MAR GREGORIOS COLLEGE OF ARTS & SCIENCE

Block No.8, College Road, Mogappair West, Chennai – 37

Affiliated to the University of Madras
Approved by the Government of Tamil Nadu
An ISO 9001:2015 Certified Institution

PG DEPARTMENT OF COMMERCE

SUBJECT NAME: QUANTITATIVE TECHNIQUES FOR BUSINESS DECISION

SUBJECT CODE: KDA2B

SEMESTER: II

PREPARED BY: PROF. B. HARISWARAN

Quantitative Techniques for Business Decisions Syllabus

Objective: To provide knowledge in quantitative methods and applications and to offer expertise in quantitative analysis

Unit I Theory of probability -probability rules -Baye's theorem -Probability distribution -Binomial, Poisson and Normal. Statistical decision theory -Decision environment -decision making under certainty and uncertainty and risk conditions -EMV, EOL and marginal analysis -value of perfect information -decision tree analysis

Unit II Sampling-Meaning of random sample -sampling methods -sampling error and standard error relationship between sample size and standard error Sampling distribution -characteristics- central limit theorem -estimating population parameters - point and interval estimates -estimating proportion, percentage and mean of population from large sample and small sample

Unit III Testing hypothesis -testing of proportions and means -large samples -small samples -one tailed and two tailed tests -testing differences between mean and proportions -errors in hypothesis testing -chi square distribution -characteristics -applications -test of independence and tests of goodness of fit -inferences -F distribution- testing of population variance- analysis of variance -one way and two way

Unit IV Correlation and regression -Simple, partial and multiple correlation -simple, partial and multiple regressions -estimation using regression line -standard error of estimate -testing significance of correlation and regression coefficients -interpreting correlation -explained variation and unexplained variation - coefficient of determination- multivariate analysis -factor, cluster and discriminant analysis

Unit V Linear programming graphic and simplex models -maximization and minimization -transportation -Assignment

UNIT-1
Theory of probability

Probability theory

A branch of mathematics concerned with the analysis of random phenomena. The outcome of a random event cannot be determined before it occurs, but it may be any one of several possible outcomes. The actual outcome is considered to be determined by change.

The probability of an event is always between 0 and 1.

• When P(A) = 0, the event is impossible.

• When P(A) = 1, the event is certain to happen.

• The total probability of happening an event and not happening an event is 1. i.e. $P(A) + P(A') = 1$.

Addition Theorem of Probability Statement:

If A and B any two events then, P(AUB) = P(A) + P(B) – P(A∩B) Proof: Two events A and B are shown in Venn-diagram. From the figure it is clear that the event A consists of two mutually exclusive events A∩B' and A∩B.

Conditional probability

This probability of occurrence of an event B when it is known that A has occurred is known as conditional probability of B under the condition that A has occurred and it is denotes by P (B/A).
• Therefore, P (B/A) = 10/60

• Thus the probability of happening an event B, when A has happened is defined as conditional probability of B under A and it is denoted by P (B/A)

Multiplication Theorem Statement:

For two events A and B prove that P(A ∩ B) = P(A) . P (B/A) OR P(A ∩ B) = P(B) . P (A/B) Proof: Suppose the sample space contains n sample points of which m are favorable to the occurrence of A and m1 are favorable to the occurrence of A ∩ B.

• If A and B are independent events, then P(A/B) = P(A) Therefore P(A) . P(B) = P (A ∩ B) This is known as the multiplication theorem.

Probability Rules

There are three main rules associated with basic probability: the addition rule, the multiplication rule, and the complement rule. You can think of the complement rule as the 'subtraction rule' if it helps you to remember it.

1.) The Addition Rule: *P(A or B) = P(A) + P(B) - P(A and B)*

If A and B are **mutually exclusive events**, or those that cannot occur together, then the third term is 0, and the rule reduces to *P(A or B) = P(A) + P(B)*. For example, you can't flip a coin and have it come up both heads and tails on one toss.

2.) The Multiplication Rule: *P(A and B) = P(A) * P(B|A) or P(B) * P(A|B)*

If A and B are **independent events**, we can reduce the formula to *P(A and B) = P(A) * P(B)*. The term independent refers to any event whose outcome is not affected by the outcome of another event. For instance, consider the second of two coin flips, which still has a .50 (50%) probability of landing heads, regardless of what came up on the first flip. What is the probability that, during the two coin flips, you come up with tails on the first flip and heads on the second flip?

Let's perform the calculations: *P = P(tails) * P(heads) = (0.5) * (0.5) = 0.25*

3.) The Complement Rule: *P(not A) = 1 - P(A)*

Do you see why the complement rule can also be thought of as the subtraction rule? This rule builds upon the mutually exclusive nature of *P(A)* and *P(not A)*. These two events can never occur together, but one of them always has to occur. Therefore *P(A) + P(not A)* = 1. For example, if the weatherman says there is a 0.3 chance of rain tomorrow, what are the chances of no rain?

Let's do the math: *P(no rain) = 1 - P(rain)* = 1 - 0.3 = 0.7

Baye's theorem in probability

In Probability, Baye's theorem is a mathematical formula, which is used to determine the conditional probability of the given event. Conditional probability is defined as the likelihood that an event will occur, based on the occurrence of a previous outcome.

The formula for Baye's theorem

The formula for Bayes theorem is:

$P(A|B) = [P(B|A). P(A)]/P(B)$

Where P(A) and P(B) are the probabilities of events A and B.

P(A|B) is the probability of event A given B
P(B|A) is the probability of event B given A.

Example:

A bag I contain 4 white and 6 black balls while another Bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags, and it is found to be black. Find the probability that it was drawn from Bag I.

Solution:

Let E1 be the event of choosing the bag I, E2 the event of choosing the bag II, and A be the event of drawing a black ball.
Then,P(E1) = P(E2) = 12
Also,P(A|E1) = P(drawing a black ball from Bag I) = 610 = 35
P(A|E2) = P(drawing a black ball from Bag II) = 37
By using Bayes' theorem, the probability of drawing a black ball from bag I out of two bags,

P(E1|A) = P(E1)P(A|E1)P(E1)P(A | E1)+P(E2)P(A|E2)
=12 × 3512 × 35 + 12 × 37 = 712

Probability distribution

The probability distribution gives the possibility of each outcome of a random experiment. It provides the probabilities of different possible occurrences.

Types of Probability Distribution

There are two types of probability distribution which are used for different purposes and various types of the data generation process.

1. Normal or Cumulative Probability Distribution
2. Binomial or Discrete Probability Distribution

Normal Distribution Examples

Since the normal distribution statistics estimates many natural events so well, it has evolved into a standard of recommendation for many probability queries. Some of the examples are:

- Height of the Population of the world

- Rolling a dice (once or multiple times)
- To judge Intelligent Quotient Level of children in this competitive world
- Tossing a coin
- Income distribution in countries economy among poor and rich
- The sizes of females shoes
- Weight of newly born babies range
- Average report of Students based on their performance

Binomial Distribution Examples

As we already know, binomial distribution gives the possibility of a different set of outcomes. In the real-life, the concept is used for:

- To find the number of used and unused materials while manufacturing a product.
- To take a survey of positive and negative feedback from the people for anything.
- To check if a particular channel is watched by how many viewers by calculating the survey of YES/NO.
- The number of men and women working in a company.
- To count the votes for a candidate in an election and many more.

Poisson Probability Distribution

The Poisson probability distribution is a discrete probability distribution that represents the probability of a given number of events happening in a fixed time or space if these cases occur with a known steady rate and individually of the time since the last event. It was titled after French mathematician Siméon Denis Poisson. The Poisson distribution can also be practiced for the number of events happening in other particularized intervals such as distance, area or volume. Some of the real-life examples are:

- A number of patients arriving at a clinic between 10 to 11 AM.
- The number of emails received by a manager between the office hours.
- The number of apples sold by a shopkeeper in the time period of 12 pm to 4 pm daily.

Statistical decision theory

Statistical Decision Theory may be defined as a body of several methods which facilitate the decision-maker to select wisely the best course of action from amongst several alternatives. In general, the problem of statistical decision theory may be stated as follows:

"Given a situation where there are several available alternative courses of action each of

which may lead to a set of mutually exclusive outcomes associated with certain probabilities, which course of action should a decision-maker take"?

The decision problem can be classified five types are:

Decision making under certainty

A condition of certainty exists when the decision-maker knows with reasonable certainty what the alternatives are, what conditions are associated with each alternative, and the outcome of each alternative.

Decision making under risk

In case of decision-making under uncertainty the probabilities of occurrence of various states of nature are not known. When these probabilities are known or can be estimated, the choice of an optimal action, based on these probabilities, is termed as decision making under risk.

Decision Making Under Uncertainty:

The process of making decision under conditions of uncertainty takes place when there is hardly any knowledge about states of nature and no objective information about their probabilities of occurrence. In such cases of absence of historical data and relative frequency, the probability of the occurrence of the particular state of nature cannot be indicated.

Decision Making Under Partial Information:

This type of situation is somewhere between the conditions of risk and conditions of uncertainty. As regards conditions of risk, we have seen that the probability of the occurrence of various states of nature are known as the basis of past experience, and in conditions of uncertainty, there is no such data available. But many situations arise where there is partial availability of data. In such circumstances, we can say that decision making is done on the basis of partial information.

Decision making under Conflict

Conflict and choice are closely related in that choice produces conflict and conflict is resolved by making a choice. The present study introduces a model (multi attribute decision field theory) that predicts a decision time pattern depending on the conflict situation.

Expected monetary value

EMV analysis is a statistical concept that calculates the average outcome when the future includes scenarios that may or may not happen. EMV for a project is calculated by multiplying the **value** of each possible outcome by its probability of occurrence and adding the products together.

Referring to the original payoff matrix, the formula for expected monetary value (EMV) is:

$$\mathbf{EMV}\ (A_i) = \mathbf{E}\ (A_i) = \sum_j p_j\ (R_{ij})$$

where *i* refers to the matrix's rows and *j* refers to the columns.

Expected Opportunity Loss

EOL Criterion is a technique used to make decisions under uncertainty, under the assumption that the probabilities of each state of nature are known. The decision made and the final state of nature (which the decision maker does not know beforehand) determines the payoff.

Recall from the savage criterion that an opportunity loss is the payoff difference between the best possible outcome under S_j and the actual outcome resulting from choosing $A_{i\ given}$ that S_j occurs. Referring now to the opportunity loss matrix, the formula for expected opportunity loss (EOL) is:

$$\mathbf{EOL}\ (A_i) = \mathbf{E}\ (A_i) = \Sigma_j p_j\ (OL_{ij})$$

Marginal Analysis

Marginal analysis is used when the number of states of nature is considerably large. Using this analysis, it is possible to locate the optimal course of action without the computation of EMV's of various actions. An order of A units is said to be optimal if the expected marginal profit of the Ath unit is non-negative and the expected marginal profit of the (A + 1)th unit is negative. Using equation (1), we can write

$(\Pi+\lambda)P(D \geq A)-\lambda \geq 0$ and .....(2)

$(\Pi+\lambda)P(D \geq A+\lambda)-\lambda < 0$..... (3)

The expected value of perfect information

The expected value of perfect information is the price that a healthcare decision maker would be willing to pay to have perfect information regarding all factors that influence which treatment choice is preferred as the result of a **cost**-effectiveness analysis.

The decision tree analysis

A decision tree is a diagram or chart that helps determine a course of action or show a statistical probability. Each branch of the decision tree represents a possible decision, outcome, or reaction. The furthest branches on the tree represent the end results of a certain decision pathway.

Unit-2

Sampling

Sampling

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

Random Sampling Definition

Random sampling is a method of choosing a sample of observations from a population to make assumptions about the population. It is also called **probability sampling**. The counterpart of this sampling is Non-probability sampling or Non-random sampling. The primary types of this sampling are simple random sampling, stratified sampling, cluster sampling, and multistage sampling. In the **sampling methods**, samples which are not arbitrary are typically called convenience samples.

Type of Random Sampling

The random sampling method uses some manner of a random choice. In this method, all the suitable individuals have the possibility of choosing the sample from the whole sample space. It is a time consuming and expensive method. The advantage of using probability sampling is that it ensures the sample that should represent the population. There are four major types of this sampling method, they are;

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling
4. Clustered Sampling

Now let us discuss its types one by one here.

Simple random sampling

In this sampling method, each item in the population has an equal and likely possibility of getting selected in the sample (for example, each member in a group is marked with a specific number). Since the selection of item completely depends on the possibility, therefore this method is called "**Method of chance Selection"**. Also, the sample size is large, and the item is selected randomly.

Systematic Random Sampling

In this method, the items are chosen from the destination population by choosing the random selecting point and picking the other methods after a fixed sample period. It is equal to the ratio of the total population size and the required population size.

Stratified Random Sampling

In this sampling method, a population is divided into subgroups to obtain a simple random sample from each group and complete the sampling process (for example, number of girls in a class of 50 strength). These small groups are called **strata**. The small group is created based on a few features in the population.

Clustered Sampling

Cluster sampling is similar to stratified sampling, besides the population is divided into a large number of subgroups (for example, hundreds of thousands of strata or subgroups). After that, some of these subgroups are chosen at random and simple random samples are then gathered within these subgroups. These subgroups are known as **clusters**. It is basically utilised to lessen the cost of data compilation.

Random Sampling Formula

If P is the probability, n is the sample size, and N is the population. Then;

- The chance of getting a sample selected only once is given by;

P = 1 – (N-1/N).(N-2/N-1)…..(N-n/N-(n-1))

Cancelling = 1-(N-n/n)

P = n/N

- The chance of getting a sample selected more than once is given by;

$P = 1-(1-(1/N))^n$

Sampling Error

Sampling error refers to differences between the sample and the population that exists only because of the observations that happened to be selected for the sample increasing the sample size will reduce this type of error.

Standard error

The standard error (SE) of a statistic is the approximate standard deviation of a statistical sample population. The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation.

Standard Error and Sample Size

The standard error of a statistic corresponds with the standard deviation of a parameter. Since it is nearly impossible to know the population distribution in most cases, we can estimate the standard deviation of a parameter by calculating the standard error of a sampling distribution. The standard error measures the dispersion of the distribution. As the sample size gets larger, the dispersion gets smaller, and the mean of the distribution is closer to the population mean. Thus, the sample size is negatively correlated with the standard error of a sample.

As the sample size gets larger, the sampling distribution has less dispersion and is more centered in by the mean of the distribution, whereas the flatter curve indicates a distribution with higher dispersion since the data points are scattered across all values.

The standard error is at the highest when the proportion is at 0.5. When conducting the experiment, if observing p getting close to 0.5(or 1-p getting close to 0.5), the standard error is increasing. To maintain the same standard error, we need to increase N, which is the sample size, to reduce the standard error to its original level.

Sampling distribution

A sampling distribution is a distribution that plots the values of a statistic for a given random sample that's part of a larger sum of data. When data scientists work with large quantities of data they sometimes use sampling distributions to determine parameters of the group of data, like what the mean or standard deviation might be. Parameters are facts about data in the form of statistical values.

Central Limit Theorem

Central limit theorem is a statistical theory which states that when the large sample size is having a finite variance, the samples will be normally distributed and the mean of samples will be approximately equal to the mean of the whole population.

In other words, the central limit theorem states that for any population with mean and standard deviation, the distribution of the sample mean for sample size N has mean μ and standard deviation $\sigma / \sqrt{n}$.

Assumptions of Central Limit Theorem

- The sample should be drawn randomly following the condition of randomization.
- The samples drawn should be independent of each other. They should not influence the other samples.
- When the sampling is done without replacement, the sample size shouldn’t exceed 10% of the total population.
- The sample size should be sufficiently large.

PARAMETRIC TESTS FOR MEANS & PROPORTIONS

Estimating Population Parameters

Population parameters like the mean, proportions, variance etc. are of great importance in significance tests and economic applications. Test based on such parameters are called parametric tests. Parametric tests enable to specify the parameters of population and the form of a concerned probability sampling distribution.

Statistical tests in which hypothesis deals with population parameters or sample statistics are called parametric tests. For example, when we want to test given population mean or population proportion, or any sample statistic, the test applied is called parametric test.

Assumptions in Parametric Tests

1. Sample Observations are independent.
2. Observations follow any sampling distribution.
3. Samples drawn are random samples.
4. Observations are made at least on interval scale

Point and Interval Estimates

A point estimate is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean. An interval estimate gives you a range of values where the parameter is expected to lie. A confidence interval is the most common type of interval estimate.

Estimation of a population mean

The most fundamental point and interval estimation process involves the estimation of a population mean. Suppose it is of interest to estimate the population means, μ, for a quantitative variable. Data collected from a simple random sample can be used to compute the sample, $\bar{x}$,where the value of $\bar{x}$ provides a point estimate of μ.

When the **sample mean** is used as a point **estimate of the population mean**, some error can be expected owing to the fact that a sample, or subset of the population, is used to compute the point estimate. The absolute value of the difference between the sample mean, $\bar{x}$,and the population earn, μ, written $|\bar{x} - \mu|$, is called the sampling error. Interval estimation incorporates a probability statement about the magnitude of the sampling error. The sampling distribution of $\bar{x}$ provides the basis for such a statement.

In the **large-sample case**, a 95% confidence interval estimate for the population mean is given by $\bar{x} \pm 1.96\sigma/\text{Square root of}\sqrt{n}$. When the population standard deviation, σ, is unknown, the sample standard deviation is used to estimate σ in the confidence interval formula. The quantity $1.96\sigma/\text{Square root of}\sqrt{n}$ is often called the margin of error for the estimate.. The quantity $\sigma/\text{Square root of}\sqrt{n}$ is the standard error, and 1.96 is the number of standard errors from the mean necessary to include 95% of the values in a normal distribution. The interpretation of a 95% confidence interval is that 95% of the intervals constructed in this manner will contain the population mean.

Thus, any interval computed in this manner has a 95% confidence of containing the population mean. By changing the constant from 1.96 to 1.645, a 90% confidence interval can be obtained. It should be noted from the formula for an interval estimate that a 90% confidence interval is narrower than a 95% confidence interval and as such has a slightly smaller confidence of including the population mean. Lower levels of confidence lead to even more narrow intervals. In practice, a 95% confidence interval is the most widely used.

Unit-3
Testing hypothesis

Hypothesis

Hypothesis is an assumption that is made on the basis of some evidence. This is the initial point of any investigation that translates the research questions into a prediction. It includes components like variables, population and the relation between the variables. A research hypothesis is a hypothesis that is used to test the relationship between two or more variables.

Null hypothesis (H_0)

The null hypothesis is a general statement that states that there is no relationship between two phenomenons under consideration or that there is no association between two groups.

Alternative hypothesis (H_1)

An alternative hypothesis is a statement that describes that there is a relationship between two selected variables in a study.

Types of hypothesis

According to the nature and situation, hypothesis may be simple or composite, parametric and non parametric, or null or alternative.

Simple and composite hypothesis

If a hypothesis is concerning sample statistic or population parameter only, it is called simple hypothesis. For example, "population standard deviation conforms to sample standard deviation" is a simple hypothesis."

Parametric and non parametric hypothesis

A hypothesis which specifies only the parameter or statistic of either the sample or population is called parametric hypothesis. If a hypothesis specifies only the form of the distribution, it is non parametric. For example, the hypothesis " Mean of the population is 2300" is a parametric hypothesis, while "population is normal" is non parametric.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides who are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Degree of freedom

While selecting items of statistical process, we have limited freedom. Degree of freedom is the number of independent choices in determining observations. In the case of individual observations, degree of freedom is total number of observations less the number of constraints. Usually is is equal to n-1.

Small sample and large sample tests

According to the number of items included in a sample, tests can be divided as small sample tests and large sample tests. If the test includes a sample of size less than 30, it is small sample test. If the size is 30 or more, it is large sample test.

Small sample tests follow student's t distribution. Large samples tests follow normal distribution. Mean tests may be conducted as large or small tests. But proportions are conducted as large sample tests on.

One tailed or two tailed tests

On the basis of location of rejection region, tests may be one tailed or two tailed. When a test examines the significance of difference of either more than a specific value or less than a specific value, rejection appears only on one side of the curve. It is called one tailed test

When test examines both more than or lower than a specific value at the same time, rejection region appears on both sides of the curve, and such test is called two tailed test. Most of tests are two tailed tests.

Example:

A sample of 300 screws has a mean length of 3.4 cm with standard deviation of 2.61 cm. Can it be regarded as a sample from a population with mean length of 3.25 cm, at α = 0.01?

Ho : No significant difference. It is sample from a population with mean length 3.25cm.

Given : Sample mean = x = 3.4 CM Population mean μ = 3.25cm σ = 2.61cm n = 300 , Z value = 1 table value @ 0.01, = 2.58

Since calculated z value is less than Z table value, difference is considered insignificant. Hypothesis is accepted. Therefore sample comes from a population with mean length 3.25 cm.

CHI-SQUARE TESTS

From a series of observation, different statistics are constructed to estimate population parameters. In general, sampling distribution of the statistic depends on parameter and form of population. The difference between distributions has been studied through constants such as mean, proportion, etc. They may not truly represent a distribution. This caused the necessity to have some index which can measure the degree of difference between actual frequencies and expected frequencies directly, without any representative value.

Thus emerged chi-square test, which is used to measure deviation of observed frequencies from expected frequencies.

Types of chi-square tests

On the basis of situation, nature and purpose of test, chi-square test may be classified as – test of independence of attributes, test of goodness of fit , test of homogeneity, and test for variance.

Test of independence of attributes

In significant testing, difference or dependence between two attributes can be studied. When we wish to test the difference of more than two proportions in terms of two attributes, chi-square test is applied. It is similar to Anova when variances between several sample groups are analyzed at a time.

For the chi square test, actual frequencies will be given in the question. Expected or theoretical frequencies have to be calculated and compared with each other to obtain measure of deviation.

Steps

1. Form null hypothesis

2. Consider observed frequencies or actual frequencies = O

3. Ascertain expected frequencies using the formula Expected Cell frequency E= × where

   E = Expected frequency A = column total B = row total AB = grand total

4. Obtain Σ ( ) for each cell frequency

5. Summate to get total chi-square value. Σ ( ) or value

6. Compare with chi-square table value at required level of significance and degree of freedom

7. Decide the fate of null hypothesis

Application of Chi square test:

The **chi-square distribution** is used in the common **chi-square** tests for goodness of fit of an observed **distribution** to a theoretical one, the independence of two criteria of classification of qualitative data, and in confidence interval estimation for a population standard deviation of a normal **distribution**.

Tests for goodness of fit:

If we have a set of frequencies of a distribution obtained by an experiment and if we are interested in knowing whether these frequencies are consistent with those which may be obtained based on some theory, then we can use chi square test of goodness of fit.

For example, if frequency distribution like Binomial or Poisson or Normal is applicable, the expected frequencies would be derived using that distribution.

Test of homogeneity

Here we have more than one sample unlike the test of independence where there is only one sample. We want to test whether these samples are homogeneous as far as a particular attribute is concerned .When there is homogeneity we conclude that the samples belong to the same population or identical population. The null hypothesis in these cases is that there is homogeneity.

The test is performed in the same manner as a test of independence. When the null hypothesis is accepted, we conclude that there is homogeneity.

F Distribution

The F Distribution is a probability distribution of the F Statistic. The F distribution is related to chi-square, because the f distribution is the ratio of two chi-square distributions with degrees of freedom ν1 and ν2 (note: each chi-square is first been divided by its degrees of freedom).

Characteristics of the F distribution

The F-distribution is either zero or positive, so there are no negative values for F. This feature of the F-distribution is similar to the chi-**square** distribution. The F-distribution is skewed to the right. Thus this probability distribution is nonsymmetrical.

Analysis of variance

ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

One Way ANOVA

As we now understand the basic terminologies behind ANOVA, let's dive deep into its implementation using a few examples.

A recent study claims that using music in a class enhances the concentration and consequently helps students absorb more information. As a teacher, your first reaction would be skepticism.

What if it affected the results of the students in a negative way? Or what kind of music would be a good choice for this? Considering all this, it would be immensely helpful to have some proof that it actually works.

To figure this out, we decided to implement it on a smaller group of randomly selected students from three different classes. The idea is similar to conducting a survey. We take three different groups of ten randomly selected students (all of the same age) from three different classrooms. Each classroom was provided with a different environment for students to study. Classroom A had constant music being played in the background, classroom B had variable music being played and classroom C was a regular class with no music playing. After one month, we conducted a test for all the three groups and collected their test scores.

Limitations of one-way ANOVA

A one-way ANOVA tells us that at least two groups are different from each other. *But it won't tell us which groups are different.* If our test returns a significant f-statistic, we may need to run a post-hoc test to tell us exactly which groups have a difference in means. Below I have mentioned the steps to perform one-way ANOVA in Excel along with a post-hoc test.

Two-Way ANOVA

Using one-way ANOVA, we found out that the music treatment was helpful in improving the test results of our students. But this treatment was conducted on students of the same age. What if the treatment was to affect different age groups of students in different ways? Or maybe the treatment had varying effects depending upon the teacher who taught the class.

Moreover, how can we be sure as to which factor(s) is affecting the results of the students more? Maybe the age group is a more dominant factor responsible for a student's performance than the music treatment.

For such cases, when the outcome or dependent variable (in our case the test scores) is affected by two independent variables/factors we use a slightly modified technique called two-way ANOVA.

That's why a two-way ANOVA can have up to three hypotheses, which are as follows:

Two null hypotheses will be tested if we have placed only one observation in each cell. For this example, those hypotheses will be:
H1: All the music treatment groups have equal mean score.
H2: All the age groups have equal mean score.

For multiple observations in cells, we would also be testing a third hypothesis:
H3: The factors are independent or the interaction effect does not exist.

An F-statistic is computed for each hypothesis we are testing.

Unit-4

Correlation and regression

CORRELATION ANALYSIS

Managers make personal and professional decision that is based on prediction of future events. To make predictions, they rely on relationships between variables. Correlation helps us to identify nature of relationship between variables, and then to determine the degree of such relation.

There are situations where there is relation between two variables and statistical analysis is necessary to study such relation. for example, a manager may want to know whether there is any relation between amount spent on research and development and sales. in this case, after identifying whether there is any relation, he may also want to know how much is the relation, what is the type of relation, etc. The quantitative technique that can be used to study such relation is the correlation analysis.

Definitions

Correlation is defined as "tendency of two or more variables to vary together directly or inversely" (Boddington). He also states that "whenever definite connection exists between two or more variables, there is said to be correlation"

Bowley defined correlation as "when two quantities are so related that fluctuations in one are in sympathy with the fluctuations of the other, or that increase or decrease of one is found in connection with increase or decrease of the other, such quantities are said to be correlated" According to M M Turtle, correlation is "an analysis of the association between two or more variables"

Thus correlation analysis is the quantitative tool used to describe the degree to which one variable is related to another variable.

Types of correlation

According to the nature of relation between variables, correlation may by positive and negative, linear or non linear, or simple, partial or multiple correlations.

Positive or negative correlation

When the values of two variables move in the same direction, correlation is said to be positive. That is, an increase in the value of one variable results in an increase in the value of the other variable also. Similarly a decrease in the value of one variable results in a decrease in the other variable.

When the value of two variables move in opposite direction, so that an increase in the value of one variable results in a decrease in the value of the other variable or vice versa, correlation is said to be negative. Generally price and supply are positively correlated, and correlation between price and demand is said to be negative.

Linear or Non Linear correlation

When the amount of change in one variable leads to a constant ratio of change in the other variable, correlation is said to be linear. For example, if price goes up by 10%, and It leads to a rise in the supply by 15% each time, there is a linear relation between price and supply. When there is linear correlation, the points plotted on a graph will give a straight line.

Correlation is said to be non linear when the amount of change in one variable is not in constant ratio to the change in the other variable. Here the ratio of change fluctuates and is never constant.

Simple, partial or multiple correlations

When there are only two variables, the correlation is said to be simple. For example, the correlation between price and demand is simple.

When one variable is related to a number of other variables, correlation is not simple. When there are three or more variables under study, at the same time, it may be multiple correlation or partial correlation.

In multiple correlations, we measure the degree of association between one variable on one side and all other variables together on the other side. The relation between yield with both rainfall and temperature is case of multiple correlations.

In partial correlation we study the relationship of one variable with one of the other variables presuming that the third or other variables remain constant. For example, we may study relation between yield and rainfall, keeping constant the effect of temperature.

Karl Pearson's coefficient of correlation

This is also known as Pearson's coefficient of correlation, and it is denoted by the symbol r. The formula for computing Pearson's Coefficient of correlation is

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Regression Equation

The **Regression Equation** is the algebraic expression of the regression lines. It is used to predict the values of the dependent variable from the given values of independent variables.

Definitions

Literally regression means 'going back'. Certain variables are found to move back and conform to progression of a related variable. In regression variables, reduce variances between them.

Regression is defined as "statistical device used to study the relationship between 2 or more variables that are related."

Regression Equations

For each regression line, there will be a regression equation. Regression equation is a mathematical relation between the dependent variable and independent variable. There are two regression equations - Y on X and X on Y.

Y on X = a +bx

Uses of Regression

The study of regression is very useful in business, economics and researches.

1. Regression helps to obtain most probable values of one variable for given values of other variable.
2. It helps to study the effect of price on supply or demand of a commodity.
3. It is widely applied in physical science where the relation is functional.
4. It is used to describe the nature of relation between 2 or more variable.
5. It reveals rate of change in one variable based on change in other variable.

Standard error of estimate

The standard error of the regression (S), also known as the standard error of the estimate, represents the average distance that the observed values fall from the regression line. Conveniently, it tells you how wrong the regression model is on average using the units of the response variable. Smaller values are better because it indicates that the observations are closer to the fitted line.

Testing significance of correlation and regression coefficients

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n together. We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

Test for Significance of Regression. The test for significance of regression in the case of multiple linear regression analysis is carried out using the analysis of variance. The test is used to check if a linear statistical relationship exists between the response variable and at least one of the predictor variables.

Coefficient of determination

The **coefficient of determination** is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1. With **linear regression**, the **coefficient of determination** is also equal to the square of the correlation between x and y scores.

Multivariate analysis

Multivariate analysis is a set of statistical techniques used for analysis of data that contain more than one variable. Multivariate analysis refers to any statistical technique used to analyze more complex sets of data.

Factor analysis

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. This technique extracts maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis.

Cluster analysis

Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects.

Discriminate analysis

Discriminate analysis is a technique that is used by the researcher to analyze the research data when the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature.

Unit-5

Linear programming, Assignment and Transportation

Linear programming Problem (LPP)

Linear programming is an optimization technique for a system of linear constraints and a linear objective function. An **objective function** defines the quantity to be optimized, and the goal of linear programming is to find the values of the variables that maximize or minimize the objective function.

This kind of problem is perfect to use linear programming techniques on.

- All of the quantifiable relationships in the problem are linear.
- The values of variables are constrained in some way.
- The goal is to find values of the variables that will maximize some quantity.

Meaning of graphical method

Graphical method of linear programming is used to solve problems by finding the highest or lowest point of intersection between the objective function line and the feasible regionon a graph.

LPP Assumptions

There are several assumptions on which the linear programming works, these are:

1. **Proportionality:** The basic assumption underlying the linear programming is that any change in the constraint inequalities will have the proportional change in the objective function.
2. **Additively:** The assumption of additively asserts that the total profit of the objective function is determined by the sum of profit contributed by each product separately.
3. **Continuity:** Another assumption of linear programming is that the decision variables are continuous. This means a combination of outputs can be used with the fractional values along with the integer values.
4. **Certainty:** Another underlying assumption of linear programming is a certainty, i.e. the parameter of objective function coefficients and the coefficients of constraint inequalities is known with certainty.
5. **Finite Choices:** This assumption implies that the decision maker has certain choices, and the decision variables assume non-negative values. The non-negative assumption is true in the sense, the output in the production problem cannot be negative.

Advantages and limitations (or)disadvantages:

LP has been considered an important tool due to following reasons:

1. LP makes logical thinking and provides better insight into business problems.
2. Manager can select the best solution with the help of LP by evaluating the cost and profit of various alternatives.
3. LP provides an information base for optimum allocation of scarce resources.
4. LP assists in making adjustments according to changing conditions.
5. LP helps in solving multi-dimensional problems.

Limitations (or) disadvantages:

1. This technique could not solve the problems in which variables cannot be stated quantitatively.
2. In some cases, the results of LP give a confusing and misleading picture.
3. LP technique cannot solve the business problems of non-linear nature.
4. The factor of uncertainty is not considered in this technique.
5. This technique is highly mathematical and complicated.

Simple steps for Graphical method

Step 1: Set the Objective function i.e. Maximization of profit and Minimization of expenses

Example Z=12X1 +16X2

Step 2: Convert Inequality into **Equality** by equation

$10X_1 + 20X_2 = 120$

$8X_1 + 8X_2 = 80$

Solve the equation taking when X1 is 0; what is X

X_1	0	?
X_2	?	0

2 ?

Step 3: Plot both equation on graph sheet

Step 4: Find the feasible region which is found under the two equations for Maximization Problems like the Region found above two equation are called Feasible region in case of Minimization Problems.

Step 5: Find all the corner values

Step 6: Apply all the Corner values in the Objective Function ie Z=12X1 +16X2

Which is **Maximum** is the solution for **Maximization Problem** and

Which is **Minimum** is the solution for **Minimization Problem** and

Simplex Method

The **Simplex Method or Simplex Algorithm** is used for calculating the optimal solution to the linear programming problem. In other words, the simplex algorithm is an iterative procedure carried systematically to determine the optimal solution from the set of feasible solutions.

Why introduce slack variable

Slack variable is a variable that is added to an inequality constraint to transform it into equality. Introducing a slack variable replaces an inequality constraint with an equality constraint and a non-negativity constraint on the slack variable.

Feasible solution meaning

A **feasible solution** is a set of values for the decision variables that satisfies all of the constraints in an optimization problem. The set of all **feasible solutions** defines the **feasible** region of the problem.

Basic feasible solution

A **basic solution** that satisfies all the constraints defining or in other words, one that lies within is called a **basic feasible solution**. let Ax=b the system of the 'm' equation with 'n' unknown variables.

LIMITATIONS OF LPP SIMPLEX METHOD

1. Simplex method Involves understanding of many conceptual technical aspects. These cannot be understood by any manager not conversant with the subject.
2. Linear programming problems need lot of expertise, time and are cumbersome. A number of steps have to be adopted to proceed in a systematic manner before one can arrive at the solution.
3. Graphic solution method has lot of applications and is relatively short and simple. However, it has limitations and cannot be applied to problems with more than two variables in the objective function.
4. Simplex method of LPP can be applied to problems with more than two variables in the objective function, the procedure adopted is complicated and long.
5. LPP does not lead to 'a unique' optimal solution. It can provide different types of solutions like feasible solution, infeasible solution, unbounded solution, degenerate solution etc.

Simplex Method steps

Step 1: Set the Objective function with introduction S_1 and S_2 i.e. Maximisation of profit and Minimization of expenses

Example $Z=12X_1 +16X_2 +0S_1+0S_2$

Step 2: Convert **Inequality** into **Equality** by equation

$10X_1 + 20X_2 + S_1 = 120$

$8X_1 + 8X_2 + S_2 = 80$

Step 3: Construct a table using the above object function and inequality Constraint values

			Key Column					
Cij	**Cj**	**12**	**16**	**0**	**0**	**Solution**	**Ratio**	
	BV	X1	**X_2 Incoming**	S_1	S_2			
0	**S_1 Out going**	10	20 Key Elemnt	1	0	120	120 / 20 = 6	**Key Row**
0	S_2	8	8	0	1	80	80 / 8 = 10	
	Zj	**0**	**0**	**0**	**0**	**0**		

12 16 0 0

Step 4: Find the Valuesof Zj

Zj=Cij x Bij ie for X_1 Column 0x10=0 + 0 x8 = 0

X_2 Column 0x20=0 + 0 x8 = 0

Step 5:Find the Optimality

For Maximisation Problems Cj- Zj $\leq$ 0
For Minimization Problem Cj- Zj $\geq$ 0

Step 6: If the Optimality is not come, go for Next Iteration Table by the following additional steps

a) bringing **incoming variable** by selecting the **Maximum Column Value** (Cj-Zj), that is **key column**

b) Find out the Ratio by = **Solution** / **Corresponding Column Value**

c) Select the least Row Ratio , that Row is called **Key Row, that Row S Value is Out Going Variable**

d) Find the Intersection of both the Row and Column, that Variable is called Key Element

e) For the Next iteration find out new Value by the following

1. in the **Key Row**, find new value by = Key Row Value

Corresponding Key Column Value

In the **Non Key Row**.,

Find the New Value =Old Value - Corr.Key Row Value x Corr.Key Column Value

Key Element

Step 7 Repeat From the Step 3

ASSIGNMENT

Introduction

The assignment problem is a special case of transportation problem wherein the number of resources (origin) equals number of activities (destinations). The capacity and demand value is exactly one unit i.e. only one unit can be supplied from each origin and each destination also requires exactly only one Unit

OBJECTIVE ASSIGNMENT PROBLEM

The **objectives** alone are considered as fuzzy. The classical **assignment problem** refers to a special class of linear programming **problems**. Linear programming is one of the most widely used decision making tool for solving real world **problems**

In shortly in objective

Objectives: The objective is to determine which origin should supply specific units to which destination.

STEPS of assignment problem

1. Deduct ROW MINIMUM from all the elements in each row for all rows.

2. From such reduced matrix, deduct COLUMN MINIMUM from all elements in each column for all the columns.

3. To find optimal solution,

   Cover MAXIMUM NUMBER OF ZEROS by drawing MINIMUM NUMBER OF VERTICAL OR HORIZONTAL LINES which should be equal to order of matrix.

   *** If the minimum number of lines are not equal to order of matrix (no optimal solution), uncovered elements are reduced by smallest element in uncoverd area and intersection element is added with such smallest element (no change in the

covered area). Repeat the same until you are getting optimal solution.

4. Make one assignment to least possible zero in each row and in each column. Subsequently, strike off other zeros found in the same row and same column.

   *Least possible zero = giving preference to row or column which has minimum zero(s)

5. Add original values of assignments located places to get the result.

Note:

1. Unbalanced problems : If the number of columns is not equal to number of rows, dummy column should be added with zero elements and vice versa.

2. Maximization case: If profit is given (usually cost is given), conversation of maximization problem into minimization problem by deduction of all elements in all rows and columns from the large value of the total elements. (Largest – All other smallest elements).

Hungarian method

The **Hungarian method** is a combinatorial optimization **algorithm** that solves the **assignment** problem in polynomial time and which anticipated later primal-dual **methods**.

What is Dummy activity?

A **dummy activity** is a simulated **activity** of sorts, one that is of a zero duration and is created for the sole purpose of demonstrating a specific relationship and path of action on the arrow diagramming method.

TRANSPORTATION

In the process of transportation of goods from one place to various distribution centers(origin) to various distribution centers(destination), transportation expenses are incurred. Some times it may be more due to random calculation. So in order to avoid increase of cost in transportation, the least cost is selected by applying a suitable method.

NORTH-WEST CORNER METHOD:

The **North-West Corner Rule** is a method adopted to compute the initial feasible solution of the transportation problem. The name North-west corner is given to this method because the basic variables are selected from the extreme left corner.

Least Cost Method

The **Least Cost Method** is another method used to obtain the initial feasible solution for the transportation problem. Here, the allocation begins with the cell which has the minimum cost.

The lower cost cells are chosen over the higher-cost cell with the objective to have the least cost of transportation.

Vogel's Approximation Method

The **Vogel's Approximation Method** or **VAM** is an iterative procedure calculated to find out the initial feasible solution of the transportation problem. Like Least cost Method, here also the shipping cost is taken into consideration, but in a relative sense.

NORTH-WEST CORNER METHOD STEPS:

1. Begin from NORTH-WEST CORNER CELL(Upper Left Hand Corner) of the transportation table.

2. ALLOT the respective Row Total or Column Total, whichever is less, in the North-West Cell .

*Side-by-side, corresponding to such row or column, mention the remaining balance to be allotted.

3. Strike off the respective Row or Column as a sign of full allotment (which has remaining balance-zero) made.

4. Select the next North-West Corner Cell and repeat the first three steps for remaining Rows and columns till possible allotments to be made.

Note: But, the total allotment made should be equal to m+n-1 in order to get feasible solution.

5. Now, add all the values found by multiplying the transportation cost with allotment made to find the total transportation cost.

LEAST COST METHOD STEPS:

1. Choose the Least Cost Cell.

2. Allot the respective Row Total or Column Table, whichever is less, in the Least Cost Cell.

* Side-by-side, corresponding to such row or column, mention the remaining balance to be allotted.

3. Strike off the respective Row or Column as a sign of full allotment (which has remaining balance-zero) made.
4. Select the next Least Cost Cell and repeat the first three steps for remaining Rows and columns till possible allotments to be made.

Note: But, the total allotment made should be equal to m+n-1 in order to get feasible solution.

5. Now, add all the values found by multiplying the transportation cost with allotment made to find the total transportation cost.

VOGEL's APPROXIMISATION METHOD (VAM) STEPS:

1. Find out the difference between two least cost in each column and in each row.

2. Select the maximum difference among them and locate the lowest cell corresponding to the maximum difference.

3. Allot the respective row total or column total whichever is less in such lowest cost cell.

4. Strike off the row or column or both as a sign of allotment fully made.

5. Repeat the first four steps for the remaining rows and columns till the possible allotments are made.

6. Now add the values found by multiplying the transportation cost with full allotment made to find out the totl transportation cost. *Generally, preference is given to minimum cost and possible maximum allotment.

MXIMISATION CASE:

If profit is given (usually cost is given), conversation of maximization problem into minimization problem by deduction of all elements in all rows and columns from the large value of the total elements. (Largest - All other smallest elements).

DEGENERACY:

In transportation problem, if total allotment is not equal to m+n-1, it is called degeneracy. In this case select the least unallotted (unallotted cell) and allot Epsilon (value close to zero or between 0 and 1) in such a least unallotted cell. The remaining steps are same as we followed earlier.

****