CBCS SCHEME

USN | | | | | | | | | | |

18CS751

Seventh Semester B.E. Degree Examination, Feb./Mar. 2022
Introduction to Big Data Analytics

Time: 3 hrs. Max. Marks: 100

Note: *Answer any FIVE full questions, choosing ONE full question from each module.*

Module-1

1 a. Define model. Explain various steps of modeling process. **(08 Marks)**

b. Explain different types of data with an example. **(10 Marks)**

c. Explain graphical model. **(02 Marks)**

OR

2 a. Explain Box plot and Histograms with a neat diagram. **(10 Marks)**

b. Explain stacked format and unstacked formats. **(05 Marks)**

c. Explain correlation and covariance. **(05 Marks)**

Module-2

3 a. Explain the following with an example.

i) Mutually exclusive events
ii) Conditional probability
iii) Equally likely events
iv) Subjective probability
v) Objective probability. **(10 Marks)**

b. Describe the summary measures of probability distribution. **(08 Marks)**

c. Assume 10, 20, 30 and 40 are possible values of random variable X, with probabilities 0.15, 0.25, 0.35 and 0.25. Find $P(X \leq 30)$. **(02 Marks)**

OR

4 a. What is a density function? Explain Normal Distribution. **(10 Marks)**

b. Explain the Microsoft excel functions for below probability distributions :

i) Normal Distribution
ii) Binomial Distribution
iii) Poisson Distribution
iv) Exponential Distribution **(10 Marks)**

Module-3

5 a. Construct the decision tree for the given data in Table Q5(a) :

		Outcome		
		O_1	O_2	O_3
Decision	D1	10	10	10
	D2	–10	20	30
	D3	–30	30	80

Table Q5(a)

Explain the various conventions used in Decision Tree. **(10 Marks)**

b. Explain Baye's Rules. **(05 Marks)**

c. Explain:

i) Utility function
ii) Exponential utility. **(05 Marks)**

OR

6 a. Explain different methods for selecting random samples. **(10 Marks)**
b. Explain the sources of estimation Errors. **(10 Marks)**

Module-4

7 a. Explain t distribution with respect to sampling. **(10 Marks)**
b. Give the applications of comparisons of means in Business. **(05 Marks)**
c. Explain sample size selection. **(05 Marks)**

OR

8 a. Explain how to find the significance of sample evidence from P-values. **(07 Marks)**
b. Explain Chi-square goodness of fit test for Normality. **(10 Marks)**
c. Explain different types of Errors. **(03 Marks)**

Module-5

9 a. Explain simple Linear regression using least square estimation. **(08 Marks)**
b. Give the characteristics of Multiple Regression. **(05 Marks)**
c. Explain Non-linear transformation for examining the variables. **(07 Marks)**

OR

10 a. How to analyze different sources of variation using ANOVA table. **(08 Marks)**
b. Give the guidelines for Including/Excluding variable in a Regression equation. **(05 Marks)**
c. Explain how to predict the value of the dependent variable for new observations. **(07 Marks)**

* * * * *

7th Semester B.E. Degree Examination Feb./Mar. 2022

18CS751 - Introduction to Big Data Analytics

Module - I

Q1 Define model. Explain various steps in modeling process

a) **Model:-** A model is an abstraction of a real problem. A model tries to capture the essence and key features of the problem without getting bogged down in relatively unimportant details.

Steps in modeling process

① **Define the problem:** A company doesnot develop a model unless it believes it has a problem. Therefore the modeling process really begins by identifying an underlying problem.

② **Collect & Summarize data.** All organizations keep track of various data on their operations, but these data are often not in the form an analyst requires. Therefore, an analyst has to gather exactly the right data and summarize the data appropriately.

③ **Develop a model** - The model can be a graphical model, an algebraic model, or a spreadsheet model. It must capture the important elements of the business problem in such a way that it is understable by all stakeholders involved.

④ **Verify the model** - Here analyst tries to determine whether the model developed in previous step is accurate representation of reality.

- check validity of the model for current situation
- Enter input parameters & checks the output.

(5) Select one or more suitable decisions: For any specific decisions the model indicates the amount of profit obtained, amount of cost incurred, the level of risk, etc. If the model is working correctly, then it can be used to see which decisions produce best outputs.

(6) Present the results to the organization: An analyst has to "sell" the model to management. The people in management may not be trained in quantative methods, so they are not always trusting complex models. This can be resolved in two ways:-

i) Relevant People are to be included throughout the company in the modeling process, so that everyone has an understanding of the model.

ii) Better to use spread sheet models whenever possible, if it is designed & documented properly.

(7) Implement the model & update it over time: The model developed by analyst is accepted by management & is implemented company wide. The model has to be updated overtime, either because of changing conditions or because the company sees more potential uses for the model.

Q1 (b) Explain different types of data with an example.

Types of Data

1) Numerical: A variable is numerical if meaningful arithmetic can be performed on it.
E.g. Age, Number of students, etc.

2) Categorical: A variable is categorical if it is represented in the form of text, not in numbers.
E.g. gender, city

3) ordinal: A categorical variable is ordinal if there is a natural ordering of its possible categories.
E.g. opinion variable can have 3 categories - agree, neutral, disagree

(3) (4) Nominal: If there is no natural ordering of categor then it is nominal variable.

Eg. gender, city.

(5) Dummy variable: A dummy variable is 0-1 coded variab for a specific category.

Eg Gender can be coded as 1 - Male & 0 - Female

Person	Age	Gender	State	Salary
1	35	1	Minnesota	65,000
2	61	0	Texas	62,000

(6) Binned variable: A binned (discretized) variable corresponds to a numerical variable that has been categorized into discrete categories. These categories are called bins.

Eg. Age can be categorized as
- Young (34 years or less)
- middle-aged (35 - 59 years)
- Elderly (60 years or older)

Person	Age	Gender	Children	Salary
1	middle-aged	1	1	.5
2	Elderly	0	0	1

(7) Discrete variable: A numerical variable is discrete if it results from count.

e.g. No. of students in college.

(8) Continuous variable: A continuous variable is the result of an essentially continuous measurement.

e.g. weight or height.

(9) Cross-sectional data: Crossectional data are data on a cross-section of a population at a distinct point in time.

e.g.

Person	Age	Gender	State	Children	Salary	Opinion
1	35	M	M	1	65,000	5
2	61	F	T	2	62,000	1

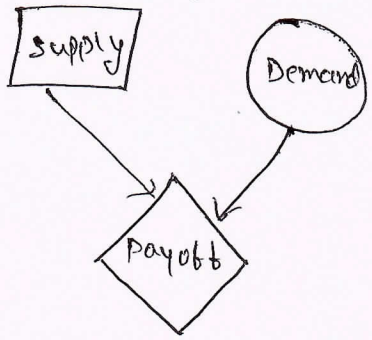
(10) Time series data: It is the data collected over time.

Quarter	Revenue
Q1-2015	1,026
Q2-2015	1,056

(8 c) graphical model.

④ Graphical models are probably the most intuitive and least quantitative type of model. They attempt to portray graphically how different elements of a problem are related.

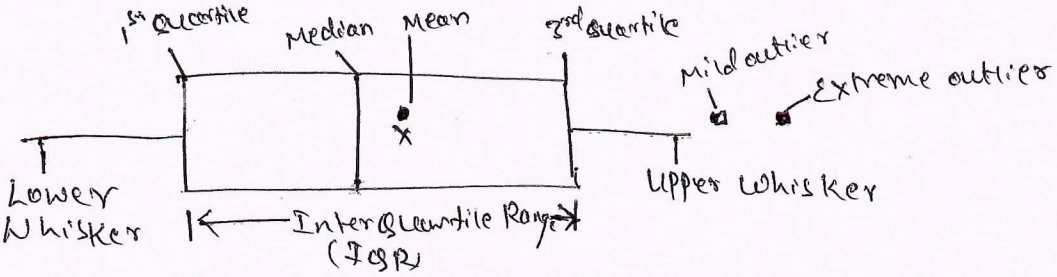
E.g. Influence Diagram:



Q2 a) Explain Box plot & Histogram with a neat diagram.

Box plots

A Box plot shows the distribution of a variable.

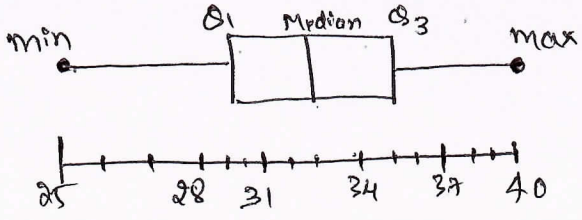


- The box itself extends, left to right from 1st quartile to 3rd quartile.
- The line inside the box is positioned at the median, and the • inside the box is positioned at the mean.
- The lines (whiskers) coming out either side of the box extend to 1.5 IQRs from the quartiles.
- More distant values, called outliers are denoted separately with small squares.

E.g.

Consider the weights of 10 boxes: 23, 28, 29, 29, 30, 34, 35, 35, 37, 38.

Median: 32; 1st quartile: 29; 3rd quartile: 35, min: 25, max: 38



- A histogram shows the distribution of a numerical variable.
- It is based on binning the variable
- It is great for showing the shape of a distribution

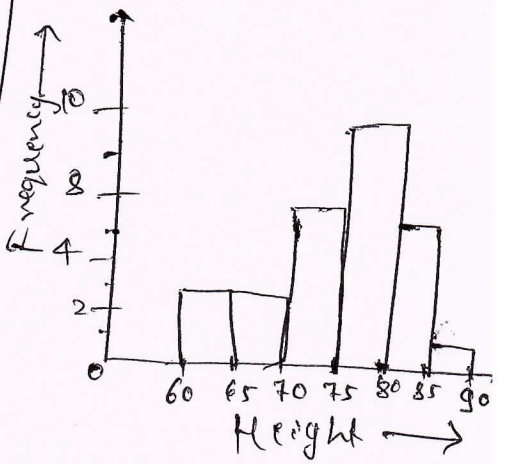
Example:

Rakesh owns a garden with 30 black Cherry trees. Each tree of a different height. the height of trees: 61, 63, 64, 66, 69, 71, 71.5, 72, 72.5, 73, 73.5, 74, 74.5, 76, 76.2, 76.5, 77, 77.5, 78, 79, 79.2, 80, 81, 82, 83, 84, 85, 87.

We can group the data as follows:

Height	No. of trees
60-65	3
66-70	3
71-75	8
76-80	10
81-85	5
86-90	1

Histogram



Q2 (b) Explain Stacked Format & Unstacked Formats.

Stacked data: The data are stacked if there are two "long" variables. There are one or more long numerical variables & another long variable that specifies which category each observation is in.

E.g.

Gender	Salary
Male	81600
Female	61600
Female	64300
Female	71900

(b) Unstacked data: The data are not stacked.

Female Salary	Male Salary
61600	81600
64300	76300
71900	60900
68200	60200

Q2 (c) Explain Correlation & Covariance

* Covariance & correlation measures the strength and direction of a linear relationship between two numerical variables.
* The two numerical variables X & Y must be paired variables.

Covariance:

Let, x_i & y_i – paired values for observation ;
n – no of observations

Covariance between X & y is given by

$$Covar(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Where,

$\bar{x}$ => Mean of X ; $\bar{y}$ => Mean of y

It is essentially an average of products of deviations from means.

Disadvantage:

Covariance is sensitive to the units in which X & y are measured.

Correlation:

- It is a unitless quantity that is unaffected by the measurement scale. It is defined by

$$Correl(X,Y) = \frac{Covar(X,Y)}{Stddev(X) \times Stdev(Y)}$$

Where stdev(X) & Stdev(Y) – standard deviations of X & Y.

* It's value is always between −1 & +1.
* correlation value '0' indicates no relationship between X & Y

(7)

Module-2

Q3

(a) Explain the following with an example:

i) Mutually exclusive events
ii) Conditional probability
iii) Equally likely events
iv) Subjective Probability
v) Objective Probability.

i) Mutually exclusive events

* The events are said to be mutually exclusive if atmost one of them can occur.
* If one of them occurs, then none of the others can occur

e.g. right & left hand turns, even & odd numbers on a die, winning & losing a game, or running & walking.

* Mutually exclusive events are also called exhaustive events.

Let $A_1, \ldots A_n$ be any 'n' events. the probability that atleast one of these events will occur is:

$$P(\text{atleast one of } A_1 \text{ through } A_n) = P(A_1) + P(A_2) + P(A_3) + \ldots + P(A$$

ii) Conditional Probability.

Let A & B be any events with probabilities P(A) & P(B) If it's said that B has occured, then the probability of A might change. the new probability A is called conditional probability of A given B. which is denoted by $P(A|B)$

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Eg. When a dice is rolled, what is the probability that the outcome is 5 given that it's odd.

(3)

iii) Equally likely events

* When the outcome of the events are equally likely, such events are called equally likely events.

e.g. flipping coins, throwing dice, drawing balls from urns

* Many probabilities, particularly in games of chance, can be calculated by using an equally likely argument.

iv) Subjective probability

* Subjective probability cannot be estimated from long-run proportions.

e.g. you think you have an 80% chance of your best friend calling today, because his/her car broke down yesterday.

* Subjective probability is where you use your opinion to find probabilities.

v) Objective probability

* Objective probabilities are those that can be estimated from long-run proportions.

* In objective probability, a person's opinion is not needed.

e.g.
* It is associated with random events like die rolls, choosing bingo balls, number coming up in lottery.

Q3 (b) Describe the summary measures of probability distribution

Summary measures of probability distribution are:

i) mean, ii) variance iii) Standard deviation

Mean:

* It is also called the expected value of X & is denoted by E[X]

* The mean is the weighted sum of the possible values, weighted by their probabilities.

(9)

$$\mu = E(x) = \sum_{i=1}^{k} V_i P(V_i)$$

V_i - values

$P(V_i)$ - Probability of values.

The mean indicates the "center" of the probability distribu

Variance:

. The variance, denoted by σ^2 or $var(x)$, is a weigh
Sum of squared deviations of the possible values from th
mean, where the weights are again the probabilities.

$$\sigma^2 = Var(x) = \sum_{i=1}^{k} (V_i - E(x))^2 P(V_i)$$

OR

$$\sigma^2 = \sum_{i=1}^{k} V_i^2 P(V_i) - \mu^2$$

Standard deviation

The standard deviation is denoted, σ or $stdev(x)$.
It is the square root of the variance.

$$\sigma = stdev(x) = \sqrt{Var(x)}$$

Q3

c) Assume 10, 20, 30 & 40 are possible values of random variable X with probabilities 0.15, 0.25, 0.35 & 0.25.
Find $P(X \leq 30)$

Solution

$$P(X \leq 30) = P(x=10) + P(x=20) + P(x=30)$$
$$= 0.15 + 0.25 + 0.35$$
$$= 0.75$$

Q4 a) What is a density function? Explain Normal Distribution (10)

A density function is denoted by $f(x)$, specifies the probability distribution of a continuous random variable x. The higher $f(x)$ is, the more likely x is. The total area between the graph of $f(x)$ and the horizontal axis, which represents the total probability is equal to 1. $f(x)$ is non-negative for all possible values of X.

Normal Distribution

* It is a continuous distribution - it is the basis of the familiar symmetric bell-shaped curve.
* Any particular normal distribution is specified by its mean & standard deviation.
* By changing the mean, the normal curve shifts to the right or left.
* By changing the standard deviation, the curve becomes more or less spread out.

Normal density

* The normal distribution is a continuous distribution with possible values ranging over the entire number line - $(-\infty$ to $+\infty)$

$$f(x) = \frac{1}{2\sqrt{\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad \text{for } -\infty < x < +\infty$$

μ & σ => mean & standard deviation of the distribution.

* There are infinitely many normal distributions. one for each pair μ & σ.

(ii) The standard normal distribution has mean 0 & standard deviation 1. It is denoted by $N(0,1)$

. It is also referred as Z-distribution

* Suppose the random variable X is normally distributed with mean μ & standard deviation σ. The random variable Z is given by.

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

The reason for standardizing is to measure variables with different means & or standard deviations on a single scale.

Z-value Interpretation

. It is the number of standard deviations to the right or the left of the mean.

If Z is positive, the original value is to the right of the mean; if Z is negative, the original score is to the left of the mean.

Q 4 (b) Explain the microsoft Excel functions for below probability distributions:

i) Normal Distribution ii) Binomial distribution
iii) Poisson Distribution iv) Exponential Distribution.

i) Normal Distribution

Two types of calculations are made with normal distributions:-

i) Finding Probabilities ii) Finding percentiles.

The functions used for normal probability calculations are NORMDIST & NORMSDIST.

(12) NORM DIST - applies to any Normal Distribution
NORMSDIST - applies only to N(0,1)

Syntax

=NORMDIST(x, μ, σ, 1)

and

=NORMSDIST(x)

where,

x => Number supplied.

μ => Mean

σ => Standard deviation

Percentile calculations that take probability & return a value are called Inverse calculations. Excel functions for these are NORMINV & NORMSINV.

Syntax

=NORMINV(P, μ, σ)

and

=NORMSINV(p)

P => probability

μ => mean

σ => standard deviation.

11) Binomial Distribution

We calculate binomal probabilities in Excel using:

=BINOMDIST(k, n, P, cum)

where,

n => number of trials

P => probability of success

K => No. of successes that you specify.

(um => It is either 0 or 1. It is '1' if you want the probability of less than or equal to k successes & '0' if you want probability of exactly K successes.

(13)

iii) POISSON Distribution

We calculate Poisson probabilities in Excel using

=POISSON(K, λ, cum)

Where,

cum ⇒ It is 0 or 1. It is '0' for P(x=k) & '1' for P(x≤k)

λ ⇒ rate

iv) Exponential Distribution

The Excel function for Exponential Distribution is

=EXPONDIST(x, λ, 1)

Where, x ⇒ given value

λ ⇒ Parameter

Module-3

Q5 (a) Construct a decision tree for the given data

	O_1	O_2	O_3
D1	10	10	10
D2	-10	20	30
D3	-30	30	80

Explain various conventions used in Decision tree.

Solution

Assume the probabilities of 3 outcomes as 0.3, 0.5 & 0.2 if decision D_2 is made & 0.5, 0.2, & 0.3 if decision D_3 is made

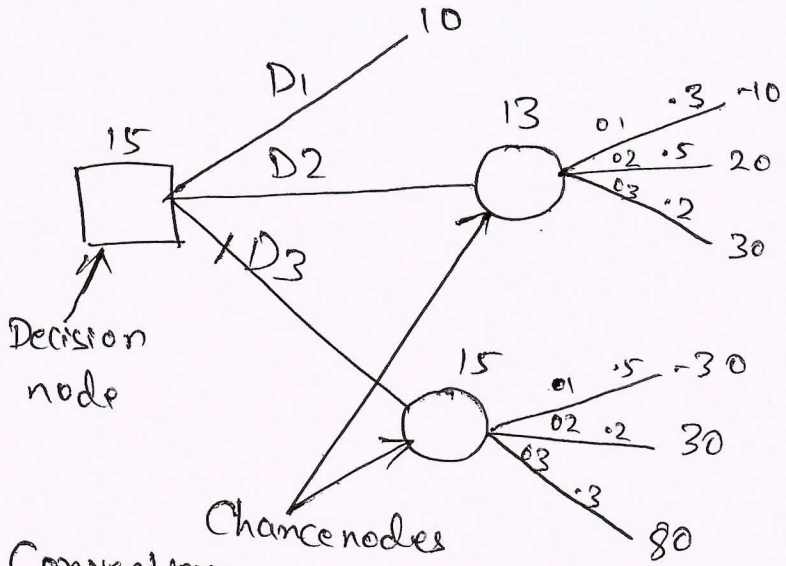
EMV for D1 = 10

EMV for D2 = -10(0.3) + 20(0.5) + 30(0.2) = 13

EMV for D3 = -30(0.5) + 30(0.2) + 80(0.3) = 15

The resulting decision tree is

(19)



Conventions

* Decision trees are composed of nodes (circles, squares & Triangles) and branches (lines).
* A decision node (a square) represents a time when the decision maker makes a decision.
* A chance node (a circle) represents a time when the result of an uncertain outcome becomes known.
* An end node indicates that the problem is completed.
* Time proceeds from left to right.
* Branches leading out of a decision node represent the possible decisions
* Probabilities are listed on chance branches.
* Monetary values are shown to the right of the end nodes.
* EMV's are calculated through a "Rolding-back" process.

Q5 (b) Explain Baye's Rule

In multistage decision tree, all chance branches toward the right of the tree are conditional on outcomes that have occured earlier to their left. Hence probabilities on these branches are of the form $P(A|B)$. In such cases, Baye's rule must be used to obtain the probabilities needed on the tree.

(15) * Let $A_1 - - - A_n$ be any outcomes. The probabilities of $A_1 - - - A_n$ be $P(A_1) - - P(A_n)$, called prior probabilities

* There are several information outcomes we might observe. Some of the outcome say B. The probabilities of B given any of A's will occur is given as

$$P(B|A_1), P(B|A_2) - - - - P(B|A_n)$$

called as likelihoods.

* An information outcome might also influence probabilities of A's. i.e. $P(A_i|B)$, which is called posterior probability of A_i.

* Bayes states that the posterior probability can be calculated using

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + - - - + P(B|A_n)P(A_n)}$$

using Law of total probability.

$$P(B) = P(B|A_1)P(A_1) + - - - + P(B|A_n)P(A_n)$$

The Bayes rule reduces to.

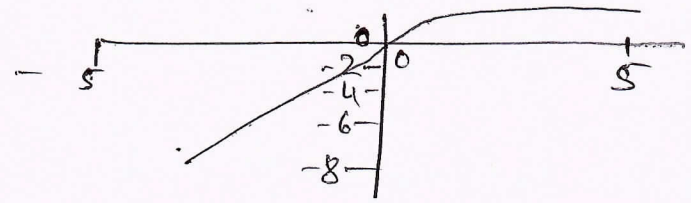
$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

Q) Explain: i) Utility function ii) Exponential utility.

Utility Function

- It is a mathematical function that transforms monetary values - payoffs & costs - into utility values.
- An individual utility function specifies the individual's preferences for various monetary payoffs & costs & it encodes the

(16) the individual attitudes toward risk.



ii) Exponential utility

An exponential utility function has only one adjustable numerical parameter called risk tolerance.

* Exponential utility functions do not capture all types of attitudes toward risk

* It is given:

$$U(x) = 1 - e^{-x/R}$$

x – Monetary value

$U(x)$ – utility of value

$R > 0$ – risk tolerance.

The risk tolerance for an exponential utility function is a single number that specifies an individual's aversion to risk. The higher the risk tolerance, the less risk averse the individual is.

Q 6 (a) Explain different methods for selecting random samples.

Various methods for selecting samples are:

1) Simple random sampling
2) Systematic sampling
3) Stratified sampling
4) Cluster sampling
5) Multistage sampling schemes.

1) Simple random Sampling

*A simple random sample of size 'n' is one where each possible sample of size 'n' has the same chance of being chosen.

(17.)

e.g. Suppose $N=5$, & five members of population are a, b, c, d & e. Also the sample size is $n=2$.

* Then possible samples are (a,b), (a,c), (a,d), (a,e), (b,c), (b,d), (b,e), (c,d), & (d,e).
* Then a simple random sample of size $n=2$ has the property that each of these 10 possible samples has th
* Sample probability, 1/10 of being chosen.
* Any member has the probability n/N of being chosen in a simple random sample.

(2) Systematic Sampling

* A systematic sample provides a convenient way to Choose the Sample.

* In general, one of the first k members is selected randomly & then every k^{th} member after this one is selected. The value 'k' is called sampling interval & equals the ration N/n. $N \Rightarrow$ population size & $n \Rightarrow$ sample si

- Suppose 250 names are to be selected from 55,000 names from a telephone book
  - Divide the population size by sample size:
    $55,000 / 250 = 220$
  - Use a random mechanism to choose a number between 1 & 220.
  - Suppose this no. is 131. Then choose the 131^{st} name and every 220^{th} name thereafter i.e. 131, 351, 571, etc.

* The key in systematic Sampling is the relationship between ordering of sampling units in the frame & the purpose of stu

(3) Stratified Sampling

- In Stratified Sampling, the population is divided into relatively homogeneous subsets called strata & then random samples are taken from each stratum.
- Separate estimates can be obtained within each stratum.

(c) * Stratified samples are typically chosen because they provide more accurate estimates of population parameters for given sampling cost

* Accuracy of the resulting population estimates can be increased by using appropriately defined strata.

④ Cluster Sampling

- In Cluster sampling, the population is separated into clusters and then a random sample of the clusters is selected.

- Cluster sampling - Process
  - Define the sampling units as clusters.
  - Then a simple random sample of clusters can be chosen.
  - Once the clusters are selected, it is typical to sample all of the population members in each selected cluster.

⑤ Multistage sampling schemes

* The cluster sampling, where a sample of clusters is chosen & then all of the sampling units within each chosen clusters are taken, is called sing-stage sampling scheme.

* Real applications are more complex than this, resulting in multistage sampling schemes.

Q 6
(b) Explain sources of estimation error.

There are two basic sources of errors that can occur when you sample randomly from a population:

1) Sampling Error.

2) Non Sampling Error.

Sampling Error: It results from "unlucky" samples. It is the inevitable result of basing an inference on a random sample rather than on the entire population.

Non Sampling Error: It can arise from variety of reasons.

i) Non Response bias: This occurs when a portion of the sample fails to respond to the survey. Anyone who has ever conducted a questionnaire, knows that the percentage of non-respondents can be quite large. Unless we are able to persuade the non-respondents to respond through a follow-up - we must guess at the amount of nonresponse bias.

ii) Nontruthful response: this results in problem when there are sensitive questions in a questionnaire. For e.g. if Question "Do you regularly use cocaine?" are asked, most people will answer "no", regardless of whether the true answer is "yes" or "no". We can get such sensitive information through 'randomized response' technique.

iii) Measurement error: This occurs when the responses to the questions do not reflect what the investigator had in mind. It might result from poorly worded questions, questions the respondents don't fully understand, questions that require respondents to supply information they don't have & so on.

iv) Voluntary response bias: This occurs when the subset of people who respond to a survey differ in some important respect from all potential respondents.

Non Sampling errors cannot be measured with probability theory. It can be controlled only by using appropriate Sampling procedure & designing good survey instruments.

MODULE 4

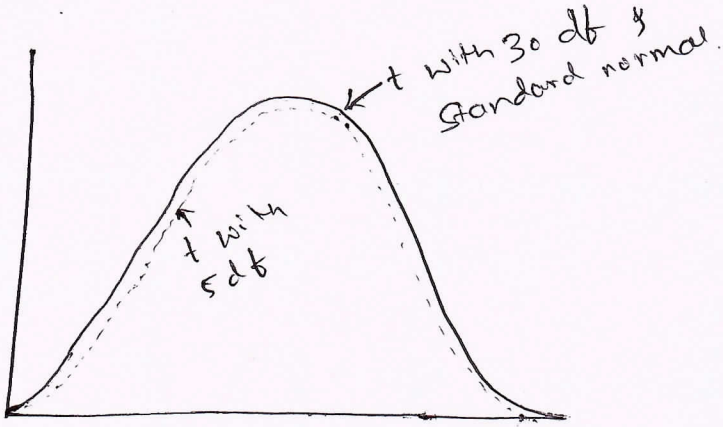
Q7 a) Explain 't' distribution with respect to sampling.

* The 't' distribution is similar to standard normal distribution. It is bell-shaped & centered at 0

* The degrees of freedom is a numerical parameter of the 't' distribution that defines the precise shape of the distribution

* The only difference between normal distribution & t distribution is that it is slightly more spread out & this increase in spread is greater for small degrees of freedom. When 'n' is large, the 't' distribution & standard normal distribution are practically indistinguishable.

* With 5 degrees of freedom, it is possible to see the increased spread in the 't' distribution. With 30 degrees of freedom, the t & standard normal curves are practically the same.



t distribution with $n-1$ degrees of freedom is given by:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$\bar{x} \Rightarrow$ Sample mean

$\mu \Rightarrow$ Population mean

$s \Rightarrow$ Sample estimate

$n \Rightarrow$ Sample size.

(b) Some applications of comparisons of means in Busine

(21) (1) Men & Women shop at a retail clothing store. The manager would like to know how much more (or less), on average, a woman spends on a typical purchase occasion than a man.

(2) Two airline companies fly similar routes. A consumer organization would like to check how much the average delay differs between the two airlines, where delay is defined as the actual arrive time at the destination minus the scheduled arrival time.

(3) A supermarket chain mails coupons for various products to a randomly selected subset of its customers in a particular city. Its other customers in this city receive no such coupons. The chain would like to check how much the avg. amount spent on these products differs between the two sets of customers over next couple of months.

(4) A computer company has a customer service center that responds to customer's questions & complaints. The center employs two types of people:

1. Formal course - Little Experience
2. No formal course - Large Experience.

The company would like to know how these two types of employees differ with respect to the average no. of customer complaints of poor service in last six months.

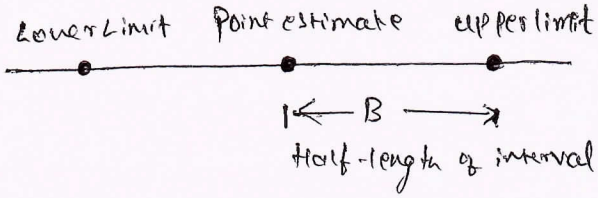
(5) A consulting company hires business students directly out of undergraduate school. The new hires all take a problem solving test. They then go through an intensive 3-month training program, after which they take another similar problem solving test. The company wants to know how much the avg. ~~salary~~ test score improves after training program.

4 T (c)

(22) Explain sample size selection

* The data in the sample directly affect the length of a confidence interval through their sample standard deviation.

* Because each confidence interval is a point estimate plus or minus some quantity, it is called half-length of the interval.



1) Sample size selection for estimation of mean

$$n = \left(\frac{z\text{-multiple} \times \sigma_{est}}{B} \right)^2$$

σ_{est} => Estimated standard deviation.

2) Sample size selection for Estimation of other parameters

i) For proportion p.

$$n = \left(\frac{z\text{multiple}}{B} \right)^2 P_{est}(1 - P_{est})$$

P_{est} => estimate of the population proportion.

ii) For difference between means

$$n = 2 \left(\frac{z\text{-multiple} \times \sigma_{est}}{B} \right)^2$$

iii) For difference between proportions

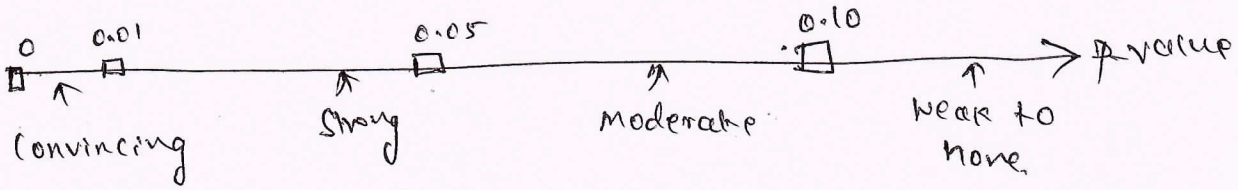
$$n = \left(\frac{z\text{-multiple}}{B} \right)^2 [P_{1est}(1 - P_{1est}) + P_{2est}(1 - P_{2est})]$$

(28) a) Explain how to find the significance of sample evidence from p-values.

- The p-value of a sample is the probability of seeing a sample with atleast as much evidence in favor of the alternative hypothesis as the sample actually observed.

The smaller the P-value, the more evidence there is in favor of the alternative hypothesis.

- P value avoids the use of significance level α & simply report how significant the sample evidence is.
- In general smaller P-value indicate more evidence in support of the alternative hypothesis. If P value is sufficiently small, then almost any analyst will conclude that rejecting the null hypothesis is the more reasonable decision.



- A p-value less than 0.01 is regarded as convincing evidence that the alternate hypothesis is true.
- A P-value between 0.01 & 0.05 indicates strong evidence in favor of the alternative hypothesis
- The interval between 0.05 & 0.1 is a gray area.
- The P-values larger than 0.10 are generally interpreted as weak evidence in support of alternative.
- * Sample evidence is statistically significant at the α level only if the p-value is less than α.

Q8 (b) Explain Chi-square goodness of fit test for Normality.

- A histogram of sample data is compared to the expected bell-shaped histogram that would be observed if the data were normally distributed with the same mean and standard deviation as in the sample.
- If the two histograms are sufficiently similar, the null

the null hypothesis of normality is accepted, otherwise it can be rejected.

- The test is based on a numerical measure of the difference between the two histograms. Let C be the number of categories in the histogram & let O_i be the observed number of observations in category i. Let E_i be the expected number of observations in category i if the population were normal with the same mean & standard deviation as in the sample. Then goodness-of measure is used as a test-statistic.

$$\chi^2\text{-value} = \sum_{i=1}^{C} (O_i - E_i)^2 / E_i$$

If the null hypothesis of normality is true, this test statistic has a Chi-square distribution with C-3 degrees of freedom.

Comments on Chi-square goodness of fit-test

i) the test does depend on which (& how many) bins we use for histogram.

ii) The test is not very effective unless the sample size is large, say atleast 80 or 100.

iii) The test tends to be too sensitive if the sample size is really large. In this case any little "bump" on the observed histogram is likely to lead to a conclusion of non-normality.

This test is often unable to distinguish between normal and non-normal distributions, & hence it often fails to reject the null hypothesis of normality when it should be rejected.

(25) Different types of errors

Type I Error: When we incorrectly reject a null hypothesis that is true.

Type II Error: When we incorrectly accept a null hypothesis that is false.

		Truth	
		H_0 is true	H_a is true
Decision	Reject H_0	~~Reject H_0~~ Type I Error	No error
	Donot Reject H_0	No error	Type II error

Module-5

Q9 a) Explain simple linear regression using least square estimation.

In simple linear regression, there is a single explanatory variable. We do so by fitting a straight line through the scatterplot of the dependent variable y versus the explanatory variable x & then basing the answers to the questions on the fitted line.

Least Square Estimation

Fundamental Equation for regression

Observed value = Fitted value + Residual.

The least squared line is the line that minimizes the sum of the squared residuals.

Any straight line can be quoted as:

$$y = a + bx$$

a ⇒ y-intercept of the line

b ⇒ slope of the line

(26) Slope in Simple Linear Regression

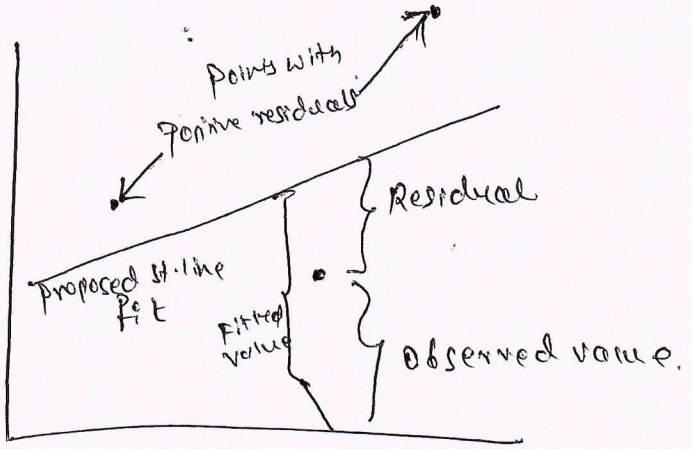
$$b = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} = r_{xy} \frac{s_y}{s_x}$$

$\bar{x}$ => Mean of x

$\bar{y}$ => Mean of y

Intercept in Simple Linear Regression

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



A fitted value is the predicted value of the dependent variable. Graphically it is the height of the line above a given explanatory value. The corresponding Residual is the difference between the actual & fitted values of the dependent variable.

Q9 b) Give the Characteristics Of Multiple Regression.

① If there are two explanatory variables, we are fitting a plane to the data in 3-D space.

② The regression equation is still estimated by the least squares method.

③ Simple regression is a special case of multiple regression.

④ There is a slope term for each explanatory variable in the equation.

(27)

(6)

of estimate and R^2 summary measures are almost same as simple regression.

Many types of explanatory variables can be included in the regression equation.

Q 9 (c)

Explain Non-linear transformation for examining the variables.

The general linear regression equation is

$$\text{Predicted } Y = a + b_1X_1 + b_2X_2 + \cdots +$$

We include non linear transformations in a regression equation because of economic considerations or curvature detected in scatterplots.

We can transform dependent variable Y or explanatory variable X_S. In such cases there are a few non linear transformations that are typically used like natural logorithm, Square root, etc. The purpose of each of these is to 'straighten out" the points in a scatter plot.

Constant Elasticity Relationships:

$$\text{Predicted } Y = aX_1^{b_1}X_2^{b_2}\cdots X_k^{b_k}$$

One property of this type of relationship is that the effect of one-unit change in any X on Y depends on the levels of other Xs, which is not true for additive relationships

$$\text{Predicted } Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

We can use linear regression to estimate the non-linear relationship by taking natural logorithms of all variables.

$$\text{Predicted } \log(Y) = \text{Log}(a) + b_1\text{Log}(X_1) + \cdots + b_k\text{Log}(X_k).$$

Q10 a) How to analyze different sources of variation using ANOVA table?

MSE is the square of the standard error of estimate i.e.

$$MSE = S_e^2 \quad (\because S_e \Rightarrow \text{standard error})$$

The Ratio of MSR to MSE is F-Ratio.

$$F\text{-ratio} = \frac{MSR}{MSE}$$

- When null hypothesis of no explanatory power is true, this F-ratio has F-distribution with K & n-k-1 degrees of freed
- The F-ratio has an associated P-value that allows to run test easily.
- Reject the null hypothesis, if the F-value in the ANOVA table is large and the corresponding P-value is small.
- An ANOVA table analyzes different sources of variation. In case of regression, the variation in question is the variation of the dependent variable y. The total variation of this variable is sum of squared deviations about mean & is labeled as SST (Sum of Squares total)

$$SST = \Sigma (y_i - \bar{y})^2$$

The ANOVA table splits this total variation into two parts, the part explained by regression equation & the part left unexplai The unexplained part is SSE (Sum of Squared Errors):

$$SSE = \Sigma e_i^2 = \Sigma (y_i - \hat{y}_i)^2$$

The F test is a formal procedure for testing whether t explained variation is large compared to unexplained variatio Each of these have an associated degrees of freedom (df).

For explained variation $df = k$.
For unexplained variation $df = n-k-1$.

$$MSR = \frac{SSR}{k}, \quad MSE = \frac{SSE}{n-k-1}$$

29

Q 10 b) Give the guidelines for Including / Excluding variable in a Regression equation.

Guide lines:

(1) Look at a variable's t-value & its associated p-value. If p value is above significance level (0.05) then, this variable is a candidate for exclusion.

(2) Check t-value of a variable is less than 1 or greater than 1 in magnitude. If less than 1, then Se will decrease if this variable is excluded from equation. If greater than 1, the opposite will occur.

(3) Look at t-values and p-values, rather than correlations, when making include / exclude decisions.

(4) When there is a group of variables that are in some sense logically related, it is good idea to include all of them or exclude all of them.

(5) Use economic & / or physical theory to decide whether to include / exclude variables, & put less reliance on t-values and / or p-values.

Q 10 c) Explain how to predict the value of the dependent variable for new observations.

Regression can be used to predict Y for a single observation or for many observations, all with same 'X' values.

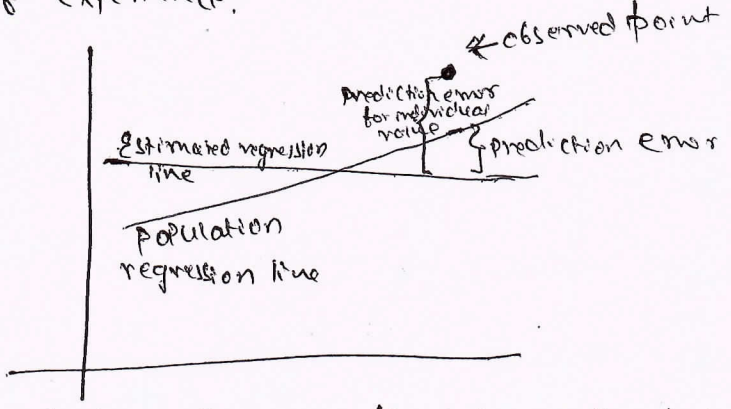
Two types of prediction problems in regression:

i) To predict the value of the dependent variable for one or more individual members of the population.

ii) To predict the mean of the dependent variable for all members of the population with certain values of the explanatory variables.

(30)

Let the dependent variable is salary & single explanatory variable is years of experience with the company. Let's suppose we want to predict either Salary for a employee with 10 years of experience or mean salary of all employees with 10 years of experience.



For each prediction problem the point prediction is the value above 10 on the estimated regression line.

It is more difficult to predict for extreme Xs than for Xs close to the mean.

Making predictions & Estimating Accuracy:

Assume single explanatory variable X. The trial value of X, be X_0 & predict the value of single Y or mean of all Ys when $X = X_0$. The point prediction is found by substituting into right side of estimated regression equation.

To measure accuracy of point predictions, calculate standard error for each prediction.

The standard error for individual prediction problem S_{ind} is given by:

$$S_{ind} = S_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \approx S_e$$

For prediction of mean, the standard error is S_{mean}

$$S_{mean} = S_e \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \approx S_e / \sqrt{n}$$

17/3/22

Dean, Academics.

(Prof. Rajnesh Kulkarni)