critical value for blocks = $f_{(m-1,\ (m-1)(k-1)),\alpha}$

Step 7 : Decision

For Treatments: If the calculated F_{0t} value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the treatment means, in atleast one pair.

For Blocks: If the calculated F_{0b} value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the block means, in at least one pair.

Example:1

A reputed marketing agency in India has three different training programs for its salesmen. The three programs are Method – A, B, C. To assess the success of the programs, 4 salesmen from each of the programs were sent to the field. Their performances in terms of sales are given in the following table.

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Test whether there is significant difference among methods and among salesmen.

Solution:

Step 1 : Hypotheses

Null Hypotheses: $H_{01} : \mu_{M1} = \mu_{M2} = \mu_{M3}$ (for treatments)

That is, there is no significant difference among the three programs in their mean sales.

H_{02} : $\mu_{S1} = \mu_{S2} = \mu_{S3} = \mu_{S4}$ **(for blocks)**

Alternative Hypotheses:

H_{11} : At least one average is different from the other, among the three programs.

H_{12} : At least one average is different from the other, among the four salesmen.

Step 2 : Data

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

$$F_{0t}(\text{treatment}) = \frac{MST}{MSE}$$

$$F_{0b}(\text{block}) = \frac{MSB}{MSE}$$

Step-5 : Calculation of the Test Statistic

	Methods			Total $x_{i.}$	$x_{i.}^2$
	A	B	C		
1	4	6	2	12	144
2	6	10	6	22	484
3	5	7	4	16	256
4	7	5	4	16	256
x_i	22	28	16	66	1140
$x_{i.}^2$	484	784	256	1524	

Squares

16	36	4
36	100	36
25	49	16
49	25	16
		$\sum\sum x_{ij}^2 = 408$

Correction Factor: $$CF = \frac{G^2}{n} = \frac{(66)^2}{12} = \frac{4356}{12} = 363$$

Total Sum of Squares: $$TSS = \sum\sum x_{ij}^2 - C.F$$
$$= 408 - 363 = 45$$

Sum of Squares due to Treatments: $$SST = \frac{\sum_{j=1}^{k} x_{.j}^2}{k} - C.F$$
$$= \frac{1140}{3} - 363$$
$$= 380 - 363 = 17$$

Sum of Squares due to Blocks:

$$SSB = \frac{\sum_{j=1}^{k} x_{.j}^2}{k} - C.F$$

$$= \frac{1524}{4} - 363$$

$$= 381 - 363$$

$$= 18$$

Sum of Squares due to Error:

$$SSE = TSS - SST - SSB$$

$$= 45 - 17 - 18 = 10$$

Mean sum of squares:

$$MST = \frac{SST}{k-1} = \frac{17}{3} = 5.67$$

$$MSB = \frac{SSB}{m-1} = \frac{18}{2} = 9$$

$$MSE = \frac{SSE}{(k-1)(m-1)} = \frac{10}{6} = 1.67$$

ANOVA Table (two-way)

Sources of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Between treatments (Programs)	17	3	5.67	$F_{ot} = \frac{5.67}{1.67} = 3.40$
Between blocks (Salesmen)	18	2	9	$F_{ob} = \frac{9}{1.67} = 5.39$
Error	10	6	1.67	
Total		11		

Step 6 : Critical values

$f_{(3,\,6),0.05} = 4.7571$ (for treatments)

$f_{(2,\,6),0.05} = 5.1456$ (for blocks)

Step 7 : Decision

(i) Calculated $F_{0t} = 3.40 < f_{(3,\,6),0.05} = 4.7571$, the null hypothesis is not rejected and we conclude that there is significant difference in the mean sales among the three programs.

(ii) Calculate $F_{0b} = 5.39 > f_{(2,\,6),0.05} = 5.1456$, the null hypothesis is rejected and conclude that there does not exist significant difference in the mean sales among the four salesmen.

Problems

1)

Four brands of flashlight batteries are to be compared by testing each brand in five flashlights. Twenty flashlights are randomly selected and divided randomly into four groups of five flashlights each. Then each group of flashlights uses a different brand of battery. The lifetimes of the batteries, to the nearest hour, are as follows.

Brand *A*	Brand *B*	Brand *C*	Brand *D*
42	28	24	20
30	36	36	32
39	31	28	38
28	32	28	28
29	27	33	25

Preliminary data analyses indicate that the independent samples come from normal populations with equal standard deviations. At the 5% significance level, does there appear to be a difference in mean lifetime among the four brands of batteries?

2. The times required by three workers to perform an assembly-line task were recorded on five randomly selected occasions. Here are the times, to the nearest minute

Hank	Joseph	Susan
8	8	10
10	9	9
9	9	10
11	8	11
10	10	9

Construct the one way ANOVA table for the data at 5% level of significance.

3.

Three drying formulas for curing glue are studied

Formula A	13	10	8	11	8	
Formula B	13	11	14	14		
Formula C	17	14	13	10	11	12

Test at 5% level of significance whether is any difference in the mean curing time of glue?

4.

The illness caused by a virus in a city concerning some restaurant inspectors is not consistent with their evaluations of cleanliness of restaurants. In order to investigate this possibility, the director has five restaurant inspectors to grade the cleanliness of three restaurants. The results are shown below.

Inspectors	Restaurants		
	I	II	III
1	71	55	84
2	65	57	86
3	70	65	77
4	72	69	70
5	76	64	85

Carry out two-way ANOVA at 5% level of significance.

5. Future group wishes to enter the frozen shrimp market. They contract a researcher to investigate various methods of groups of shrimp in large tanks. The researcher suspects that temperature and salinity are important factor influencing shrimp yield and conducts a two-way analysis of variance with their levels of temperature and salinity. That is each combination of yield for each (for identical gallon tanks) is measured. The recorded yields are given in the following chart:

Categorical variable Salinity (in pp)

Temperature	700	1400	2100	Total	Mean
60^0 F	3	5	4	12	4
70^0 F	11	10	12	33	11
80^0 F	16	21	17	54	18
Total	30	36	33	99	11

Compute the ANOVA table for the model.

Thank you

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session – 11

Correlation & Regression

IMP Note to Self

Contact Session 11

Contact Session 11: Module 5: Prediction & Forecasting

Contact Session	List of Topic Title	Reference
CS - 11	Correlation & regression	T1& T2
HW	Problems on correlation and regression	T1& T2
Lab	Regression – Analysis of model summary	Lab 5

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist correlation (i.e., association) between two (or more) variables?
If yes, of what degree?

Q2: Is there any cause and effect relationship between the two variables (in case of bivariate population) or one variable in one side and two or more variables on the other side (in case of multivariate population)?
If yes, of what degree and in which direction?

To find solutions to the above questions, two approaches are known.
Correlation Analysis
Regression Analysis

Correlation Analysis

Example: Wage Data

A large data regarding the wages for a group of employees from the western region of a country is given.

In particular, we wish to understand the following relationships:

- *Employee's age and wage:* How wages vary with ages?
- *Calendar year and wage:* How wages vary with time?
- *Employee's age and education:* Whether wages are anyway related with employees' education levels?

- Case I. Wage versus Age

- *Employee's age and wage:* How wages vary with ages?

Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

➢ Case II. Wage versus Calendar

➢ *Wage and calendar year:* **How wages vary with years?**

Interpretation: ?

➢ Case III. Wage versus Age

Wage and education level: **Whether wages vary with employees' education levels?**

Interpretation: ?

Covariance

The statistical technique used for studying the existence and extent of relationship among 2 or more variables is through Covariance

$$\text{Covariance}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n - 1}$$

- So in bivariate and multivariate case, the statistical technique used for studying the existence of relationship between all variables is through **Covariance.**
- Covariance signifies the direction of the linear relationship between the two variables.
- By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative effect on the value of the other variable).

- The values of covariance can be any number between the two opposite infinities. Also, it's important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.
- The upper and lower limits for the covariance depend on the variances of the variables involved. These variances, in turn, can vary with the scaling of the variables. Even a change in the units of measurement can change the covariance. Thus, covariance is only useful to find the direction of the relationship between two variables and not the magnitude.

If X_1 and X_2 are two variables, the covariance between X_1 and X_2 is defined as

$$\mathrm{Cov}(X_1, X_2) = \frac{\sum_{i=1}^{n} (X_1 - \overline{X}_1)(X_2 - \overline{X}_2)}{n-1}$$

If $X_1, X_2, \ldots, X_n$ are n variables, then the Covariance between these n variables is the Variance – Covariance matrix given by

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & . & . & . & \sigma_{12} \\ \sigma_{21} & o_{22} & . & . & . & \sigma_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_{n1} & o_{n2} & . & . & . & o_{n1} \end{bmatrix}$$

where

$$\sigma_{ij} = \begin{cases} \text{Variance, for } i = j \\ \text{Covariance, for } i \neq j \end{cases}$$

All diagonal elements in the variance-covariance matrix are variances and off diagonal elements are covariances.

Σ is a symmetric matrix.

- Even though the covariance explains the relationship between two or more variables, the units the measurement of variable is attached with the value of the covariance. If they have different unit of measurements, based on the sign it is possible to say whether they are related or not, but the degree (strength) of relationship is not possible to decide based on the value of covariance due to different units of measurement.

Example 1 :

Find the covariance between age (months) and weight (kgs) for the following data

Age (years)	7	6	8	5	6	9
Weight (kgs)	12	8	12	10	11	13

Solution -1

Age (yrs) (X)	Weight (kgs) (Y)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$
5	20	-2	-3	6
4	18	-3	-5	15
7	23	0	0	0
10	32	3	9	27
8	26	1	3	3
8	25	1	2	2
$\sum X = 42$	$\sum X = 144$			$\sum (X-\bar{X})(Y-\bar{Y}) = 53$

$$\mathrm{Cov}(X,Y) = \frac{53}{5} = 10.6 \text{ years.kgs} \rightarrow \text{Interpret this answer !}$$

Covariance based on Probability distribution.

If X and Y are two random variables with probability mass function p(x)/ probability density function f(x), then the covariance between X and Y

$$\mathbf{Cov\ (X, Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - E(X)E(Y)}$$

where $E(X) = \mu_X$ and $E(Y) = \mu_Y$ are the means of X and Y.

- $\mathbf{Covariance(x, x) = Var(x)}$

- $\mathbf{Covariance(x, \boldsymbol{y}) = Covariance(y, x)}$

Application of Covariance.

1. Multivariate linear regression:

- If there a dependent variable and n independent variables, to find out the relationship between these (n+1) variables, a variance – covariance matrix will be helpful.

2. Time series:

- In the given series of data over the time, it helps in finding the autocorrelation at different time lags. For example, if X_{t1} , X_{t2} , ... , X_{tn} denote the time series data over a given period, then the autocovariance

Application of Covariance. at

lag h is given by

$$\gamma_X(t+h, t) = \mathrm{Cov}\,(X_{t+h}, X_t)$$

For example, if $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ denote the time series data over a given period, then the auto-covariance.

The auto-covariance at lag h is then given by

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n-|h|}(X_{t+|h|} - \overline{X})(X_t - \overline{X}), -n < h < n.$$

Application of Covariance.

With the help of auto-covariance, the autocorrelation is given by

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}$$

Karl Pearson's Correlation coefficient

- To overcome this, Karl Pearson's suggestion was to divide the covariance of the variables by their respective standard deviations. The ratio of covariance to the product of standard deviations is called Karl Pearson's product moment correlation, which is free of unit of measurement.
- Karl Pearson's product moment correlation returns a numerical value, popularly known as **Coefficient of Correlation**, denoted by **'r'**
- Correlation coefficient is a statistic that assesses the **strength** and **direction of relationship** of two continuous variables

Formula for computing Karl Pearson's correlation coefficient (r) is standardized covariance (unit less)

$$r = \frac{\text{Covariance}(x, y)}{\sqrt{\text{Var}(x)}\sqrt{\text{Var}(Y)}}$$

$$r = \frac{\dfrac{\sum_{i=1}^{n}(X-\overline{X})(Y-\overline{Y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(X-\overline{X})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(Y-\overline{Y})^2}{n-1}}} = \frac{\sum_{i=1}^{n}(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y-\overline{Y})^2}}$$

Karl Pearson's Correlation coefficient

Or alternatively we can use

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}, x = (X - \overline{X}), y = (Y - \overline{Y})$$

Simple Correlation coefficient

- If the sign is +ve this means the relation is direct (an increase in one variable is associated with an increase in the other variable and a decrease in one variable is associated with a decrease in the other variable).
- If the sign is -ve this means an inverse or indirect relationship (an increase in one variable is associated with a decrease in the other).
- The value of r ranges between (-1) and (+1), i.e., $-1 \leq r \leq +1$
- The value of r denotes the strength of the association as illustrated by the following diagram.

Strong
Intermediate
Weak
Weak
Intermediate
Strong
- 1
- 0.75
- 0.25
0
0.25
0.75
+ 1
indirect
Direct
Perfect - ve correlation
No relation
Perfect +ve correlation

Simple Correlation coefficient

- r = - 1 ⇒ Perfect negative correlation
- - 0.99 to - 0.76 ⇒ Strong negative correlation
- - 0.75 to - 0.26 ⇒ Intermediate (Moderate) negative correlation
- -0.25 to 0 ⇒ Weak negative correlation
- r = 0 ⇒ Zero correlation
- 0 to 0.25 ⇒ Weak positive correlation
- 0.26 to 0.75 ⇒ Intermediate (Moderate) positive correlation
- 0.76 to 0.99 ⇒ Strong positive correlation
- r = = 1 ⇒ Perfect positive correlation

Example

Serial No	Age (yrs)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Solution

Sl no.	Age (yrs) (X)	Wt (Kg) (Y)	XY	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	$\sum X=41$	$\sum Y=66$	$\sum XY= 461$	$\sum X^2= 291$	$\sum Y^2= 742$

Karl Pearson's Correlation coefficient

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^{!} - (\sum X)^{!}}\sqrt{n\sum Y^{!} - (\sum Y)^{!}}}$$

$$r = \frac{6 * 461 - 41 * 66}{\sqrt{6 * 291 - (41)^{!}}\sqrt{6 * 742 - (66)^{!}}}$$

r = 0.759 (Positive (Direct) strong correlation)

Karl Pearson's Correlation coefficient

Example: Relationship between anxiety and test Scores

Anxiety (X)	Test score (Y)	X^2	Y^2	XY
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum X = 32$	$\sum Y = 32$	$\sum X^2 = 230$	$\sum Y^2 = 204$	$\sum XY = 129$

Correlational Analysis ➡ Karl Pearson's Correlation coefficient

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^{!} - (\sum X)^{!}}\sqrt{n\sum Y^{!} - (\sum Y)^{!}}}$$

$$r = \frac{6*129 - 32*32}{\sqrt{\left(6*230 - (32)^2\right)\left(6*204 - (32)^2\right)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -0.94$$

r = - 0.94 (Negative (Indirect) strong correlation)

Example : A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

Solution :

In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus *Y* = birth weight and *X* = gestational age.

For the given data, it can be shown the following

$$\overline{X} = \frac{\Sigma X}{n} = \frac{652.1}{17} = 38.4.$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{49{,}334}{17} = 2902.$$

$$s_x^2 = \frac{\Sigma(X-\overline{X})^2}{n-1} = \frac{159.45}{16} = 10.0.$$

$$s_y^2 = \frac{\Sigma(Y-\overline{Y})^2}{n-1} = \frac{7{,}767{,}660}{16} = 485{,}578.8.$$

$$r^* = \frac{\Sigma(X_i-\overline{X})(Y_i-\overline{Y})}{s_x.s_y} = 0.82$$

Conclusion: The sample's correlation coefficient indicates a **strong positive correlation** between Gestational Age and Birth Weight.

Homework Problems

Q1) Consider the following data.

Expenditure (in 1000 Rs)	30	70	40	20	10	40	10	20
Actual Sales(in 1000 Rs)	10	16	19	14	17	16	13	18

For the above data,

i) Find coefficient of correlation between expenditure and actual sales. **[Ans = 0.8108]**

ii) Find rank correlation between expenditure and actual sales. **[Ans = - 0.2424]**

Q2) Find coefficient of correlation between X & Y.

X	79	29	45	61	24	38	33	52	65	63	82	50
Y	15	12	12	6	6	7	3	10	12	13	13	14

Ans = r= 0.59

Q3) Check whether X and Y are having any relation.

X	12	14	15	17	17	16	16	15	14	14
Y	52	52	57	62	67	67	67	62	62	52

Ans = r = 0.7866

Regression

x	1	2	3	4	5	6
y	1	4	9	16	25	36

Regression

x Income (Lakhs)	25	30	20	15	18	22
y Exp (lakhs)	6	8	5	7	4	5

when $x = 32$, then

$y = ?$

100
80
60
40
30
20
10
1 2 3 4 5 6
Linear

Non-linear

CORRELATION	REGRESSION
Measuring strength or degree of the relationship between two variables	Having an algebraic equation between two variables
No estimation	Estimation
Both variables are independent	One is dependent variable and the other is independent variable

Method of Least squares

y : Dependent Variable

x : Independent Variable

predictor variable

response Variable

Method of Least squares

Method of Least squares

$$S(\beta_0, \beta_1) = \underbrace{\sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_i)^2}$$

we need to choose β_0 and β_1 which minimizes the error.

Method of Least squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow 2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right) (-1)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right) (2)(-x_i)$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get β_0 & β_1 which minimizes error.

Linear regression –

$$y = \beta_0 + \beta_1 x$$

$$\left.\begin{aligned} \Sigma y &= \beta_0 n + \beta_1 \Sigma x \\ \Sigma xy &= \beta_0 \Sigma x + \beta_1 \Sigma x^2 \end{aligned}\right\}$$

Normal equations.

$$Y = \beta_0 + \beta_1 X$$

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

Example :-

company	Advt Expt	Sales Revenue
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

$$y = a + bx$$

$$\Sigma y = an + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Example :-

$y = \beta_0 + \beta_1 x$

Sales Revenue y	Advt. expt. x	x^2	xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
$\Sigma 40$	$\Sigma 56$	$\Sigma 524$	$\Sigma 373$

$\Sigma y = n\beta_0 + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$

$\Rightarrow 40 = 8\beta_0 + 56\beta_1$

$373 = 56\beta_0 + 524\beta_1$

on solving

$\beta_0 = 0.072$

$\beta_1 = 0.704$

$\therefore y = (0.072) + (0.704)x$

ie $y = (0.072) + (0.704)x$

when $x = 0.075$, then

$$y = (0.072) + (0.704)(0.075)$$

$$= 0.1248 \simeq 12.48\%$$

Example:

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

x	y	xy	x^2
1	0.5	0.5	1
2	1	2	4
4	2	8	16
0	0	0	0
$\Sigma = 7$	$\Sigma\ 3.5$	$\Sigma\ 10.5$	$\Sigma\ 21$

$$y = \beta_0 + \beta_1 x$$

$$\Sigma y = n\beta_0 + \beta_1 \Sigma x$$

$$\Sigma xy = \beta_0 \Sigma x_1 + \beta_1 \Sigma x^2$$

$$3.5 = 4\beta_0 + \beta_1 (7)$$

$$10.5 = 7\beta_0 + \beta_1 (21)$$

on solving these

$$\beta_0 = 0$$

$$\beta_1 = 0.5$$

i.e $y = 0 + (0.5)x$

$$\boxed{\text{when } x = 5, \quad y = (0.5)5 = 0.25}$$

Assumptions of the Linear Regression

Assumptions about the Error

- $E(\varepsilon_i) = 0$ for $i = 1, 2, \ldots, n$.
- $\sigma(\varepsilon_i) = \sigma_\varepsilon$ where σ_ε is unknown.
- The errors are independent, that is, the error in the *i*th observation is independent of the error observed in the *j*th observation.
- The ε_i are normally distributed (with mean 0 and standard deviation σ_ε).

$$y_i = \hat{y}_i + (y_i - \hat{y}_i)$$

observed — Fit — Deviation from fit

subtracting $\bar{y}$ from both sides

$$(y_i - \bar{y}) = (\hat{y} - \bar{y}) + (y_i - \hat{y}_i)$$

Deviation from mean — Deviation due to fit — Residual

RSS $\longrightarrow$ Residual sum of squares

$$= \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i) \right)^2$$

TSS $\longrightarrow$ $\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2$ mean of respective variables

$$R^2 = 1 - \frac{RSS}{TSS}$$

Overfitting

Which of these two models would be a better fit to the data?

Multicollinearity

- It refers to the phenomenon of having related predictor variables in the input dataset.
- In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated, due to which the presence of that variable in the model is redundant. You drop some of these related independent variables as a way of dealing with multicollinearity.
- **Multicollinearity affects:**

Interpretation:

- Does "change in Y, when all others are held constant" apply?

Inference:

- Coefficients swing wildly, signs can invert
- p-values are, therefore, not reliable

Dealing with Multicollinearity

- Two basic ways of dealing with multicollinearity
- Looking at **pairwise correlations**
  - Looking at the correlation between different pairs of independent variables

Checking the **Variance Inflation Factor** (VIF)

- Sometimes pairwise correlations aren't enough
- Instead of just one variable, the independent variable might depend upon a combination of other variables
- VIF calculates how well one independent variable is explained by all the other independent variables combined

Variance Inflation Factor (VIF)

- The VIF is given by:

$$VIF_i = \frac{1}{1-{R_i}^2}$$

where *'i'* refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.

- The common heuristic we follow for the VIF values is:

> 10: Definitely high VIF value and the variable should be eliminated.

> 5: Can be okay, but it is worth inspecting.

< 5: Good VIF value. No need to eliminate this variable.

Example:-

x_0	Size	No of rooms	No of floors	Age of home	price Lakhs
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000
	x_1	x_2	x_3	x_4	y

Multiple Linear Regression

The data consists of n observations on a dependent or response variable Y and p predictor or explanatory variables

$X_1, X_2, \ldots, X_p$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

β_i's are regression coefficients.

normal equations

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2$$

Multiple Linear

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\Sigma y = \beta_0 n + \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2 + \beta_3 \Sigma x_3 + \beta_4 \Sigma x_4$$

$$\Sigma x_1 y = \beta_0 \Sigma x_1 + \beta_1 \Sigma x_1^2 + \beta_2 \Sigma x_1 x_2 + \beta_3 \Sigma x_1 x_3 + \beta_4 \Sigma x_1 x_4$$

$$\Sigma x_2 y = \beta_0 \Sigma x_2 + \beta_1 \Sigma x_1 x_2 + \beta_2 \Sigma x_2^2 + \beta_3 \Sigma x_2 x_3 + \beta_4 \Sigma x_2 x_4$$

$$\Sigma x_3 y = \beta_0 \Sigma x_3 + \beta_1 \Sigma x_1 x_3 + \beta_2 \Sigma x_2 x_3 + \beta_3 \Sigma x_3^2 + \beta_4 \Sigma x_3 x_4$$

$$\Sigma x_4 y = \beta_0 \Sigma x_4 + \beta_1 \Sigma x_1 x_4 + \beta_2 \Sigma x_2 x_4 + \beta_3 \Sigma x_3 x_4 + \beta_4 \Sigma x_4^2$$

solve $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

normal equations

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2$$

Non - Linear Regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\sum y = \beta_0 n + \beta_1 \sum x + \beta_2 \sum x^2$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2 + \beta_2 \sum x^3$$

$$\sum x^2 y = \beta_0 \sum x^2 + \beta_1 \sum x^3 + \beta_2 \sum x^4$$

1) $Y = \alpha x^{\beta}$

2) $Y = \alpha e^{\beta x}$

3) $Y = \alpha + \beta \log x$

4) $Y = \dfrac{\alpha}{\alpha x - \beta}$

Other regressions

just a look

Suppose $y = ae^{bx}$

$$\underbrace{\log y}_{Y} = \underbrace{\log a}_{A} + \underbrace{b \log x}_{X}$$

ie $Y = A + bX$

$$\Sigma Y = An + b\Sigma X \longrightarrow (1)$$

$$\Sigma XY = A\Sigma X + b\Sigma X^2 \longrightarrow (2)$$

Hence, we get

$$y = ae^{bx}$$

R – Squared vs Adjusted R - Squared

- In multiple regression, adjusted R – squared is better metric than R – squared asses the goodness of fit of the model

- R – squared always increases if additional variables are added into model , even if they are not related to the dependent variable

Lasso & Ridge

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- OLS estimation: $$\min SSE = \sum \left(Y - \hat{Y} \right)^2$$
- LASSO estimation: $$\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|$$
- Ridge regression estimation: $$\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j^{\,2} \right|$$

IMP Note to Self

Thanks

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Sessions 13, 14 & 15

Time Series Analysis

(26th August, 2nd & 9th Sep 2023)

Time Series

Time Series → Definition

A time series is defined as a set of observations on a variable generated sequentially in time. The measurement of the variable may be made continuously or at discrete (equally spaced) intervals.

Often a variable continuous in time is measured at regular intervals and this produces a discrete series of data.

Time Series ➡ Definition

Time series may be represented as a set of data (Y_1, Y_2, Y_3, …, Y_n), Y_t denoting the value of the variable Y at time t.

Example of continuous variable:

-Temperature on a chemical reactor

- Level of tide at a particular site

- Amplitude of an electrical signal

Time Series → Definition

A discrete time series data may be generated from the accumulation of data over a period of time.

Example of discrete variable:

- Monthly sales
- Daily rainfall
- Production of a crop over different years in a country

Statistical Methods for Data Science

Time Series → Graphs

Fig 14.1 U.S. Population at ten year intervals, 1790-1980.
[*Source*. Brockwell and Davis, 1990.]

Time Series ➡ Graphs

Fig 14.2 District-wise production of rice for four districts in West Bengal.

Time Series ➡ Graphs

Fig 14.3 Short-term movement to and from Australia.
[*Source*: Immigration update, December, Quarter 1991, Bureau of Immigration Research, Statistics Section, Australia.]

Statistical Methods for Data Science

Time Series ➡ Graphs

Fig 14.5 Frequency of homicide victims by month, 1961-1980 (excluding manslaughters and infanticides).
[*Source*: McLeod, MacNeil and Bhattacharyya (1985).]

- **Time Series Models:**
  - A **time series** is a set of observations Y_t , each one being recorded at a specific time *t.* A ***discrete-time*** *time series* is one in which the set T_0 of times at which observations are made is a discrete set, as is the case, for example, when observations are made at fixed time intervals.
  - ***Continuous - time*** *time series* are obtained when observations are recorded continuously over some time interval, e.g., when T_0 [0, 1].

Time Series → Decomposition Models

This is an important technique for all types of time series analysis, especially for seasonal adjustment. It seeks to construct, from an observed time series, a number of component series (that could be used to reconstruct the original by additions or multiplications) where each of these has a certain characteristic or type of behaviour. For example, time series are usually decomposed into:

Time Series → Decomposition of Time Series as Components

Time Series ➡ Decomposition view

- Assess the trend component by smoothing or curve fitting (regression with time)
- Assess/ account for seasonality i.e. deseasonalize the data
  - For Additive Adjustment: Look at all periods of a given type (eg. First Qtr periods where data is quarterly, or a all August period where the data is monthly) & compute an average deviation of the actual values from the smooth or fitted values in those periods. The average can then be added to the trend to adjust for seasonality.
  - For multiplicative adjustment: Instead of calculating the average deviation, compute an average ratio, also called seasonal indices, of the actual values to the smooth or fitted values in those periods. The indices are then used as multiplier to adjust for seasonality.
- Forecast by projecting trend component into the future and then adding or multiplying the seasonal component, as the case may be according to your chosen model

T_t: the trend component at time *t*, which reflects the long-term progression of the series (secular variation). A trend exists when there is a persistent increasing or decreasing direction in the data. **The trend component does not have to be linear**

Trend:

- Persistent, overall upward or downward pattern
- Due to population, technology etc.
- Several years duration

Time Series → Seasonal variation

S_t: the seasonal component at time *t*, reflecting seasonality (seasonal variation). A seasonal pattern exists when a time series is influenced by seasonal factors. Seasonality occurs over a fixed and known period (e.g., the quarter of the year, the month, or day of the week)

Time Series → Seasonal variation

The seasonal variation of a time series is a pattern of change that **recurs** regularly over time.

Seasonal variations are usually due to the differences between seasons and to festive occasions such as Hindu, Christian or Muslim festivals.

Examples include:

- Air conditioner sales in Summer
- Heater sales in Winter
- Flu cases in Winter
- Airline tickets for flights during school vacations

Time Series → Seasonal variation

Seasonality in a time series can be identified by **regularly spaced peaks and troughs** which have a consistent direction and approximately the same magnitude every year, relative to the trend. The following diagram depicts a strongly seasonal series. There is an obvious large seasonal increase in December retail sales in New South Wales due to Christmas shopping. In this example, the magnitude of the seasonal component increases over time, as does the trend.

Statistical Methods for Data Science

Time Series → Seasonal variation

Time Series ➡ Time Series Components

- Regular pattern of up & down fluctuations
- Due to weather, customs etc.
- Occurs within one year

Seasonal Movement

Seasonal movement refers to regular periodic fluctuations that occur in each time period – yearly, monthly, daily. Some examples are speciality cards for Valentine's Day, monthly travel passes and off-peak heating.

Seasonal variations greatly impact on the outcomes of recorded data and often belie the underlying trend. Businesses need to identify the seasonal impact:

- So that a measurement (index) can be used to adjust the expected outcome.

- In order to recognise the direction of the underlying trend.

Time Series → Cyclical variation

C_t: the cyclical component at time *t*, which reflects repeated but non-periodic fluctuations. The duration of these fluctuations is usually of at least two years

Time Series → Cyclical variation

Cyclical variations also have recurring patterns but with a longer and more **erratic time scale** compared to Seasonal variations.

The name is quite misleading because these cycles can be far from regular and it is usually impossible to predict just how long periods of expansion or contraction will be.

There is no guarantee of a regularly returning pattern.

Time Series → Cyclical variation

Example include:

- Floods
- Wars
- Changes in interest rates
- Economic depressions or recessions
- Changes in consumer spending

Cyclical:

- Repeating up & down movements above trend line
- Due to interactions of factors influencing economy
- Usually 2-10 years duration

Time Series ➡ Cyclical variation

Cyclical Movement

This reflects the level of business activity and economic movement over time by fluctuating patterns, known as the economic cycle. These variations measure periods of expansion and contraction in industry and the economy. Their regularity and intensity are not predictable, however certain economic indicators contribute to their existence – level of investment, confidence in the economy, GDP, trade indexes and government policy.

Time Series ➡ Cyclical variation

This chart represents an economic cycle, but we know it doesn't always go like this. The timing and length of each phase is not predictable.

Time Series → Irregular variation

I_t: the irregular component (or "noise") at time *t*, which describes random, irregular influences. It represents the residuals or remainder of the time series after the other components have been removed.

Time Series → Irregular variation

An irregular (or random) variation in a time series occurs over varying (usually short) periods.

It follows no pattern and is by nature unpredictable.

It usually occurs randomly and may be linked to events that also occur randomly.

Irregular variation cannot be explained mathematically.

Time Series → Irregular variation

If the variation cannot be accounted for by secular trend, season or cyclical variation, then it is usually attributed to irregular variation. Example include:

- Sudden changes in interest rates
- Collapse of companies
- Natural disasters
- Sudden shift s in government policy
- Dramatic changes to the stock market
- Effect of Middle East unrest on petrol prices.

Statistical Methods for Data Science

Time Series ➡ Irregular variation

Irregular:

- Erratic, unsystematic, ‘residual’ fluctuations
- Due to random variation or unforeseen events like Union strike, War
- Short duration & nonrepeating
- Rapid changes or bleeps in the data caused by short term unanticipated and non-recurring factors. Irregular fluctuations can happen as often as day to day

© 1984-1994 T/Maker Co.

Time Series → Irregular variation

The irregular component results from short term fluctuations in the series which are neither systematic nor predictable. In a highly irregular series, these fluctuations can dominate movements, which will mask the trend and seasonality. The graph above is of a highly irregular time series.

http://www.youtube.com/watch?v=ca0rDWo7IpI
good you tube video

Irregular Movements

These patterns refer to random variations that impact greatly on the level of business activity, often called natural variation. Some examples are extreme weather patterns (flood, fire, cyclone), extreme business variation (stock market crash, drop in $A), political climate (sudden elections, wars, death of a leader), and industry changes (pilot strikes, waterside strikes.). The resulting patterns will exert a great pressure on the predicted underlying trends and for this reason must be accounted for when planning for the future. However, the irregular movements are unpredictable.

Hence a time series using an additive model can be thought of as:

$$y_t = T_t + C_t + S_t + I_t$$

Where as the multiplicative model can be thought of as:

$$y_t = T_t \times C_t \times S_t \times I_t$$

An additive model would be used when the variations around the trend do not vary with the level of the time series.

Multiplicative model would be appropriate if the trend is proportional to the level of the time series.

Sometimes the trend and cyclical components are grouped into one, called the trend-cycle component. The trend-cycle component can just be referred to as the "trend" component, even though it may contain cyclical behaviour. For example, a seasonal decomposition of time series plot decomposes a time series into seasonal, trend and irregular components using loess and plots the components separately, whereby the cyclical component (if present in the data) is included in the "trend" component plot.

Forecasting is the process of making **predictions of the future** based on **past and present data** and most commonly by analysis of trends. A commonplace example might be estimation of some variable of interest at some specified future date.

Forecasting is required in many situations:

(a) deciding whether to build another power generation plant in the next five years requires forecasts of future demand

(b) scheduling staff in a call centre next week requires forecasts of call volumes

(c) stocking an inventory requires forecasts of stock requirements.

Time Series → Applications of Forecasting

- Forecasts can be required several years in advance

Example – in case of capital investments

- only a few minutes beforehand

Example - telecommunication routing

Whatever the circumstances or time horizons involved, forecasting is an important aid to effective and efficient planning.

- Some things are easier to forecast than others. The time of the sunrise tomorrow morning can be forecast precisely.
- On the other hand, tomorrow's iPhone sale cannot be forecast with any accuracy.

Time Series → Definition of Forecasting

- The predictability of an event or a quantity depends on several factors including:
  - how well we understand the factors that contribute to it;
  - how much data is available;
  - whether the forecasts can affect the thing we are trying to forecast.

- Forecasts of electricity demand can be highly accurate because all three conditions are usually satisfied.
- forecasting currency exchange rates, only one of the conditions is satisfied: there is plenty of available data. However, we have a limited understanding of the factors that affect exchange rates.

Forecasting:

is about predicting the future as accurately as possible, given all of the information available, including historical data and knowledge of any future events that might impact the forecasts

Planning:

is a response to forecasts and goals. Planning involves determining the appropriate actions that are required to make your forecasts match your goals.

Goals:

are what you would like to have happen. Goals should be linked to forecasts and plans, but this does not always occur. Too often, goals are set without any plan for how to achieve them, and no forecasts for whether they are realistic.

Short-term Forecasts

Short-term forecasts generally involve some form of scheduling which may include for example the seasons of the year for planning purposes.

The cyclical and seasonal factors are more important in these situations.

Short-term Forecasts

Such forecasts are usually prepared every 6 months or on a more frequent basis.

These forecasts are needed for the scheduling of personnel, production and transportation. As part of the scheduling process, forecasts of demand are often also required.

Medium-term Forecasts

Medium-term forecasts are generally prepared for planning, scheduling, budgeting and resource requirements purposes.

Medium-term forecasts are needed to determine future resource requirements, in order to purchase raw materials, hire personnel, or buy machinery and equipment.

Medium-term Forecasts

The trend factor, as well as the cyclical component, plays a key role in the medium-term forecast as the year to year variations in traffic growth are an important element in the planning process

Long-term Forecasts

Long-term forecasts are used mostly in connection with strategic planning to determine the level and direction of capital expenditures and to decide on ways in which goals can be accomplished.

The trend element generally dominates long term situations and must be considered in the determination of any long-run decisions.

Long-term Forecasts

The methods generally found to be most appropriate in long-term situations are econometric analysis and life-cycle analysis. Long-term forecasts are used in strategic planning. Such decisions must take account of market opportunities, environmental factors and internal resources.

Long-term Forecasts

It is also important that since the time span of the forecast horizon is long, forecasts should be calibrated and revised at periodic intervals (every two or three years depending on the situation).

Time Series → The basic steps in a forecasting task

Step 1: Problem definition

Step 2: Gathering information

Step 3: Preliminary (exploratory) analysis

Step 4: Choosing and fitting models

Step 5: Using and evaluating a forecasting model

Time Series → Types of forecasting methods

Forecasting Methods

Qualitative

- Personal Opinion or Judgment
- Panel consensus
- Delphi Method
- Market Research

Quantitative

Time Series

- Smoothing methods
- Exponential smoothing
- Trend projection method

Causal

- Regression Trend Analysis

Time Series → Qualitative forecasting

- **Personal Opinion**
  - Individuals forecasts future based on their own judgment or opinion without any formal model
  - Such assessments are relatively reliable and accurate

Time Series → Qualitative forecasting

- **Panel consensus**
  - Reduces the prejudice and ignorance that may arise in the individual judgment
  - Possible to develop consensus among group of individuals
  - Panel of individuals are encouraged to share information, opinions, and assumptions, if any, to predict future value of some variable under study

Delphi Method

The Delphi method was invented by Olaf Helmer and Norman Dalkey of the Rand Corporation in the 1950s for the purpose of addressing a specific military problem. The method relies on the key assumption that forecasts from a group are generally more accurate than those from individuals.

Delphi Method

The aim of the Delphi method is to construct consensus forecasts from a group of experts in a structured iterative manner. A facilitator is appointed in order to implement and manage the process. The Delphi method generally involves the following stages:

Delphi Method

- A panel of experts is assembled.
- Forecasting tasks/challenges are set and distributed to the experts.
- Experts return initial forecasts and justifications. These are compiled and summarised in order to provide feedback.

Delphi Method

- Feedback is provided to the experts, who now review their forecasts in light of the feedback. This step may be iterated until a satisfactory level of consensus is reached.
- Final forecasts are constructed by aggregating the experts' forecasts.

Delphi Method

- Each stage of the Delphi method comes with its own challenges. In what follows, we provide some suggestions and discussions about each one of these.

- **Market Research Method**
  - This method is used to collect based on well-defined objectives and assumptions about future value of variable.
  - Questionnaire is used to gather data and prepared summary of responses.
  - This method produces a narrow range of forecasts rather than a single view of future.

Time Series → Qualitative forecasting

Type	Characteristics	Strengths	Weaknesses
Executive opinion	A group of managers meet & come up with a forecast	Good for strategic or new-product forecasting	One person's opinion can dominate the forecast
Market research	Uses surveys & interviews to identify customer preferences	Good determinant of customer preferences	It can be difficult to develop a good questionnaire
Delphi method	Seeks to develop a consensus among a group of experts	Excellent for forecasting long-term product demand, technological	Time consuming to develop

Time Series
Trend?
No
Yes
Smoothing Methods
Trend Models
Moving Average
Exponential Smoothing
Linear
Quadratic
Exponential
Auto-Regressive

Modeling Time Series

Methods for forecasting a time series fall into two general classes: *smoothing methods* and *regression-based modeling methods*.

Although the smoothing methods do not explicitly use the time series components, it is a good idea to keep them in mind. The regression models explicitly estimate the components as a basis for building models.

Time Series ➡ Fitting a trend to Time Series

A linear trend between time series values Y_t and the time t is given by the equation $Y_t = a + bt + e$.

For the pair of values of (Y_t, t), a & b as estimated by the principles of least squares are

$$\hat{b} = \frac{\sum xY_t}{\sum x^2}$$

and

$$\hat{a} = \frac{\sum Y_t}{n}$$

$$where\ for\ odd\ numbers$$

$$x = \frac{t - middle\ value\ of\ t}{interval\ (h)}$$

and for even numbers

$$x = \frac{t - (mean\ of\ middle\ two\ values\ of\ t)}{\frac{1}{2}\,interval\ (h)}$$

Time Series ➡ Fitting a trend to Time Series

The following data relates to production of crops (in tons) in a place from 2016 -2020. Fit a linear trend $Y_t = a + bt + e$ for the data given below:

	Year (t)	Production (tons) (Y_t)	x	x^2	xY_t
	2016	35	-2	4	-70
	2017	55	-1	1	-55
	2018	79	0	0	0
	2019	80	1	1	80
	2020	40	2	4	80
Sum	10090	289	0	10	35

$$\boldsymbol{x = \frac{t - 2018}{1} = t - 2018}$$

$$\hat{b} = \frac{\sum xY_t}{\sum x^2} = \frac{35}{10} = 3.5$$

$$\hat{a} = \frac{\sum Y_t}{n} = \frac{289}{5} = 57.8$$

$$\hat{Y}_t = \hat{a} + \hat{b}\,x = 57.8 + 3.5x$$

The forecasted value Y_t *for the year 2021 is*

$$\hat{Y}_t = \hat{a} + \hat{b}\,x = 57.8 + 3.5\,(3) = 68.3$$

Time Series ➡ Fitting a trend to Time Series

The following data relates to Sale (in '000) in a Mall from 2016 -2021. Fit a linear trend $Y_t = a + bt + e$ for the data given below:

	Year (t)	Sales ('000) (Y_t)	x	x^2	xY_t
	2016	23	-5	25	-115
	2017	27	-3	9	-81
	2018	31	-1	1	-31
	2019	34	1	1	34
	2020	22	3	9	66
	2021	29	5	25	145
Sum	**12111**	**166**	**0**	**70**	**18**

$$\hat{b} = \frac{\sum xY_t}{\sum x^2} = \frac{18}{70} = 0.2571$$

$$\hat{a} = \frac{\sum Y_t}{n} = \frac{12111}{6} = 27.6667$$

$$\hat{Y}_t = \hat{a} + \hat{b}\,x = 27.6667 + 0.2571x$$

The forecasted value Y_t for the year 2022 is

$$\boldsymbol{x = \frac{t - (2018 + 2019)/2}{1/2} = 2(t - 2018.5)}$$

$$\boldsymbol{\hat{Y}_t = \hat{a} + \hat{b}\,x = 27.6667 + 0.2571\,(7) = 29.4664}$$

Time Series ➡ Fitting a trend to Time Series

A trend need not be a linear, it may even be quadratic. Let the quadratic trend between time series values Y_t and the time t is given by the equation

$Y_t = a + bt + ct^2 + e.$

For the pair of values of (Y_t, t), a, b and c as estimated by the principles of least squares is given by the system of three normal equations as follows:

$$\sum Y_t = na + b\sum x + c\sum x^2$$

$$\sum xY_t = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2Y_t = a\sum x^2 + b\sum x^3 + c\sum x^4$$

As there are three equations in three unknown quantities, any of the numerical techniques for solving the system of equations can be made use like Gauss – Elimination method.

Note: x is defined as in case of fitting linear trend

Time Series ➡ Smoothing Techniques – Moving Averages

- Appropriate for a time series with a horizontal pattern ie., the data that are stationary.
  - Moving Average (the average of the most recent k data values forms the forecast for the next period)

$$F_{t+1} = \frac{\sum \text{most recent k data values}}{k}$$

$$F_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-k+1}}{k}$$

Statistical Methods for Data Science

Time Series → Simple Moving Average

Following data shows production volume (in '000 tones). Compute 3 and 5-year moving average

Period (t)	Sales (Y_t)	3yr MA sum	3yr MA sum	5yr MA sum	MA sum
1	5.3				
2	4.4	15.1	5.03		
3	5.4	15.6	5.20	26.5	5.30
4	5.8	16.8	5.60	26	5.20
5	5.6	16.2	5.40	27.2	5.44
6	4.8	16	5.33	27.4	5.48
7	5.6	16	5.33	27	5.40
8	5.6	16.6	5.53	27.9	5.58
9	5.4	17.5	5.83	28.2	5.64
10	6.5	17	5.67	28.4	5.68
11	5.1	17.4	5.80	27.8	5.56
12	5.8	15.9	5.30	28.6	5.72
13	5	17	5.67	27.7	5.54
14	6.2	16.8	5.60	29.3	5.86
15	5.6	18.5	6.17	28.7	5.74
16	6.7	17.5	5.83	29.2	5.84
17	5.2	17.4	5.80	28.8	5.76
18	5.5	16.5	5.50	28.3	5.66
19	5.8	16.4	5.47	27.4	5.48
20	5.1	16.7	5.57	28.9	5.78
21	5.8	17.6	5.87	28.6	5.72
22	6.7	17.7	5.90	28.8	5.76
23	5.2	17.9	5.97	29.5	5.90
24	6	17	5.67		
25	5.8				

Time Series → Simple Moving Average

- The idea behind the moving averages is that observations which are nearby in time are also likely to be close in value.
- The average of the points near an observation will provide a reasonable estimate of the trend-cycle at that observation.
- The average eliminate some of the randomness in the data, and leaves a smooth trend-cycle component.

Time Series ➡ Simple Moving Average

- Appropriate for a time series with a horizontal pattern ie., the data that are stationary.
  - Moving Average (the average of the most recent k data values forms the forecast for the next period)

$$F_{t+1} = \frac{\sum \text{most recent k data values}}{k}$$

$$F_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-k+1}}{k}$$

Statistical Methods for Data Science

Time Series ➡ Centered – Moving Averages

Following data shows production volume (in '000 tones). Compute 3-year moving average for all available years

Year	Production	3 - year moving total	3 - year moving average
1995	21		
1996	22	66	22.00
1997	23	70	23.33
1998	25	72	24.00
1999	24	71	23.67
2000	22	71	23.67
2001	25	73	24.33
2002	26	78	26.00
2003	27	79	26.33
2004	26		

Statistical Methods for Data Science

Time Series → Centered – Moving Averages

Following data shows production volume (in '000 tones). Compute 4-year centered moving average

Year	Production(Y)	4-yearly moving average	4-year centered moving average
1995	21.0		
1996	22.0		
		22.75	
1997	23.0		23.13
		23.50	
1998	25.0		23.50
		23.50	
1999	24.0		23.75
		24.00	
2000	22.0		24.13
		24.25	
2001	25.0		24.63
		25.00	
2002	26.0		25.50
		26.00	
2003	27.0		26.38
		26.75	
2004	26.0		26.75
2005	28.0		

Statistical Methods for Data Science

Time Series ➡ Smoothing Techniques – Moving Averages

The following data represent 15 quarters of manufacturing capacity utilization

Quarter/Year	Utilization (%)
1/2003	82.50
2/2003	81.30
3/2003	81.30
4/2003	79.00
1/2004	76.60
2/2004	78.00
3/2004	78.40
4/2004	78.00
1/2005	78.80
2/2005	78.70
3/2005	78.40
4/2005	80.00
1/2006	80.70
2/2006	80.70
3/2006	80.80

Observe the 2,3 & 4 quarter moving average (drawn using Excel)

Thank you

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Sessions 14 Time Series Analysis

(2nd Sep 2023)

Time Series

Calculate the exponentially smoothed series F_t from the original time series Y_t as follows:

$$F_2 = Y_1$$

$$F_3 = \alpha Y_2 + (1 - \alpha)F_2$$

$$F_4 = \alpha\, Y_3 + (1 - \alpha)F_3$$

$$F_{t+1} = \alpha\, Y_t + (1 - \alpha)F_t$$

Time Series ➡ Smoothing Techniques – Exponential

- Exponential (Weighted average of all the past time series values with exponentially decreasing importance in the forecast)

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t$$

- where $F_t = \alpha Y_{t-1} + (1-\alpha)F_{t-1}$
- α=1 ⇒ Forecast value is just the previous actual value. A naïve forecast.
- α=0 ⇒ Current Forecast value is just the previous forecast value.

Quar/Yr	Utilization(%)	F_t (α=0.2)	F_t (α=0.4)	F_t (α=0.7)
Q1/2017	82.50	82.50	82.50	82.50
Q2/2017	81.30	82.50	82.50	82.50
Q3/2017	81.30	82.26	82.02	81.66
Q4/2017	79.00	82.07	81.73	81.41
Q1/2018	76.60	81.45	80.64	79.72
Q2/2018	78.00	80.48	79.02	77.54
Q3/2018	78.40	79.99	78.61	77.86
Q4/2018	78.00	79.67	78.53	78.24
Q1/2019	78.80	79.34	78.32	78.07
Q2/2019	78.70	79.23	78.51	78.58
Q3/2019	78.40	79.12	78.59	78.66
Q4/2019	80.00	78.98	78.51	78.48
Q1/2020	80.70	79.18	79.11	79.54
Q2/2020	80.70	79.69	80.14	81.05
Q3/2020	80.80	80.33	81.25	82.35
Q4/2006		**80.23**	**81.05**	**82.04**

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t$$

Quar/Yr	Utilization(%)	F_t (α=0.1)	F_t (α=0.2)	F_t (α=0.3)	F_t (α=0.4)	F_t (α=0.5)	F_t (α=0.6)	F_t (α=0.7)
Q1/2017	82.50	82.50	82.50	82.50	82.50	82.50	82.50	82.50
Q2/2017	81.30	82.50	82.50	82.50	82.50	82.50	82.50	82.50
Q3/2017	81.30	82.38	82.26	82.14	82.02	81.90	81.78	81.66
Q4/2017	79.00	82.27	82.07	81.89	81.73	81.60	81.49	81.41
Q1/2018	76.60	81.94	81.45	81.02	80.64	80.30	80.00	79.72
Q2/2018	78.00	81.41	80.48	79.70	79.02	78.45	77.96	77.54
Q3/2018	78.40	81.07	79.99	79.19	78.61	78.23	77.98	77.86
Q4/2018	78.00	80.80	79.67	78.95	78.53	78.31	78.23	78.24
Q1/2019	78.80	80.52	79.34	78.67	78.32	78.16	78.09	78.07
Q2/2019	78.70	80.35	79.23	78.71	78.51	78.48	78.52	78.58
Q3/2019	78.40	80.18	79.12	78.70	78.59	78.59	78.63	78.66
Q4/2019	80.00	80.01	78.98	78.61	78.51	78.49	78.49	78.48
Q1/2020	80.70	80.01	79.18	79.03	79.11	79.25	79.40	79.54
Q2/2020	80.70	80.18	79.69	79.83	80.14	80.47	80.78	81.05
Q3/2020	80.80	80.45	80.33	80.75	81.25	81.69	82.05	82.35
Q4/2006		**80.33**	**80.23**	**80.61**	**81.05**	**81.45**	**81.78**	**82.04**
MSE		**6.36**	**4.76**	**3.94**	**3.41**	**3.07**	**2.85**	**2.71**

Quar/Yr	Utilization(%)	F_t (α=0.8)	F_t (α=0.9)
Q1/2017	82.50	82.50	82.50
Q2/2017	81.30	82.50	82.50
Q3/2017	81.30	81.54	81.42
Q4/2017	79.00	81.35	81.31
Q1/2018	76.60	79.47	79.23
Q2/2018	78.00	77.17	76.86
Q3/2018	78.40	77.83	77.89
Q4/2018	78.00	78.29	78.35
Q1/2019	78.80	78.06	78.03
Q2/2019	78.70	78.65	78.72
Q3/2019	78.40	78.69	78.70
Q4/2019	80.00	78.46	78.43
Q1/2020	80.70	79.69	79.84
Q2/2020	80.70	81.30	81.51
Q3/2020	80.80	82.58	82.76
Q4/2006		**82.25**	**82.42**
MSE		**2.62**	**2.56**

- To try to determine which one of these forecasting methods gives the “best” forecast, we evaluate the “success” of each method using a ***performance measure***
- Each performance measure begins by evaluating the ***forecast errors*** Δ_t for each time period t given by

Time Series → Measurement of forecast error

What was forecasted to happen at time t using the forecast technique

Forecast Error at time t

$$\Delta_t = y_t - F_t$$

What actually happened at time t

Time Series → Performance Measures

Mean Square Error (MSE)

$$MSE = \frac{\sum (\Delta_t)^2}{n}$$

Mean Absolute Percent Error (MAPE)

$$MAPE = \frac{\sum \frac{|\Delta_t|}{y_t}}{n} 100$$

Mean Absolute Deviation (MAD)

$$MAD = \frac{\sum |\Delta_t|}{n}$$

Largest Absolute Deviation (LAD)

$$LAD = Max\,|\Delta_t|$$

- Done the analysis on the previous problem for 3 different values of α
- For α=0.7, it is better
- The exponential smoothing does not have a significant initial delay like the Moving Av. where the initial k values are not smoothed
- Formula can be

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t$$

Quar/Yr	Utilization(%)	F_t (α=0.2)	F_t (α=0.4)	F_t (α=0.7)
Q1/2017	82.50	82.50	82.50	82.50
Q2/2017	81.30	82.50	82.50	82.50
Q3/2017	81.30	82.26	82.02	81.66
Q4/2017	79.00	82.07	81.73	81.41
Q1/2018	76.60	81.45	80.64	79.72
Q2/2018	78.00	80.48	79.02	77.54
Q3/2018	78.40	79.99	78.61	77.86
Q4/2018	78.00	79.67	78.53	78.24
Q1/2019	78.80	79.34	78.32	78.07
Q2/2019	78.70	79.23	78.51	78.58
Q3/2019	78.40	79.12	78.59	78.66
Q4/2019	80.00	78.98	78.51	78.48
Q1/2020	80.70	79.18	79.11	79.54
Q2/2020	80.70	79.69	80.14	81.05
Q3/2020	80.80	80.33	81.25	82.35
Q4/2006		**80.23**	**81.05**	**82.04**
MSE		**4.76**	**3.41**	**2.71**

Time Series ➡ Exponential Smoothing Techniques

1. Select an exponential smoothing constant, α, between 0 and 1. Remember that **small values of α give less weight to the current value of the series** and **yield a smoother series**. **Larger choices of α** assign **more weight to the current value of the series** and **yield a more variable series.**

Statistical Methods for Data Science

Time Series ➡ Smoothing Techniques – Exponential

- Done the analysis on the previous problem for 3 different values of α
- For α=0.7, it is better
- The exponential smoothing does not have a significant initial delay like the Moving Av. where the initial k values are not smoothed
- Formula can be rearranged as

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t$$

Quarter/ Year	Utilization (%)	F_t (α=0.2)	F_t (α=0.4)	F_t (α=0.7)
Q1/2017	82.50			
Q2/2017	81.30	82.5	82.5	82.5
Q3/2017	81.30	82.26	82.02	81.66
Q4/2017	79.00	81.61	80.81	79.80
Q1/2018	76.60	80.61	79.13	77.56
Q2/2018	78.00	80.09	78.68	77.87
Q3/2018	78.40	79.75	78.57	78.24
Q4/2018	78.00	79.40	78.34	78.07
Q1/2019	78.80	79.28	78.52	78.58
Q2/2019	78.70	79.16	78.59	78.66
Q3/2019	78.40	79.01	78.52	78.48
Q4/2019	80.00	79.21	79.11	79.54
Q1/2020	80.70	79.51	79.75	80.35
Q2/2020	80.70	79.75	80.13	80.60
Q3/2020	80.80	79.96	80.40	80.74
	MSE	2.70	1.04	0.26

$$\Delta_t = Y_t - F_t$$

Quar/Yr	Utilization (%)	F_t (α=0.2)	Δ_t^2(α=0.2)
Q1/2017	82.50	82.50	0.00
Q2/2017	81.30	82.26	0.92
Q3/2017	81.30	82.07	0.59
Q4/2017	79.00	81.45	6.00
Q1/2018	76.60	80.48	15.05
Q2/2018	78.00	79.99	3.96
Q3/2018	78.40	79.67	1.61
Q4/2018	78.00	79.34	1.80
Q1/2019	78.80	79.23	0.18
Q2/2019	78.70	79.12	0.18
Q3/2019	78.40	78.98	0.34
Q4/2019	80.00	79.18	0.67
Q1/2020	80.70	79.49	1.46
Q2/2020	80.70	79.73	0.94
Q3/2020	80.80	79.94	0.74
	MSE		**2.297**

$$\Delta_t = Y_t - F_t$$

Quar/Yr	Utilization (%)	F_t (α=0.4)	Δ_t^2(α=0.4)
Q1/2017	82.50	82.5	0.00
Q2/2017	81.30	82.02	0.52
Q3/2017	81.30	81.73	0.18
Q4/2017	79.00	80.64	2.69
Q1/2018	76.60	79.02	5.86
Q2/2018	78.00	78.61	0.37
Q3/2018	78.40	78.53	0.02
Q4/2018	78.00	78.32	0.10
Q1/2019	78.80	78.51	0.08
Q2/2019	78.70	78.59	0.01
Q3/2019	78.40	78.51	0.01
Q4/2019	80.00	79.11	0.79
Q1/2020	80.70	79.74	0.92
Q2/2020	80.70	80.13	0.32
Q3/2020	80.80	80.4	0.16
	MSE		**0.803**

$$\Delta_t = Y_t - F_t$$

Quar/Yr	Utilization (%)	F_t (α=0.7)	Δ_t^2(α=0.7)
Q1/2017	82.50	82.50	0.00
Q2/2017	81.30	81.66	0.13
Q3/2017	81.30	81.41	0.01
Q4/2017	79.00	79.72	0.52
Q1/2018	76.60	77.54	0.88
Q2/2018	78.00	77.86	0.02
Q3/2018	78.40	78.24	0.03
Q4/2018	78.00	78.07	0.00
Q1/2019	78.80	78.58	0.05
Q2/2019	78.70	78.66	0.00
Q3/2019	78.40	78.48	0.01
Q4/2019	80.00	79.54	0.21
Q1/2020	80.70	80.35	0.12
Q2/2020	80.70	80.6	0.01
Q3/2020	80.80	80.74	0.00
	MSE		**0.133**

Statistical Methods for Data Science

Time Series ➡ Exponential Smoothing Techniques

- Done the analysis on the previous problem for 3 different values of α
- For α=0.7, it is better
- The exponential smoothing does not have a significant initial delay like the Moving Average, where the initial k values are not smoothed
- Formula can be rearranged as

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t$$

Quarter/ Year	Utilization (%)	F_t (α=0.2)	F_t (α=0.4)	F_t (α=0.7)
Q1/2017				
Q2/2017	81.30	82.26	82.02	81.66
Q3/2017	81.30	82.07	81.73	81.41
Q4/2017	79.00	81.45	80.64	79.72
Q1/2018	76.60	80.48	79.02	77.54
Q2/2018	78.00	79.99	78.61	77.86
Q3/2018	78.40	79.67	78.53	78.24
Q4/2018	78.00	79.34	78.32	78.07
Q1/2019	78.80	79.23	78.51	78.58
Q2/2019	78.70	79.12	78.59	78.66
Q3/2019	78.40	78.98	78.51	78.48
Q4/2019	80.00	79.18	79.11	79.54
Q1/2020	80.70	79.49	79.74	80.35
Q2/2020	80.70	79.73	80.13	80.60
Q3/2020	80.80	79.94	80.40	80.74
	MSE	2.30	0.80	0.13

Statistical Methods for Data Science

Time Series ➡ Smoothing Techniques – Moving Averages

- The Moving Average basically filters out rapid fluctuations .i.e. high frequency noise. Thus it acts as a low-pass filter.
- We can choose the index of the MA filter by trial and error. Comparing the Forecast Accuracy for index selection.
- For our eg., we find that MSE for the 2-Qtr filter is better

Quarter/Year	Utilization (%)	2-Qtr MA	3-Qtr MA	4-Qtr MA
1/2003	82.50			
2/2003	81.30	81.90	81.70	
3/2003	81.30	81.30	80.53	
4/2003	79.00	80.15	78.97	81.03
1/2004	76.60	77.80	77.87	79.55
2/2004	78.00	77.30	77.67	78.73
3/2004	78.40	78.20	78.13	78.00
4/2004	78.00	78.20	78.40	77.75
1/2005	78.80	78.40	78.50	78.30
2/2005	78.70	78.75	78.63	78.48
3/2005	78.40	78.55	79.03	78.48
4/2005	80.00	79.20	79.70	78.98
1/2006	80.70	80.35	80.47	79.45
2/2006	80.70	80.70	80.73	79.95
3/2006	80.80	80.75		80.55
	MSE	0.33	0.84	1.42

Statistical Methods for Data Science

Time Series ➡ Smoothing Techniques – Moving Averages

- The Moving Average basically filters out rapid fluctuations .i.e. high frequency noise. Thus it acts as a low-pass filter.
- We can choose the index of the MA filter by trial and error. Comparing the Forecast Accuracy for index selection.
- For our eg., we find that MSE for the 2-Qtr filter is better

Quarter/Year	Utilization (%)	2-Qtr MA	3-Qtr MA	4-Qtr MA
Q1/2017	82.50			
Q2/2017	81.30	81.90	81.70	
Q3/2017	81.30	81.30	80.53	
Q4/2017	79.00	80.15	78.97	81.03
Q1/2018	76.60	77.80	77.87	79.55
Q2/2018	78.00	77.30	77.67	78.73
Q3/2018	78.40	78.20	78.13	78.00
Q4/2018	78.00	78.20	78.40	77.75
Q1/2019	78.80	78.40	78.50	78.30
Q2/2019	78.70	78.75	78.63	78.48
Q3/2019	78.40	78.55	79.03	78.48
Q4/2019	80.00	79.20	79.70	78.98
Q1/2020	80.70	80.35	80.47	79.45
Q2/2020	80.70	80.70	80.73	79.95
Q3/2020	80.80	80.75		80.55
	MSE	0.33	0.84	1.42

Statistical Methods for Data Science

Time Series → Smoothing Techniques – Moving Averages

Quarter/	Utilization	2-Qtr	Abs(Δt)	$(\Delta_t)^2$	3-Qtr	Abs(Δt)	$(\Delta_t)^2$	4-Qtr	Abs(Δt)	$(\Delta_t)^2$
Q1/17	82.5									
Q2/17	81.3	81.9	0.6	0.36	81.7					
Q3/17	81.3	81.3	0	0.00	80.53	0.4	0.16			
Q4/17	79	80.15	1.15	1.32	78.97	1.53	2.34	81.03	2.03	4.12
Q1/18	76.6	77.8	1.2	1.44	77.87	2.37	5.62	79.55	2.95	8.70
Q2/18	78	77.3	0.7	0.49	77.67	0.13	0.02	78.73	0.73	0.53
Q3/18	78.4	78.2	0.2	0.04	78.13	0.73	0.53	78	0.4	0.16
Q4/18	78	78.2	0.2	0.04	78.4	0.13	0.02	77.75	0.25	0.06
Q1/19	78.8	78.4	0.4	0.16	78.5	0.4	0.16	78.3	0.5	0.25
Q2/19	78.7	78.75	0.05	0.00	78.63	0.2	0.04	78.48	0.22	0.05
Q3/19	78.4	78.55	0.15	0.02	79.03	0.23	0.05	78.48	0.08	0.01
Q4/19	80	79.2	0.8	0.64	79.7	0.97	0.94	78.98	1.02	1.04
Q1/20	80.7	80.35	0.35	0.12	80.47	1	1.00	79.45	1.25	1.56
Q2/20	80.7	80.7	0	0.00	80.73	0.23	0.05	79.95	0.75	0.56
Q3/20	80.8	80.75	0.05	0.00		0.07	0.00	80.55	0.25	0.06
	Sum		**5.85**	4.64		**8.39**	10.94		**10.43**	17.11
	MSE			**0.33**			**0.84**			**1.43**
	MAD		**0.42**			**0.65**			**0.87**	

Time Series ➡ Which performance should be used?

- Choice of the modeler
- **Advantages**
  - MSE gives greater weight to larger deviations (which could result from outliers)
  - MAD gives less weight to larger deviations
  - MAPE gives less overall weight to a large deviation if the time series value is large
  - LAD tells us if all deviations fall below some threshold value
- Although we illustrate all 4 techniques, in general focus will be on MSE and MAD.

Statistical Methods for Data Science

Time Series ➡ Smoothing Techniques – Moving Averages

Quarter/	Utilization	2-	Abs(Δt)	$(\Delta_t)^2$	3-	Abs(Δt)	$(\Delta_t)^2$	4-	Abs(Δt)	$(\Delta_t)^2$
Q1/17	82.5									
Q2/17	81.3	81.9	0.6	0.36	81.7					
Q3/17	81.3	81.3	0	0.00	80.5	0.4	0.16			
Q4/17	79	80.2	1.15	1.32	79	1.53	2.34	81	2.03	4.12
Q1/18	76.6	77.8	1.2	1.44	77.9	2.37	5.62	79.6	2.95	8.70
Q2/18	78	77.3	0.7	0.49	77.7	0.13	0.02	78.7	0.73	0.53
Q3/18	78.4	78.2	0.2	0.04	78.1	0.73	0.53	78	0.4	0.16
Q4/18	78	78.2	0.2	0.04	78.4	0.13	0.02	77.8	0.25	0.06
Q1/19	78.8	78.4	0.4	0.16	78.5	0.4	0.16	78.3	0.5	0.25
Q2/19	78.7	78.8	0.05	0.00	78.6	0.2	0.04	78.5	0.22	0.05
Q3/19	78.4	78.6	0.15	0.02	79	0.23	0.05	78.5	0.08	0.01
Q4/19	80	79.2	0.8	0.64	79.7	0.97	0.94	79	1.02	1.04
Q1/20	80.7	80.4	0.35	0.12	80.5	1	1.00	79.5	1.25	1.56
Q2/20	80.7	80.7	0	0.00	80.7	0.23	0.05	80	0.75	0.56
Q3/20	80.8	80.8	0.05	0.00		0.07	0.00	80.6	0.25	0.06
	Sum		**5.85**	4.64		**8.39**	10.94		**10.43**	17.11
	MSE			**0.33**			**0.84**			**1.43**
	MAD		**0.42**			**0.65**			**0.87**	

Time Series ➡ Smoothing Techniques – Exponential

- Objective is to estimate the overall time series as a combination of long term trend and seasonality
  - Additive Model: $X_t = F_t + S_t + \varepsilon_t$
  - Multiplicative Model: $X_t = F_t \times S_t \times \varepsilon_t$
  - Multiplicative model can be converted to additive model by taking log. Other models are also possible
- The trend and seasonality can be decomposed using smoothing and regression methods
- Exponential smoothing is suitable with constant variance and no seasonality. Recommended for short-term forecast.
- Another method for stationarizing the time series is by doing transformation .eg. Differencing

Time Series ➡ Holt Double Exponential Smoothing Technique

- The simple exponential smoothing forecast is a weighted sum of actual value with **α** and the forecast for the same period with (1-**α**). At **α** =1, it is identical to the naïve forecast.
- However the simple exponential smoothing forecasts lag the actual time series values when the time series has a trend.
- The trend makes it difficult for forecast to ‘keep up’ since there is no account of changes in level from period to period.
- The Holt double exponential smoothing method adds the ability to account for the trend component the time series in addition to the level

Time Series ➡ Holt Double Exponential Smoothing Technique

- Holt double exponential is a simple exponential smoothing with a added trend components that helps fix the lag issues
- Weights are used to smooth the level and trend.
- Holt smoothed value calculation for the level is exactly same calculation as in simple exponential smoothing
- The trend component is a weighted difference between the current smoothed level and the prior period smoothed level plus prior trend.
- Finally, the Holt double exponential smoothing forecast for the next period is a sum of smoothed level and smoothed trend of the current period

Statistical Methods for Data Science

Time Series → Holt Double Exponential Smoothing Technique

Y_t
Current actual level
α

+

$L_{t-1} + T_{t-1}$
Prior forecast level
$1-\alpha$

→

Current smooth level
L_t

+

$L_t - L_{t-1}$
Current smoothed level – Prior smoothed level
β

+

T_{t-1}
Prior smoothed trend
$1-\beta$

→

Current smooth trend
T_t

→

Forecast value
$F_{t+1}=L_t+T_t$

Time Series ➡ Holt Double Exponential Smoothing Technique

Current smoothed level: $L_t = \boldsymbol{\alpha} Y_t + (1 - \boldsymbol{\alpha})(L_{t-1} + T_{t-1})$

Current smoothed Trend: $T_t = \boldsymbol{\beta}(L_t - L_{t-1}) + (1 - \boldsymbol{\beta})\, T_{t-1}$

Forecast next period: $F_{t+1} = L_t + T_t$

Forecast next k period: $F_{t+\mathrm{k}} = L_t + kT_t$

Level smoothing factor: $0 \leq \boldsymbol{\alpha} \leq 1$

Level smoothing factor: $0 \leq \boldsymbol{\beta} \leq 1$

Time Series ➡ Holt Double Exponential Smoothing Technique

Current smoothed level: $L_t = \boldsymbol{\alpha} Y_t + (1 - \boldsymbol{\alpha})(L_{t-1} + T_{t-1})$

Level smoothing constant

Current data value

Prior period smoothed level

Prior period smoothed trend

Simple exponential smoothing

Time Series ➡ Holt Double Exponential Smoothing Technique

When $\boldsymbol{\alpha} = 0$ ➡ $L_t = (\mathbf{0})Y_t + (1 - (\mathbf{0}))(L_{t-1} + T_{t-1})$

$$L_t = L_{t-1} + T_{t-1}$$

$$L_t = F_t$$

Smoothed level = Current forecast

Time Series ➡ Holt Double Exponential Smoothing Technique

When $\boldsymbol{\alpha} = 1$ ➡ $L_t = (\mathbf{1})Y_t + (1 - (\mathbf{1}))(L_{t-1} + T_{t-1})$

$$L_t = Y_t$$

Smoothed level = Current actual level of the data

Time Series ➡ Holt Double Exponential Smoothing Technique

Current smoothed trend: $T_t = \boldsymbol{\beta}(L_t - L_{t-1}) + (1 - \boldsymbol{\beta})T_{t-1}$

Trend smoothing constant

Current smoothed level

Prior period smoothed level

Prior period smoothed trend

Time Series ➡ Holt Double Exponential Smoothing Technique

When $\boldsymbol{\beta}$= 0 ➡

$$T_t = \boldsymbol{\beta}(L_t - L_{t-1}) + (1 - \boldsymbol{\beta})T_{t-1}$$

$$T_t = (\mathbf{0})(L_t - L_{t-1}) + (1 - (0))T_{t-1}$$

$$T_t = T_{t-1}$$

Smoothed trend = Trend from the last period

Time Series ➡ Holt Double Exponential Smoothing Technique

When $\boldsymbol{\beta}$= 1 ➡ $T_t = \boldsymbol{\beta}(L_t - L_{t-1}) + (1 - \boldsymbol{\beta})T_{t-1}$

$$T_t = (1)(L_t - L_{t-1}) + (1 - (\mathbf{1}))T_{t-1}$$

$$T_t = L_t - L_{t-1}$$

Current smoothed trend = Current smoothed level – prior smoothed level

Time Series ➡ Holt Double Exponential Smoothing Technique

Trend smoothing constant (β)	Level smoothing constant (α): 0	Level smoothing constant (α): 1
1	Forecast all the same (flat line) of Y_t	Smoothed level is the same as actual level. Trend is the full difference between smoothed level
0	Forecast all the same (flat line) of Y_t	Same as Naïve fore cast, no trend component

Statistical Methods for Data Science

Time Series ➡ Holt Double Exponential Smoothing Technique

Time Series ➡ Holt Double Exponential Smoothing Technique

Example: Year & GDP

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
GDP	1451	1499	1686	1764	1879	1948	2013	2090	2173	2286	2404
Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	
GDP	2465	2501	2597	2689	2688	2576	2530	2513	2503	2396	

Statistical Methods for Data Science

Time Series ➡ Holt Double Exponential Smoothing Technique

Example: Year & GDP

Year	GDP	Smoothed L_t at α=0.8	Smoothed T_t at β=0.4	Forecast
		$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1})$	$T_t = \beta (L_t - L_{t-1}) + (1-\beta)T_{t-1}$	$F_{t+1} = L_t + T_t$
2000	1451	1451	0	
2001	1499	0.8*1499+(1-0.8)*(1451+0)=1489	0.4*(1489-1451)+(1-0.4)*0=15	1451+0 = 1451
2002	1686	0.8*1686+(1-0.8)*(1489+15)=1650	0.4*(1650-1489)+(1-0.4)*15=73	1489+15=1505
2003	1764	0.8*1764+(1-0.8)*(1650+73)=1756	0.4*(1756-1650)+(1-0.4)*73=86	1650+73=1723
2004	1879	0.8*1879+(1-0.8)*(1756+86)=1872	0.4*(1872-1756)+(1-0.4)*86=98	1756+86=1842
2005	1948	0.8*1948+(1-0.8)*(1872+98)=1952	0.4*(1952-1872)+(1-0.4)*98=91	1872+98=1970
2006	2013			1952+91=2044
. . .	. . .	. . .	. . .	. . .
2018	2513	0.8*2513+(1-0.8)*(2546-28)=2514	0.4*(2514-2546)+(1-0.4)*(-28)=-30	1872-28=2518
2019	2503	0.8*2503+(1-0.8)*(2514-30)=2499	0.4*(2499-2514)+(1-0.4)*(-30)=-24	2514-30=2484
2020	2396	0.8*2396+(1-0.8)*(2499-24)=2412	0.4*(2412-2499)+(1-0.4)*(-24)=-49	2499-24=2475
Forecast for the year 2021				2412-49=2363

Steps in calculation

M: No. of periods/ year (Quarterly (M=4), Monthly (M=12))

Initial seasonal factors:

$$S_1 = \frac{Y_1}{Average\,(Y_1, Y_2, \ldots, Y_M)}, \quad S_2 = \frac{Y_2}{Average\,(Y_1, Y_2, \ldots, Y_M)},$$

. . .

$$S_M = \frac{Y_M}{Average\,(Y_1, Y_2, \ldots, Y_M)}$$

Time Series ➡ Holts-Winters Triple Exponential Smoothing

Steps in calculation

M: No. of periods/ year (Quarterly (M=4), Monthly (M=12))

Initial level:

$$L_{13} = \frac{Y_{13}}{S_1}$$

Initial trend:

$$T_{13} = \frac{Y_{13}}{S_1} - \frac{Y_{12}}{S_{12}}$$

Time Series ➡ Holts-Winters Triple Exponential Smoothing

Level formula

$$L_t = \alpha \frac{Y_t}{S_{t-M}} + (1-\alpha)(L_{t-1} + T_{t-1})$$

Trend formula

$$T_t = \beta(L_t + L_{t-1}) + (1-\beta)T_{t-1}$$

Seasonal factor formula

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-M}$$

Forecast

$$F_{t+1} = (L_t + T_t)S_{t-M+1}$$

Time Series ➡ Holts-Winters Triple Exponential Smoothing

Example on demand data

Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Year	2011	2011	2011	2011	2012	2012	2012	2012	2013	2013	2013	2013	2014	2014
Quarter	1	2	3	4	1	2	3	4	1	2	3	4	1	2
Demand	362	385	432	341	382	409	498	387	473	513	582	474	544	582
Period	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Year	2014	2014	2015	2015	2015	2015	2016	2016	2016	2016				
Quarter	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Demand	681	557	628	707	773	592	627	725	854	661				

Period	Year	Quarter	Demand	Level	Trend	Seasonal	Forecast	|Error|	α	β	γ
1	2011	1	362			0.952632			0.2	0.3	0.15
2	2011	2	385			1.013158		Initialization of Seasonal factors			
3	2011	3	432			1.136842					
4	2011	4	341			0.897368					
5	2012	1	382	400.9945	20.99448	0.952632		Initialization of L5 and T5			
6	2012	2	409	418.3288	19.89644	1.007839	427.541436	18.54144			
7	2012	3	498	438.1913	19.88625	1.136789	498.192926	0.192926			
8	2012	4	387	452.7143	18.27726	0.89099	411.064349	24.06435			
9	2013	1	473	476.0971	19.80893	0.958761	448.681393	24.31861			
10	2013	2	513	498.5268	20.59516	1.011018	499.793507	13.20649			
11	2013	3	582	517.6912	20.16593	1.134904	590.13223	8.13223			
12	2013	4	474	536.6842	19.81407	0.889821	479.225172	5.225172			
13	2014	1	544	558.6784	20.4681	0.961006	533.5489	10.4511			
14	2014	2	582	578.4487	20.25875	1.010286	585.527621	3.527621			
15	2014	3	681	598.9761	20.33935	1.13521	679.475612	1.524388			
16	2014	4	557	620.646	20.73852	0.890966	551.080127	5.919873			
17	2015	1	628	643.804	21.46436	0.963173	616.374301	11.6257			
18	2015	2	707	672.175	23.53636	1.016515	672.111514	34.88849			
19	2015	3	773	692.7554	22.64957	1.132303	789.778262	16.77826			
20	2015	4	592	705.2134	19.59211	0.88324	637.401508	45.40151			
21	2016	1	627	710.0391	15.16219	0.951154	698.113019	71.11302			
22	2016	2	725	722.8053	14.44339	1.014493	737.17786	12.17786			
23	2016	3	854	740.642	15.46136	1.135416	834.789143	19.21086			
24	2016	4	661	754.5588	14.99799	0.882156	667.82101	6.82101	k		
25							871.371647		1		
26							692.950477		2		
27							760.498109		3		
28							826.356023		4		

Example on seasonal sale of houses

Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Year	2015	2015	2015	2015	2015	2015	2015	2015	2015	2015	2015	2015	2016	2016	2016	2016
Month	January	February	March	April	May	June	July	August	September	October	November	December	January	February	March	April
Date	01-01-2015	01-02-2015	01-03-2015	01-04-2015	01-05-2015	01-06-2015	01-07-2015	01-08-2015	01-09-2015	01-10-2015	01-11-2015	01-12-2015	01-01-2016	01-02-2016	01-03-2016	01-04-2016
# of houses sold	131646	176938	204653	214838	246787	277672	255973	266235	230914	213793	180825	210587	154183	168895	202480	237613
Period	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Year	2016	2016	2016	2016	2016	2016	2016	2016	2017	2017	2017	2017	2017	2017	2017	2017
Month	May	June	July	August	September	October	November	December	January	February	March	April	May	June	July	August
Date	01-05-2016	01-06-2016	01-07-2016	01-08-2016	01-09-2016	01-10-2016	01-11-2016	01-12-2016	01-01-2017	01-02-2017	01-03-2017	01-04-2017	01-05-2017	01-06-2017	01-07-2017	01-08-2017
# of houses sold	276346	288312	264873	261003	235482	238741	210884	200432	162196	193105	217340	267999	275747	293088	276851	266860
Period	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Year	2017	2017	2017	2017	2018	2018	2018	2018	2018	2018	2018	2018	2018	2018	2018	2018
Month	September	October	November	December	January	February	March	April	May	June	July	August	September	October	November	December
Date	01-09-2017	01-10-2017	01-11-2017	01-12-2017	01-01-2018	01-02-2018	01-03-2018	01-04-2018	01-05-2018	01-06-2018	01-07-2018	01-08-2018	01-09-2018	01-10-2018	01-11-2018	01-12-2018
# of houses sold	240859	235874	218218	209316	155722	197765	224567	252527	260426	301841	274278	263149	256927	226363	205678	216753
Period	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	
Year	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2020	2020	2020	
Month	January	February	March	April	May	June	July	August	September	October	November	December	January	February	March	
Date	01-01-2019	01-02-2019	01-03-2019	01-04-2019	01-05-2019	01-06-2019	01-07-2019	01-08-2019	01-09-2019	01-10-2019	01-11-2019	01-12-2019	01-01-2020	01-02-2020	01-03-2020	
# of houses sold	143530	195802	221326	240711	264338	323832	271417	272680	251292	236077	246599	234858	158800	211854	230158	

Period	Year	Month	Actual values	# of houses sold	Calculation		Seasonal	
1	2015	January	Y1	131646	Y1/Average (Y1, Y2, ..., Y12)	131646/217571.8	0.60506936	S1
2	2015	February	Y2	176938	Y2/Average (Y1, Y2, ..., Y12)	176938/217571.8	0.81323977	S2
3	2015	March	Y3	204653	Y3/Average (Y1, Y2, ..., Y12)	204653/217571.8	0.94062304	S3
4	2015	April	Y4	214838	Y4/Average (Y1, Y2, ..., Y12)	214838/217571.8	0.98743518	S4
5	2015	May	Y5	246787	Y5/Average (Y1, Y2, ..., Y12)	246787/217571.8	1.13427869	S5
6	2015	June	Y6	277672	Y6/Average (Y1, Y2, ..., Y12)	277672/217571.8	1.27623186	S6
7	2015	July	Y7	255973	Y7/Average (Y1, Y2, ..., Y12)	255973/217571.8	1.17649925	S7
8	2015	August	Y8	266235	Y8/Average (Y1, Y2, ..., Y12)	266235/217571.8	1.2236653	S8
9	2015	September	Y9	230914	Y9/Average (Y1, Y2, ..., Y12)	230914/217571.8	1.06132345	S9
10	2015	October	Y10	213793	Y10/Average (Y1, Y2, ..., Y12)	213793/217571.8	0.98263217	S10
11	2015	November	Y11	180825	Y11/Average (Y1, Y2, ..., Y12)	180825/217571.8	0.83110514	S11
12	2015	December	Y12	210587	Y12/Average (Y1, Y2, ..., Y12)	210587/217571.8	0.9678968	S12

Statistical Methods for Data Science

Time Series → Holts-Winters Triple Exponential Smoothing

Level	Seasonal
S1	0.605069
S2	0.81324
S3	0.940623
S4	0.987435
S5	1.134279
S6	1.276232
S7	1.176499
S8	1.223665
S9	1.061323
S10	0.982632
S11	0.831105
S12	0.967897

Period	Year	Month	# of houses	Level	Trend	Seasonal	Forecast	Error
12	2015	December	**210587**					
13	2016	January	**154183**	**254819**	**37247**	0.605069		
14	2016	February	168895	283627	36403	0.791464	237519	-68624

$$L_{13} = \frac{Y_{13}}{S_1} = \frac{154183}{0.605069} = 254819$$

$$T_{13} = \frac{Y_{13}}{S_1} - \frac{Y_{12}}{S_{12}} = \frac{154183}{0.605069} - \frac{210587}{0.967897} = 37247$$

Period	Year	Month	Date	# of houses sold	Level	Trend	Seasonal	Forecast	Error
13	2016	January	01-01-2016	154183	254819	37247	0.605069		
14	2016	February	01-02-2016	168895	283627	36403	0.791464	237519	-68624
15	2016	March	01-03-2016	202480	309554	35355	0.911971	301028	-98548
16	2016	April	01-04-2016	237613	334482	34313	0.959731	340575	-102962
17	2016	May	01-05-2016	276346	356278	33061	1.098416	418316	-141970
18	2016	June	01-06-2016	288312	372996	31427	1.225905	496887	-208575
19	2016	July	01-07-2016	264873	386494	29634	1.127382	475803	-210930
20	2016	August	01-08-2016	261003	395845	27606	1.167234	509202	-248199

53	2019	May	01-05-2019	264338	241299	-4314	1.041209	248171	16167	
54	2019	June	01-06-2019	323832	241218	-3891	1.177702	274759	49073	
55	2019	July	01-07-2019	271417	238952	-3728	1.076898	254022	17395	
56	2019	August	01-08-2019	272680	236468	-3604	1.106219	258983	13697	
57	2019	September	01-09-2019	251292	235441	-3346	0.981184	226254	25038	
58	2019	October	01-10-2019	236077	234962	-3059	0.91525	210117	25960	
59	2019	November	01-11-2019	246599	240361	-2214	0.80385	180692	65907	
60	2019	December	01-12-2019	234858	240942	-1934	0.891856	210200	24658	
61	2020	January	01-01-2020	158800	242951	-1540	0.578649	136310	22490	
62	2020	February	01-02-2020	211854	245654	-1116	0.757977	180183	31671	
63	2020	March	01-03-2020	230158	246861	-883	0.866844	210197	19961	k
64	2020	April	01-04-2020					256114		1
65	2020	May	01-05-2020					288648		2
66	2020	June	01-06-2020					262990		3
67	2020	July	01-07-2020					269174		4
68	2020	August	01-08-2020					237882		5
69	2020	September	01-09-2020					221089		6
70	2020	October	01-10-2020					193469		7
71	2020	November	01-11-2020					213862		8
72	2020	December	01-12-2020					138246		9

Period	Year	Month	Date	# of houses sold	Level	Trend	Seasonal			
1	2015	January	01-01-2015	131646	S1		0.605069			
2	2015	February	01-02-2015	176938	S2		0.81324			
3	2015	March	01-03-2015	204653	S3		0.940623			
4	2015	April	01-04-2015	214838	S4		0.987435			
5	2015	May	01-05-2015	246787	S5		1.134279			
6	2015	June	01-06-2015	277672	S6		1.276232			
7	2015	July	01-07-2015	255973	S7		1.176499			
8	2015	August	01-08-2015	266235	S8		1.223665	α	0.1	
9	2015	September	01-09-2015	230914	S9		1.061323	β	0.1	
10	2015	October	01-10-2015	213793	S10		0.982632	γ	0.1	
11	2015	November	01-11-2015	180825	S11		0.831105			
12	2015	December	01-12-2015	210587	S12		0.967897	Forecast	Error	
13	2016	January	01-01-2016	154183	254819	37247	0.605069			
14	2016	February	01-02-2016	168895	283627	36403	0.791464	237519	-68624	
15	2016	March	01-03-2016	202480	309554	35355	0.911971	301028	-98548	
16	2016	April	01-04-2016	237613	334482	34313	0.959731	340575	-102962	
17	2016	May	01-05-2016	276346	356278	33061	1.098416	418316	-141970	
18	2016	June	01-06-2016	288312	372996	31427	1.225905	496887	-208575	
19	2016	July	01-07-2016	264873	386494	29634	1.127382	475803	-210930	
20	2016	August	01-08-2016	261003	395845	27606	1.167234	509202	-248199	
21	2016	September	01-09-2016	235482	403293	25590	1.013581	449418	-213936	
22	2016	October	01-10-2016	238741	410291	23731	0.942557	421434	-182693	
23	2016	November	01-11-2016	210884	415993	21928	0.798689	360717	-149833	
24	2016	December	01-12-2016	200432	414837	19619	0.919423	423862	-223430	
25	2017	January	01-01-2017	162196	417817	17955	0.583382	262876	-100680	
26	2017	February	01-02-2017	193105	416594	16038	0.758671	344898	-151793	
27	2017	March	01-03-2017	217340	413200	14094	0.873373	394547	-177207	
28	2017	April	01-04-2017	267999	412489	12614	0.928729	410088	-142089	
29	2017	May	01-05-2017	275747	407697	10873	1.056209	466940	-191193	
30	2017	June	01-06-2017	293088	400621	9078	1.176473	513127	-220039	
31	2017	July	01-07-2017	276851	393287	7437	1.085038	461888	-185037	
32	2017	August	01-08-2017	266860	383514	5716	1.120094	467739	-200879	
33	2017	September	01-09-2017	240859	374070	4200	0.976612	394516	-153657	
34	2017	October	01-10-2017	235874	365468	2920	0.912842	356541	-120667	
35	2017	November	01-11-2017	218218	358871	1968	0.779627	294227	-76009	
36	2017	December	01-12-2017	209316	347522	636	0.887712	331764	-122448	
37	2018	January	01-01-2018	155722	340035	-176	0.57084	203109	-47387	
38	2018	February	01-02-2018	197765	331941	-968	0.742382	257842	-60077	
39	2018	March	01-03-2018	224567	323588	-1706	0.855435	289063	-64496	
40	2018	April	01-04-2018	252527	316885	-2206	0.915547	298941	-46414	
41	2018	May	01-05-2018	260426	307868	-2887	1.035179	332367	-71941	
42	2018	June	01-06-2018	301841	300139	-3371	1.159393	358801	-56960	
43	2018	July	01-07-2018	274278	292369	-3811	1.070346	322004	-47726	
44	2018	August	01-08-2018	263149	283196	-4347	1.101006	323212	-60063	
45	2018	September	01-09-2018	256927	277272	-4505	0.971613	272327	-15400	
46	2018	October	01-10-2018	226363	270288	-4753	0.905306	248993	-22630	
47	2018	November	01-11-2018	205678	265363	-4770	0.779172	207018	-1340	
48	2018	December	01-12-2018	216753	258951	-4934	0.882645	231331	-14578	
49	2019	January	01-01-2019	143530	253758	-4960	0.570318	145003	-1473	
50	2019	February	01-02-2019	195802	250293	-4811	0.746373	184703	11099	
51	2019	March	01-03-2019	221326	246807	-4678	0.859567	209994	11332	
52	2019	April	01-04-2019	240711	244208	-4470	0.92256	221680	19031	
53	2019	May	01-05-2019	264338	241299	-4314	1.041209	248171	16167	
54	2019	June	01-06-2019	323832	241218	-3891	1.177702	274759	49073	
55	2019	July	01-07-2019	271417	238952	-3728	1.076898	254022	17395	
56	2019	August	01-08-2019	272680	236468	-3604	1.106219	258983	13697	
57	2019	September	01-09-2019	251292	235441	-3346	0.981184	226254	25038	
58	2019	October	01-10-2019	236077	234962	-3059	0.91525	210117	25960	
59	2019	November	01-11-2019	246599	240361	-2214	0.80385	180692	65907	
60	2019	December	01-12-2019	234858	240942	-1934	0.891856	210200	24658	
61	2020	January	01-01-2020	158800	242951	-1540	0.578649	136310	22490	
62	2020	February	01-02-2020	211854	245654	-1116	0.757977	180183	31671	
63	2020	March	01-03-2020	230158	246861	-883	0.866844	210197	19961	k

Time Series ➡ Holts-Winters Triple Exponential Smoothing

Trend	Seasonality		
	No	Linear	Ratio
No			
Linear			
Ratio			

Time Series ➡ Smoothing Techniques – Exponential

No trend

No seasonality

Time Series ➡ Smoothing Techniques – Exponential

No trend

Linear seasonality

Time Series ➡ Smoothing Techniques – Exponential

No trend

Ratio seasonality

Time Series ➡ Smoothing Techniques – Exponential

Linear trend

No seasonality

Time Series ➡ Smoothing Techniques – Exponential

Linear trend

Linear seasonality

Time Series ➡ Smoothing Techniques – Exponential

Linear trend

Ratio seasonality

Time Series ➡ Smoothing Techniques – Exponential

Ratio trend

No seasonality

Time Series ➡ Smoothing Techniques – Exponential

Ratio trend

Linear seasonality

Time Series ➡ Smoothing Techniques – Exponential

Ratio seasonality

Ratio trend

Time Series ➡ Measurement of forecast error

What was forecasted to happen at time t using the forecast technique

Forecast Error at time t

$$\Delta_t = Y_t - F_t$$

What actually happened at time t

Statistical Methods for Data Science

Time Series ➡ Decomposition view

Thank you

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Sessions 15
Time Series Analysis
(9th Sep 2023)

Time Series

Time Series ➡ Definition of Stochastic Process

- Consider the sequence of random variables Y_0, Y_1, Y_2, … which forms a family of random variables **$\{Y_t, t \in T\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$**

where Ω - Sample space

$\mathcal{F}$ - all collection of subsets A of Ω and

P – probability measure

- T – index set

Index set:

- The index set **T** is a collection of **all time functions** that can result from random experiment, usually the index **T** denote time. Y_t are **independent and identically distributed (iid) random variables** with mean μ_t and variance σ_t^2. Each realization of this process gives an ensemble or a data set.
- iid rvs: if each random variable has the same probability distribution as the others and all are mutually independent

Time Series ➡ Definition of Stochastic Process

- Examples of index sets:

  (1) T = (-∞, ∞) or T = [0, ∞). In this case Y_t is a continuous time stochastic process.

  (2) T = {0, ±1, ±2, ….} or T = {0, 1, 2, …}. In this case Y_t is a discrete time stochastic process.

- We use uppercase letter $\{Y_t\}$ to describe the process. A time series, $\{Y_t\}$ is a realization or sample function from a certain process.

- We use information from a time series to estimate parameters and properties of process $\{Y_t\}$.

Time Series ➡ Definition of Stochastic Process

- Y_t represents random quantity at time t

- In general, the value Y_t might depend on the quantity Y_{t-1} at time t-1, or even the value Y_s for other times $s < t$.

Time Series ➡ Probability distribution of the process

- For any stochastic process with index set *T*, its probability distribution function is uniquely determined by its finite dimensional distributions.
- The *k* dimensional distribution function of a process is defined by

$$F_{Y_{t_1},Y_{t_2},\dots,Y_{t_k}} = P(Y_{t_1} \le y_{t_1}, Y_{t_2} \le y_{t_2}, \dots, Y_{t_k})$$

for any $t_1, t_2, \dots, t_k \in T$ and any real numbers $y_1, \dots, y_k$.

- The distribution function tells us everything we need to know about the process $\{Y_t\}$.

In stochastic time series,

$$Y_t,\ t \in Z = \{0, \pm 1, \pm 2, \ldots\}$$

is a family of random variables, Y_t denoting the value of the characteristic of interest at time t.

Thus, $\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$ is seen as a realized value of the random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ with joint probability density function $f_{\mathbf{Y}}(y)$.

The joint distribution function of a finite random variables

$$\{Y_{t_1}, \ldots Y_{t_n}\},\ t_1 < t_2 < \ldots < t_n$$

from the collection $\{Y_t, t \in T\}$ is

$$F_{Y_{t_1}, \ldots Y_{t_n}}(y_1, \ldots, y_n) = P[Y_{t_1} \leq y_{t_1}, \ldots, P[Y_{t_n} \leq y_{t_n}],$$

$$(y_{t_1}, \ldots, y_{t_n}) \in R^n$$

Time Series ➡ Stationary Stochastic Process in Time series

A special kind of stochastic process is based on the assumption that the process is in a particular state of equilibrium. This type of assumption is called stationarity.

A stochastic process is strictly stationary if its properties are unaffected by a change of origin.

Time Series ➡ Moments of Stochastic Processes

- We can describe a stochastic process via its moments, ie., $E(Y_t)$, $E(Y^2_t)$, $E(Y_t, Y_s)$, etc. We often use the first two moments.
- The mean function of the process is $E(Y_t) = \mu_t$.
- The variance function of the process is $V(Y_t) = \sigma^2_t$.

Time Series ➡ Moments of Stochastic Processes

- The Covariance function between Y_t, and Y_s is

$$\text{Cov}\,(Y_t, Y_s) = E((Y_t - \mu_t)\,(Y_s - \mu_s))$$

- The Correlation function between Y_t, and Y_s is

$$\rho\,(Y_t, Y_s) = \frac{Cov\,(Y_t, Y_s)}{\sqrt{\sigma_t^2}\sqrt{\sigma_s^2}}$$

- These moments are functions of time

Thus, a time series stochastic process $\{Y_{t_1}, Y_{t_2}, \ldots Y_{t_k}\}$ is said to be strictly stationary, if the joint distribution of k observations $Y_{t_1}, \ldots Y_{t_k}$ made at time $t_1, \ldots, t_k$ is same as that of k+h observations $Y_{t_{1+h}}, Y_{t_{2+h}}, \ldots Y_{t_{k+h}}$ made at time points $t_{1+h}, \ldots, t_{k+h}$ for any h. That is,

$$F_{Y_{t_1}, \ldots Y_{t_k}}(y_{t_1}, \ldots, y_{t_k}) = F_{Y_{t_{1+h}}, Y_{t_{2+h}}, \ldots Y_{t_{k+h}}}(y_{t_{1+h}}, Y_{t_{2+h}}, \ldots, y_{t_{k+h}})$$

Time Series ➡ Stationary Stochastic Process in Time series

- If $\{Y_t\}$ is a strictly stationary process and $E(Y^2_t) < \infty$ then, the mean function is a constant and the variance function is also a constant.
- Moreover, for a strictly stationary process with first two moments finite, the covariance function, and the correlation function depends only on the time difference s.
- A trivial example of a strictly stationary process is a sequence of independent and identically distributed (iid) random variables.

Time Series ➡ Stationary Stochastic Process in Time series

That is, $F_{Y_{t_1}, \ldots Y_{t_k}}(y_{t_1}, \ldots, y_{t_k}) = F_{Y_{t_{1+h}}, \ldots Y_{t_{k+h}}}(y_{t_{1+h}}, \ldots, y_{t_{k+h}})$

for all possible non-empty finite distinct sets

$(t_1, \ldots, t_k)$ and $(t_{1+h}, \ldots, t_{k+h})$ in the index set T and all $y_{t_1}, \ldots, y_{t_k}$ in the range of random variables Y_t.

Note: For reliable prediction to be made, the time series should be stationary (No systematic change such as seasonality, trend i.e. only random fluctuations)

Time Series ➡ Stochastic Process

- Strict stationarity is too strong of a condition in practice. It is often difficult assumption to assess based on an observed time series $Y_1,\ldots,Y_k$.
- In time series analysis we often use a weaker sense of stationarity in terms of the moments of the process.
- A process is said to be *n*th-order weakly stationary if all its joint moments up to order *n* exists and are time invariant, i.e., independent of time origin.

Time Series → Stochastic Process

- For example, a second-order weakly stationary process will have constant mean and variance, with the covariance and the correlation being functions of the time difference along.
- A strictly stationary process with the first two moments finite is also a second-ordered weakly stationary. But a strictly stationary process may not have finite moments and therefore may not be weakly stationary.

When k = 1, strict stationary implies that the pdf $f(y_t)$ is the same for all t, is say f(y). The stochastic process f(y) has a constant mean

$$\mu = E(Y_t) = \int_{-\infty}^{\infty} yf(y)dy$$

and a constant variance

$$\sigma^2 = E(Y_t - \mu)^2 = \int_{-\infty}^{\infty} (y - \mu)^2 \mathrm{f(y)dy}$$

provided the mean and variance exists.

The mean (μ) and variance (σ^2) are estimated as

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\hat{\sigma}_y^2 = S_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2$$

Thus, a stationary process remains at equilibrium about a common mean value.

However, in industry, business (ex. Stock price) and economics many time series are better represented as non-stationary and in particular, having no natural mean.

Time Series ➡ Stationary Stochastic Process in Time series

Stationary time series

Non-stationary time series

Statistical Methods for Data Science

Time Series ➡ Stationary Stochastic Process in Time series

Fig 15.3 The Wolfer sunspot numbers, 1770–1869.
(Source: Box and Jenkins, 1976)

Time Series ➡ Stationary Stochastic Process in Time series

However, many non-stationary time series can be so modified that the reduced to time series obeys the original series.

The two main component which cause lack of stationarity are trend and seasonality. In fitting the stationary time series model, we therefore, assume that the trend and seasonality have been eliminated from the original series.

Time Series ➡ Autocovariance

- Stationarity implies that the joint distribution of pair of observations y_{t_1}, y_{t_2} viz.,

$$f(y_{t_1}, y_{t_2}) = f(y_{t_1+h}, y_{t_2+h})$$

for any integer h. That is, the joint distribution of any pair of observations on time points which differ by a constant quantity is the same for such pairs.

Thus

$$f(y_t, y_{t+h}) = f(y_{t_1}, y_{t_1+h}) = f(y_{t_2}, y_{t_2+h})$$

have the same distribution.

The form of $f(y_t, y_{t+h})$ can be inferred by plotting values of (y_t, y_{t+h}); $t = 1, 2, \ldots.$, i.e., values of y_t separated by lag h.

Time Series ➡ Autocovariance

- Let $\{Y_t\}$ be a Time Series with $E\{Y_t^2\} < \infty$. The mean function of $\{Y_t\}$ is $\mu_y(t) = E(Y_t)$.
- The covariance function of $\{Y_t, Y_{t+h}\}$ is
- $\gamma_h = \mathrm{Cov}(Y_t, Y_{t+h}) = E[(Y_t-\mu)(Y_{t+h}-\mu)]$
- In general

$$\gamma_y(r, s) = \mathrm{Cov}(Y_r, Y_s) = E[\{Y_r-\mu_y(r)\}\{Y_s-\mu_y(s)\}]$$

for all integers r and s

Time Series ➡ Autocovariance

- Clearly, with h=0, in $\gamma_h = \mathrm{Cov}\,(Y_t, Y_{t+h}) = E[(Y_t-\mu)(Y_{t+h}-\mu)]$

$$\gamma_0 = \mathrm{Cov}\,(Y_t, Y_{t+0}) = E[(Y_t-\mu)(Y_{t+0}-\mu)] = \sigma^2_y$$

$$|\gamma_h| \leq \gamma_0, \forall\, h = 1, 2, \ldots$$

because, $|Cov(Y_{t+h}, Y_t)| \leq \sqrt{V(Y_{t+h})V(Y_t)}$

Time Series ➡ Autocovariance function

- Let $\{Y_t\}$ be a stationery Time Series.

- The Autocovariance Function (ACVF) of $\{Y_t\}$ at lag h is

$$\gamma_y(t+h, t) = \mathrm{Cov}\,(Y_{t+h}, Y_t).$$

The Autocorrelation Function (ACF) of $\{Y_t\}$ at lag h is

$$\rho_h = \frac{\gamma_h}{\gamma_0} = \frac{\mathrm{Cov}(Y_{t+h}, Y_t)}{\sigma^2}, \quad \boldsymbol{\rho_0 = 1}$$

Time Series ➡ Autocorrelation

- Correlation: measure of relationship between two variables are related
  - Direction of the relationship (Negative, Zero, and/ or positive
  - Degree/ Extent of relationship

Time Series ➡ Autocorrelation

- The autocorrelation measures the degree of relationship of the same variable between the observation at the current time period and the observations at the prior time periods.
- It measures how the lagged version of the value of a variable is related to the original version of it in time series.

The value of r_k can be written as

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}$$

where T is the length of time series.

Time Series ➡ Autocorrelation

- Autocorrelation between (Y_t, Y_{t-1})

$$r_1 = \frac{\sum(Y_t - \bar{Y})\,(Y_{t-1} - \bar{Y})}{\sum(Y_t - \bar{Y})^2}$$

- Autocorrelation between (Y_t, Y_{t-2})

$$r_2 = \frac{\sum(Y_t - \bar{Y})\,(Y_{t-2} - \bar{Y})}{\sum(Y_t - \bar{Y})^2}$$

– – – –

- Autocorrelation between (Y_t, Y_{t-k})

$$r_k = \frac{\sum(Y_t - \bar{Y})\,(Y_{t-k} - \bar{Y})}{\sum(Y_t - \bar{Y})^2}$$

- There are several autocorrelation coefficients, corresponding to each panel in the lag plot. For example, r_1 measures the relationship between Y_t and Y_{t-1}, r_2 measures the relationship between Y_t and Y_{t-2}, and so on rk measures the relationship between Y_t and Y_{t-k}.

Year	2011				2012			
Quarter	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Retail quarterly sale (million) (Y_t)	147772	154400	166188	170202	173264	175371	184957	186395
Year	2013				2014			
Quarter	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Retail quarterly sale (million) (Y_t)	191130	191213	195749	198262	199980	209566	212529	213754
Year	2015				2016			
Quarter	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Retail quarterly sale (million) (Y_t)	222124	224372	229871	236260	238044	244076	244944	245799
Year	2017				2018			
Quarter	Q1	Q2	Q3	Q4	Q1	Q2		
Retail quarterly sale (million) (Y_t)	251125	254557	255223	264793	265064	277282		

Time	Value	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10
1	22										
2	24	22									
3	25	24	22								
4	25	25	24	22							
5	28	25	25	24	22						
6	29	28	25	25	24	22					
7	34	29	28	25	25	24	22				
8	37	34	29	28	25	25	24	22			
9	40	37	34	29	28	25	25	24	22		
10	44	40	37	34	29	28	25	25	24	22	
11	51	44	40	37	34	29	28	25	25	24	22
12	48	51	44	40	37	34	29	28	25	25	24
13	47	48	51	44	40	37	34	29	28	25	25
14	50	47	48	51	44	40	37	34	29	28	25
15	51	50	47	48	51	44	40	37	34	29	28

	Autocorrelation
Lag 1	0.83174
Lag 2	0.65632
Lag 3	0.49105
Lag 4	0.27864
Lag 5	0.03103
Lag 6	-0.1653
Lag 7	-0.3037
Lag 8	-0.401
Lag 9	-0.4582
Lag 10	-0.4505

Month/ Year	Retail quarterly sale (million)	Lag1	Lag2	Lag3	Lag4	Lag5	Lag6
Q1 - 2011	147772						
Q2 - 2011	154400	147772					
Q3 - 2011	166188	154400	147772				
Q4 - 2011	170202	166188	154400	147772			
Q1 - 2012	173264	170202	166188	154400	147772		
Q2 - 2012	175371	173264	170202	166188	154400	147772	
Q3 - 2012	184957	175371	173264	170202	166188	154400	147772
Q4 - 2012	186395	184957	175371	173264	170202	166188	154400
Q1 - 2013	191130	186395	184957	175371	173264	170202	166188
Q2 - 2013	191213	191130	186395	184957	175371	173264	170202
Q3 - 2013	195749	191213	191130	186395	184957	175371	173264
Q4 - 2013	198262	195749	191213	191130	186395	184957	175371
Q1 - 2014	199980	198262	195749	191213	191130	186395	184957
Q2 - 2014	209566	199980	198262	195749	191213	191130	186395
Q3 - 2014	212529	209566	199980	198262	195749	191213	191130
Q4 - 2014	213754	212529	209566	199980	198262	195749	191213
Q1 - 2015	222124	213754	212529	209566	199980	198262	195749
Q2 - 2015	224372	222124	213754	212529	209566	199980	198262
Q3 - 2015	229871	224372	222124	213754	212529	209566	199980
Q4 - 2015	236260	229871	224372	222124	213754	212529	209566
Q1 - 2016	238044	236260	229871	224372	222124	213754	212529
Q2 - 2016	244076	238044	236260	229871	224372	222124	213754
Q3 - 2016	244944	244076	238044	236260	229871	224372	222124
Q4 - 2016	245799	244944	244076	238044	236260	229871	224372
Q1 - 2017	251125	245799	244944	244076	238044	236260	229871
Q2 - 2017	254557	251125	245799	244944	244076	238044	236260
Q3 - 2017	255223	254557	251125	245799	244944	244076	238044
Q4 - 2017	264793	255223	254557	251125	245799	244944	244076
Q1 - 2018	265064	264793	255223	254557	251125	245799	244944
Q2 - 2018	277282	265064	264793	255223	254557	251125	245799

1.00000
0.80000
0.60000
0.40000
0.20000
0.00000
-0.20000
0.87413
0.77038
0.67408
0.58895
0.49961
0.40745
0.33113
0.25055
0.16761
0.08688
0.00769
-0.06849
-0.14020
Lag 1
Lag 2
Lag 3
Lag 4
Lag 5
Lag 6
Lag 7
Lag 8
Lag 9
Lag 10
Lag 11
Lag 12
Lag 13

- Trend: Autocorrelation for small lags tend to be large and positive

- Seasonal: Autocorrelation will be larger for smaller seasonal lags (at multiple seasonal frequency)

- Trend: Autocorrelation for small lags tend to be large and positive

- Seasonal: Autocorrelation will be larger for smaller seasonal lags (at multiple seasonal frequency)

Seasonal series

Time Series ➡ White Noise

Time series that show no autocorrelation are called **white noise.** That is, if the variables are independent and identically distributed with a mean of zero. This means that all variables have the same variance (σ^2) and each value has a zero correlation with all other values in the series.

Time Series ➡ White Noise

A white noise called **white** when it has the same intensity at every frequency. Its name is derived by analogy to light, which is **called** "**white**" when it contains all visible frequencies **noise.**

The term **noise**, in this context, came from signal processing where it was used to refer to unwanted electrical or electromagnetic energy that degrades the quality of signals and data

Time Series ➡ White Noise

For white noise series, we expect each autocorrelation to be close to zero. Of course, they will not be exactly equal to zero as there is some random variation. For a white noise series, we expect 95% of the spikes in the ACF to lie within $\pm 2/\sqrt{T}$ where T is the length of the time series.

It is common to plot these bounds on a graph of the ACF (the blue dashed lines above). If one or more large spikes are outside these bounds, or if substantially more than 5% of spikes are outside these bounds, then the series is probably not white noise

- White Noise series
- Time series that shows no autocorrelation
- 95% of ACF spikes should be within $\pm\frac{2}{\sqrt{T}}$
- If more than 5% of ACF spikes are outside $\pm\frac{2}{\sqrt{T}}$, then the series is not white noise.

Example: If T = 50, then

$$\pm\frac{2}{\sqrt{T}} = \pm\frac{2}{\sqrt{50}} = \pm 2.8$$

For T = 200, then

$$\pm\frac{2}{\sqrt{T}} = \pm\frac{2}{\sqrt{200}} = \pm 2.8$$

Statistical Methods for Data Science

Time Series ➡ Autocorrelation

Time Series ➡ Test for Autocorrelation: Ljung-Box test

- A statistical test of whether any group of autocorrelation of a time series are different from zero
- A test for the overall randomness based on a number of lags
- **H_0:** The series is random or white noise or independent and identically distributed (iid)
- **H_1:** The series exhibits serial/ autocorrelation (non-random)
- A small P-value ($P < 0.05$) ➡ Reject H_0 ➡ the series not white noise
- A large P-value ($P > 0.05$) ➡ Fail to Reject H_0 ➡ the series white noise

- **Note: Normally, P-value will be very large for white noise series**

Statistical Methods for Data Science

Time Series → Autocorrelation

Time	Value	lag 1	lag 2	lag 3	lag 4	lag 5
1	22					
2	24	22				
3	25	24	22			
4	25	25	24	22		
5	28	25	25	24	22	
6	29	28	25	25	24	22
7	34	29	28	25	25	24
8	37	34	29	28	25	25
9	40	37	34	29	28	25
10	44	40	37	34	29	28
11	51	44	40	37	34	29
12	48	51	44	40	37	34
13	47	48	51	44	40	37
14	50	47	48	51	44	40
15	51	50	47	48	51	44

Lag 1	0.8317
Lag 2	0.6563
Lag 3	0.4911
Lag 4	0.2786
Lag 5	0.0310

Time Series ➡ Partial Autocorrelation Function (PACF)

- The correlation between the observation at two time points given that we consider both observations are correlated to observations at other time periods.
  - Example: Today's stock price can be correlated to the day before yesterday and yesterday can also be correlated to the day before yesterday. Then PACF of yesterday is the "real" correlation between today and yesterday after taking out the influence of the day before yesterday.

Time Series ➡ Partial – Autocorrelation Function (PACF)

- The partial correlation between two variables is a conditional correlation taking into account their dependence on all other remaining variables
  - Eg. A third order (lag) partial autocorrelation is

$$\frac{\mathrm{Cov}(X_t, X_{t-3} | X_{t-1}, X_{t-2})}{\sqrt{\mathrm{Var}(X_t | X_{t-1}, X_{t-2})\mathrm{Var}(X_{t-3} | X_{t-1}, X_{t-2})}}$$

- The first order PACF & ACF are same

Statistical Methods for Data Science

Time Series ➡ Correlogram

ACF for d_premium
+- 1.96/T^0.5
0.4
0.2
0
-0.2
-0.4
0
20
40
60
80
100
lag

PACF for d_premium
+- 1.96/T^0.5
0.4
0.2
0
-0.2
-0.4
0
20
40
60
80
100
lag

Time Series ➡ Correlogram

The Autocorrelation plot for the data is:

Time Series ➡ Autoregressive (AR) models

- An **autoregressive** (**AR**) **model** is a representation of some type of random process; used to describe certain time-varying processes. The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic (random) error term; thus the model is in the form of a stochastic difference equation.

Time Series ➡ Autoregressive (AR) models

- The Autoregressive model of order p denoted by AR (p) is given by

$$Y_t = c_t + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + \varepsilon_t$$

where ϕ_1, ϕ_2, ϕ_3, ... ϕ_p are the parameter of the model; c is constant and ε_t is the **white noise**.

Introduction

Table 1

| | | |

Regression Statistics	Results
Multiple R	0.972056
R Square	0.944892
Adjusted R Square	0.941813
Standard Error	67.17728
Observations	190

	Coefficients	Standard Error	t Stat	P-value
Intercept	-301.643	57.0362	-5.28862	3.56E-07
lag 1	-0.27739	0.074602	-3.71825	0.000268
lag 2	-0.55811	0.076787	-7.2683	1.09E-11
lag 3	0.356679	0.086535	4.12177	5.74E-05
lag 4	**0.013823**	**0.090524**	**0.152697**	**0.878809**
lag 5	-0.14453	0.089015	-1.62366	0.106209
lag 6	-0.02001	0.078551	-0.25476	0.7992
lag 7	0.048186	0.074732	0.644783	0.519894
lag 8	-0.01276	0.072079	-0.17699	0.859713
lag 9	-0.04807	0.057523	-0.83569	0.404441
lag 10	-0.04469	0.057081	-0.78286	0.434743

ANOVA

SV	df	SS	MS	F-value	P-value
Regression	10	13850518	1385052	306.9172	5.2E-107
Residual	179	807788.9	4512.787		
Total	189	14658307			

Regression Statistics	Results
Multiple R	0.965633
R Square	0.932447
Adjusted R Square	0.93066
Standard Error	73.23764
Observations	195

ANOVA

SV	df	SS	MS	F-value	P-value
Regression	5	13992985	2798597	521.7611	1.6E-108
Residual	189	1013749	5363.752		
Total	194	15006735			

	Coefficients	Standard Error	t Stat	P-value
Intercept	-288.389	34.1692	-8.44002	8.17E-15
lag 1	-0.11143	0.06172	-1.80543	0.072598
lag 2	-0.66093	0.061539	-10.7399	2.62E-21
lag 3	0.388821	0.072508	5.362419	2.38E-07
lag 5	-0.12262	0.061361	-1.99838	0.047109

$Y_t = b_0+b_1X_{t-1}+b_2X_{t-2}+b_3X_{t-3}+b_5X_{t-5}$

$Y_t = -288.389 - 0.11143 *X_{t-1} - 0.66093*X_{t-2} + 0.388821*X_{t-3} - 0.12262*X_{t-5}$

Regression Statistics	Results
Multiple R	0.965633
R Square	0.932447
Adjusted R Square	0.93066
Standard Error	73.23764
Observations	195

ANOVA

SV	df	SS	MS	F-value	P-value
Regression	5	13992985	2798597	521.7611	1.6E-108
Residual	189	1013749	5363.752		
Total	194	15006735			

	Coefficients	Standard Error	t Stat	P-value
Intercept	-288.389	34.1692	-8.44002	8.17E-15
Lag 1	-288.389	34.1692	-8.44002	8.17E-15
Lag 2	-0.11143	0.06172	-1.80543	0.072598
Lag 3	-0.66093	0.061539	-10.7399	2.62E-21
Lag 4	0.388821	0.072508	5.362419	2.38E-07
Lag 5	-0.11281	0.061297	-1.84034	0.067286

$Y_t = b_0+b_1X_{t-1}+b_2X_{t-2}+b_3X_{t-3}+b_4X_{t-4}+b_5X_{t-5}$

$Y_t = -288.389 - 0.11143 *X_{t-1} - 0.66093*X_{t-2} + 0.388821*X_{t-3} - 0.11281*X_{t-4} - 0.12262*X_{t-5}$

Period (t)	Sales (Y_t)	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7
1	5.30							
2	4.40	5.30						
3	5.40	4.40	5.30					
4	5.80	5.40	4.40	5.30				
5	5.60	5.80	5.40	4.40	5.30			
6	4.80	5.60	5.80	5.40	4.40	5.30		
7	5.60	4.80	5.60	5.80	5.40	4.40	5.30	
8	5.60	5.60	4.80	5.60	5.80	5.40	4.40	5.30
9	5.40	5.60	5.60	4.80	5.60	5.80	5.40	4.40
10	6.50	5.40	5.60	5.60	4.80	5.60	5.80	5.40
11	5.10	6.50	5.40	5.60	5.60	4.80	5.60	5.80
12	5.80	5.10	6.50	5.40	5.60	5.60	4.80	5.60
13	5.00	5.80	5.10	6.50	5.40	5.60	5.60	4.80
14	6.20	5.00	5.80	5.10	6.50	5.40	5.60	5.60
15	5.60	6.20	5.00	5.80	5.10	6.50	5.40	5.60
16	6.70	5.60	6.20	5.00	5.80	5.10	6.50	5.40
17	5.20	6.70	5.60	6.20	5.00	5.80	5.10	6.50
18	5.50	5.20	6.70	5.60	6.20	5.00	5.80	5.10
19	5.80	5.50	5.20	6.70	5.60	6.20	5.00	5.80
20	5.10	5.80	5.50	5.20	6.70	5.60	6.20	5.00
21	5.80	5.10	5.80	5.50	5.20	6.70	5.60	6.20
22	6.70	5.80	5.10	5.80	5.50	5.20	6.70	5.60
23	5.20	6.70	5.80	5.10	5.80	5.50	5.20	6.70
24	6.00	5.20	6.70	5.80	5.10	5.80	5.50	5.20
25	5.80	6.00	5.20	6.70	5.80	5.10	5.80	5.50
		5.80	6.00	5.20	6.70	5.80	5.10	5.80
			5.80	6.00	5.20	6.70	5.80	5.10
				5.80	6.00	5.20	6.70	5.80
					5.80	6.00	5.20	6.70
						5.80	6.00	5.20
							5.80	6.00
								5.80

Lag	Intercept	Coefficient
1	8.502	-0.486
2	4.814	0.161
3	6.408	-0.121
4	6.617	-0.159
5	6.246	-0.094
6	3.381	0.421
7	5.816	-0.017

Time Series ➡ Test for stationarity using Dickey-Fuller test

- **Consider the first order autoregressive model**

$$Y_t = \mu + \phi_1 Y_{t-1} + \epsilon_t \text{ ---------- } (1)$$

where μ is constant, ϕ_1 is the regression coefficient and ϵ_t is the random error with zero mean and constant variance

If Y_t is a non-stationary process, then it is essential to test ϕ_1 whether it has unitary root or not.

Hence, the hypothesis to be tested is

H_0: ϕ_1=1 (Unitary root) against H_1: ϕ_1 < 1 (No unitary root)

Time Series ➡ Test for stationarity using Dickey-Fuller test

- **By rewriting the model (1)**

$$Y_t - Y_{t-1} = \mu + (\phi_1 - 1)\, Y_{t-1} + \epsilon_t. \text{ ---------(2)}$$

Denoting $\Delta Yt = Y_t - Y_{t-1}$ the model (2) may be expressed as

$$\Delta Y_t = \mu + \delta Y_{t-1} + \epsilon_t \text{ ---------- (3)}$$

Hence, the hypothesis to be tested is

H_0: $\delta = 0$ against H_1: $\delta < 0$

$$\textit{The test} - \textit{statistic}: \; t_{\hat{\delta}} = \frac{\hat{\delta}}{SE(\hat{\delta})}$$

$$\textit{The test} - \textit{statistic}: \; t_{\hat{\delta}} = \frac{\hat{\delta}}{SE(\hat{\delta})}$$

Will follow Dickey – Fuller distribution

Reject H0, if $t_{\hat{\delta}}$ < DF – CV (P ≤ 0.05)

and

Fail to reject H0, if $t_{\hat{\delta}}$ > DF – CV (P > 0.05)

Time Series ➡ Test for stationarity using Dickey-Fuller test

- **Consider the 'p' order autoregressive model**

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \cdots + \phi_p Y_{t-p} + \epsilon_t \text{ ---------- (1)}$$

where μ is constant, ϕ_1, ϕ_2 , ϕ_3 , …, ϕ_p are the regression coefficients and ϵ_t is the random error with zero mean and constant variance

If Y_t is a non-stationary process, then it is essential to test ϕ_1, ϕ_2 , ϕ_3 , …, ϕ_p whether they haveunitary root or not.

Hence, the hypothesis to be tested is

H_0: ϕ_i=1 (Unitary root) against H_1: ϕ_i < 1 (No unitary root) for all i = 1, 2, 3, …, p

Time Series ➡ Test for stationarity using Dickey-Fuller test

- **By rewriting the model (1)**

$$Y_t - Y_{t-1} = \mu + (\phi_1 - 1)\, Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \cdots + \phi_p Y_{t-p} + \epsilon_t \text{ -------(2)}$$

Denoting $\Delta Yt = Y_t - Y_{t-1}$ the model (2) may be expressed as

$$\Delta Y_t = \mu + \delta Y_{t-1} + \sum_{i=2}^{p} \beta_i Y_{t-i} + \epsilon_t \text{ --------}$$

Hence, the hypothesis to be tested is

H_0: $\delta = 0$ against H_1: $\delta < 0$

$$\boldsymbol{The\ test - statistic:\ t_{\hat{\delta}} = \frac{\hat{\delta}}{SE(\hat{\delta})}}$$

$$\boldsymbol{The\ test - statistic:\ t_{\hat{\delta}} = \frac{\hat{\delta}}{SE(\hat{\delta})}}$$

Will follow Dickey – Fuller distribution

Reject H0, if $t_{\hat{\delta}}$ < DF (CV) (P ≤ 0.05)

and

Fail to reject H0, if $t_{\hat{\delta}}$ > DF (CV) (P > 0.05)

Time Series ➡ Test for stationarity using Dickey-Fuller test

$$\Delta Y_t = \mu + \delta Y_{t-1} + \sum_{i=2}^{p} \beta_i Y_{t-i} + \epsilon_t \text{ ---------- (3)}$$

Like testing for δ, there is a need to test other parameters β_i

H_0: $\beta_i = 0$ against H_1: $\beta_i < 0$ for all i = 1, 2, 3, …, p

$$\textit{The test} - \textit{statistic}: t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Hence, the hypothesis to be tested is

H_0: $\delta = 0$ against H_1: $\delta < 0$

$$\textit{The test} - \textit{statistic}: t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Will follow Student's t – distribution with p-1 degrees of freedom

Reject H0, if $t_{\hat{\beta}_i}$ > t (CV) (P ≤ 0.05) and

Fail to reject H0, if $t_{\hat{\delta}}$ < t (CV) (P > 0.05)

Time Series ➡ Moving Average (MA) Model

- **Moving-average model**: The **moving-average model** specifies that the output variable depends linearly on the **current and various past** values of a stochastic term.

- Moving average model of order q (MA(q)):

$$Y_t = w_t + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_q Y_{t-q} + \varepsilon_t$$

Time Series ➡ Moving Average (MA) Model

- **Moving-average model** of order q (MA(q)):

$$X_t = w_t + \sum_{i=1}^{q} \theta_i w_{t-q} + \varepsilon_t$$

where: $\theta_1, \theta_2, \ldots, \theta_q$ are constants with $\theta_q \neq 0$; and w_t is **Gaussian white noise** $w_t(0, \sigma^2_w)$.

Note: **Gaussian noise**, named after Carl Friedrich **Gauss**, is **statistical noise** having a probability density function (PDF) equal to that of the normal distribution, which is also known as the **Gaussian** distribution. In other words, the values that the **noise** can take on are **Gaussian** distributed.

- **Autoregressive–moving-average model**. In the statistical analysis of time series Auto-Regressive Moving Average (ARMA) models provide a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the **autoregression** (AR) and the second for the **moving average** (MA).

Time Series ➡ Autoregressive Moving Average (ARMA) Model

- The AR and MA models dynamics can be combined into what is called an *autoregressive moving-average (ARMA) model*.
- The ARMA (1, 1) is $Y_t = \phi_0 + \phi_1 Y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$
  - ϕ_0 – Constant term in AR model
  - ϕ_1 – Coefficient associated with Y_{t-1} in AR model
  - θ_1 – Coefficient associated with ϵ_{t-1} in MA model
  - ϵ_{t-1} – Error while measuring Y_{t-1} in MA model
  - ϵ_t – Error while measuring Y_t in MA model

Time Series ➡ Autoregressive Moving Average (ARMA) Model

- The predicted value for the ARMA (1, 1) is

$$\hat{Y}_t = \phi_0 + \phi_1 Y_{t-1} + \theta_1 \epsilon_{t-1}$$

- Here ϕ_0, ϕ_1, and ϵ_{t-1} are all known except ϵ_t – the error of the current time
- With the help of ACF which helps to identify the MA order and PCAF helps to identify the AR model the ARMA (p, q) model is given by

Time Series ➡ Autoregressive Moving Average (ARMA) Model

- The ACF and PACF are one of the many ways to use to decide the order of ARMA model

ACF

MA order is 2

PACF

AR order is 1

Time Series ➡ Autoregressive Moving Average (ARMA) Model

- Higher order ARMA processes involve additional lags of X and epsilon.

- The ARMA (p, q) is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p}$$
$$+ \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}, \quad or$$

$$\Phi(L) y_t = \Theta(L) \varepsilon_t$$

Time Series ➡ Autoregressive Integrated Moving Average (ARIMA) Model

- ARIMA is an acronym that stands for Auto-Regressive Integrated Moving Average. Specifically,
  - AR ➡ Auto-regression: A model that uses the dependent relationship an observation and some number of **lagged observations**
  - I ➡ Integrated: The use of **differencing** of raw observations in order to make the time series stationary
  - MA ➡ Moving Average: A model that uses the dependency between an observation and **residual error** from a moving average model applied to lagged observations

Non-stationary – Mean not same

Stationary – Mean same

The Stationary model now is $Y_k = \sum_{i=1}^{k-l} Z_{k-i} + a_l$

Time Series ➡ Autoregressive Integrated Moving Average (ARIMA) Model

- Let Y_t, Y_{t-1}, Y_{t-2}, …, Y_{t-p} are the 'p' non-stationary time series data.
- Define $Z_t = Y_{t+1} - Y_t$ be the first order differencing between two successive values.
- The Z_t will now be transformed into stationary data

- Each of these are explicitly specified in the model as a parameter.
- Note that AR and MA are two widely used **linear models** that work on stationary time series and I is a **pre-processing to stationarize** time series if needed.

(a)
goog200
500
450
400
0
50
100
150
200
Day
(b)
diff(goog200)
60
40
20
0
0
50
100
150
200
Day
(c)
strikes
6000
5000
4000
1950
1955
1960
1965
1970
1975
1980
Year
(d)
hsales
80
60
40
1975
1980
1985
1990
1995
Year
(e)
eggs
300
200
100
1900
1920
1940
1960
1980
Year
(f)
pigs
100000
75000
50000
1980
1985
1990
1995
Year
(g)
lynx
6000
4000
2000
0
1820
1840
1860
1880
1900
1920
Year
(h)
beer
180
160
140
120
1991
1992
1993
1994
1995
Year
(I)
elec
16000
12000
8000
4000
1960
1970
1980
1990
Year

Time Series ➡ Autoregressive Integrated Moving Average (ARIMA) Model

- **Rationale** - The first task is to provide a reason why we're interested in a particular model, as quants. Why are we introducing the time series model? What effects can it capture? What do we gain (or lose) by adding in extra complexity?
- **Definition** - We need to provide the full mathematical definition (and associated notation) of the time series model in order to minimise any ambiguity

- **Second Order Properties** - We will discuss (and in some cases derive) the second order properties of the time series model, which includes its mean, its variance and its autocorrelation function
- **Correlogram** - We will use the second order properties to plot a correlogram of a realisation of the time series model in order to visualise its behaviour.

- **Simulation** - We will simulate realisations of the time series model and then fit the model to these simulations to ensure we have accurate implementations and understand the fitting process.

Time Series ➡ Choice of p, d and q

- Look at autocorrelation graph of data (will help if MA model is appropriate).
- Look at partial autocorrelation graph of data (will help if MR model is appropriate).
- Look at extended autocorrelation chart of data (will help if a combination of MA and AR models is needed).

Time Series ➡ Choice of p, d and q

- Try Akaike's information criterion (AIC) on a set of models and investigate the models with the lowest AIC values
- Try the Schwartz Bayesian information criterion (BIC) and investigate the models with the lowest BIC values
- All the above criterion to choose p, d and q are available as package in R.

Time Series ➡ SARIMA model

- Autoregressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting.
- Although the method can handle data with a trend, it does not support time series with a seasonal component.
- An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called Seasonal Autoregressive Integrated Moving Average model (SARIMA).

Time Series ➡ SARIMAX model

- SARIMAX(Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) is **an updated version of the ARIMA model**. ARIMA includes an autoregressive integrated moving average, while SARIMAX includes seasonal effects and eXogenous factors with the autoregressive and moving average component in the model.

Time Series ➡ VAR and VARMAX models

- VAR models (vector autoregressive models) are used for multivariate time series. The structure is that each variable is a linear function of past lags of itself and past lags of the other variables in the system.
- The VARMAX procedure **enables you to model the dynamic relationship both between the dependent variables and also between the dependent and independent variables**. VARMAX models are defined in terms of the orders of the autoregressive or moving-average process (or both).

Reference

- Peter J Brockwell and Richard A Davis. Introduction to Time Series and Forecasting, 2/e, Springer
- Douglas C. Montgomery, and Cheryl L. Jennings, Murat Kulahcin. *Introduction to* Time Series Analysis and Forecasting, 2/e, Wiley

Thank you

2.15%
13.6%
34.1%
34.1%
13.6%
2.15%
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$
68.2%
95.4%
99.7%

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session 15

(Maximum Likelihood Estimation & Gaussian Mixture Model)

16th and 17th Sep 2023

IMP Note to Self

Maximum Likelihood Estimation (MLE)

Estimation is the process of estimating unknown true values of population parameters using their corresponding best sample statistics (good estimators) in an optimum manner.

An estimator is said to be a good if it is

- unbiased,
- consistent,
- efficient and
- sufficient while estimating its parameter.

Maximum Likelihood Estimation (MLE)

- Method of Maximum Likelihood Estimation is the best and most popular one among all methods to obtain an almost good or best estimator for a population parameter.
- It is a method of obtaining an estimator which most (maximum) likely estimates the true value of the parameter i.e., finding an estimator that can give most likely nearer value for the unknown true value of parameter.
- The corresponding estimator is called maximum likelihood estimator (MLE).

Maximum Likelihood Estimation (MLE)

Suppose we have a random sample $x_1, x_2, .., x_n$ whose assumed probability distribution depends on some unknown parameter θ.

Ex:

1) For Binomial unknown parameters are n, p.

2) For Poisson unknown parameter is λ.

3) For Normal unknown parameters are $\mu \; and \; \sigma^2$.

Our goal is to find good estimate of θ(population parameter) using sample and which can be done with the help of MLE.

Maximum Likelihood

- It is observed that a good estimate of unknown parameter θ would be the value of θ that maximizes the probability
- i.e. the likelihood of getting the data we observed (this is reason, why we called as likelihood function)

Maximum Likelihood function

- Let $x_1, x_2, .., x_n$ be i.i.d. random variables drawn from some probability distribution that depends on some unknown parameter θ.
- The goal of MLE to maximize likelihood function

$$L(\theta) = f(x_1, x_2, \ldots, x_n \mid \theta)$$
$$= f(x_1|\theta) * f(x_2|\theta) \ldots f(x_n|\theta)$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

Maximum Likelihood Estimation (MLE)

- The maximum likelihood estimate (MLE) of θ is that value of θ that maximizes likelihood(θ).

It is defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i / \theta)$$

For maximization, we have

$$\frac{dL}{d\theta} = 0 \; ; \quad \frac{d^2L}{d\theta^2} < 0$$

Maximum Likelihood Estimation (MLE)

If L and $\log_e L$ are not differentiable or integrable or principle of maxima-minima fails then in such case direct method of finding the estimator of the parameter which maximizes L or $\log_e L$ is applied using order statistic principle empirically.

Maximum Likelihood Estimation (MLE)

MLEs are:

- Consistent
- Efficient
- Sufficient
- MLEs May (or may not) be unbiased
- MLEs are Asymptotically normally distributed
- Asymptotically tend to have least variance.

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}, \frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}, \frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Solution: Here the distribution is the binomial distribution with $n = 100$.

$$P\left(H = 61 \mid p = \frac{1}{3}\right) = \binom{100}{61}\left(\frac{1}{3}\right)^{61}\left(\frac{2}{3}\right)^{39} \approx 9.6 \times 10^{-9}$$

$$P\left(H = 61 \mid p = \frac{1}{2}\right) = \binom{100}{61}\left(\frac{1}{2}\right)^{61}\left(\frac{1}{2}\right)^{39} = 0.007$$

$$P\left(H = 61 \mid p = \frac{2}{3}\right) = \binom{100}{61}\left(\frac{2}{3}\right)^{61}\left(\frac{1}{3}\right)^{39} = 0.040$$

p.m.f.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$0 \le p \le 1$$
$$x = 0,1,2,\ldots,n;$$

Since $P\left(H = 61 \mid p = \frac{2}{3}\right)$ is maximum and hence MLE is $p = \frac{2}{3}$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Solution: Since the distribution follow is Binomial distribution, with parameter p. Here $n = 100$. The likelihood function (MLE) is

$$P(H = 61|p) = \binom{100}{61} p^{61}(1-p)^{39}$$

For maximization

$$\frac{d}{dp} P(H = 61|p) = 0$$

$$\Rightarrow \binom{100}{61}\left[61p^{60}(1-p)^{39} - 39p^{61}(1-p)^{38}\right] = 0$$

$$\Rightarrow p^{60}(1-p)^{38}(61 - 100p) = 0$$

$$\Rightarrow p = 0, \frac{61}{100}, 1$$

Thus, the likelihoods are

$$P(H = 61|p = 0) = 0$$

$$P\left(H = 61\middle|p = \frac{61}{100}\right) = \binom{100}{61}\left(\frac{61}{100}\right)^{61}\left(\frac{39}{100}\right)$$

$$P(H = 61|p = 1) = 0$$

Since $P\left(H = 61\middle|p = \frac{61}{100}\right)$ is maximum and hence $p = \frac{61}{100}$ is the MLE.

Maximum Likelihood for a Binomial distribution

- Suppose we wish to find the maximum likelihood estimate (MLE) of θ for a Binomial distribution,

$$p_k(k,\theta) = nC_k \theta^k (1-\theta)^{n-k}$$

$$\log p_k(k,\theta) = \log(nC_k) + k\log(\theta) + (n-k)\log((1-\theta)$$

$$\frac{\partial \log p_k(k,\theta)}{\partial \theta} = 0 \Rightarrow 0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

$$k - k\theta = n\theta - k\theta \Rightarrow \theta = \frac{k}{n}$$

Maximum Likelihood Example 3:

Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X=0]=(1-p), P[X=1]=p. Find the maximum likelihood estimate of p.

Soln :

Maximum Likelihood Example 3:

Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X=0]=(1-p), P[X=1]=p. Find the maximum likelihood estimate of p.

Soln :

L(p)=P[X=0] P[X=1] P[X=0] P[X=0] P[X=1] P[X=0]

=(1-p) p (1-p) (1-p) p (1-p)

$=(1-p)^3p^2$.

Log L(p)=log[$(1-p)^3p^2$.]=log[$(1-p)^3$]+log[($p^{2)}$.]=3log(1-p)+2logp

$$\frac{\partial LogL(p)}{\partial p} = 0$$ means, $$\frac{-3}{1-p} + \frac{2}{p} = 0 \Rightarrow \frac{-3p+2-2p}{p(1-p)} = 0 \Rightarrow p = 2/5$$

That is , there is 1/3 chance to observe this sample if we believe the population to be Binomial distributed .

MLE for Binomial distribution parameter P

Let $X_1, X_2, \ldots, X_N \in R$ be samples obtained from a Binomially Distribution.

Binomial Distribution is used to model 'x' successes in 'n' Bernoulli trials. Its p.d.f. is given by:

$$f(x) = \frac{n!}{x!(n-x)!} p^{x} (1-p)^{n-x}$$

The likelihood function L(p) is given by:

$$L(p) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{N} \frac{n!}{x_i!(n-xi)!} p^{x_i} (1-p)^{n-x_i}$$

The log-likelihood is:

$$lnL(p) = \sum_{i=1}^{N} ln(n!) - \sum_{i=1}^{N} ln(x_i!) - \sum_{i=1}^{N} ln(n-x_i!) + \sum_{i=1}^{N} xi.ln(p) + (n - \sum_{i=1}^{N} xi).ln(1-p)$$

Setting its derivative with respect to p to zero,

$$\frac{d}{dp} lnL(p) = \frac{1}{p} . \sum_{i=1}^{N} xi - \frac{1}{1-p} \sum_{i=1}^{N} (n - xi) = 0$$

which implies,

$$\frac{1}{p} . \sum_{i=1}^{N} xi = (\frac{1}{1-p})(N.n - \sum_{i=1}^{N} xi)$$

giving,

$$\hat{p} = \frac{1}{N}(\frac{\sum_{i=1}^{N} x_i}{n}) = \frac{1}{N}(\frac{X1}{n} + \frac{X2}{n} + \ldots + \frac{XN}{n})$$

which is the maximum likelihood estimate.

MLE for Poisson Distribution Parameter

Let $X_1, X_2, \ldots, X_n \in R$ be a random sample from a Poisson distribution

The p.d.f. of a Poisson Distribution is :

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \Bigg| \text{; where x =0,1,2,...}$$

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-\lambda n} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i} \Bigg|$$

The log-likelihood is:

$$lnL(\lambda) = -\lambda n + \sum_{i=1}^{n} xi.\, ln(\lambda) - ln(\prod_{i=1}^{n} xi) \Bigg|$$

Setting its derivative with respect to λ to zero, we have:

$$\frac{d}{d\lambda} lnL(\lambda) = -n + \sum_{i=1}^{n} xi. \frac{1}{\lambda} = 0$$

giving,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X}$$

which is the maximum likelihood estimate

MLEs for Normal Distribution Parameters

Let $X_1, X_2, \cdots, X_n$ be a random sample from a normal distribution with unknown mean μ and variance σ^2. Find maximum likelihood estimators of mean μ and variance σ^2.

In finding the estimators, the first thing we'll do is write the probability density function as a function of $\theta_1 = \mu$ and $\theta_2 = \sigma^2$:

$$f(x_i;\theta_1,\theta_2) = \frac{1}{\sqrt{\theta_2}\sqrt{2\pi}}\exp\left[-\frac{(x_i-\theta_1)^2}{2\theta_2}\right]$$

for $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. We do this so as not to cause confusion when taking the derivative of the likelihood with respect to σ^2. Now, that makes the likelihood function:

$$L(\theta_1,\theta_2) = \prod_{i=1}^{n} f(x_i;\theta_1,\theta_2) = \theta_2^{-n/2}(2\pi)^{-n/2}\exp\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right]$$

and therefore the log of the likelihood function:

$$\log L(\theta_1,\theta_2) = -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum(x_i-\theta_1)^2}{2\theta_2}$$

Now, upon taking the partial derivative of the log likelihood with respect to θ_1, and setting to 0, we see that a few things cancel each other out, leaving us with:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} = \frac{-\cancel{2} \sum (x_i - \theta_1)\cancel{(-1)}}{\cancel{2}\,\theta_2} \overset{\text{SET}}{\equiv} 0$$

Now, multiplying through by θ_2, and distributing the summation, we get:

$$\sum x_i - n\theta_1 = 0$$

Now, solving for θ_1, and putting on its hat, we have shown that the maximum likelihood estimate of θ_1 is:

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

Now for θ_2. Taking the partial derivative of the log likelihood with respect to θ_2, and setting to 0, we get:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{\sum (x_i - \theta_1)^2}{2\theta_2^2} \overset{\text{SET}}{\equiv} 0$$

Multiplying through by $2\theta_2^2$:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} = \left[-\frac{n}{2\theta_2} + \frac{\sum (x_i - \theta_1)^2}{2\theta_2^2} \overset{set}{\equiv} 0\right] \times 2\theta_2^2$$

Hence the MLEs for the parameters mean and variance of the normal model are respectively:

we get:

$$-n\theta_2 + \sum(x_i - \theta_1)^2 = 0$$

And, solving for θ_2, and putting on its hat, we have shown that the maximum likelihood estimate of θ_2 is:

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X} \text{ and } \hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

MLEs for the parameters of Uniform distribution

Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from a Uniform population with probability function

$f(X, a, b) = 1/(b-a)$ for $a <= X <= b$.

The likelihood function is expressed as follows:

$L = f(x_1, a, b)\, f(x_2, a, b) \ldots f(x_n, a, b)$

$= 1/(b-a) \times 1/(b-a) \times \ldots \times 1/(b-a)$

$= 1/(b-a)^n = (b-a)^{-n}$

Log L = -n Log (b-a)

d Log L/da = 0

n/(b-a) = 0

n = 0 (Contradiction)

Also, d Log L/db = 0

-n/(b-a) = 0

-n = 0 (Contradiction)

Therefore, principle of maxima fails to give MLEs.

Now, we use order statistic principle. By ordering the sample observations, we obtain

$a <= x_{(1)} < x_{(2)} < \ldots < x_{(n)} <= b$ since $a <= X <= b$

Where $x_{(1)} = \text{Min}\{x_1, x_2, \ldots, x_n\}$ and $x_{(n)} = \text{Max}\{x_1, x_2, \ldots, x_n\}$

As $x_{(1)}$ and $x_{(n)}$ fall nearest to a and b respectively in the inside interval [a, b] then $x_{(1)}$ and $x_{(n)}$ are considered as MLEs for the respective parameters a and b using order statistic principle.

EM alternates between performing an expectation E-step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization M-step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E-step. The parameters found on the M-step are then used to begin another E-step, and the process is repeated until convergence.

Gaussian Mixture Model

- Suppose Company A share price is normally distributed with mean price Rs. 100 and standard deviation of price Rs.2 with 1000 sample points
- Company B share price is normally distributed with mean price Rs. 100 and standard deviation of price Rs.2 with 800 sample points
- Now, both samples are mixed, then we obtain Normal (Gaussian) mixture model.

So after mixing the processes together, we have the dataset that we see on the plot. We can notice 2 peaks: around 90 and 100, but for many of the points in the middle of the peaks it is ambiguous to which distribution they were drawn from. So how should we approach this problem?

Gaussian Distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Mixtures of Gaussians

- Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$

Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

➢ Find parameters through EM (Expectation Maximization) algorithm

Probabilistic version: Mixtures of Gaussians

Single Gaussian

Mixture of two Gaussians

Gaussian Mixture Model

- Data set

$$D = \{\mathbf{x}_n\} \qquad n = 1, \ldots, N$$

- Consider first a single Gaussian
- Assume observed data points generated independently

$$p(D|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Viewed as a function of the parameters, this is known as the *likelihood function*

The Gaussian Distribution

- Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

mean covariance

Gaussian Mixture Model

- K-dimensional binary random variable z having a 1-of-K representation in which a particular element z_k is equal to 1 and all other elements are equal to 0.
- The values of z_k therefore satisfy $z_k \in \{0,1\}$
- K possible states for the vector z according to which element is nonzero.
- Joint distribution p(x,z) in terms of a marginal distribution p(z) and a conditional distribution p(x|z),
- Marginal distribution over z is specified in terms of the mixing coefficients π_k, such that $p(z_k = 1) = \pi_k$

Fitting the Gaussian Mixture

- We wish to invert this process – given the data set, find the corresponding parameters:
  - mixing coefficients
  - means
  - covariances
- If we knew which component generated each data point, the maximum likelihood solution would involve fitting each component to the corresponding cluster

- Problem: the data set is unlabelled
- We shall refer to the labels as *latent* (= hidden) variables

Gaussian Mixture Model

- Linear super-position of Gaussians

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Normalization and positivity require

$$\sum_{k=1}^{K} \pi_k = 1 \qquad 0 \leqslant \pi_k \leqslant 1$$

- Can interpret the mixing coefficients as prior probabilities

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k)$$

Gaussian Mixture Model

- z uses a 1-of-K representation, we can also write this distribution in the form

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- Conditional distribution of x given a particular value for z is a Gaussian

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

- Joint distribution is given by p(z)p(x|z), and the marginal distribution of x is then obtained by summing the joint distribution over all possible states of z to given

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Gaussian Mixture Model

- Conditional probability of z given x
- use $\gamma(z_k)$ to denote $p(z_k = 1 \mid x)$, whose value can be found using Bayes' theorem

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- π_k as the prior probability of $z_k = 1$, and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed x.

Maximum Likelihood

Log of likelihood function:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- Maximizing the log likelihood function for a Gaussian mixture model turns out to be a more complex problem than for the case of a single Gaussian.

- The difficulty arises from the presence of the summation over k that appears inside the logarithm, so that the logarithm function no longer acts directly on the Gaussian.

GMM Problems and Solutions

- How to maximize the log likelihood
  - solved by expectation-maximization (EM) algorithm
- How to avoid singularities in the likelihood function
  - solved by a Bayesian treatment
- How to choose number *K* of components
  - also solved by a Bayesian treatment

Expectation Maximization (EM) Algorithm

- Conditions for MLE: Setting the derivatives of ln p(X|π,μ,Σ)

in with respect to the means μk of the Gaussian components to zero, we obtain

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

rearranging we obtain:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

where we have defined $$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

Expectation Maximization (EM) Algorithm

- μ_k for the kth Gaussian component is obtained by taking a weighted mean of all of the points in the data set
- Weighting factor for data point x_n is given by the posterior probability $\gamma(z_{nk})$ that component k was responsible for generating xn
- If we set the derivative of ln $p(X|\pi,\mu,\Sigma)$ with respect to Σk to 0, and follow a similar line of reasoning, making use of the result for the maximum likelihood solution for the covariance matrix of a single Gaussian

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

Expectation Maximization (EM) Algorithm

- Maximize ln p(X|π,μ,Σ) with respect to the mixing coefficients π_k with constraint

$$\sum_{k=1}^{K} \pi_k = 1 \qquad 0 \leqslant \pi_k \leqslant 1$$

- Using a Lagrange multiplier and maximizing the following quantity

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

which gives

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$$

- If we now multiply both sides by πk and sum over k, we find λ = −N.
- Rearranging we obtain

$$\pi_k = \frac{N_k}{N}$$

Expectation Maximization (EM) Algorithm

- We first choose some initial values for the means, covariances, and mixing coefficients.
- Then we alternate between the following two updates that we shall call the E step and the M step
- In the expectation step, or E step, we use the current values for the parameters to evaluate the posterior probabilities,
- We then use these probabilities in the maximization step, or M step, to re-estimate the means, covariances, and mixing
- In practice, the algorithm is deemed to have converged when the change in the log likelihood function, or alternatively in the parameters, falls below some threshold

EM Algorithm – Informal Derivation

- The solutions are not closed form since they are coupled
- Suggests an iterative scheme for solving them:
  - make initial guesses for the parameters
  - alternate between the following two stages:
    - E-step: evaluate responsibilities
    - M-step: update parameters using ML results
- Each EM cycle guaranteed not to decrease the likelihood

Initialization E step M step

Figures from Chris Bishop

EM algorithm for GMM

1. Initialize the means μk, covariances Σk and mixing coefficients πk, and evaluate the initial value of the log likelihood.
2. **E step:** Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

EM algorithm for GMM

3. **M step:** Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \qquad \text{where} \quad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}$$

IMP Note to Self

Thanks

Introduction to Statistical Methods

BITS Pilani

Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

Course No: DSECL ZC413 / AIML ZC418
Course Title: ISM
WEBINAR 1 : 06.06.2023

Topics - Webinar

- Descriptive Statistics
  - Measures of Central Tendency
  - Measures of Variability
- Probability
  - Probability – Introduction and Basics
  - Conditional probability

Measures of Central Tendency

A psychologist wrote a computer program to simulate the way a person responds to a standard IQ test. To test the program, he gave the computer 15 different forms of a popular IQ test and computed its IQ from each form

IQ Values:

134	136	137	138	138	143	144	144	145	146	146	146
	147	148	153								

Find the following Statistical measures:

i. Mean, median, and mode

ii. Rang, Variance and standard deviation

iii. Five point summary of the data.

iv. The interquartile range.

v. Identify potential outliers, if any.

vi. Construct and interpret a boxplot

Measures of Central Tendency

2 3 4 5
Average.

Arranging the data in ascending order:

M

134, 136, 137, 138, 138, 143, 144, 144, 145, 146, 146, 146, 147, 148, 153

N=15

N=

S.No	Formula	Solution
Mean	$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$	μ = 143
Median	$p = \frac{n+1}{2}$ th $= \frac{15+1}{2} = 8^{th}$ item	Median = 144
Mode	The mode is the value or values that occur most frequently in the data set. A data set can have more than one mode, and it can also have no mode	Mode = 146

x	f
134	1
136	1
⋮	
144	2
146	3

Mode

144 3
146 3
148 3

Unimodal

147 3

146, 147

Bimodal.

Multimodal

Mode, Bimodal, and Multimodal

A given set of data may have one or more than one Mode. A set of numbers with one Mode is unimodal, a set of numbers having two Modes is bimodal, a set of numbers having three Modes is trimodal, and any set of numbers having four or more than four Modes is known as multimodal.

A set of values having a single unique mode is said to be unimodal, one with two modes is called bimodal, and one with three modes is called trimodal.

Bimodal Mode – A set of data including two modes is identified as a bimodal model. This indicates that there are two data values that possess the highest frequencies. For example, the mode of data set B = { 8, 12, 12, 14, 15, 19, 17, 19} is 12 and 19 as both 12 and 19 are repeated twice in the given set.

No mode:
If no number in a set of numbers occurs more than once, that set has no mode: 3, 6, 9, 16, 27, 37, 48.

Population $\mu =$

Sample $\bar{x}$

Measures of Central Tendency

Range	$\text{Range} = x_n - x_1$	Minimum = 134 Maximum =153 Range R = 19
Variance	For a Population $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$ For a Sample $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	Variance = 26 $x \mid (x-\mu)^2$ Σx $\mu =$
Standard deviation	For a Population $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$ For a Sample $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	Standard Deviation = 5.09901951

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

$$= \frac{\sum_{i=1}^{N}(X^2 - 2\mu X + \mu^2)}{N}$$

$$= \frac{\sum X^2}{N} - \frac{2\mu \sum X}{N} + \frac{N\mu^2}{N}$$

$$= \frac{\sum X^2}{N} - 2\mu^2 + \mu^2$$

$$\sigma^2 = \frac{\sum X^2}{N} - \mu^2$$

$\sum_{i=1}^{N} \mu^2$

cons.

$\sum x_i^2$

$$= \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2$$

x	x^2
$\sum x$	$\sum x^2$

Measures of Central Tendency

Min

134 136 137 138 138 143 144 144 145 146 146 146 147

148 153

Max

Q1 M Q3

- **Minimum**.

$\frac{1}{4}$ N^{th}

- **Q1** (the first quartile, or the 25% mark).
- **Median**.
- **Q3** (the third quartile, or the 75% mark).
- **Maximum**.

Measures of Central Tendency

Quartiles	Quartiles separate a data set into four sections. The median is the second quartile Q_2. It divides the ordered data set into higher and lower halves. The first quartile, Q_1, is the median of the lower half not including Q_2. The third quartile, Q_3, is the median of the higher half not including Q_2.	Quartiles: Q_1 --> 138 Q_2 — Mead --> 144 Q_3 --> 146
Interquartile range	IQR=Q3−Q1	Interquartile Range IQR =8
Potential outliers, if any.	$\text{Upper Fence} = \overset{146}{Q_3} + \boxed{1.5 \times \overset{8}{IQR}} = 158$ $\text{Lower Fence} = Q_1 - \boxed{1.5 \times IQR} = 126$	none

Major outliers .. $Q_3 + 3(IQR)$

Measures of Central Tendency – Box plot

Population size: 15
Median: 144
Minimum: 134
Maximum: 153
First quartile: 138
Third quartile: 146
Interquartile Range: 8
Outliers: none

(126, 158)

M = middle $\frac{N+1}{2}$ th.

$Q_1 = \frac{1}{4}(N+1)$

$Q_3 = \frac{3}{4}$

Q1 , Median (Q2) and Q3

Example of Quartiles (n is odd)

Suppose the distribution of math scores in a class of 19 students in ascending order is:

- 59, 60, 65, 65, 68, 69, 70, 72, 75, **75**, 76, 77, 81, 82, 84, 87, 90, 95, 98

First, mark down the median, Q2, which in this case is the 10th value: 75.

Q1 is the central point between the smallest score and the median. In this case, Q1 falls between the first and fifth score: 68. (Note that the median can also be included when calculating Q1 or Q3 for an odd set of values. If we were to include the median on either side of the middle point, then Q1 will be the middle value between the first and 10th score, which is the average of the fifth and sixth score—(fifth + sixth)/2 = (68 + 69)/2 = 68.5).

Q3 is the middle value between Q2 and the highest score: 84. (Or if you include the median, Q3 = (82 + 84)/2 = 83).

Now that we have our quartiles, let's interpret their numbers. A score of 68 (Q1) represents the first quartile and is the 25th percentile. 68 is the median of the lower half of the score set in the available data—that is, the median of the scores from 59 to 75.

Q1 tells us that 25% of the scores are less than 68 and 75% of the class scores are greater. Q2 (the median) is the 50th percentile and shows that 50% of the scores are less than 75, and 50% of the scores are above 75. Finally, Q3, the 75th percentile, reveals that 25% of the scores are greater and 75% are less than 84.

Q1 , Median (Q2) and Q3 (n is even)

Consider the following statistical summary of a dataset. Write at least three useful observations as a part of data pre – processing. **[Midsem Sep 2023]**

	Nr	Cells	QValue	Fat	Protein
count	969.000000	969.000000	969.000000	969.000000	969.000000
mean	7.074303	358.284830	90.016584	3.620279	3.300196
std	4.759793	344.324223	4.998924	0.349956	0.136071
min	1.000000	0.000000	62.650000	2.240000	2.750000
25% Q1	3.000000	157.000000	87.570000	3.410000	3.220000
50%	6.000000	283.000000	90.750000	3.610000	3.310000
75% Q3	10.000000	476.000000	93.400000	3.820000	3.380000
max	20.000000	5226.000000	100.000000	5.420000	3.840000

I. Direct Observation

- Mea

Range

- Coeff. Var = Sd/Mean → Protein
- Range → Protein
- $Q_3 - Q_1$

Contin….

Observation 1: The variation in Protein is found very low when compared with re
maining using coefficient of variation = sd/mean

Observation 2: The range in Protein is found very low when compared with remain
ing using range = max-Min

Observation 3: The middle range in Protein is found very low when compared with
remaining using quartile range = Q3(75%)-Q1(25%)

Etc……

Basic Probability

In a sample of 100 nails, 18 of them have ~~only~~ defective heads, 12 contain defective tail ends and 9 contain both defectives.

What is the probability that a nail selected at random is either a defective head or a defective tail?

A B $A \cap B$

$A \cup B$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{18}{100} + \frac{12}{100} - \frac{9}{100}$$

$$= 0.21$$

$P(A-B) \neq P(A) - P(B)$

Not

Basic Probability

If $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/5$ then find

a). $P(A \cup B)$

b). $P(A^c \cap B)$

c). $P(A \cap B^c)$ $\rightarrow$ $P(A) - P(A \cap B)$

d). $P(A^c \cap B^c)$

$P(A \cup B)^c$

$= 1 - P(A \cup B)$

e). $P(A^c \cup B^c)$

f). $P((A \cup B)^c)$

$P(A^c \cap B) = P(B - (A \cap B)) = P(B) - P(A \cap B)$

$P(A^c)P(B)$

Basic Probability

If $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/5$ then find

a). $P(A \cup B)$ b). $P(A^c \cap B)$ c). $P(A \cap B^c)$

d). $P(A^c \cap B^c)$ e). $P(A^c \cup B^c)$ f). $P((A \cup B)^c)$

If $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/5$ then find

a). $P(A \cup B)$ b). $P(A^c \cap B)$ c). $P(A \cap B^c)$

d). $P(A^c \cap B^c)$ e). $P(A^c \cup B^c)$ f). $P((A \cup B)^c)$

Solution

a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/3 - 1/5 = 19/30 = 0.6333$

b) $P(A^c \cap B) = P(B - A) = P(B - (A \cap B)) = P(B) - P(A \cap B) = 1/3 - 1/5 = 2/15 = 0.1333$

c) $P(A \cap B^c) = P(A - B) = P(A - (A \cap B)) = P(A) - P(A \cap B) = 1/2 - 1/5 = 3/10 = 0.3$

d) $P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - 19/30 = 11/30 = 0.3667$

e) $P(A^c \cup B^c) = P((A \cap B)^c) = 1 - P(A \cap B) = 1 - 1/5 = 4/5 = 0.8$

f) $P((A \mid\mid B)^c) = 1 - P(A \mid\mid B) = 1 - 19/30 = 11/30 = 0.3667$

Basic Probability

Consider the following data from the industry. Let D be the set of software developers in a company . Some subsets of D and some of probability that a random employee will belong to the set are represented below. Find the missing term in the table with justification.

Try to identify inconsistencies also (if exists and validate them). **[Midsem Jan 2023]**

Set in words	**Symbol**	**Probability**
All software developers	D	?
All software testing engineers	T	?
All software developers ~~and~~ / or software testing engineers	$D \cup T$	0.25
All software developers who are not software testing engineers	$D - T$	0.05
All software testing engineers who are not software developers	$T - D$	0.10
All software developers and software testing engineers	$T \cap D$	?

$D - (D \cap T)$

$D \cup T^c$

$P(D - T) =$

$P(D - (D \cap T)) = 0.05$

$P(D) - P(D \cap T) = 0.05$

$P(T) - P(D \cap T) = 0.10$

$[P(D) + P(T) - P(D \cap T)] - P(D \cap T) = 0.15$

$0.25 - P(D \cap T) = 0.15$

$P(D \cap T) = 0.1$

$P(D) = 0.15$

$P(T) = 0.2$

Basic Probability

A political leader has submitted his nomination to compete in two different electoral constituencies namely A1 and A2. The probability of wining in constituency A1 and A2 is 0.80 and 0.65 respectively. The probability of losing at least one of the constituencies is 0.35. What will be the probability that he will win in one of the constituencies?

A B

- A –

$P(A) = 0.8$, A^c

$P(B) = 0.65$ B^c

$P(A^c \cup B^c) = 0.35$

$P((A \cap B)^c) = 0.35$

$\boxed{P(A \cap B) = 0.65}$

$A \cup B$

$P(A^c \cup B^c) = (A \cap B)^c$

$A^c \cup B^c = (A \cap B)^c$

$P(A^c) = 1 - P(A)$

Required $= P(A \cup B) - P(A \cap B)$

$= 0.8 - 0.65$

$= 0.15$ //

Basic Probability

A political leader has submitted his nomination to compete in two different electoral constituencies namely A1 and A2. The probability of wining in constituency A1 and A2 is 0.80 and 0.65 respectively. The probability of losing at least one of the constituencies is 0.35. What will be the probability that he will win in one of the constituencies?

Assume that A, B be the events defined as follows:

A : "Winning in constituency A1"

B : "Winning in constituency A2"

Given:

P(A) = 0.80, P(B) = 0.65

and $P(\bar{A} \cup \bar{B}) = 0.35$

Now, $\because P(\bar{A} \cup \bar{B}) = 0.35$

$\therefore P(\overline{A \cap B}) = 0.35$

$\Rightarrow 1 - P(A \cap B) = 0.35$

$\Rightarrow \boxed{P(A \cap B) = 0.65}$

Then, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

= 0.80 + 0.65 – 0.65

$\therefore P(A \cup B) = 0.80$

Then, $P(He\ will\ win\ in\ one\ of\ the\ constituencies) = P(A \cup B) - P(A \cap B)$

$= 0.80 - 0.65$

$\therefore P(He\ will\ win\ in\ one\ of\ the\ constituencies) = 0.15$

$P(He\ will\ win\ in\ constituency\ A1\ ONLY) = P(A) - P(A \cap B)$

= 0.80 – 0.65

= 0.15

$P(He\ will\ win\ in\ constituency\ A2\ ONLY) = P(B) - P(A \cap B)$

= 0.65 – 0.65

= 0

Independent events

Comment on the statement:

"If two events are mutually exclusive, then they are independent also and vice versa"

$P(A \cap B) = 0$

$\downarrow$

$P(A) \times P(B) = 0$

either $P(A) = 0$ or $P(B) = 0$

Two independent events cannot be mutually exclusive events - unless one or both events have a probability of zero (meaning one of the events is impossible).

Independent events

Let A and B be the two possible outcomes of an experiment and suppose P(A) = 0.4, P(B) = p and P(AUB) = 0.7

(i) For what choice of ‘p’ are A and B mutually exclusive?

(ii) For what choice of ‘p’ are A and B independent?

Independent events

No common element

Let A and B be the two possible outcomes of an experiment and suppose P(A) = 0.4, P(B) = p and P(AUB) = 0.7

(i) For what choice of 'p' are A and B mutually exclusive?

element

(ii) For what choice of 'p' are A and B independent?

Prob

~~P(A~~

(i) If A and B are mutually exclusive then $P(A \cap B) = 0$

Thus $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ becomes

$0.7 = 0.4 + P(B) - 0$

$\therefore P(B) = 0.7 - 0.4 = 0.3$

(ii) If A and B are independent then $P(A \cap B) = P(A).P(B)$

Thus $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ becomes

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$$

$$\begin{aligned} 0.7 &= 0.4 + P - 0.4 \cdot P \\ &= 0.4 + P(1 - 0.4) \\ &= 0.4 + 0.6P \\ 0.6P &= 0.7 - 0.4 = 0.3 \\ P &= \frac{0.3}{0.6} = \frac{1}{2} = 0.5 \end{aligned}$$

$\therefore P(B) = 0.5$

Answer. (i) p=0.3, (ii) p=0.5

Dependent events – Conditional Probability

If P (A) =1/2, P (B) =1/3, P (A/B) = 1/6 find i). P(B/A) and ii). P(AUB / A)

A, B

$P(B) > 0$

$P(A/B)$

$\{P(\text{pass}/\text{Teach}) = 1.00$

$P(\text{pass}) = 0.75$

A & B depend.

- $P(A/B) = \dfrac{P(A \cap B)}{P(B)}$

- A & B independent

$P(A/B) = P(A)$

- $\boxed{P(A)}$

$P(B/A) = \dfrac{P(B \cap A)}{P(A)}$

$P(A/B) = \dfrac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A/B)\,P(B)$

$P(B \cap A) = \frac{1}{6} \times \frac{1}{3}$

$P[\text{pass}/\text{height}] =$

$P(\text{pass})$

$P(A \cup B/A) = \dfrac{P[(A \cup B) \cap A]}{P(A)} = \dfrac{P(A)}{P(A)} = 1$

Conditional Probability

Consider the following data. **[Midsem Jan 2023]**

A: the event that a patient selected is with high B.P

B: the event that a patient selected is diabetic

C: the event that a patient selected is with cancer

Events	A	B	C	A and B	B and C	A and C	(A and B and C)
Probability	0.35	0.40	0.45	0.22	0.25	0.22	0.15

i) If a patient is selected at random, what is the probability that he suffering from at least one of these health issues?

$P(A \cup B \cup C) =$

i) What is the probability that the patient selected is with B.P given that he is with either of the remaining two diseases?

$P(A / B \cup C)$

iii) What is the probability that the patient is diabetic given that he is suffering from at least one of the diseases mentioned?

$P(B / A \cup B \cup C)$

$P(A/B)$ given already occured

Conditional Probability

F A

In an online shopping survey, 30% of persons made shopping in Flipkart, 45% of persons made shopping in Amazon and 5% made purchases in both. If a person is selected at random, find

i) the probability that he makes shopping in at least one of two companies $P(F \cup A)$

ii).the probability that he makes shopping in Amazon given that he already made shopping in Flipkart. $P(A/F)$

iii).the probability that the person will not make shopping in Flipkart given that he already made purchase in Amazon.

$P[F^c/A]$

Conditional Probability

In an online shopping survey, 30% of persons made shopping in Flipkart, 45% of persons made shopping in Amazon and 5% made purchases in both. If a person is selected at random, find

i) the probability that he makes shopping in at least one of two companies

3rd law. 0.25

ii).the probability that he makes shopping in Amazon given that he already made shopping in Flipkart.

iii).the probability that the person will not make shopping in Flipkart given that he already made purchase in Amazon.

Solution: Given $P(F) = 30\% = 0.30$

$P(A) = 45\% = 0.45$

$P(F \cap A) = 5\% = 0.05$

i) $P(F \cup A) = P(F) + P(A) - P(F \cap A)$

$= 0.30 + 0.45 - 0.05 = 0.7$

ii) $P(A \mid F) = \frac{P(A \cap F)}{P(F)}$

$= \frac{0.05}{0.30} = 0.167$

iii) $P(F' \mid A) = \frac{P(F' \cap A)}{P(A)}$

$P(F' \cap A) = P(A) - P(A \cap F)$

$= 0.45 - 0.05$

$= 0.40$

$P(F' \mid A) = \frac{0.40}{0.45} = 0.88$

100 %

$P(A \cup B) = P(A) + P(B)$

Total Probability

An investor made 45% of investment in stocks, 15% in mutual funds and the rest in his personal business. There are 25%, 10% and 20% of chances of obtaining profit in stocks, mutual funds and personal business respectively. What is the probability that the investor gets profit?

$P(PS) = P(P/S) = 0.25$

$P(P/MF) = 0.10$

$P(P/PB) = 0.2$

$P(\text{Profit})$

$= P(PS \text{ or } PMF \text{ or } PPB)$

$= P(PS) \ldots$

$P(S) = 0.45$

$P(MF) = 0.15$

$P(PB) = 0.4$

$P(PS) = P(S)\, P(P/S)$

An investor made 45% of investment in stocks, 15% in mutual funds and the rest in his personal business. There are 25%, 10% and 20% of chances of obtaining profit in stocks, mutual funds and personal business respectively.

a) What is the probability that the investor gets profit.

b) Given that the investor has got some profit, what are the probabilities that he got it from stocks, mutual funds and business.

Solution Q2:

Given $P(S) = 45\% = 0.45$, $P(M) = 15\% = 0.15$, $P(B) = 0.4$

$P(P/S) = 0.25$, $P(P/M) = 0.1$, $P(P/B) = 0.2$

a) $P(P) = P(S)P(P/s)+P(M)P(P/M)+P(B)P(P/B) = 0.2075$

b) $P(S/P) = P(S)P(P/s)/P(P) = 0.542169$

$P(M/P) = P(M)P(P/M)/P(P) = 0.072289$

$P(B/P) = P(B)P(P/B)/P(P) = 0.385542$

$P(X \cap Y) = P(X/Y)P(Y)$

$= P(Y/X)P(X)$

$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}$

Total Probability

Of 1000 car parts produced, it is known that 350 are produced in one plant, 450 parts in a second plant, and 200 parts in a third plant. Also, it is known that the probabilities are 0.18, 0.21, and 0.11 that the parts will be defective if they are produced in the first, second and third plants respectively. What is the probability that a randomly picked part from this batch is defective?

$P(B) = P(BA_1 \text{ or } BA_2 \text{ or } BA_3)$

$= P(BA_1) + \cdots$

$= P(A_1) \times P(B/A_1)$

$=$

$$P(B) = P(A_1)\,P(B|A_1) + P(A_2)\,P(B|A_2) + P(A_3)\,P(B|A_3) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_2 \cap B) = 0.1795$$

An insurance company insured 1,000 taxi drivers, 2,000 car drivers and 3,000 truck drivers. The probability of their accident is 0.05, 0.15 and 0.1 respectively. One of the insured persons meets with an accident. What is the probability that he is a truck driver?

Exercise

1. If P (A) =1/2, P (B) =1/3, P (A/B) = 1/6 find i). P(B/A) ii). P(B/A') iii).P(AUB / A) iv).P(B/A) .
2. There are three events A, B and C. The probability of occurrence of at least one of them is 0.23. Using the probabilities given in the following Venn diagram, find the probability of the event A.

3. A manufacturing company produces certain types of output by 4 machines i.e A, B, C and D. Machine A produces 15%, Machine B produces 30 % and Machine C produces 30% of daily production. Based on experience it is observed that 1% of the output by Machine A is defective. Similarly, the defectives by other machines are 2% ,3% and 4% respectively. An item is drawn at random and found to be defective. Is it possible to find the defective item is produced bv which Machine? If so, find it.

Exercise

4. A manufacturer has three machine operators A, B and C. The first operator A produce 1% defective items, whereas the other two operators B and C produce 5% and 7% defective items respectively. A is on the job for 50% of the time, B is on the job for 30% of the time. A defective item is produced, what is the probability that it was produced by A, B, C? Based on this write your observations.

5. Consider the following data related to the employees, who are on travel. 40% check work email, 20% use cell phone to stay connected to work, 25% bring laptop with them, 23% check both work email and use cell phone to stay connected, and 50% neither check work email nor use a cell phone to stay connected nor bring a laptop. In addition, 88 out of every 100 who bring a laptop also check work email, and 70 out of every 100 who use a cell phone to stay connected also bring a laptop.

i) What is the probability that a randomly selected traveller who checks work email also uses a cell phone to stay connected?

ii) What is the probability that someone who brings a laptop on vacation also uses a cell phone to stay connected?

iii) If the randomly selected traveler checked work email and brought a laptop, what is the probability that he/she uses a cell phone to stay connected?

THANK YOU!

Introduction to Statistics

BITS Pilani
Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

Course No: DSECL ZC413
Course Title: ISM
WEBINAR 2

Topics - Webinar

- Bayes' and Naïve Bayes Classifiers
- Discrete Random Variable
- Continuous Random Variable
- Probability distribution and density function
- Mean Variance and Standard deviation
- Joint Probability function
- Binomial Distribution

Topics - Webinar

- Poisson Distribution
- Normal Distribution
- Uniform Distribution

Conditional Probability

$\frac{{}^{50}C_1}{} = \frac{50}{} = 1$

Example: Suppose three coins tossed simultaneously. Let E and F be two events which are defined as follows:

E- At least Two heads appear.

F-First Coin shows tail.

a) Find $P(E)$. b) If it is given that the first coin shows tail then what is $P(E)$?

Solution Let S be a sample space.

$\therefore S = \{HHH, HHT, THH, HTH, TTH, HTT, THT, TTT\}$

$n(S) = 8$,

$E = \{HHH, HHT, THH, HTH\}$

$F = \{TTT, THH, THT, TTH\}$

(a) $P(E) = \frac{n(E)}{n(S)} = \frac{4}{8} = \frac{1}{2}$.

(b) $P(E|F) = \frac{n(E \cap F)}{n(F)} = \frac{P(E \cap F)}{P(F)} = \frac{1}{4}$.

Bayes' Theorem

Example: Bag I contains 3 red and 4 black balls and while another Bag II contains 5 red and 6 black balls. One ball is drawn at random from one of the bags and it is found to be red. Find the probability that it was drawn from Bag II.

Solution

$$P(II \mid R) = \frac{P(II)\,P(R \mid II)}{P(R)} = \frac{\frac{1}{2} \times \frac{5}{11}}{\frac{1}{2} \times \frac{3}{7} + \frac{1}{2} \times \frac{5}{11}}$$

Bayes' Theorem

Example: Given three identical boxes I,II and III, each containing two coins . In box I both coins are gold coins, in box II, both are silver coins and in the box III there is one gold and one silver coin. A person chooses a box at random and takes out a coin. If the coin is gold, what is the probability that the other coin in the box is also of gold.

Solution

Prob of selecting gold coin

$$= P(G) = \frac{1}{3}\times 1 + \frac{1}{3}\times 0 + \frac{1}{3}\times\frac{1}{2}$$

$$= \frac{1}{3} + 0 + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

From getting other coin also gold, we have to select first coin from bag I.

$$\therefore \text{Required prob} = P(I|G) = \frac{P(I)\,P(G|I)}{P(G)} = \frac{\frac{1}{3}\times 1}{1/2} = \frac{2}{3}.$$

Bayes' Theorem

Example : From the past record it is known that the chance of hot temperature , mild and cool temperature on a particular day is 0.4,0.3 and 0.3 respectively. The chance of playing golf is depending on temperature of that day. If temperature is hot then chance of playing golf is 0.2, for mild temperature chance is 0.45 and for cool temperature chance is 0.6. If the team plays golf on a particular day then find the probability that the day was hot.

Solution

$0.08 + 0.135 + 0.18$

Prob of playing golf game

$$P(Y) = (0.4)\times(0.2) + (0.3)(0.45) + (0.3)(0.6) = 0.395$$

$$P(H|Y) = \frac{P(H)\,P(Y|H)}{P(Y)} = \frac{0.4\times 0.2}{0.395} = \frac{0.08}{0.395}$$

Naive Bayes Classifiers

Consider a fictional dataset that describes the weather conditions for playing a game of golf. Given the weather conditions, each tuple classifies the conditions as fit("Yes") or unfit("No") for playing golf.

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Naive Bayes Classifiers

If Given that the weather conditions are "Rainy outlook", "Temperature is hot", "high humidity" and "no wind" then predict that the golf match will be played or not under such condition.

Solution

$$P(Y / R, H, HH, F)$$

$$= P(R/Y)\, P(H/Y)\, P(HH/Y)\, P(F/Y) \cdot P(Y)$$

$$P(N / R, H, HH, F) = P(R/N)\, P(H/N) \cdots P(N)$$

Naive Bayes Classifiers

$P(H/Y) = \frac{P(H \cap Y)}{9} = \frac{2}{9}$

$E = \{2, 4, 6\}$

$F = \{1, 3, 5\}$

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9 $\frac{2}{9}$	2/5
Total	9	5	100%	100%

$= \frac{P(R \cap Y)}{9}$

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

$E \cap F = \phi.$

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

$\{HH, HT, TH, TT\}$

$E = \{HH\}$

$F = \{TT\}$

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

$E \cap F = \phi.$

Naive Bayes Classifiers

Suppose today = (Sunny, Hot, Normal, False)

So, probability of playing golf is given by:

$$P(Yes|today) = \frac{P(SunnyOutlook|Yes)P(HotTemperature|Yes)P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes)}{P(today)}$$

and probability to not play golf is given by:

$$P(No|today) = \frac{P(SunnyOutlook|No)P(HotTemperature|No)P(NormalHumidity|No)P(NoWind|No)P(No)}{P(today)}$$

Since, P(today) is common in both probabilities, we can ignore P(today) and find proportional probabilities as:

$$P(Yes|today) \propto \frac{2}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.0141$$

$P(A)+P(A^c)=1$

And

$$P(No|today) \propto \frac{3}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

Now, since

$$P(Yes|today) + P(No|today) = 1$$

These numbers can be converted Into a probability by making the sum equal to 1 (normalization):

$$P(Yes|today) = \frac{0.0141}{0.0141+0.0068} = 0.67 \quad \text{and} \quad P(No|today) = \frac{0.0068}{0.0141+0.0068} = 0.33$$

Since $P(Yes|today) > P(No|today)$.

So, prediction that golf would be played is 'Yes'.

Discrete Random Variable & their Probability Distribution

1. Suppose a random variable X takes on the values $-4,2,3,7$ with respective probabilities $\frac{k+2}{10}, \frac{2k-3}{10}, \frac{3k-4}{10}, \frac{k+1}{10}$. Find the distribution and expected value of X.

$\Rightarrow$ Here $f(-4) = \frac{k+2}{10}$, $f(2) = \frac{2k-3}{10}$, $f(3) = \frac{3k-4}{10}$, $f(7) = \frac{k+1}{10}$

since $f(x)$ is prob mass function

$\therefore \sum f(x) = 1$

$$\frac{k+2+2k-3+3k-4+k+1}{10} = 1$$

$$\frac{7k-4}{10} = 1 \Rightarrow k = 2.$$

$\therefore$ Distribution is $f(-4) = \frac{4}{10}$, $f(2) = \frac{1}{10}$, $f(3) = \frac{2}{10}$, $f(7) = \frac{3}{10}$

The expected value of $X = \sum x f(x) = -4 \times \frac{4}{10} + 2 \cdot \frac{1}{10} + 3 \times \frac{2}{10} + 7 \times \frac{3}{10}$

$$= \frac{-16+2+6+21}{10} = \frac{13}{10} = 1.3$$

Discrete Random Variable & their Probability Distribution

2. Let X be a random variable with following distribution

x	1	3	4	5
$f(x)$	0.4	0.1	0.2	0.3

Find i) the mean ii) variance iii) standard deviation iv) $E(3X+2)$ v) $E(X^2)$ vi) $E(2^x)$

$\Rightarrow$ i) the mean $= \sum x f(x) = 1\times\frac{4}{10} + 3\times\frac{1}{10} + 4\times\frac{2}{10} + 5\times\frac{3}{10}$

$= \frac{4+3+8+15}{10} = \frac{30}{10} = 3.$

ii) variance $= \sigma^2 = E(x^2) - (E(x))^2$ —— (1)

$E(x^2) = \sum x^2 f(x) = 1\times\frac{4}{10} + 9\times\frac{1}{10} + 16\times\frac{2}{10} + 25\times\frac{3}{10}$

$= \frac{4+9+32+75}{10} = 12$

$\therefore \sigma^2 = 12 - (3)^2 = 12 - 9 = 3$

iii) $S.D = \sqrt{3}$.

iv) $E(3x+2) = 3E(x)+2 = 3\times3+2 = 11.$

v) $E(x^2) = 12$

vi) $E(2^x) = \sum 2^x f(x) = 2^1\times\frac{4}{10} + 2^3\times\frac{1}{10} + 2^4\times\frac{2}{10} + 2^5\times\frac{3}{10}$

$= \frac{8}{10} + \frac{8}{10} + \frac{32}{10} + \frac{96}{10} = \frac{144}{10}.$

$32 \times 3 = 96$, $96 + 48 = 144$

Discrete Random Variable & their Probability Distribution

3. A fair coin is tossed until a head or five tails occur. Find the expected number of tosses of the coin.

$\Rightarrow$ Sample Space $= \{H, TH, TTH, TTTH, TTTTH, TTTTT\}$

X be a random which counts number of tosses

$\therefore$ X assigns $\{1, 2, 3, 4, 5\}$ values to sample space.

The pmf $f(x)$ is, $f(1) = \frac{1}{2}$, $f(2) = \frac{1}{4}$, $f(3) = \frac{1}{8}$, $f(4) = \frac{1}{16}$

$$f(5) = \frac{1}{32} + \frac{1}{32} = \frac{1}{16}.$$

$\therefore$ Expected number of tosses $= \sum x f(x)$

$$= 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + 5 \cdot \frac{1}{16}$$

$$= \frac{8 + 8 + 6 + 4 + 5}{16} = \frac{31}{16} \approx 2$$

Discrete Random Variable & their Probability Distribution

4. A box contains 8 light bulbs of which 3 are defective. A bulb is selected from the box and tested. If it is defective, another bulb is selected and tested, until a non-defective bulb is chosen. Find the expected number of bulbs chosen.

$\Rightarrow$ If we write D for defective and N for non-defective then sample space S has four elements.

$$S = \{N, DN, DDN, DDDN\}$$

Let x be a random variable which represent the number of bulbs chosen.

$\therefore$ x assigns $\{1, 2, 3, 4\}$ values to sample points.

and pmf is, $f(1) = \frac{5}{8}$, $f(2) = \frac{3}{8} \cdot \frac{5}{7} = \frac{15}{56}$, $f(3) = \frac{3}{8} \cdot \frac{2}{7} \times \frac{5}{6} = \frac{5}{56}$

$$f(4) = \frac{3}{8} \cdot \frac{2}{7} \cdot \frac{1}{6} \cdot \frac{5}{5} = \frac{1}{56}$$

$$\therefore E(x) = \sum x f(x) = 1 \cdot \frac{5}{8} + 2 \cdot \frac{15}{56} + 3 \cdot \frac{5}{56} + 4 \cdot \frac{1}{56} = \frac{3}{2}$$

Discrete Random Variable & their Probability Distribution

5. A game is played by two person A and B as follows:

"A throws two dice. If the sum is 7 he wins \$3 from B. If the sum is 8 he loses \$2 and if sum is 3 he loses \$4, otherwise no money changes hand. Is this a fair game?"

$\Rightarrow$ Let x be a random variable which represent the gain by person A.

$\therefore x$ assigns $\{0, 3, -2, -4\}$ to sample space.

Let f be prob distribution function.

$\therefore f(3) = P\{$getting sum 7 after throwing two dice$\}$

$= P\{(1,6), (6,1), (2,5), (5,2), (3,4), (4,3)\} = \frac{6}{36}$.

$f(-2) = P\{$getting sum 8$\}$

$= P\{(2,6), (6,2), (4,4), (3,5), (5,3)\} = \frac{5}{36}$.

$f(-4) = P\{$getting sum 3$\}$

$= P\{(1,2), (2,1)\} = \frac{2}{36}$

$f(0) = P\{$getting sum other than 7, 8, and 3$\} = \frac{23}{36}$.

$\therefore E(x) = 0 \times \frac{23}{36} + 3 \times \frac{6}{36} - 2 \times \frac{5}{36} - 4 \times \frac{2}{36} = 0$

$\therefore$ Game is fair.

Discrete Random Variable & their Probability Distribution

6. A linear array EMPLOYEE has n elements. Suppose NAME appears randomly in the array, and there is a linear search to find the location K of NAME, that is, to find K such that EMPLOYEE[K]=NAME. Let X denote the number of comparisons in the linear search. Find the expected value of X.

$\Rightarrow$ Let X denote number of comparisons. Since NAME appears randomly in any position in the array with the same prob $\frac{1}{n}$. we have $X = 1, 2, 3, \ldots n$ each probability with $1/n$.

Hence $E(x) = 1 \cdot \frac{1}{n} + 2 \cdot \frac{1}{n} + \cdots + n \cdot \frac{1}{n} = \frac{n+1}{2}$

Continuous Random Variable & their Probability Distribution

7. Let X be random variable with probability density function $f(x) = \frac{1}{2}e^{-|x|}$ for all $x \in \mathbb{R}$. If $Y = X^2$ then find cumulative distribution function of Y.

$$\Rightarrow \quad F(Y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq x \leq \sqrt{y})$$

$$= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2} e^{-|x|} dx$$

But $e^{-|x|}$ is an even function

$$\therefore F(Y) = 2\int_0^{\sqrt{y}} \frac{1}{2} e^{-x} dx = \int_0^{\sqrt{y}} e^{-x} dx = -e^{-x}\Big|_0^{\sqrt{y}}$$

$$= 1 - e^{-\sqrt{y}}$$

$$\therefore F(Y) = \begin{cases} 1 - e^{-\sqrt{y}} & y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Continuous Random Variable & their Probability Distribution

8. Let X be random variable with probability density function $f(x) =$

$$\begin{cases} x^2\left(2x+\frac{3}{2}\right) & 0 < x \le 1 \\ 0 & otherwise \end{cases}$$

If $y = \frac{2}{x} + 3$ then find $var(y)$.

$$\Rightarrow \quad Var(Y) = 4\,Var\left(\frac{1}{x}\right) = 4\left\{E(1/x^2) - (E(1/x))^2\right\}$$

$$E(1/x^2) = \int_{-\infty}^{\infty} \frac{1}{x^2} f(x)\,dx = \int_0^1 \frac{1}{\cancel{x^2}} \cancel{x^2}\left(2x+\frac{3}{2}\right)dx$$

$$= \frac{2x^2}{2} + \frac{3}{2}x \Big|_0^1 = \frac{5}{2}.$$

$$E(1/x) = \int_0^1 \frac{1}{x} x^2\left(2x+\frac{3}{2}\right)dx$$

$$= \int_0^1 \left(2x^2 + \frac{3}{2}x\right)dx = \frac{2x^3}{3} + \frac{3}{2}\frac{x^2}{2}\Big|_0^1$$

$$= \frac{2}{3} + \frac{3}{4}\Big|_0^1 = \frac{14}{12} = \frac{7}{6}$$

$$\therefore Var(Y) = 4\left\{\frac{5}{2} - \left(\frac{7}{6}\right)^2\right\} = 4\left\{\frac{5}{2} - \frac{49}{36}\right\}.$$

Continuous Random Variable & their Probability Distribution

9. The distribution of amount of gravel (in tons) sold by a particular construction supply company in a given week is a continuous random variable X with probability density function

$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & 0 \le x \le 1 \\ 0 & otherwise \end{cases}$. What is the cumulative distributive function of sales for any X? How do you use this to find the probability that $X < 0.25$? What about the probability that X is greater than 0.75? What about $p(0.25 < X < 0.75)$.

$\Rightarrow$ Let $F(x)$ be cdf, $\therefore F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\,dx$

If $x < 0$, then $F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{x} 0\,dx = 0.$

If $0 \le x < 1$ then $F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{x} \frac{3}{2}(1-x^2)\,dx$

$= \frac{3}{2}\left(x - \frac{x^3}{3}\Big|_0^x\right) = \frac{3}{2}\left(x - \frac{x^3}{3}\right)$ —— (1)

If $x \geq 1$ then $F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{1} \frac{3}{2}(1-x^2)\,dx + \int_{1}^{x} 0\,dx$

$= 1.$

(b) $P(X < 0.25) = F(0.25) = \frac{3}{2}\left(0.25 - \frac{(0.25)^3}{3}\right)$ from (1)

(c) $P(X > 0.75) = \int_{0.75}^{\infty} f(x)\,dx = \int_{3/4}^{1} \frac{3}{2}(1-x^2)\,dx$

$$= \frac{3}{2}\left(x - \frac{x^3}{3}\Big|_{3/4}^{1}\right)$$

$$= \frac{3}{2}\left\{1 - \frac{1}{3} - \frac{3}{4} + \frac{(3/4)^3}{3}\right\}$$

$$= \frac{3}{2}\left\{\frac{2}{3} - \frac{3}{4} + \frac{27}{64 \times 3}\right\}.$$

(d) $P(0.25 < X < 0.75) = \int_{0.25}^{0.75} f(x)\,dx = \int_{1/4}^{3/4} \frac{3}{2}(1-x^2)\,dx$

$$= \frac{3}{2}\left(x - \frac{x^3}{3}\Big|_{1/4}^{3/4}\right) = \frac{3}{2}\left\{\frac{3}{4} - \frac{1}{3}\left(\frac{3}{4}\right)^3 - \frac{1}{4} + \frac{1}{3}\left(\frac{1}{4}\right)^3\right\}$$

$$= \frac{3}{2}\left\{\frac{3}{4} - \frac{9}{64} - \frac{1}{4} + \frac{1}{64 \times 3}\right\}$$

Continuous Random Variable & their Probability Distribution

10. If cumulative distribution function is $F(x) = \begin{cases} 0 & x \le 0 \\ (1-e^{-x})^2 & x > 0 \end{cases}$. Find probability density function and $p(1 < X < 2)$.

$\Rightarrow$ Let $f(x)$ be density function of x.

$$\therefore f(x) = \frac{d}{dx} F(x) = \begin{cases} 0 & x \le 0 \\ 2(1-e^{-x})e^{-x} & x > 0 \end{cases}$$

$$= \begin{cases} 0 & x \le 0 \\ 2(e^{-x} - e^{-2x}) & x > 0 \end{cases}$$

$$\therefore p(1 < x < 2) = 2\int_1^2 (e^{-x} - e^{-2x})\,dx = 2\left\{-e^{-x} + \frac{e^{-2x}}{2}\right|_1^2\}$$

$$= 2\left\{-e^{-2} + \frac{e^{-4}}{2} + e^{-1} - \frac{e^{-2}}{2}\right\}$$

$$= 2\left\{-\frac{3e^{-2}}{2} + \frac{e^{-4}}{2} + e^{-1}\right\}$$

Joint Probability Distribution

11. Let X denote the number of times a photocopy machine will malfunction: 0, 1, 2, or 3 times, on any given month. Let Y denote the number of times a technician is called on an emergency call. The joint p.m.f. $p(x, y)$ is presented in the table below:

X

Y	0	1	2	3
0	0.15	0.30	0.05	0
1	0.05	0.15	0.05	0.05
2	0	0.05	0.10	0.05

Find a) $P(Y > X)$ b) Marginal probability function of X and Y. c) $P(X \leq 2, Y \geq 1)$.

(a) $p(Y>X) = p(0,1) + p(0,2) + p(1,2)$

$= 0.05 + 0 + 0.05 = 0.1$

(b) Marginal prob function of $X = f_1(x) = \sum_y f(x,y)$

$\therefore f_1(0) = f(0,0) + f(0,1) + f(0,2) = 0.15 + 0.05 + 0 = 0.2$

$f_1(1) = f(1,0) + f(1,1) + f(1,2) = 0.30 + 0.15 + 0.05 = 0.5$

$f_1(2) = f(2,0) + f(2,1) + f(2,2) = 0.05 + 0.05 + 0.10 = 0.2$

$f_1(3) = f(3,0) + f(3,1) + f(3,2) = 0 + 0.05 + 0.05 = 0.1$

Marginal prob function of $Y = f_2(y) = \sum_x f(x,y)$.

$f_2(0) = f(0,0) + f(1,0) + f(2,0) + f(3,0) = 0.15 + 0.30 + 0.05 + 0 = 0.5$

$f_2(1) = f(0,1) + f(1,1) + f(2,1) + f(3,1) = 0.05 + 0.15 + 0.05 + 0.05 = 0.3$

$f_2(2) = f(0,2) + f(1,2) + f(2,2) + f(3,2) = 0 + 0.05 + 0.10 + 0.05 = 0.2$

(c) $P(X \leq 2, Y \geq 1) = f(0,1) + f(0,2) + f(1,1) + f(1,2) + f(2,1) + f(2,2)$

$= 0.30 + 0.05 + 0.15 + 0.05 + 0.05 + 0.10$

$= 0.7$

Joint Probability Distribution

12. Let the joint density function for (X, Y) be $f(x, y) = \begin{cases} \frac{c(x+y)}{3} & 0 < x < 2, 0 < y < 1 \\ 0 & ohrtwise \end{cases}$

Find a) the constant c b) $P(X > Y)$ c) Marginal density function of X and Y.

d) Are X and Y are independent?

(a) since $f(x,y)$ is a density function

$$\therefore \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx = 1 \;;\; \int_0^2\int_0^1 \frac{c(x+y)}{3}\,dy\,dx = 1$$

$$\therefore \frac{c}{3}\int_0^2 \left.xy + \frac{y^2}{2}\right|_0^1 dx = 1 \Rightarrow \frac{c}{3}\int_0^2 \left(x + \frac{1}{2}\right) dx = 1$$

$$\Rightarrow \frac{c}{3}\left\{\frac{x^2}{2} + \frac{x}{2}\right|_0^2\} = 1$$

$$\Rightarrow \frac{c}{3}\left\{\frac{4}{2} + \frac{2}{2}\right\} = 1 \Rightarrow c = 1.$$

(b) $P(X > Y) = 1 - P(X \le Y)$

$$= 1 - \int_0^1\int_x^1 \frac{x+y}{3}\,dy\,dx$$

$$= 1 - \frac{1}{3}\int_0^1 xy + \frac{y^2}{2}\Big|_x^1 dx = 1 - \frac{1}{3}\int_0^1 x + \frac{1}{2} - \frac{3x^2}{2}\, dx$$

$$= 1 - \frac{1}{3}\left\{ \frac{x^2}{2} + \frac{x}{2} - \frac{3x^3}{6} \Big|_0^1 \right\}$$

$$= 1 - \frac{1}{3}\left\{ \frac{1}{2} + \cancel{\frac{1}{2}} - \cancel{\frac{1}{2}} \right\}$$

$$= 1 - \frac{1}{6} = \frac{5}{6}.$$

marginal density function of $X = f_1(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$

$$= \int_0^1 \frac{x+y}{3}\, dy$$

$$= \frac{1}{3}\left\{ xy + \frac{y^2}{2} \Big|_0^1 \right\} = \begin{cases} \frac{1}{3}\left(x + \frac{1}{2}\right) & 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

marginal density function of $Y = f_2(y)$

$$= \int_{-\infty}^{\infty} f(x,y)\, dx = \frac{1}{3}\int_0^2 x + y\, dx$$

$$= \frac{1}{3}\left\{ \frac{x^2}{2} + xy \Big|_0^2 \right\} = \begin{cases} \frac{1}{3}(2 + 2y) & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(d) consider $f_1(x) \cdot f_2(y) = \frac{1}{3}\left(\frac{2x+1}{2}\right) \cdot \frac{(2+2y)}{3}$

$$\neq \frac{x+y}{3} = f(x,y)$$

$\therefore$ X and Y are not independent.

Joint Probability Distribution

13. Dick and Jane have agreed to meet for lunch between noon (0:00 p.m.) and 1:00 p.m. Denote Jane's arrival time by X, Dick's by Y, and suppose X and Y are independent with probability density functions $f_X(x) = \begin{cases} 3x^2 & 0 \le x \le 1 \\ 0 & otherwise \end{cases}$, $f_Y(y) = \begin{cases} 2y & 0 \le y \le 1 \\ 0 & otherwise \end{cases}$

Find the probability that Jane arrives before Dick.

$\Rightarrow$ Here it is given that X and Y are independent

$$\therefore f(x,y) = f_X(x) f_Y(y)$$

$$= \begin{cases} 6x^2y & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$P(\text{Jane arrives before Dick}) = P(X<Y)$$

$$= \int_0^1 \int_0^x 6x^2y \, dy \, dx$$

$$= \int_0^1 \frac{6x^2y^2}{2} \Big|_0^x dx = 3\int_0^1 x^2 \cdot [x^2 - 0] dx$$

$$= 3\int_0^1 x^4 dx = 3\left\{ \frac{x^5}{5} \Big|_0^1 \right\} = \frac{3}{5}.$$

Binomial Distribution

14. If from six to seven in the evening one telephone line in every five is engaged in a conversation: what is the probability that when 10 telephone numbers are chosen at random, only two are in use?

$\Rightarrow$ Let x be a random variables which represant the number of telephone lines in use.

This is Binomally distributed random variable because telephone line either in use or not in use.

Let the telephone line in use will be success.

$\therefore$ Prob (success) $= \frac{1}{5}$.

$n = 10$, $p(x=2) = \binom{10}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^8$.

Binomial Distribution

15. It has been determined that 5% of drivers checked at a road stop show traces of alcohol and 10% of drivers checked do not wear seat belts. In addition, it has been observed that the two infractions are independent from one another. If an officer stops five drivers at random:

a) Calculate the probability that exactly three of the drivers have committed any one of the two offenses.

b) Calculate the probability that at least one of the drivers checked has committed at least one of the two offenses.

$\Rightarrow$ Let A be an event that driver traces of alcohol.

B be an event that driver do not wear seat bell.

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = (0.05) + 0.01 - (0.05)(0.01)$

$= 0.145.$

Here success is driver commited any one the two offenses.

$\therefore\ p(\text{success}) = 0.145.$

(a). $n = 5,\ p = 0.145,\ q = 0.855$

$$p(x=3) = {}^5C_3 (0.145)^3 (0.855)^2 = 0.0223$$

(b) $p(x \geq 1) = 1 - p(x=0) = 1 - {}^5C_0 (0.145)^0 (0.855)^5.$

$$= 0.543.$$

Binomial Distribution

16. A student takes an 18-question multiple choice exam, with four choices per question. Suppose one of the choice is obviously incorrect and the students make an "educated" guess of the remaining choices. Find the expected number E of correct answers and the standard deviation σ.

$\Rightarrow$ Let x be a random variable which count the number of correct answers.

Here success is giving correct answer to the question.

$p(\text{success}) = \frac{1}{3}$ {because one of the choice is obviously incorrect}.

As x is binomally distributed

$\therefore$ Expected value of $x = E(x) = np = 18 \times \frac{1}{3} = 6$

and $S.D = \sqrt{npq} = \sqrt{18 \times \frac{1}{3} \times \frac{2}{3}} = \sqrt{4} = 2$.

Poisson Distribution

17. Suppose 300 misprints are distributed randomly throughout a book of 500 pages. Find the probability P that a given page contains: (a) Exactly 2 misprints (b) 2 or more misprints.

$\Rightarrow$ We view the number of misprints on one page as the number of success in a seq of Bernoullis trials.

Here $n = 300$, prob of getting a mistake on a given page $= \frac{1}{500}$.

$\therefore \lambda = np = \frac{300}{500} = 0.6$.

As number of misprints are more, so we can solve this problem by Poisson distribution.

(a) $p(x=2) = \frac{(0.6)^2 e^{-0.6}}{2!} = 0.0988$.

(b) $p(x \geq 2) = 1 - \{p(0) + p(1)\}$

$= 1 - \left\{\frac{(0.6)^0 e^{-0.6}}{0!} + \frac{(0.6)^1 e^{-0.6}}{1!}\right\}$

$= 0.122$.

Poisson Distribution

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

18. The number of industrial injuries per working week in a particular factory is known to follow a Poisson distribution with mean 0.5. Find the probability that (a) in a particular week there will be: (i) less than 2 accidents, (ii) more than 2 accidents; (b) in a three-week period there will be no accidents.

$\Rightarrow$ Here $\lambda = 0.5$

Let x be a random variable which count number of accidents in industry which is Poisson ally distr

① $p(x<2) = p(x=0) + p(x=1)$

$$= \frac{e^{-0.5}(0.5)^0}{0!} + \frac{e^{-0.5}(0.5)^1}{1!}$$

② $p(x>2) = 1 - \{p(0) + p(1) + p(2)\}$

$$= 1 - \left\{\frac{e^{-0.5}(0.5)^0}{0!} + \frac{e^{-0.5}(0.5)^1}{1!} + \frac{e^{-0.5}(0.5)^2}{2!}\right\}$$

③ In a three week period, the avg of number of accidents $= 3 \times \lambda$

$$\therefore p(x=0) = \frac{e^{-(3\times\lambda)}(1.5)^0}{0!} = (e^{-\lambda})^3 = (e^{-0.5})^3.$$

Normal Distribution

19. Suppose the weights of 2000 male students are normally distributed with $\mu = 155lb$ and standard deviation $\sigma = 20\ lb$. Find the number of students with weights: a) not more than 100lb b) between 150 and 175 lb (inclusive) c) greater than or equal to 200lb.

$\Longrightarrow$ Let X be random variables which measure wt of students is normally distributed with

$\mu = 155, \quad \sigma = 20.$

(a) $p(X \leq 100) = ?$

Let $z = \dfrac{100 - 155}{20} = \dfrac{-55}{20} = -2.75$

$\therefore p(X \leq 100) = p(z \leq -2.75)$

$= 0.5 - \phi(2.75)$

$= 0.5 - 0.4970 = 0.03$

$\therefore$ num of students ht not more than 100 lb $= 0.03 \times 2000$
$= 60.$

(b) $p(150 \leq X \leq 175)$

std unit of 150 $= \dfrac{150 - 155}{20} = \dfrac{-5}{20} = -0.25.$

std unit of 175 $= \frac{175-155}{20} = 1$.

$\therefore P(150 \le x \le 175) = P(-0.25 \le z \le 1)$

$= \phi(0.25) + \phi(1)$

$= 0.0987 + 0.3413$

$= 0.44$.

$\therefore$ number of students with ht between 150 and 175 lb $= 0.44 \times 2000$

$= 880$.

(c) $P(x \geq 200) = ?$

std unit of 200 $= \frac{200-155}{20} = \frac{45}{20} = 2.25$.

$P(x \geq 200) = P(z \geq 2.25)$

$= 0.5 - \phi(2.25)$

$= 0.5 - 0.4878$

$= 0.0122$.

$\therefore$ Number of students with ht more than 200 lb $= 2000 \times 0.0122$

$= 24.4 \approx 24$

Standard Normal Distribution Table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995

Gilles Cazelais. Typeset with LaTeX on April 20, 2006.

Normal Distribution

20. A fair die is tossed 720 times. Find the probability that the face 6 will occur: Between 100 and 125 times b) more than 130 times c)less than 110 times.

$\Rightarrow$ This is problem of Binomial distribution with large number of trials, and

$$\text{mean} = \mu = np = 720 \times \frac{1}{6} = 120, \quad \sigma = \sqrt{npq} = 10$$

Here np is also large, so we can solve this problem by using normal distribution.

(a) $BP(100 \le x \le 125) = NP(99.5 \le x \le 125.5)$.

$\therefore$ std unit of $99.5 = \dfrac{99.5 - 120}{10} = -2.05$

std unit of $125 = \dfrac{125 - 120}{10} = 0.5$

$\therefore NP(99.5 \le x \le 125.5) = NP(-2.05 \le z \le 0.5)$

$= \phi(2.05) + \phi(0.5) = 0.4798 + 0.1915$

$= 0.6713$.

(b) $BP(X \geq 130) = NP(X \geq 129.5)$

std unit of $129.5 = \dfrac{129.5 - 120}{10}$

$= 0.95$

$NP(X \geq 129.5) = NP(Z \geq 0.95)$

$= 0.5 - \phi(0.95)$

$= 0.5 - 0.3289 = 0.1711.$

(c) $BP(X \leq 110) = NP(X \leq 110.5)$

std unit of $110.5 = \dfrac{110.5 - 120}{10} = -0.95.$

$NP(X \leq 110.5) = NP(Z \leq -0.95)$

$= 0.5 - \phi(0.95)$

$= 0.5 - 0.3289$

$= 0.1711.$

Standard Normal Distribution Table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4997	.4997

Gilles Cazelais. Typeset with LaTeX on April 20, 2006.

Thank you

IMP Note to Self

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Webinar:3
Testing of Hypothesis

Hypothesis

A statement which is yet to be proved/ established or a statement on the parameter(s) of the Probability distribution to be tested

Null Hypothesis

Hypothesis of no difference or neutral or may be due to Sampling variation

Alternative Hypothesis

Hypothesis of difference which is yet to be proved/ established

Steps involved in Testing of Hypothesis

State null and alternative hypotheses

↓

Specify the level of significance 'α'

↓

Define the probability distribution the data follows

↓

Compute the test statistic based defined population

↓

Define the rejection criteria/critical regional

↓

Conclusion

Rejection criteria

Summary of One- and Two-Tail Tests

One-Tail Test (left tail)	One-Tail Test (right tail)	Two-Tail Test (Either left or right tail)
$H_0: \mu = \mu_0$ vs $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ vs $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$

Example -1

1. A medical research team is investigating the benefits of a new surgical treatment. One of the claims is that the mean recovery time for patients after the new treatment is less than 96 hours. How would you write the null and alternative hypotheses if

a. you are on the research team and want to support the claim?

b. you are on an opposing team and want to reject the claim?

Solution-1

1. To answer the question, first think about the context of the claim. Because you want to support this claim, make the alternative hypothesis state that the mean recovery time for patients is less than 96 hours. So, $H_a: \mu < 96$ hours. Its complement, $\mu \geq 96$ hours, would be the null hypothesis.

   $H_0: \mu \geq 96$

   $H_a: \mu < 96$ (Claim)

2. First think about the context of the claim. As an opposing researcher, you do not want the recovery time to be less than 96 hours. Because you want to reject this claim, make it the null hypothesis. So, $H_0: \mu \leq 96$ hours. Its complement, $\mu > 96$ hours, would be the alternative hypothesis.

   $H_0: \mu \leq 96$ (Claim)

   $H_a: \mu > 96$

Example -2

$n = 32$ | $\sigma = 0.19$

$\bar{x} = 12.9$

$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{12.9 - 13}{0.19/\sqrt{32}}$

innovate achieve lead

. In auto racing, a pit stop is where a racing vehicle stops for new tires, fuel, repairs, and other mechanical adjustments. The efficiency of a pit crew that makes these adjustments can affect the outcome of a race. A pit crew claims that its mean pit stop time (for 4 new tires and fuel) is less than 13 seconds. A random selection of 32 pit stop times has a sample mean of 12.9 seconds and a standard deviation of 0.19 second. Is there enough evidence to support the claim at 1% level of significance. ~~Use a *P*-value~~

-2.98

$.0014$

$H_0: \mu \geq 13$

$H_1: \mu < 13$ (claim)

P)

$\alpha = 0.01$

2.331

5% (1.645)

Z)

$-2.98 \rightarrow -2.9 + 0.08$

$\Rightarrow .0014$

The table shows cumulative probabilities for the standard normal curve.

Cumulative probabilities for NEGATIVE z-values are shown first. SCROLL DOWN to the 2nd page for POSITIVE z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Cumulative probabilities for POSITIVE z-values are shown below.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Solution-2

The claim is "the mean pit stop time is less than 13 seconds." So, the null and alternative hypotheses are

H_0: $\mu \geq 13$ seconds and H_a: $\mu < 13$ seconds. (Claim)

The level of significance is $\alpha = 0.01$. The standardized test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$ Because $n \geq 30$, use the z-test.

$$\approx \frac{12.9 - 13}{0.19/\sqrt{32}}$$ Because $n \geq 30$, use $\sigma \approx s = 0.19$. Assume $\mu = 13$.

$$\approx -2.98.$$

In Table 4 in Appendix B, the area corresponding to $z = -2.98$ is 0.0014. Because this test is a left-tailed test, the P-value is equal to the area to the left of $z = -2.98$. So, $P = 0.0014$. Because the P-value is less than $\alpha = 0.01$, you should decide to reject the null hypothesis.

Left-Tailed Test

Interpretation There is enough evidence at the 1% level of significance to support the claim that the mean pit stop time is less than 13 seconds.

Example: 3

3. The National Institute of Diabetes and Digestive and Kidney Diseases reports that the average cost of bariatric (weight loss) surgery is about $22,500. You think this information is incorrect. You randomly select 30 bariatric surgery patients and find that the average cost for their surgeries is $21,545 with a standard deviation of $3015. Is there enough evidence to support your claim at 5% level of significance Use a *P*-value.

Solution: 3

The claim is "the mean is different from \$22,500." So, the null and alternative hypotheses are

$$H_0: \mu = \$22{,}500$$

and

$$H_a: \mu \neq \$22{,}500. \text{ (Claim)}$$

The level of significance is $\alpha = 0.05$. The standardized test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$ Because $n \geq 30$, use the z-test.

$$\approx \frac{21{,}545 - 22{,}500}{3015/\sqrt{30}}$$ Because $n \geq 30$, use $\sigma \approx s = 3015$. Assume $\mu = 22{,}500$.

$$\approx -1.73.$$

In Table 4, the area corresponding to $z = -1.73$ is 0.0418. Because the test is a two-tailed test, the P-value is equal to twice the area to the left of $z = -1.73$. So,

$$P = 2(0.0418)$$
$$= 0.0836.$$

Because the P-value is greater than α, you should fail to reject the null hypothesis.

Interpretation There is not enough evidence at the 5% level of significance to support the claim that the mean cost of bariatric surgery is different from \$22,500.

Example: 4

4. Employees at a construction and mining company claim that the mean salary of the company's mechanical engineers is less than that of one of its competitors, which is $68,000. A random sample of 30 of the company's mechanical engineers has a mean salary of $66,900 with a standard deviation of $5500. At 5% level of significance test the employees' claim.

Solution: 4

The claim is "the mean salary is less than \$68,000." So, the null and alternative hypotheses can be written as

$$H_0: \mu \geq \$68{,}000 \quad \text{and} \quad H_a: \mu < \$68{,}000. \text{ (Claim)}$$

Because the test is a left-tailed test and the level of significance is $\alpha = 0.05$, the critical value is $z_0 = -1.645$ and the rejection region is $z < -1.645$. The standardized test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Because $n \geq 30$, use the z-test.

$$\approx \frac{66{,}900 - 68{,}000}{5500/\sqrt{30}}$$

Because $n \geq 30$, use $\sigma \approx s = 5500$. Assume $\mu = 68{,}000$.

$$\approx -1.10.$$

The graph shows the location of the rejection region and the standardized test statistic z. Because z is not in the rejection region, you fail to reject the null hypothesis.

5% Level of Significance

Interpretation There is not enough evidence at the 5% level of significance to support the employees' claim that the mean salary is less than \$68,000.

Example -5

5. The U.S. Department of Agriculture claims that the mean cost of raising a child from birth to age 2 by husband-wife families in the United States is $13,120.A random sample of 500 children (age 2) has a mean cost of $12,925 with a standard deviation of $1745. At 10 % level of significance is there enough evidence to reject the claim?

Solution: 5

The claim is "the mean cost is \$13,120." So, the null and alternative hypotheses are

$$H_0: \mu = \$13{,}120 \quad \text{(Claim)}$$

and

$$H_a: \mu \neq \$13{,}120.$$

Because the test is a two-tailed test and the level of significance is $\alpha = 0.10$, the critical values are $-z_0 = -1.645$ and $z_0 = 1.645$. The rejection regions are $z < -1.645$ and $z > 1.645$. The standardized test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Because $n \geq 30$, use the z-test.

$$\approx \frac{12{,}925 - 13{,}120}{1745/\sqrt{500}}$$

Because $n \geq 30$, use $\sigma \approx s = 1745$. Assume $\mu = 13{,}120$.

$$\approx -2.50.$$

The graph shows the location of the rejection regions and the standardized test statistic z. Because z is in the rejection region, you should reject the null hypothesis.

Interpretation There is enough evidence at the 10% level of significance to reject the claim that the mean cost of raising a child from birth to age 2 by husband-wife families in the United States is \$13,120.

5% Level of Significance

Testing of Hypothesis

6. It is claimed that sports-car owners drive on the average 17000 kms per year. A consumer firm believes that the average milage is probably higher. To check, the consumer firm obtained information from randomly selected 40 sports-car owners that resulted in a sample mean of 17352 kms with a population standard deviation of 1348 kms. At what can be concluded about this claim at

   a. 5% level of significance (Critical value is 1.645)

   b. 1% level of significance (Critical value is 2.331)

Example

H_0 → The average milage of sports-car as claimed and the sample average milage may be same

$$H_0 : \mu = \mu_0 = 17000$$

H_1 → The average milage of sports-car as claimed may be higher than the sample average milage

$$H_1 : \mu > \mu_0 = 17000$$

(a) At 5% level of significance with critical value 1.645

$$Z = \frac{17352 - 17000}{\frac{1348}{\sqrt{40}}} = 1.650$$

Critical value for

$\alpha = 0.05$ is 1.645

Since $Z = 1.650 > 1.645$,

Reject H_0 and Accept H_1

Hypothesis to test

H_0: $\mu = 17000$

vs

H_1: $\mu > 17000$

(b) At 1% level of significance with critical value 2.331

$$Z = \frac{17352 - 17000}{\frac{1348}{\sqrt{40}}} = 1.650$$

Hypothesis to test

H_0:μ=17000

vs

H_1:μ>17000

Critical value for

α = 0.01 is 2.331

Since Z =1.624 < 2.330,

Accept H_0 Reject H_1

Testing of Hypothesis

7. Jeffrey, as an eight-year old, established a mean time of 16.43 seconds for swimming the 25-yard freestyle, with a standard deviation of 0.8 seconds. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for 15 25-yard freestyle swims. For the 15 swims, Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds. Conduct a hypothesis test using a preset $\alpha = 0.05$.

calculated test statistic is 2.08, meaning that the sample mean is 2.08 standard deviations away fr
the hypothesized mean of 16.43.

$$t_c = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{16 - 16.43}{.8/\sqrt{15}} = -2.08$$

Testing of Hypothesis

8. It is claimed that sports-car owners drive on the average 18580 kms per year. A consumer firm believes that the average milage is probably higher. To check, the consumer firm obtained information from randomly selected 10 sports-car owners that resulted in a sample mean of 17352 kms with a sample standard deviation of 2012 kms. What can be concluded about this claim at

a. 5% level of significance

b. 1% level of significance

H_0 ⇨ The average milage of sports-car as claimed and the sample average milage may be same

$$H_0 : \mu = \mu_0 = 18580$$

H_1 ⇨ The average milage of sports-car as claimed may be higher than the sample average milage

$$H_1 : \mu > \mu_0 = 18580$$

(a) At 5% level of significance with critical value 1.645

$$|t| = \frac{|17352 - 18580|}{\frac{2012}{\sqrt{10}}} = 1.929$$

95% CI for μ is

$$\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = [16184.91, 18519.09]$$

$$P - value = 0.0428$$

Hypothesis to test

H_0:μ=18580 vs H_1:μ>18580

Critical value for α = 0.05 is 1.833 for 9 degree of freedom

Since | t | =1.929 > 1.833, Reject H_0 and Accept H_1

Testing of Hypothesis

9. The target thickness for silicon wafers used in a certain type of integrated circuit is 245mm. A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18mm and a sample standard deviation of 3.60mm. Does this data suggest that true average wafer thickness is something other than the target value?

Discussion

- Let μ = true average wafer thickness
- Given that μ = 245 , n = 50, $\bar{x} = 246.18$ and s = 3.60.
- Null hypothesis H_0 : $\mu = 245$
- i.E true average wafer thickness is same the target value

- Alternate hypothesis H_1: $\mu \neq 245$ (Two tailed test)
- true average wafer thickness is something other than the target value

- Level of significance: $\propto = 0.05$
- Here the tailed test is Two tailed test,
- Tabulated value of Z at 5% LOS is 1.96 (Z_{tab}).

Test Statistic: $Z_{cal} = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}} = \dfrac{246.18 - 245}{3.60/\sqrt{50}} = 2.32$

Here $|Z_{cal}| = |2.32| > Z_{tab} = 1.96$ at α = 5 % LOS.

Decision : we reject the Null hypothesis H_0 at α = 5 % LOS.

i.e, we accept H_1

i.e, true average wafer thickness is something other than the target value.

Testing of Hypothesis

10. Two types of new cars produced in INDIA are tested for petrol mileage. One sample consisting of 42 cars averaged 15 kmpl while the other sample consisting of 80 cars averaged 11.5 kmpl with population variance as ${\sigma_1}^2 = 2$ and ${\sigma_2}^2 = 1.5$ respectively. Test whether there is any significance difference in the petrol consumption of these two types of cars.

- Given that $n_1 = 42, n_2 = 80, \bar{x}_1 = 15, \bar{x}_2 = 11.5, {\sigma_1}^2 = 2 \; and \; {\sigma_2}^2 = 1.5$
- Null hypothesis H_0 : $\mu_1 = \mu_2$
- i.e, there is no significance difference in the petrol consumption of these two types of cars.

- Alternate hypothesis H_1: $\mu_1 \neq \mu_2$ (Two tailed test)
- i.e, there is significance difference in the petrol consumption of these two types of cars.
- Level of significance: $\propto = 0.05$

- Here the tailed test is Two tailed test,
- Tabulated value of Z at 5% LOS is 1.96 (Z_{tab}).

Test Statistic: $Z_{cal} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} = \frac{15-11.5}{\sqrt{\frac{2}{42} + \frac{1.5}{80}}} = 13.587$

Here $|Z_{cal}| = |13.587| > Z_{tab} = 1.96$ at $\alpha = 5$ % LOS.

Decision : we reject the Null hypothesis H_0 at $\alpha = 5$ % LOS.

i.e, we accept . H_1

i.e, there is significance difference in the petrol consumption of these two types of cars.

Testing of Hypothesis

11. A random sample of 150 recent donations at a certain blood bank reveals that 82 were type A blood. Does this suggest that the actual percentage of type A donations differs from 40%, the percentage of the population having type A blood? Carry out a test of the appropriate hypotheses using a significance level of .01.

Given that P = 0.4 , Q = I – P = 0.6, n = 150

p is the proportion of type A blood donations,

$$p = \frac{82}{150} = 0.546$$

Null hypothesis H_0 : P = 0.4

(actual percentage of type A donations not differs from 40%)

Alternate hypothesis H_1: P ≠ 0.4 (Two tailed test)

Level of significance ∝ = 0.01

Here the tailed test is Two tailed test,

Tabulated value of Z at 1% LOS is 2.58 (Z_{tab}).

Test Statistic: $$Z_{cal} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.546 - 0.4}{\sqrt{\frac{(0.4)(0.6)}{150}}} = 3.65$$

Here $|Z_{cal}| = |3.65| > Z_{tab} = 2.58$ at α = 1 % LOS.

Decision : we reject the Null hypothesis H_0 at α = 1 % LOS.

i.e, we accept H_1.

actual percentage of type A donations differs from 40%

Testing of Hypothesis

12. **In the population, the average IQ is 100 with a standard deviation of 15. A team of scientists want to test a new medication to see if it has either a positive or negative effect on intelligence, or not effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect intelligence?**

Solution : 12

Using the data given in the equation we would have the following:

$\mu_0 = 100$

$\sigma = 15$

$n = 30$

$\bar{x}_n = 140$

Plugging the values into the formular we have:

$$z = \frac{140 - 100}{15}\sqrt{30} = 14.606$$

Step 4: Look up the values of z (called the critical value) from statistical tables.

You can access statistical table from here. Statistical Table

From the table: we get a value of 1.96

Step 5: Draw a conclusion

In this case the tested statistic value of z calculated is more than the critical value obtained from statistical tabl

14.606 > 1.96

Therefore we reject the null hypothesis.

This means, from the question, that the medication administered does not affect intelligence.

You can get video lessons from my YouTube Channel below. Also subscribe for regular updates.

Kindson The Genuis Lessons

Thank you

IMP Note to Self

Introduction to Statistical Methods

Team ISM

BITS
Pilani

Pilani|Dubai|Goa|Hyderabad

innovate achieve lead

BITS Pilani

Webinar-4

Correlation Analysis

(Correlation, Regression and Maximum likelihood estimation)

(24th August 2023)

Contact session	List of topic	References
Webinar-4	Correlation, Regression Analysis and Maximum likelihood estimation.	

CORRELATION Analysis

1.1 Definition of Correlation

1.2Types of Correlation

1.3 Methods of studying Correlation

Definition of Correlation

- Correlation is a statistical tool which is used to know the relationship between two or more than two variables.
- In bivariate frequency distribution f(x,y) if a change in one variable x creates or effects any change in another variable y, then the two variables are set to be correlated.

Types of Correlation

- Positive or Negative Correlation
- Simple, Partial and Multiple correlation
- Linear and non linear correlation

Positive or Negative Correlation

- If a change in the variable X creates any change in the variable Y in the same direction then X and Y are said to be Positive correlated.
- If one variable increases the second variable will also increases correspondingly.

Examples: 1. Income and expenditure

2. Distance travelled by train and train fare

3.Age and sickness

4. Advertising and sales

If a change in one variable creates change in other variable, but in opposite direction then two variables are said to be negatively correlated. i.e if one variable increases, then the other decreases or vice versa.

Examples: 1.Price and demand

2. Demand for sweaters and the day temperature.

Cont…

Scatter Plots and Types of Correlation

x = SAT score
y = GPA

Positive Correlation–as x increases, y increases

Cont…

Scatter Plots and Types of Correlation

x = hours of training (horizontal axis)
y = number of accidents (vertical axis)

Negative Correlation–as x increases, y decreases

Simple, partial and multiple correlation

- The distinction between simple, partial and multiple correlation is based on the number of variables studied.
- When only two variables are studied, it is a simple correlation

  Example: Intelligent quotient and marks secured.

  - When 3 or more variables are studied then it is a multiple correlation.

  Example: yield of rice per acre and both the amount of rainfall and the amount of fertilizers used.

  - In partial correlation we recognize more than two variables but consider only two variables are influencing each other and ignore other variables by keeping them as constant.

  Example: Mileage of car and car type can only be studied by ignoring road conditions, driving skills etc..

Linear and Non-linear(curvi linear correlation)

- If the change in one variable causes the change in other variable in constant ratio then it is called linear correlation.
- If such variables are plotted on a graph paper, plotted points will fall on a straight line.

example: Amount of raw material required and number of finished goods.

When the amount of change in one variable is not in a constant ratio to the amount of change in other variable then it is called non-linear correlation.

Example: Amount of rainfall and yield of crop.

Methods of studying Correlation

There are various methods to know whether the two variables are correlated or not. They are

- Scatter diagram
- Karl Pearson's coefficient of correlation

Scatter Diagram

To know the correlation between any two variables , scatter diagrams are used.

Procedure: given n pair of values (x1,y1) (x2,y2)…(xn,yn) of two variables X and Y.

- Take the independent variable on x axis and dependent variable on y axis. Then plot the n points on the graph.
- The diagram of dots thus obtained is called scatter diagram.
  - If the points reveal any upward or downward trend the variables are said to be correlated , otherwise un-correlated.
  - If the points are very closed to each other, a good amount of correlation exists.
  - Upward trend indicates +ve correlation and downward trend indicates –ve correlation.
  - Limitation: By this method we cannot establish the exact degree of correlation between the variables.

Scattered diagram- Some specific cases

$$-1 \leq Correlation(x, y) \leq +1$$

Karl Pearson's coefficient of correlation

The extent relationship between the variables calculated with the help of statistical techniques known as correlation coefficient. It was developed by Karl Pearson. It is denoted by r(x,y) or r_{xy}

It always varies between +1 or -1

When r= +1 then it denotes perfect +ve correlation

r= 0 denotes no correlation

r= -1 perfect –ve correlation

Correlation coefficent between two random variables X and Y is as follows:

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}$$

$$= \frac{E(X,Y) - E(X)E(Y)}{\sqrt{E(X^2) - [E(X)]^2} . \sqrt{E(\mathrm{Y}^2) - [E(\mathrm{Y})]^2}}$$

$$= \frac{\frac{1}{N} \sum (X_i - \bar{X})(\mathrm{Y}_i - \bar{Y})}{\sqrt{\frac{1}{N} \sum (X_i - \bar{X})^2} \sqrt{\frac{1}{N} \sum (\mathrm{Y}_i - \bar{Y})^2}}$$

$$= \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}$$

Properties of the coefficient of correlation

1. The coefficient of correlation lies between -1 and +1.
2. Correlation is independent of change of origin and scale.
3. The coefficient of correlation is the geometric mean of two regression coefficients

$$r_{xy} = \sqrt{b_{xy} b_{yx}}$$

1. The degree of relationship between two variables is symmetric

$$r_{xy} = r_{yx}$$

Calculate karl pearson's coefficient of correlation for the following data

X	9	8	7	6	5	4	3	2	1
Y	15	16	14	13	11	12	10	8	9

Sol: Correlation coefficient

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

X	Y	X^2	Y^2	XY
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$

Therefore, N=9, $\sum XY = 597$, $\sum X = 45$, $\sum X^2 = 285$, $\sum Y = 108$, $\sum Y^2 = 1356$

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{9*597 - 45*108}{\sqrt{9*285 - (45)^2}\sqrt{9*1356 - (108)^2}}$$

$$= \frac{513}{\sqrt{540}\sqrt{540}} = \frac{513}{540}$$

$$= +0.95$$

There is a high degree of +ve correlation between variables X and Y.
i.e if x increases y also increases

Calculate karl pearson's coefficient of correlation for the following data

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

X	Y	$u = x - \bar{x}$	$v = y - \bar{y}$	u^2	v^2	uv
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
$\sum x = 544$	$\sum y = 552$	$\sum u = 0$	$\sum v = 0$	$\sum u^2 = 36$	$\sum v^2 = 44$	$\sum uv = 24$

$$\bar{x} = \frac{\sum x}{n} = \frac{544}{8} = 68 \qquad \bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69$$

$$r_{xy} = \frac{n\sum uv - (\sum u)(\sum v)}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$$

$$= \frac{8(24) - 0}{\sqrt{8(36) - 0}\sqrt{8(44) - 0}}$$

$$= 0.603$$

sales	15	18	25	27	30	35
Advertising expenditure	50	65	82	95	110	120

Find the correlation coefficient for the above data

Regression Analysis

The term regression literally means stepping back towards the average

Definition: Regression analysis is a mathematical measure of the average relationship between 2 or more variables in terms of original units of the data.

- In regression analysis there are 2 variables.
- The variable whose values are influenced or can be predicted is called dependent variable and the variable which influences or used for prediction is called independent variable.
- In regression analysis the dependent variable is known as regressed or explained variable, while the independent variable is known as regressor or predictor or explanator.

Lines of Regression

- If we plot the points of a bivariate distribution in the scatter diagram the points will cluster round a curve called curve of regression.
- If the curve is straight line it is called line of regression or linear regression otherwise the regression is said to be non linear regression.
- Line of regression of y on x is the line which gives the best estimate for the value of y for any specified value of x.
- Line of regression of x on y is the line which gives the best estimate for the value of x for any specified value of y.

Types of Regression

- Simple Regression: the regression analysis confined to the study of only two variables at a time is called as simple regression
- Multiple Regression: The regression analysis for studying more than two variables at a time is termed as Multiple Regression.
- Utility of Regression Test: Regression lines are useful in prediction of values of one variable for a specified value of other variable.
- Examples:

  1.When price and demand are related, we can estimate the future demand for a specified price.

  2. When crop yield depend on amount of rain fall then regression test can predict crop yield for a particular amount of rain fall.

Cont…

- Let x and y are two variables for regression analysis then there will be two regression lines.
- Regression line of x on y = $$x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$
$$x-\bar{x}=b_{xy}(y-\bar{y})$$
- Regression line of y on x= $$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$
$$y-\bar{y}=b_{yx}(x-\bar{x})$$
- The intersection of two lines x on y and y on x gives the mean values of x series and y series. i.e., and $\bar{x}$ $\bar{y}$

Properties of Regression coefficients

- Correlation coefficient is the geometric mean of the regression coefficients. $r = \sqrt{b_{xy}.b_{yx}}$

- If one regression coefficient is > unity then the other must be < unity.
- The arithmetic mean of the two regression coefficient is > the correlation coefficient.

Problems

- The folowing table gives the normal weight of kids during the first 8 years of life;

Age	0	1	2	3	5	6	7	8
Weight	5	7	8	10	15	17	20	22

i. Obtain two lines of regression
ii. Estimate the weight of kid at the age of 4 years

solution : we know that regression line y on x is $y-\bar{y}=b_{yx}(x-\bar{x})$

Where $\bar{x},\bar{y}$ are means

$$b_{yx}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$ is regression coefficient.

Line of regression x on y is $x-\bar{x}=b_{xy}(y-\bar{y})$

$$b_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(y_i-\bar{y})^2}$$

N=no of observations.

$$\bar{x}=\frac{\sum x}{N}=\frac{32}{8}=4$$

$$\bar{y}=\frac{\sum y}{N}=\frac{104}{8}=13$$

Age(x)	Weight(y)	$(x-\bar{x})=x-4$	$(y-\bar{y})=y-13$	$(x-\bar{x})^2=(x-4)^2$	$(y-\bar{y})^2=(y-13)^2$	$(x-\bar{x})(y-\bar{y})=(x-4)(y-13)$
0	5	-4	-8	16	64	32
1	7	-3	-6	9	36	18
2	8	-2	-5	4	25	10
3	10	-1	-3	1	9	3
5	15	1	2	1	4	2
6	17	2	4	4	16	8
7	20	3	7	9	49	21
8	22	4	9	16	81	36
Total=32	104	0	0	60	284	130

From the table $\sum x=32, \sum y=104, \sum(x-\bar{x})=0, \sum(y-\bar{y})=0, \sum(x-\bar{x})^2=60$

$$\sum(y-\bar{y})^2=284, \sum(x-\bar{x})(y-\bar{y})=130$$

$$b_{yx} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{130}{60} = 2.167$$

$$b_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} = \frac{130}{284} = 0.458$$

Line of regression y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 13 = 2.167(x - 4)$$
$$y - 13 = 2.167x - 8.668$$
$$y = 2.167x + 5.668$$

Line of regression x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$x - 4 = 0.458(y - 13)$$
$$x - 4 = 0.458y - 5.954$$
$$x = 0.458y - 1.954$$

Weight of a kid at the age of 4 is estimating y when x=4
We use y on x regression line.

$$y = 2.167x + 5.668$$

$$y = 2.167(4) + 5.668$$

$$y = 14.336$$

The weight of a kid at the age of 4is 14.336.

Problem: The following are the regression equations. Find the regression line x on y and y on x and also find

i. $\bar{x}$ And $\bar{y}$
ii. The regression coefficients
iii. Correlation Coefficients

$$8x - 10y + 66 = 0$$

$$40x - 18y = 214$$

Solution:

$$8x - 10y + 66 = 0 \quad \text{.....(1)}$$

$$40x - 18y = 214 \quad \text{.....(2)}$$

i. By solving two regression lines we get $\bar{x}$ And $\bar{y}$

$\bar{x}$ = 13 And $\bar{y}$ =17

ii. Rewrite the equations of regression to get the form x on y and y on x
From eq-(1)

$$10y = 8x + 66 \Rightarrow y = \frac{8}{10}x + \frac{66}{10}$$

$$y = a + bx \Rightarrow y = \frac{66}{10} + \frac{8}{10}x$$

$$b_{yx} = \frac{8}{10} = \frac{4}{5}$$

Equation (2) $40x - 18y = 214$

$$\therefore x = \frac{18}{40}y + \frac{214}{40}$$

$$x = a + by \Rightarrow x = \frac{214}{40} + \frac{18}{40}y$$

$$\therefore b_{xy} = \frac{18}{40} = \frac{9}{20}$$

Regression coefficient x on y $= \frac{9}{20} = b_{xy}$

Regression coefficient y on x $= \frac{4}{5} = b_{yx}$

iii)Correlation coefficient $r = \pm\sqrt{b_{xy}.b_{yx}} = \pm\sqrt{\frac{9}{20}.\frac{4}{5}} = \pm\sqrt{\frac{9}{25}} = \pm 0.6$

Since both regression coefficients are positive ,r must be positive there fore r=0.6

Least square Method

Algebraically method-:

I. Least Square Method-:

The regression equation of X on Y is :

$$X = a + bY$$

Where,

X=Dependent variable

Y=Independent variable

The regression equation of Y on X is:

$$Y = a + bX$$

Where,

Y=Dependent variable

X=Independent variable

And the values of a and b in the above equations are found by the method of least of Squares-reference . The values of a and b are found with the help of normal equations given below:

(I)

$$\sum X = na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

(II)

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Cont…

Example1-:From the following data obtain the two regression equations using the method of Least Squares.

X	3	2	7	4	8
Y	6	1	8	5	9

Solution-:

X	Y	XY	X^2	Y^2
3	6	18	9	36
2	1	2	4	1
7	8	56	49	64
4	5	20	16	25
8	9	72	64	81
$\sum X = 24$	$\sum Y = 29$	$\sum XY = 168$	$\sum X^2 = 142$	$\sum Y^2 = 207$

Cont…

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Substitution the values from the table we get

29=5a+24b……(i)
168=24a+142b
84=12a+71b……..(ii)

Multiplying equation (i) by 12 and (ii) by 5

348=60a+288b……(iii)
420=60a+355b……(iv)

By solving equation(iii)and (iv) we get

a=0.66 and b=1.07

Cont…

By putting the value of a and b in the **Regression equation Y on X** we get

Y=0.66+1.07X

Now to find the regression equation of X on Y ,
The two normal equation are

$$\sum X = na + b\sum Y$$
$$\sum XY = a\sum Y + b\sum Y^2$$

Substituting the values in the equations we get

24=5a+29b……(i)
168=29a+207b……..(ii)

Multiplying equation (i)by 29 and in (ii) by 5 we get

a=0.49 and b=0.74

Substituting the values of a and b in the **Regression equation X and Y**

X=0.49+0.74Y

- Fit a straight line to the following data

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.5

- Show that the line of fit of the following data is given by y=0.7x+11.28

x	0	5	10	15	20	25
y	12	15	17	22	24	30

Non linear least square approximation:

Parabola: let the equation of parabola to be fit be $y = a + bx + cx^2 .....(1)$

The normal equations are

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 .....(2)$$

Solving these equations for a,b,c and substituting in (1) we get required parabola of best fit.

Problem

1)Fit the second degree parabola to the following data

x	0	1	2	3	4
y	1	5	10	22	38

sol: let the equation of the parabola be

$$y = a + bx + cx^2 \ldots\ldots\ldots\ldots(1)$$

The normal equations are

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \ldots\ldots\ldots(2)$$

x	y	xy	x^2	x^2y	x^3	x^4
0	1	0	0	0	0	0
1	5	5	1	5	1	1
2	10	20	4	40	8	16
3	22	66	9	198	27	81
4	38	152	16	608	64	256
$\sum x = 10$	$\sum y = 76$	$\sum xy = 152$	$\sum x^2 = 30$	$\sum x^2 y = 851$	$\sum x^3 = 100$	$\sum x^4 = 354$

Sub these values in the normal equations

$$76 = 5a + 10b + 30c$$

$$243 = 10a + 30b + 100c$$

$$851 = 30a + 100b + 354c$$

Solving above a=1.42, b=0.26, c=2.221

Sub a, b, c in (1)

$$y = 1.42 + 0.26x + 2.21x^2$$ Is the required parabola of the best fit.

Power curve: The power curve is $y = ax^b \ldots\ldots(1)$

Taking log on both sides

$$\log_{10} y = \log_{10} a + b \log_{10} x$$

$$Y = A + bX \ldots\ldots\ldots\ldots(2)$$

$$where, Y = \log_{10} y, A = \log_{10} a, X = \log_{10} x$$

(2) Is a linear equation in X and Y
The normal equations are

$$\sum Y = nA + b\sum X$$

$$\sum XY = A\sum X + b\sum X^2$$

Solving for A and b and substitute in (2).

Problem: Fit a curve $y = ax^b$ to the following data

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.10	6.80	7.50

Sol: $y = ax^b$ .....(1)

Taking log on both sides

$$\log_{10} y = \log_{10} a + b \log_{10} x$$

$$Y = A + bX \text{.....}(2)$$

$$where, Y = \log_{10} y, A = \log_{10} a, X = \log_{10} x$$

This is a linear eq in X and Y
The normal eq s are

$$\sum Y = nA + b \sum X$$

$$\sum XY = A \sum X + b \sum X^2 \qquad \text{.....}(3)$$

x	X=log x	y	Y=log y	XY	X^2
1	0	2.98	0.4742	0	0
2	0.3010	4.26	0.6294	0.1894	0.0906
3	0.4771	5.21	0.7168	0.3420	0.2276
4	0.6021	6.10	0.7853	0.4728	0.3625
5	0.6990	6.80	0.8325	0.5819	0.4886
6	0.7782	7.50	0.8751	0.6810	0.6056
TOTAL	$\sum X = 2.8574$		$\sum Y = 4.3133$	$\sum XY = 2.2671$	$\sum X^2 = 1.7749$

Substituting the above values in (3)

4.3133 = 6A+2.8574 b

2.2671=2.8574 A +1.7749 b

Solving the equations for A and b

A = 0.4739 And b = 0.5143

a = $(10)^A$ = $(10)^{0.4739}$ = 2.978

Substituting a and b in (1)

$$y = 2.978x^{0.5143}$$

Exponential curve: $1)\, y = ae^{bx}, 2)\, y = ab^{x}$

1)The exponential curve $y = ae^{bx}$.....(1)

Taking log on both sides

$$\log_{10} y = \log_{10} a + bx \log_{10} e$$

$$Y = A + Bx.....(2)$$

$$where, Y = \log_{10} \mathrm{y}, A = \log_{10} a, \mathrm{B} = b \log_{10} e$$

Equation (2) is linear equation in x and Y
So the normal equations are given by

$$\sum Y = nA + B \sum x$$

$$\sum xY = A \sum x + B \sum x^{2}$$

Problem: Fit acurve of the form $y = ae^{bx}$ to the following data

x	0	5	8	12	20
y	3.0	1.5	1.0	0.55	0.18

Sol: Let $y = ae^{bx}$.....(1)

Taking log on both sides

$$\log_{10} y = \log_{10} a + bx \log_{10} e$$

$$Y = A + Bx.....(2)$$

$$where, Y = \log_{10} \mathrm{y}, A = \log_{10} a, \mathrm{B} = b \log_{10} e$$

Equation (2) is linear equation in x and Y
So the normal equations are given by

$$\sum Y = nA + B\sum x$$

$$\sum xY = A\sum x + B\sum x^2 \ldots\ldots\ldots\ldots(3)$$

x	y	$Y = \log_{10} y$	x^2	xY
0	3.0	0.4771	0	0
5	1.5	0.1761	25	0.8805
8	1.0	0	64	0
12	0.55	-0.2596	144	-3.1152
20	0.18	-0.7447	400	-14.894
$\sum x = 45$		$\sum Y = -0.3511$	$\sum x^2 = 633$	$\sum xY = -17.1287$

Sub these values in (3) we get

$$5A + 45B = -0.3511$$

$$45A + 633B = -17.1287$$

Solving above for A and B
A=0.4815 and B=-0.0613

$$A = \log_{10} a = 0.4815$$

$$\therefore a = (10)^A = (10)^{0.4815} = 3.0304$$

$$B = b \log_{10} e = -0.0613$$

$$\therefore b = \frac{-0.0613}{0.4343} = -0.1411$$

Sub a and b values in (1)

$$y = 3.0304e^{-0.1411x}$$

Maximum Likelihood Estimation (MLE)

- Suppose we have a random sample $x_1, x_2, .., x_n$ whose assumed probability distribution depends on some unknown parameter θ.
- Ex:

1) For Binomial unknown parameters are n, p.

2) For Poisson unknown parameter is λ.

Our goal is to find good estimator of θ(population parameter) using sample and which can be done with the help of MLE.

Maximum Likelihood function

- Let $x_1, x_2, .., x_n$ be i.i.d. random variables drawn from some probability distribution that depends on some unknown parameter θ.
- The goal of MLE to maximize likelihood function

$$L(\theta) = f(x_1, x_2, \ldots, x_n \mid \theta)$$

$$= f(x_1|\theta) f(x_2|\theta) \ldots f(x_n|\theta)$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

- The maximum likelihood estimate (MLE) of θ is that value of θ that maximizes likelihood(θ).

It is defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i / \theta)$$

For maximization, we have

$$\frac{dL}{d\theta} = 0 \; ; \quad \frac{d^2L}{d\theta^2} < 0$$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}, \frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Solution: Here the distribution is the binomial distribution with $n = 100$.

$$P\left(H = 61 \mid p = \frac{1}{3}\right) = \binom{100}{61}\left(\frac{1}{3}\right)^{61}\left(\frac{2}{3}\right)^{39} \approx 9.6 \times 10^{-9}$$

$$P\left(H = 61 \mid p = \frac{1}{2}\right) = \binom{100}{61}\left(\frac{1}{2}\right)^{61}\left(\frac{1}{2}\right)^{39} = 0.007$$

$$P\left(H = 61 \mid p = \frac{2}{3}\right) = \binom{100}{61}\left(\frac{2}{3}\right)^{61}\left(\frac{1}{3}\right)^{39} = 0.040$$

p.m.f.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$0 \le p \le 1$$

$$x = 0,1,2,\ldots,n;$$

Since $P\left(H = 61 \mid p = \frac{2}{3}\right)$ is maximum and hence MLE is $p = \frac{2}{3}$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Solution: Since the distribution follow is Binomial distribution, with parameter p. Here $n = 100$. The likelihood function (MLE) is

$$P(H = 61|p) = \binom{100}{61} p^{61}(1-p)^{39}$$

For maximization

$$\frac{d}{dp} P(H = 61|p) = 0$$

$$\Rightarrow \binom{100}{61}\left[61p^{60}(1-p)^{39} - 39p^{61}(1-p)^{38}\right] = 0$$

$$\Rightarrow p^{60}(1-p)^{38}(61 - 100p) = 0$$

$$\Rightarrow p = 0, \frac{61}{100}, 1$$

Thus, the likelihoods are

$$P(H = 61|p = 0) = 0$$

$$P\left(H = 61\middle|p = \frac{61}{100}\right) = \binom{100}{61}\left(\frac{61}{100}\right)^{61}\left(\frac{39}{100}\right)^{3}$$

$$P(H = 61|p = 1) = 0$$

Since $P\left(H = 61\middle|p = \frac{61}{100}\right)$ is maximum and hence $p = \frac{61}{100}$ is the MLE.

Maximum Likelihood for a Binomial distribution

- Suppose we wish to find the maximum likelihood estimate (MLE) of θ for a Binomial distribution,

$$p_k(k,\theta) = nC_k \theta^k (1-\theta)^{n-k}$$

$$\log p_k(k,\theta) = \log(nC_k) + k\log(\theta) + (n-k)\log((1-\theta)$$

$$\frac{\partial \log p_k(k,\theta)}{\partial \theta} = 0 \Rightarrow 0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

$$k - k\theta = n\theta - k\theta \Rightarrow \theta = \frac{k}{n}$$

Let $X_1, X_2, \ldots, X_n \in R$ be a random sample from a Poisson distribution

The p.d.f. of a Poisson Distribution is :

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \Bigg| \text{; where } x = 0, 1, 2, \ldots$$

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-\lambda n} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i} \Bigg|$$

The log-likelihood is:

$$lnL(\lambda) = -\lambda n + \sum_{i=1}^{n} xi.\, ln(\lambda) - ln(\prod_{i=1}^{n} xi) \Bigg|$$

Setting its derivative with respect to λ to zero, we have:

$$\frac{d}{d\lambda} lnL(\lambda) = -n + \sum_{i=1}^{n} xi. \frac{1}{\lambda} = 0$$

giving,

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X}$$

which is the maximum likelihood estimate

Thank you