Unit - 1

Introduction

In this Unit, we shall study of the probability distributions that are used most prominentlyin statistical theory and application. We shall also study their parameter that is the quantities that are constants for particular distributions but that can take on different values for different members of families of distributions of the same kind. We shall introduce number of discrete probability distributions that have been successfully applied in a wide variety of decision situations. The purpose of this Unit is to show the types of situations in which these distributions can be applied.

It may be mentioned that a theoretical probability distribution gives us a law according to which different values of the random variable are distributed with specified probabilities according to some definite law which can be expressed mathematically. It is possible to formulate such laws either on the basis of given conditions (a prior consideration) or on the basicof the results (a posterior inference) of an experiment. This unit is devoted to the study of univariate discrete and continuous distributions like,Binomial, Poisson and Normal distributions

Binomial Experiment

A binomial distribution can be used under the following condition:

(i) Any trail with two possible outcomes that is any trail result in a success or failure.

(ii) The number of trials n is finite and independent,when n is number of trial.

(iii) a probability of success is the same in each trial. i.e., p is the constant.

Applications of Binomial Distribution

1. The quality control measures and sampling process in industries to classify the itemsare defective or non-defective.
2. Medical applications as a success or failure of a surgery and cure or non cure of apatient.
3. Military application as a hit a target or miss a target

The binomial distribution formula is for any random variable X, given by;

$$P(x:n,p) = {}^{n}C_{x}\, p^{x}\, (1-p)^{n-x}$$

Or

$$P(x:n,p) = {}^{n}C_{x}\, p^{x}\, (q)^{n-x}$$

Where,

n = the number of experiments

x = 0, 1, 2, 3, 4, …

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

The binomial distribution formula can also be written in the form of n-Bernoulli trials, where ${}^{n}C_{x}$ = n!/x!(n-x)!. Hence,

$P(x:n,p) = n!/[x!(n-x)!].p^{x}.(q)^{n-x}$

Binomial Distribution Mean and Variance:

For a binomial distribution, the mean, variance and standard deviation for the given number of success are represented using the formulas

Mean, $\mu = np$

Variance, $\sigma^2 = npq$

Standard Deviation $\sigma = \sqrt{(npq)}$

Where p is the probability of success

q is the probability of failure, where q = 1-p

Properties of Binomial Distribution

The properties of the binomial distribution are:

- There are two possible outcomes: true or false, success or failure, yes or no.
- There is 'n' number of independent trials or a fixed number of n times repeated trials.
- The probability of success or failure varies for each trial.
- Only the number of success is calculated out of n independent trials.
- Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.

Poisson distribution

Poisson distribution is a theoretical discrete probability and is also known as the Poisson distribution probability mass function. It is used to find the probability of an independent event that is occurring in a fixed interval of time and has a constant mean rate. The Poisson distribution probability mass function can also be used in other fixed intervals such as volume, area, distance, etc. A Poisson random variable will relatively describe a phenomenon if there are few successes over many trials. The Poisson distribution is used as a limiting case of the binomial distribution when the trials are large indefinitely. If a Poisson distribution models the same binomial phenomenon, λ is replaced by np. Poisson distribution is named after the French mathematician Denis Poisson.

Poisson distribution Formula:

Poisson distribution formula is used to find the probability of an event that happens independently, discretely over a fixed time period, when the mean rate of occurrence is constant over time. The Poisson distribution formula is applied when there is a large number of possible outcomes. For a random discrete variable X that follows the Poisson distribution, and λ is the average rate of value, then the probability of x is given by: $f(x) = P(X=x) = (e^{-\lambda} \lambda^x)/x!$

Where

- x = 0, 1, 2, 3...
- e is the Euler's number(e = 2.718)
- λ is an average rate of the expected value and λ = variance, also λ>0

Poisson Distribution Mean and Variance

For Poisson distribution, which has λ as the average rate, for a fixed interval of time, then the mean of the Poisson distribution and the value of variance will be the same. So for X following Poisson distribution, we can say that λ is the mean as well as the variance of the distribution.

Hence: E(X) = V(X) = λ

where

- E(X) is the expected mean
- V(X) is the variance
- λ > 0

Properties of Poisson Distribution

The Poisson distribution is applicable in events that have a large number of rare and independent possible events. The following are the properties of the Poisson Distribution. In the Poisson distribution,

- The events are independent.
- The average number of successes in the given period of time alone can occur. No two events can occur at the same time.
- The Poisson distribution is limited when the number of trials n is indefinitely large.
- mean = variance = λ
- np = λ is finite, where λ is constant.
- The standard deviation is always equal to the square root of the mean μ.
- The exact probability that the random variable X with mean μ =a is given by $P(X= a) = \mu^a / a! e^{-\mu}$
- If the mean is large, then the Poisson distribution is approximately a normal distribution.

Applications of Poisson distribution:

There are various applications of the Poisson distribution. The random variables that follow a Poisson distribution are as follows:

- To count the number of defects of a finished product
- To count the number of deaths in a country by any disease or natural calamity
- To count the number of infected plants in the field
- To count the number of bacteria in the organisms or the radioactive decay in atoms
- To calculate the waiting time between the events.

Remarks:

- The formula for Poisson distribution is $f(x) = P(X=x) = (e^{-\lambda} \lambda^x)/x!$.
- For the Poisson distribution, λ is always greater than 0.
- For Poisson distribution, the mean and the variance of the distribution are equal.

Normal Distribution

In probability theory and statistics, the **Normal Distribution**, also called the **Gaussian Distribution**, is the most significant continuous probability distribution. Sometimes it is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. Furthermore, it can be used to approximate other probability distributions, therefore supporting the usage of the word 'normal 'as in about the one, mostly used.

Normal Distribution Definition

The Normal Distribution is defined by the probability density function for a continuous random variable in a system. Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx.

$f(x) \geq 0 \; \forall \; x \in (-\infty, +\infty)$

And $\int_{-\infty}^{+\infty} f(x) = 1$

Normal Distribution Formula

The probability density function of normal or gaussian distribution is mean "μ" and standard deviation "σ", the normal distribution formula is given by:

$$\mathbf{f(x) = (1/\sqrt{(2\pi\sigma^2)})\ (e^{[-(x-\mu)^2/2\sigma^2)}}.$$

Where,

- x is the variable
- μ is the mean
- σ is the standard deviation

Normal Distribution Curve

The random variables following the normal distribution are those whose values can find any unknown value in a given range. For example, finding the height of the students in the school. Here, the distribution can consider any value, but it will be bounded in the range say, 0 to 6ft. This limitation is forced physically in our query.

Whereas, the normal distribution doesn't even bother about the range. The range can also extend to $-\infty$ to $+\infty$ and still we can find a smooth curve. These random variables are called Continuous Variables, and the Normal Distribution then provides here probability of the value lying in a particular range for a given experiment. Also, use the normal distribution calculator to find the probability density function by just providing the mean and standard deviation value.

Normal Distribution Standard Deviation

Generally, the normal distribution has any positive standard deviation. We know that the mean helps to determine the line of symmetry of a graph, whereas the standard deviation helps to know how far the data are spread out. If the standard deviation is smaller, the data are somewhat close to each other and the graph becomes narrower. If the standard deviation is larger, the data are dispersed more, and the graph becomes wider. The standard deviations are used to subdivide the area under the normal curve. Each subdivided section defines the percentage of data, which falls into the specific region of a graph.

Using 1 standard deviation, the Empirical Rule states that,

- Approximately 68% of the data falls within one standard deviation of the mean. (i.e., Between Mean- one Standard Deviation and Mean + one standard deviation)
- Approximately 95% of the data falls within two standard deviations of the mean. (i.e., Between Mean- two Standard Deviation and Mean + two standard deviations)
- Approximately 99.7% of the data fall within three standard deviations of the mean. (i.e., Between Mean- three Standard Deviation and Mean + three standard deviations)

Normal Distribution Properties

Some of the important properties of the normal distribution are listed below:

- In a normal distribution, the mean, median and mode are equal.(i.e., Mean = Median= Mode).
- The total area under the curve should be equal to 1.
- The normally distributed curve should be symmetric at the centre.
- There should be exactly half of the values are to the right of the centre and exactly half of the values are to the left of the centre.
- The normal distribution should be defined by the mean and standard deviation.
- The normal distribution curve must have only one peak. (i.e., Unimodal)
- The curve approaches the x-axis, but it never touches, and it extends farther away from the mean.

Characteristics of Normal Distribution

Normal Distribution has the following characteristics that distinguish it from the other forms of probability representations:

- **Empirical Rule**: In a normal distribution, 68% of the observations are confined within -/+ one standard deviation, 95% of the values fall within -/+ two standard deviations, and almost 99.7% of values are confined to -/+ three standard deviations.
- **Bell-shaped Curve**: Most of the values lie at the center, and fewer values lie at the tail extremities. This results in a bell-shaped curve.
- **Mean and Standard Deviation**: This data representation is shaped by mean and standard deviation.
- **Equal Central Tendencies**: The mean, median, and mode of this data are equal.
- **Symmetric**: The normal distribution curve is centrally symmetric. Therefore, half of the values are to the left of the center, and the remaining values appear on the right.
- **Skewness and Kurtosis**: Skewness is the the symmetry. The skewness for a normal distribution is zero. Kurtosis studies the tail of the represented data. For a normal distribution, the kurtosis is 3.
- **Total Area = 1**: The total value of the standard deviation, i.e., the complete area of the curve under this probability function, is one. Also, the entire mean is zero.

Unit – 2

Testing of Hypothesis

The estimate based on sample values do not equal to the true value in the population due to inherent variation in the population. The samples drawn will have different estimates compared to the true value. It has to be verified that whether the difference between the sample estimate and the population value is due to sampling fluctuation or real difference. If the difference is due to sampling fluctuation only it can be safely said that the sample belongs to the population under question and if the difference is real we have every reason to believe that sample may not belong to the population under question.

The probability of the difference between sample estimate and the true taken with standard error compared with observed difference is very small then there is significant difference between the sample estimate and the population value. That is, the probability is less for the difference between sample estimate and the true (or population) not due to sampling fluctuation but due to real difference. If the probability is very large then there is no significant difference between the sample estimate and the true value. The difference so obtained is due to sampling fluctuation only.

A statistical hypothesis is some statement or assertion about a population or equivalently about the probability distribution characteristing a population which we want to

verify on the basis of information available from a sample.

Simple and Composite Hypothesis:

When a hypothesis specifies all the parameters of a probability distribution, it is known as simple hypothesis. The hypothesis specifies all the parameters, i.e μ and σ of a normal distribution.

Example: The random variable x is distributed normally with mean μ=0 & SD=1 is a simple hypothesis. The hypothesis specifies all the parameters (μ & σ) of a normal distributions.

If the hypothesis specific only some of the parameters of the probability distribution, it is known as composite hypothesis. In the above example if only the μ is specified or only the σ is specified it is a composite hypothesis.

Test of Statistical Hypothesis:

A test of statistical hypothesis is a two action decision problem after the experimental sample value has been obtained. The two action being acceptance rejection of the hypothesis under consideration.

Null Hypothesis: In hypothesis, testing a decision maker should not be motivated by prospects of profit or loss resulting from the acceptance or rejection of the hypothesis, ie., neutral or general statement about the population parameter is known as null hypothesis.

Alternative Hypothesis: it is desirable to reject the hypothesis based on statistical test in other words, the general statement which is opposite to be null hypothesis stated is known as alternative hypothesis.

STEPS INVOLVED IN TESTING OF HYPOTHESIS:

1. Explicit knowledge of the nature of population distribution and the parameter of interest (i.e) the parameter about which the hypothesis are setup
2. Setting up the null hypothesis H_0 and the alternative hypothesis H_1 in terms of the range of parameter values each ones embodies.
3. The choice of a suitable statistic called the test statistic which will be reflecting upon the probability of H_0 and H_1.
4. Partitioning the set of possible values of the test statistic into two disjoint sets w and $\overline{w}$ and framing the following test.
   i) Reject H_0 if the value of test statistic falls in w (critical region)
   ii) Accept H_0 if falls in $\overline{w}$ (acceptance region)
5. After framing the above obtain experimental sample observation, compute the appropriate test statistic and take actions accordingly.

Steps involved in statistical test of significance:

A statistical test of significance is a statistical test of hypothesis using the following procedure.

1. Formulation of hypothesis:

The hypothesis to be test is taken as null hypothesis H_0. Normally when one parameter is involved the hypothesis is –there is no significant difference between the hypothetical value of the parameter and corresponding statistical value from the sample‖. When two parameters are involved, the null hypothesis is –there is no significant difference between statistic obtained from two sample‖. The alternative hypothesis is normally two sided and just opposite of null hypothesis.

2. Choosing the level of significance:

α = level of significance

= P [Type I error]

= size of critical region

α value is fixed at low level usually it is fixed as 5% or 1%.

3. Selecting statistic & finding its distribution:

Let t be a statistic such that E(t)= ϕ where ϕ is the parameter of the distribution. We must find standard error of t which is the standard deviation of the sampling distribution of the statistic.

$$\text{test statistic} = \frac{t - E(t)}{SE(t)}$$

Find the distribution of test statistic which may be normal, t, x^2 or F distribution.

Critical region is { |test statistic| ≥ critical values }

A.R = { |test statistic| < critical values}.

If the value of test statistic ≥ critical value H_1 is rejected. If the value of test statistic less than critical value then there is no reason to reject H_0 at level α. Accordingly, inferences can be drawn.

LARGE SAMPLE TEST:

Any statistical test based on the assumption that the sample size n is large ($n \to \infty$) is called asymptotic test. We know that as $n \to \infty$ any statistic irrespective of the parent population from which sample is drawn follows Normal Distribution (Central limit theorem).

Hence any statistic follows Normal Distribution as $n \to \infty$ the test based on such a statistic is called asymptotic test. Any statistical test based on exact distribution of a statistical under consideration is called exact test. Here, there is no assumption on the sample size most of the statistical test uses t-distribution, x^2 distribution and F-distribution which are exact distribution of statistic. Hence test based on t, F, x^2 distributions are called exact test. Sometimes the statistic may also follow Normal Distribution and in such cases, it is also an exact test.

Z Test

Z test is a statistical test that is conducted on data that approximately follows a normal distribution. The z test can be performed on one sample, two samples, or on proportions for hypothesis testing. It checks if the means of two large samples are different or not when the population variance is known. A z test can further be classified into left-tailed, right-tailed, and two-tailed hypothesis tests depending upon the parameters of the data. In this article, we will learn more about the z test, its formula, the z test statistic, and how to perform the test for different types of data using examples.

- A z test is a test that is used to check if the means of two populations are different or not provided the data follows a normal distribution. For this purpose, the null hypothesis and the alternative hypothesis must be set up and the value of the z test statistic must be calculated. The decision criterion is based on the z critical value.

- A z test is conducted on a population that follows a normal distribution with independent data points and has a sample size that is greater than or equal to 30. It is used to check whether the means of two populations are equal to each other when the population variance is known. The null hypothesis of a z test can be rejected if the z test statistic is statistically significant when compared with the critical value.

Steps to perform Z-test:

- First, identify the null and alternate hypotheses.
- Determine the level of significance ($\propto$).

- Find the critical value of z in the z-test using
- Calculate the z-test statistics. Below is the formula for calculating the z-test statistics.

  $Z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ where standard deviation is known.

  $Z = (\bar{x} - \mu)/(s/\sqrt{n})$ where standard deviation is not known.

  Where, $\bar{x}$ is sample mean

  μ is Population mean

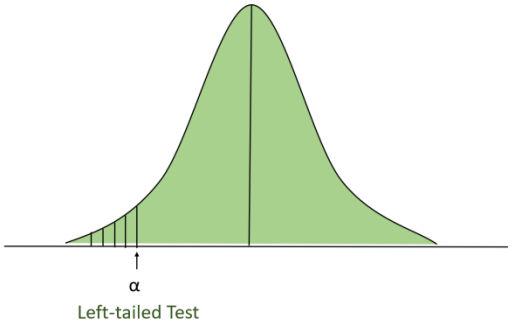
  σ is population SD

  S is the sample SD and

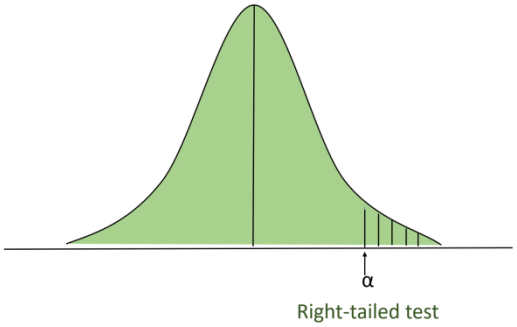
  n is sample size
- Now compare with the hypothesis and decide whether to reject or not to reject the null hypothesis

Types of z test:

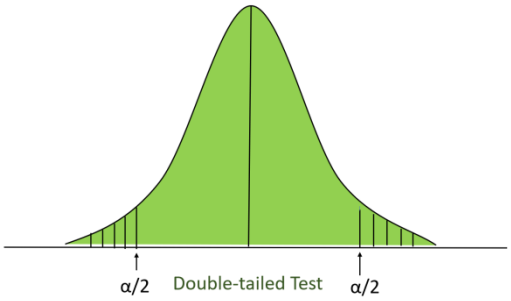
- **Left-tailed Test:** In this test, our region of rejection is located to the extreme left of the distribution. Here our null hypothesis is that the claimed value is less than or equal to the mean population value.



- **Right-tailed Test:** In this test, our region of rejection is located to the extreme right of the distribution. Here our null hypothesis is that the claimed value is less than or equal to the mean population value.



- **Two-tailed test:** In this test, our region of rejection is located to both extremes of the distribution. Here our null hypothesis is that the claimed value is equal to the mean population value.



Assumptions of Parametric data:

In order for the results of parametric tests to be valid, the following four assumptions should be met:

1. Normality – Data in each group should be normally distributed.

2. Equal Variance – Data in each group should have approximately equal variance.

3. Independence – Data in each group should be randomly and independently sampled from the population.

4. No Outliers – There should be no extreme outliers.

T test:

- A T-test is a statistical method of comparing the means or proportions of two samples gathered from either the same group or different categories.

- It is aimed at hypothesis testing, which is used to test a hypothesis pertaining to a given population.
- It is the difference between population means and a hypothesized value.
- One-sample, two-sample, paired, equal, and unequal variance are the types of T-tests users can use for mean comparisons.

Assumptions:

The test runs on a set of assumptions, which are as follows:

- The measurement scale used for such **hypothesis testing** follows a set of continuous or ordinal patterns. The accounted parameters and variants influencing the samples and surrounding the groups are based on the standard consideration.
- The tests are completely based on random sampling. As no individuality is maintained in the samples, the reliability is often questioned.
- When the data is plotted with respect to the T-test distribution, it should follow a **normal distribution** and bring about a bell-curved graph.
- For a clearer **bell curve**, the **sample size** needs to be bigger.
- The variance should be such that the **standard deviations** of the samples are almost equal.

One-Sample T-Test

While performing this test, the mean or average of one group is compared against the set average, which is either the theoretical value or means of the population. For example, a teacher wishes to figure out the average height of the students of class 5 and compare the same against a set value of more than 45 kgs.

The teacher first randomly selects a group of students and records individual weights to achieve this. Next, she finds out the mean weight for that group and checks if it meets the standard set value of 45+. The formula used to obtain one-sample t-test results is:

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

- T = t-statistic
- $\bar{x}$= mean of the group
- μ= theoretical mean value of the population
- s = standard deviation of the group
- n = sample size

Independent Two-Sample T-Test

This is the test conducted when samples from two different groups, species, or populations are studied and compared. It is also known as an independent T-test. For example, if a teacher wants to compare the height of male students and female students in class 5, she would use the independent two-sample test.

The **T-test formula** used to calculate this is:

$$t = \frac{\bar{x_1} - \bar{x_2}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$DF = n_1 + n_2 - 2$$

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Where,

- x1 – x2 = means of samples from two different groups or populations
- n1 – n2 = respective sample sizes
- sp = standard deviation or common variance of two samples

Paired Sample T-Test

This hypothesis testing is conducted when two groups belong to the same population or group. The groups are studied either at two different times or under two varied conditions. The formula used to obtain the t-value is:

$$t_{calc} = \frac{\bar{d}}{s_d / \sqrt{n}}$$

Unit – 3

Non- Parametric test

Chi-square test:

A Chi-square test is performed to determine if there is a difference between the theoretical population parameter and the observed data.

- Chi-square test is a non-parametric test where the data is not assumed to be normally distributed but is distributed in a chi-square fashion.
- It allows the researcher to test factors like a number of factors like the goodness of fit, the significance of population variance, and the homogeneity or difference in population variance.
- This test is commonly used to determine if a random sample is drawn from a population with mean μ and the variance σ^2.

Uses of chi square test:

Chi-square test is performed for various purposes, some of which are:

1. This method is commonly used by researchers to determine the differences between different categorical variables in a population.
2. A Chi-square test can also be used as a test for goodness of fit. It enables us to observe how well the theoretical distribution fits the observed distribution.
3. It also works as a test of independence where it enables the researcher to determine if two attributes of a population are associated or not.

Formula for chi square test:

Chi-square test is symbolically written as χ^2 and the formula of chi-square for comparing variance is given as:

$$\chi^2 = \frac{\sigma s^2}{\sigma p^2}(n-1)$$

where σs^2 is the variance of the sample,

σp^2 is the variance of the sample.

Similarly, when chi-square is used as a non-parametric test for testing the goodness of fit or for testing the independence, the following formula is used:

$$\chi^2 = \sum \frac{(Oij - Eij)^2}{Eij}$$

Where *Oij* is the observed frequency of the cell in the i^{th} row and j^{th} column,

Eij is the expected frequency of the cell in the i^{th} row and j^{th} column.

Conditions for the chi square test:

For the chi-square test to be performed, the following conditions are to be satisfied:

1. The observations are to be recorded and collected on a random basis.
2. The items in the samples should all be independent.
3. The frequencies of data in a group should not be less than 10. Under such conditions, regrouping of items should be done by combining frequencies.
4. The total number of individual items in the sample should also be reasonably large, about 50 or more.
5. The constraints in the frequencies should be linear and not containing squares or higher powers.

SIGN TEST:

Sign test is preferred under the following situations

- Population density function is unknown
- Sample observations are paired
- Different pairs are observed under different variances and so paired –t test cannot be applied.
- Measurements are such that $d_i = x_i - y_i$ can be expressed as positive or negative sign.
- Variables are continuous and d_i's are independent.

Procedure:

Null hypothesis:

Two populations have identical distribution

ie., $f_x(.)=f_y(.)$,$P[(X-Y)>0]=1/2$, $P[(X-Y)<0]=1/2$

Alternative Hypothesis:

Two populations have different distribution

$$f_x(.) \neq f_y(.), \quad P[(X-Y)<0] \neq 1/2$$

Level of significance:

$\alpha = 0.05/\ 0.01/$ *anyother specificvalue*

Test Statistic

$E(U)=np=n(1/2)=n/2$

$V(U)=npq=n(1/2)(1/2)=n/4$

When the sample is large, we can have normal approximation

$$Z_0 = \frac{U-E(U)}{\sqrt{V(U)}} = \frac{2U-n}{\sqrt{n}} \sim N(0,1)$$

Critical Value:

When $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$ and $\alpha = 0.01$, $Z_{\alpha/2} = 2.58$

Inference:

If $|Z_0| > Z_{\alpha/2}$ we reject H_0 otherwise there is no reason to reject H_0

Note: $d_i=x_i-y_i$ is no sign attached and the pair is omitted from the sample size n and the reduced sample will have n-1 observation.

MANN-WHITENEY U TEST (RANK SUM TEST)

Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample.

Unlike the t-test it does not require the assumption of normal distributions. It is nearly as efficient as the t-test on normal distributions.

A Wilcoxon signed-rank test is a nonparametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution. A Wilcoxon rank sum test is a nonparametric test that can be used to determine whether two independent samples were selected from populations having the same distribution.

PROCEDURE

Null hypothesis

The populations have the same density function i.e., H_0: $f_x(.)=g_y(.)$

Alternative Hypothesis

The populations have the same density function i.e., H_0: $f_x(.)\neq g_y(.)$
Level of significance:

$\alpha = 0.05/\ 0.01/$ *anyother specificvalue*
Test statistic:

Combine the two sample and assign rank

R= sum of the ranks in second sample

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where, n_1=size of the first sample
n_2=size of the second sample

Inference:

If $|Z_0| > Z_{\alpha/2}$ we reject H_0 otherwise there is no reason to reject H_0

KRUSKAL-WALLIS TEST

The Kruskal-Wallis test is a nonparametric (distribution free) test, and is used when the assumptions of one-way ANOVA are not met. Both the Kruskal-Wallis test and one-way ANOVA assess for significant differences on a continuous dependent variable by a categorical independent variable (with two or more groups). In the ANOVA, we assume that the dependent variable is normally distributed and there is approximately equal variance on the scores across groups. However, when using the Kruskal-Wallis Test, we do not have to make any of these assumptions. Therefore, the Kruskal-Wallis test can be used for both continuous and ordinal-level dependent variables. However, like most non-parametric tests, the Kruskal-Wallis Test is not as powerful as the ANOVA.

Assumptions

1. We assume that the samples drawn from the population are random.

2. We also assume that the observations are independent of each other.

3. The measurement scale for the dependent variable should be at least ordinal.

Null hypothesis:

Null hypothesis assumes that the samples (groups) are from identical populations.

Alternative hypothesis:

Alternative hypothesis assumes that at least one of the samples (groups) comes from a different population than the others.

Level of significance:

α = 0.05/ 0.01/ *anyother specificvalue*

Kruskal-Wallis Formula

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

Where, n = sum of sample sizes for all samples,

c = number of samples,

Rj = sum of ranks in the jth sample, nj = size of the jth sample

Inference:

$H_{cal} > \chi^2(c-1)df$, we reject null hypothesis otherwise there is no reason to reject the null hypothesis.

Unit - 4

Statistical Decision Theory: Decision Types, Decision Framework and Decision Criteria

Every individual has to make some decisions or others regarding his every day activity. The decisions of routine nature do not involve high risks and are consequently trivial in nature. When busi-ness executives make decisions, their decisions affect other people like consumers of the product, shareholders of the business unit, and employees of the organisation.

Such decisions which affect other people in society involve a very careful and objective analysis of their consequences. The statistician's task is to split a decision problem in its simple components and study whether any or some of them are amenable to scientific treatment and therefore he tries to bring out a method by which these components can be woven into coherent and consistent decision of the problem as a whole.

He puts in an effort to detect if there is a behaviour pattern which is relevant to a particular decision process and whether it is consistent enough to be expressed in the form of a rule. The best way of finding out if there is any consistency is by fixing certain standards forejudging a particular situation.

These standards are fixed, based on past experiences or on the knowledge about past events. The business decision maker can make his work easier with the assistance of some standards and tools. Here the statistician's task is to evolve such standards and tools of measurement.

Decision Types:

The decision problems can be classified into five types and they are:

1. Decision Making Under Certainty:

There are a few problems where the decision maker gets almost complete information so that he knows all the facts about the state of nature and again which state of nature would occur and also the consequences of the state of nature. In such a situation, the problem of decision making is simple because the decision maker has only to choose the strategy which will give him maximum pay-off in terms of utility.

In cases where the strategy rows are normally very large and it is impossible even to list them, the technique of operational research like linear and non-linear programming and geometric

programming would have to be used to achieve the optimal strategy.

2. Decision Making Under Risk:

A problem of this kind arises when the state of nature is unknown, but based on the objective or empirical evidence, we can possibly assign probabilities to various states of nature. In a number of problems on the basis of historical data and past experience, we are able to assign probabilities to various states of nature. In such cases, the pay-off matrix is of immense help for reaching an optimal decision by assigning probabilities to various states of nature.

3. 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.

Such situations arise when a new product is introduced or a new plant is set up. Of course, even in such cases some market surveys are conducted and relevant information is gathered though it is not generally sufficient to indicate a probability figure for the occurrence of a particular state of nature.

4. 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.

5. Decision Making Under Conflict:

A condition of conflict is supposed to occur when we are dealing with rational opponent rather than the state of nature. The decision maker, therefore, has to choose a strategy taking into consideration the action or counter-action of his opponent. Brand compe-tition, military weapons, market place, etc. are problems which come under this category. The strategy choice is done as the basis of game theory where a decision maker anticipates the action of the opponent and then determines his own strategy.

Logical Decision Framework:

The main purpose of studying decision theory is to put the problem into a suitable logical frame-work. It includes identification of the problem. Personal perception and innovativeness are two essential things for the identification of the problem, and then generating alternative course of action and finally evolving criteria for evaluating the different alternatives to arrive at the best choice of action.

The basic components of a decision situation are the following:

1. Acts:

There are many alternative courses of action in any decision problem. But only some relevant alternatives need be considered. For instance, the business firm may decide to market its goods within the state or within the country or beyond the boundaries of the country. Here, there are three alternatives. There may be more such alternatives. The final choice of any one will depend upon the pay-offs from each strategy.

2. States of Nature:

There are those possible events or the states of nature which are uncertain but are vital for the choice of any one of the alternative acts. For example, the radio dealer does not know how many radios he will be able to sell. There is an element of uncertainty about it and for this reason he cannot decide how many radios to buy. This uncertainty is known as the state of nature or the state of the world.

3. Outcomes:

There is an outcome of the combination of each of the likely acts and possible states of nature. This is otherwise known as conditional value. The outcome has not much significant unless we calculate the pay-offs in terms of monetary gain or loss for each outcome. Thus outcome refers to the result of the combination of an act and each of the states of nature.

4. Pay-off:

The pay-off deals with the monetary gain or loss from each of the outcomes. It can be also in terms of cost-saving or lime-saving but the expression of pay-off should always be in quan-titative terms to help precise analysis. Therefore where the value of output is expressed directly in terms of gain expressed in money it is called pay-off. The calculation of pay-off or utility of each outcome has to be carefully done.

5. Expected Values of Each Act:

In practical business situation, there is risk and uncertainty. In the case of risk, the probability of each state of nature is known, and in uncertainty, it is unknown. Therefore, each likely outcome of an act has to be appraised with reference to the probability of occur-rence.

The expected value of a given act can be calculated by the following formula:

$$\sum_{j=1}^{n} PO_{ij} = P_1O_{i1} + P_2O_{i2} + \ldots\ldots\ldots P_nO_{in}.$$

Where P_1 toP_n refers to event probabilities of events E_1 to E_n and O_{ij}, the pay-offs of the outcome with the combination of each event and act. The expected value of each alternative is thus calculated with reference to probability assigned to each state of nature.

Choice of Decision Criteria:

The nature of the decision criteria would depend upon the type of the decision situation as follows:

1. Under Conditions of Certainty:

Under this condition; there is one pay-off for each strategy. The pay-off represents the degree of achievements of the objective, hence the largest pay-off is chosen and the corresponding strategy is selected.

2. Under Conditions of Risk:

Under condition of risk, there would be more than one state of nature but the probabilities of their occurrence are known on the basis of past experience. The strategy which gives the maximum pay-off is selected.

3. Under Conditions of Uncertainty:

Under conditions of uncertainty, we do not have a set of probabilities for the state of nature. Therefore, for each alternative only pay-offs or utilities are known. But nothing is known about the likelihood of each state of nature. The problem becomes more complex and the personality of the decision maker plays an important role in the selection of the strategy.

The following criteria are generally followed:

(1) Maximin Criterion:

The strategy which gives the highest minimum pay-off will be chosen. The basic rationale behind this criterion is that pessimism is not irrational under the state of uncertainty. The idea is to avoid the worst. In this criterion the motive of self preservation is considered.

(2) Maximax Criterion:

If the decision maker is an optimist by nature, he would always think that the state of nature would be the best from his point of view. He will find out the expected pay-off of all the strategies and pick up the strategy which gives the maximum pay-off out of all the strategies. He will always think that the state of nature would be favourable.

(3) Minimax Regret Criterion:

When the criterion is in terms of cost or regret then the deci-sion maker would choose the strategy in which the maximum regret or the cost is the lowest. The regrets have to be calculated for each act with reference to the best pay-off of the various alternative acts.

(4) Laplace Criterion:

Under this criterion when under conditions of uncertainty there is com-plete ignorance about the probability of the occurrences of state of nature, it is assumed that the prob-ability of the occurrence of each state of nature is the same. After this, the strategy which maximises the expected pay-off is chosen.

(5) Subjective Expected Utility Criterion:

Under this criterion not only the knowledge gath-ered from past experience but also the judgment of the decision maker is taken into account in assign-ing probabilities to the states of nature. In this criterion, therefore, the expected value will be calculated by taking into account the posterior probabilities in regard to the state of nature in place of prior prob-abilities given.

Unit – 5
Statistical Quality Control

Basics in Statistical Quality Control (SQC)

A quality control system performs inspection, testing and analysis to ensure that the quality of the products produced is as per the laid down quality standards. It is called "Statistical Quality Control". The statistical techniques are employed to control, improve and maintain quality or to solve quality problems. Statistics is the collection, organisation, analysis, interpretation and presentation of the data. It is based on law of large numbers and mathematical theory of probability. It is just one of the many tools necessary to solve quality problems it takes into account the existence of variation. Building an information system to satisfy the concept of „prevention" and „control" and improving upon product quality, requires statistical thinking.

SQC is systematic as compared to guess-work of haphazard process inspection and the mathematical statistical approach neutralizes personal bias and uncovers poor judgement. SQC consists of three general activities:

1. Systematic collection and graphic recording of accurate.
2. Analysing the data.
3. Practical engineering or management action, if the information obtained indicates significant deviations from the specified limits.

Modern techniques of SQC and acceptance sampling have an important part to play in the improvement of quality, enhancement of productivity, creation of consumer confidence and development of industrial economy of the country.

Relying itself on probability theory, statistical quality control plays an important role in total quality control. The following statistical tools are generally used for the purpose of exercising control, improvement of quality, enhancement of productivity, creation of consumer confidence and development of the country.

Frequency distribution

Frequency distribution is a tabulation or tally of the number of times a given quality characteristic occurs within the samples. Graphic representation of frequency distribution will show:

(a) Average quality
(b) Spread of quality
(c) Comparison with specific requirements
(d) Process capability

Control chart

Control chart is a graphical representation of quality characteristics, which indicates whether the process is under control or not.

1. Acceptance sampling

Acceptance sampling is the process of evaluating a portion of the product/material in

a lot for the purpose of accepting or rejecting the lot on the basis of conforming or not conforming to a quality specification. It reduces the time and cost of inspection and exerts more effective pressure on quality improvement than it is possibly by 100 percent inspection.

It is used when assurance is desired for the quality of material/products either produced or received.

2. Analysis of the data

It includes special methods, which include such techniques as the analysis of tolerance, correlation, analysis of variance, analysis for engineering design, problem solving technique to cause of troubles.

Statistical methods can be used in arriving at proper specification limits of products, in designing the products, in the purchase of raw material, semi-finished and finished products, manufacturing processes, inspection, packaging, sales and also after sales services.

Benefits of Statistical Quality Control

1. **Efficiency:** The use of SQC ensures rapid and efficient inspection at a minimum cost.
2. **Reduction of Scrap:** It uncovers the cause of excessive variability in manufactured produced – forecasting trouble before rejections occur and reducing the amount of spoiled work.
3. Moreover, the use of acceptance sampling in SQC, exerts more effective pressure for quality improvement than is possible by 100% inspection.
4. **Easy detection of faults:** In SQC after plotting the control charts $\bar{X}, R, p, c, u, np$. When the points fall above the upper control limits or below the lower control limit it is an indication of deterioration in quality, necessary corrective action is then taken. On the other hand, with 100% inspection, unwanted variations in quality may be detected at a stage when large amount of defective products have already been produced.
5. **Adherence to specification:** So long as a statistical control continues specifications can be accurately predicted for future, by which it is possible to assess whether the production processes are capable of producing the products with the given set of specifications.
6. Increases, output and reduces wasted machine and man hours.
7. Efficient utilization of personnel, machines and materials resulting in higher productivity.
8. Better customer relations through general improvement in product and higher share of the market.
9. SQC has provided a common language that may be used, by all three groups (designers, production personnel and inspectors) in arriving at a rational solution of mutual problems.
10. Elimination of bottlenecks in the process of manufacturing.
11. Point out when and where 100 percent inspection, sorting or screening is required.
12. Creating quality awareness in employees.

However, it should be emphasized that SQC is not a panacea for assuring product quality. It simply furnished “perspective facts” upon which intelligent management and engineering action can be based. Without such action, the method is ineffective. Even the application of standard procedures without adequate study of the process is extremely dangerous.

Meaning and Scope of Statistical Quality Control

Quality has become one of the most important consumer decision factors in the selection among competing products and services. The traditional definition of quality is based on the viewpoint that products and services must meet the requirements of those who use them, that is customer‟s risk).

Quality means fitness for use.

Quality is inversely proportional to variability.

There are two general aspects of fitness for use

1. Quality of Design
2. Quality of conformance

All products and services are produced in various in grades or levels of quality are international and consequently, the appropriate technical term in Quality of Design.

For example, all automobiles have as their basis objective providing safe transportation for the consumer. However, automobiles differ with respect to size, appointments, appearance and performance. These differences between the types of automobiles.

The quality of conformance is how well the product conforms to the specifications required by the design. Quality of conformance is influenced by a number of factors, including the choice of manufacturing process, the training and supervision of the workforce, the type of quality assurance system used (process control, tests, inspection activates etc.), the extent to which these quality assurance producers are followed and the modification of the workforce to achieve quality.

Dimensions of Quality

Garvin (1987) provides an eight components or dimensions of quality.

1. Performance
2. Reliability
3. Durability
4. Serviceability
5. Aesthetics
6. Features
7. Perceived Quality
8. Conformance to Standards.

1. Performance (Will the product do the intended job?)

Potential customers evaluate a product to determine if it will perform certain specific functions and determine how well it performs then.

2. *Reliability (How often does the product fail?)*

Complex products, such as many applications, automobiles or airplanes will require some repair over their service life. For industry in which the customer‟s view of quality is greater impacted by the reliability dimension of quality.

3. *Durability (How long does the product last?)*

This is the effective service life of the product. Customers want products that perform satisfactory over a long period of time.

4. *Serviceability (How easy is it to repair the products?)*

There are many industries in which the customer‟s view of quality is directly influenced by how quickly and economically a repair or routine maintenance activity can be accomplished.

5. *Aesthetics (What does the product look like?)*

This is the visual appeal of the product, often taking into account factors such as style, colour, shape and other censoring features.

6. *Features (What does the product do?)*

Usually customers associate high quality with products that have added features.

7. *Perceived Quality (What is the reputation of the company or its product?)*

Customers ready on the past reputation of the company concerning quality of its products. This reputation is divert influenced by failures of the products. This reputation is visible to the public and by how the customer is treated when a quality related problem with the product is reported.

8. *Conformance to Standards (Is the product made exactly as these designer intended?)*

A high quality product exactly meets the requirements of customers. Manufactured parts that do not exactly meet the designer‟s requirements can cause significant quality provides when they are used as the components of a more complex assembly.

Quality Improvement

Quality improvements are the reduction of variability in process and products.

Quality Characteristic

Every product possesses a number of elements that jointly describe what the uses or consumer thinks of as quality. There parameters are often called quality characteristic.

Quality Characteristic may be of several types

1. Physical: length, weight, voltage, viscosity
2. Sensory: taste, appearance, colour
3. Time Orientation: reliability, durability, serviceability

Quality Engineering

Quality engineering is the set of operational, managerial and engineering activities that a company uses to ensure that the quality characteristics of a product are at the nominal

of required levels.

A values of a measurements that corresponds to the desired values for that quality characteristic.

Note

Most organisations find it difficult and expressive to provide the customer with products that have quality characteristics that are always identical from unit to unit that match customer‟s expectations. A major season for this is variability there is a certain amount of variability in every product consequently. No two products are ever identical. So this variability can be described by some statistical methods. These methods plays a central role inequality improvements efforts.

- In the application of statistical method to quality. Engineering it is fairly typical to classify data on quality characteristic as either attributes or variables.
- Variable data are usually continuous measurements such as length voltage or viscosity.
- Attributes data are usually discrete data often taking the form of counts.

Continuous data involve counts (integer) for example, number of admissions, number of patients waiting, number of defective items.

Upper Specification Limit (USL)

Quality characteristic are often calculate relative to specifications for a manufactured product the specifications are the desired measurements for the quality characteristic on the components in the final product.

The larger allowable value for a quality characteristic is called USL.

Lower Specification Limit (LSL)

The smallest allowable value for a quality characteristic is called LSL.

Non-Conforming Product

A Quality Product is called non-conforming product.

Non-Conformity

A specific type of failure in the product is called a non-conformity.

Defective

A non-conforming product is considered which are non-conforming product is considered defective

Defects

If it has one or more defects, which are non-conformities that are series enough to significantly affect the safe or effective use of the product.

Quality Product

A quality product is defined as a product that meets the needs of the market place.

Quality and Improving Quality

Quality and improving quality has become an important business strategy for many

organization, manufactures, distributors, transportation companies, financial services organizations, health care providers and governments agencies.

Non-conforming unit

A non-conforming unit is a unit of product that does not satisfy one or more of the specifications for that product.

Difference between defect and defective

An item is said to be defective if it fails to conform to the specifications in any of the characteristic.

Each characteristic that does not meet the specifications is a defect.

An item is defective if it contains at least one defect. For example, if a casting contains undesirable hard spots, blow holes etc., the casting is defective and the hard spots, blow holes etc. are the defects, which make the casting defective.

The np chart applies to the number of defectives in subgroups of constant size. Whereas c chart applies to the number of defects in a subgroup of constant size.

Control Chart

A control chart is an important aid or statistical device used for the study and control of the repetitive processes. It was developed by A. Shewhart and it is based upon the fact that variability does exist in all the repetitive process.

A control chart is a graphical representation of the collected information. The information may pertain to measured quality characteristics of samples.

Uses of Control Charts

The most important use of a control chart is to improve the process. We have found that, generally,

1. Most processes do not operate in a state of statistical control.
2. Consequently, the routine and attentive use of control charts will identify assignable causes. If these caused can be eliminated from the process, variability will be reduced and the process will be improved.
3. The control chart will only detect assignable causes. Managements, operator and engineering action will usually be necessary to eliminate the assignable causes.

Reasons for the Control Charts are popular in Industries

1. Control charts are a proven technique for improving productivity.
2. Control charts are effective in defect prevention
3. Control charts prevent unnecessary process adjustment
4. Control charts provide diagnostic information
5. Control charts provide information about process capability.

Types of Control Charts

Basically control charts are classified into two types.

1. Variable Control Charts
2. Attribute Control Charts

Variable Control Chart

Variable control chart mainly consist of three charts namely

4. Mean (Average) control chart ($\bar{X}$)
5. Range control chart (R)
6. Standard deviation control chart (σ)

Attribute Control Chart

Attribute control chart mainly consist of three charts which are

1. Fraction defective chart (p)
2. Chart for defects (c)
3. Chart of number of defectives (np or d)

Construction of Average and Range $(\bar{X}$ and $\bar{R})$ Charts

The construction of Average and Range charts are based on measurements of produced goods. The measurements may be length, breath, area or volume. The selection of samples or subgroups is very essential. We select N samples in which each sample has n subgroups.

Let Xij be the jth observation of the ith sample (i=1,2,…,N; j=1,2,...n). From the measurable data we have to calculate sample statistics such as mean ($\bar{X}_i$) , Range (Ri) andStandard deviation (Si) of ith sample,

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \qquad (1)$$

$$R_i = \max_j X_{ij} - \min_j X_{ij} \qquad (2)$$

$$s_i = \sqrt{\frac{\sum_{i=1}^{n}(X_{ij} - X_i)^2}{n}} \qquad (3)$$

By using above statistics, we have to compute their averages

$$\bar{\bar{X}} = \frac{\sum \bar{X}_i}{N} \qquad (4)$$

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i \qquad (5)$$

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i \qquad (6)$$

After finding the averages of the statistics, we have to frame the control limits for framing a limits we have to calculate the value of 3σ, where σ is the standard deviations of the universe. The standard error of i^{th} subgroup is defined

$$SE(\overline{X}_i)=\frac{\sigma}{\sqrt{n}}, i=1,2,...,N \quad (7)$$

From the sampling distribution of range,

$$E(R)=\overline{R}=d_2\sigma \quad (8)$$

$$\Rightarrow \sigma = \frac{\overline{R}}{d_2} \quad (9)$$

where d_2 is a constant relative to the subgroup size.

Control chart for average is constructed when,

(i) μ and σ are unknown
(ii) μ and σ are known
(iii) R is unknown

Case (i): μ and σ are unknown

Suppose the population mean (μ) and population standard deviation (σ) are not given we have to calculate the sample mean and sample standard deviation or sample range.

Let $\overline{X}_1, \overline{X}_2, ..., \overline{X}_N$ be averages of subgroups. The average of averages is estimated by using the formula given in equation (4). The required 3σ control limits is defined as

$$E(\overline{X}_i) \pm 3SE(\overline{X}_i) = \overline{\overline{X}} \pm \frac{3\sigma}{\sqrt{n}} \quad \text{(using (7))}$$

$$= \overline{\overline{X}} + \frac{3}{\sqrt{n}}\frac{\overline{R}}{d_2} \quad \text{(using equation (9))}$$

$$= \overline{\overline{X}} + \left(\frac{3}{\sqrt{n}}d_2\right)\overline{R}$$

$$= \overline{\overline{X}} \pm A_2\overline{R} \quad (10)$$

Here A_2 is also a constant which is obtained from a table containing the factors for control charts and it depends on its subgroup size n.

The equation (10) is re-written as

$$UCL = \overline{\overline{X}} + A_2\overline{R} \tag{11}$$

$$LCL = \overline{\overline{X}} - A_2\overline{R} \tag{12}$$

By using equation (4), we draw a horizontal line parallel to x - axis and it represents the central line of the chart. $\therefore CL = \overline{\overline{X}}$. Similarly using equations (11) and (12), we draw dotted horizontal lines and they represent upper control limit and lower control limit respectively.

Using subgroup averages (equation (1)) we plot the points and infer that weather the process is in control or out of control. If the process is out of control do the necessary steps and draw the process is in control.

Case (ii): μ and σ are known

In the case of known population constants we have to calculate sample subgroup means and standard deviation. We have

$$E(\overline{X}_i) \pm 3SE(\overline{X}_i) = \mu \pm 3\frac{\sigma}{\sqrt{n}}$$

$$= \mu \pm \frac{3}{\sqrt{n}}\sigma$$

$$= \mu \pm A\sigma \tag{13}$$

where A is a constant depends on n. The required control limits are,

$$UCL = \mu + A\sigma \tag{14}$$

$$LCL = \mu - A\sigma \tag{15}$$

Case (iii): R is unknown

In the case of unknown range we have to construct $\overline{X}$ chart by using another statistic called standard deviation. The standard deviation of i^{th} subgroup s_i (given in (3)) is calculated and also the average of standard deviation is computed by using

$$\overline{S} = \frac{1}{N}\sum_{i=1}^{N} s_i$$

We know that the relation between average of the sample standard deviations and population standard deviations.

$$\overline{S} = C_2\sigma$$

$$\Rightarrow \sigma = \frac{\overline{S}}{C_2} \tag{16}$$

Control limits are defined as

$$E(\bar{X}_i) \pm 3SE(\bar{X}_i)$$

$$= \bar{\bar{X}} \pm 3\frac{\sigma}{\sqrt{n}}$$

$$= \bar{\bar{X}} \pm \frac{3}{\sqrt{n}}\frac{\bar{S}}{C_2}$$

$$= \bar{\bar{X}} \pm \left(\frac{3}{\sqrt{n}C_2}\right)\bar{S}$$

$$= \bar{\bar{\mathrm{X}}} \pm \mathrm{A}_1\bar{\mathrm{S}} \tag{17}$$

The required control limits are

$$\mathrm{UCL} = \bar{\bar{\mathrm{X}}} + \mathrm{A}_1\bar{\mathrm{s}} \tag{18}$$

$$\mathrm{LCL} = \bar{\bar{\mathrm{X}}} - \mathrm{A}_1\bar{\mathrm{s}} \tag{19}$$

$$\mathrm{CL} = \bar{\bar{\mathrm{X}}} \tag{20}$$

here A_1 is also a constant depends on n. By using the value of $\bar{\bar{X}}$, we draw a horizontal line parallel to x-axis and it is named as central line of the chart (20). The upper and lower control limits are drawn are dotted horizontal lines by using the equations (18) and (19) respectively.

The values of subgroup averages are plotted in the chart and verified the process is in control or not.

Control Limits for R Chart

Let X_{ij} be j^{th} observation in i^{th} subgroup (i=1, 2, …, N; j=1, 2, …, n). We have to find range for each subgroup. The range of i^{th} subgroup is,

$$R_i = Max(X_{ij}) - Min(X_{ij}) \quad (i = 1,2,...,N)$$

The average of ranges is

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i$$

Control limits for R chart are defined as

$$E(R) \pm 3SE(R)$$

$$= \bar{\mathrm{R}} \pm 3\mathrm{C}\bar{\mathrm{R}}$$

$$= (1 \pm 3C)\bar{R}$$

Here

$$\mathrm{E(R)} = \bar{\mathrm{R}}$$

$$SE(R) = \sigma R$$

We know that, $\sigma R = C.\ E(R)$

$$= C\overline{R}$$

$$SE(R) = \sigma R = C\overline{R}$$

Hence, the limits are

$$UCL = (1 + 3C)\overline{R} = D_4\overline{R}$$

and

$$LCL = (1 - 3C)\overline{R} = D_3\overline{R}$$

Here D_3 and D_4 are constants taken from the table containing the factors of control charts defending on the subgroup size.

If the subgroup size is less than 7, D_3 becomes zero. In this case we obtain only upper control limit. As in the case of $\overline{X}$ chart, we draw central line by using the value of $\overline{R}$ as a bold horizontal line and the upper control limit is a drawn as dotted horizontal line using in the value of $D_4\overline{R}$. After drawing in central line and control limit we plot the values of ranges. Finally we concluded that the process is in control or not. Suppose a subgroup size $n \geq 7$, lower control limit is also drawn as dotted horizontal line.

Control Chart for Standard Deviation σ Chart

Let X_{ij} be j^{th} observation of the i^{th} sample. Let S and σ be the standard deviations of sample and population. Control chart for standard deviation is constructed under the condition that when σ is unknown and σ is known.

Case (i): σ is unknown

Suppose the population standard deviation is not known, we have to calculate sample standard deviation for constructing standard deviation chart. Let S_i be the standard deviation of i^{th} subgroup

$$S_i = \sqrt{\frac{\sum (X_{ij} - \overline{X}_i)^2}{n}}$$

The average of the standard deviation is obtained as

$$\overline{S} = \frac{1}{N}\sum_{i=1}^{N} S_i$$

We know that, the relation between population standard deviation and average of the sample standard deviation is

$$\overline{S} = C_2\sigma$$

The control limits of standard deviation chart are

$$= E(S) \pm 3SE(S)$$

$$= \bar{S} \pm 3(C_3\sigma)$$

$$= \bar{S} \pm 3C_3\left(\frac{\bar{S}}{C_2}\right)$$

$$= \bar{S} \pm \left(3\frac{C_3}{C_2}\right)\bar{S}$$

Hence,

$$UCL = \left(1+3\frac{C_3}{C_2}\right)\bar{S} = B_4\bar{S}$$

$$LCL = \left(1-3\frac{C_3}{C_2}\right)\bar{S} = B_3\bar{S}$$

After drawing central line and control limits, plot the points by using subgroup standard deviation and draw suitable conclusion. If the value of lower control limit is negative, we take the value as zero. Here the values of D_3 and D_4 are taken from the pre-assigned table depending on subgroup size n.

Case (ii): σ is known

Suppose the population standard deviation is known, there is no need to compute sample statistic, now we define the following:

$E(S) = C_2\sigma$ and

$SE(S) = C_3\sigma$.

The control limits of standard deviation chart are

$$= E(S) \pm 3SE(S)$$

$$= C_2\sigma \pm 3C_3\sigma$$

$$= (C_2 \pm 3C_3)\sigma$$

Hence,

$$UCL = (C_2 + 3C_3)\sigma = B_2\sigma$$

$$LCL = (C_2 - 3C_3)\sigma = B_1\sigma$$

$$Central\ Line = C_2\sigma.$$

2.1. Introduction

Average and Range charts are very powerful statistical techniques to point out the troubles in the production process. Variable charts are drawing based on measurable units. They do not help to study the quality characteristics of the products. For analysing the quality characteristics of the products, Shewhart has established another set of control charts which are called attribute control charts and they are given below:

1. p – chart: control chart for fraction defective
2. np – chart: control chart for number of defectives
3. c – chart: control chart for number of defects
4. u – chart: control chart for number of defects in variable sample size.

2.2. Control chart for fraction defectives (p – chart)

In production process, inspection is carried out for identified conformity and non-conformity units. Let d be the number of non-conformity is a sample of size n. Let p be the sample proportional defective and it is defined as the ratio of the number of non-conformity units to the total number of units inspected. That is,

$$P = \frac{d}{n}$$

The corresponding population proportion is taken as P according to Binomial law the number of non-conformities,

$$d \sim B(n, P)$$
$$\Rightarrow E(d) = nP$$

and v(d) = nPQ, where Q = 1-P for constructing control limits,

$$\mathrm{v(p)} = \frac{1}{\mathrm{n}^2}\,\mathrm{v(d)}$$

$$= \frac{1}{n^2} nPQ$$

$$v(p) = \frac{PQ}{n}$$

Control limits are defined as

$$E(p) = E\left(\frac{d}{n}\right)$$

$$= \frac{1}{n} E(d)$$

$$= \frac{1}{n} np$$

$$E(p) = P$$

$$E(p) \pm 3SE(p)$$

$$= P \pm 3\sqrt{\frac{PQ}{n}}$$

$$= P \pm A\sqrt{PQ}, \quad \text{where} \quad A = \frac{3}{\sqrt{n}}$$

Case (i): Standards are known

Let P' be the value of P then the control limits are,

$$P' \pm A\sqrt{P'(1-P')} \ \ and \ \ CL = P'$$

Case (ii): Standards are unknown

(a) Sample size varies:

Consider k samples of different sizes $n_1, n_2, \ldots, n_k$ the sample units are inspected and the number of defective units are obtained as $d_1, d_2, \ldots, d_k$ respectively. The fraction defective is defined as,

$$P_i = \frac{d_i}{n_i}, \quad i = 1,2,\ldots,k$$

The average with the fraction defectives in computed as

$$\overline{P} = \frac{\sum p_i}{k}$$

In other words, $\overline{p}$ is computed as,

$$\bar{p} = \frac{\sum d_i}{\sum n_i} = \frac{\sum n_i p_i}{\sum n_i}$$

$\bar{p}$ indicates the central line. The control limits are

$$\bar{p} \pm 3\sqrt{\bar{p}(1-\bar{p})/n_i}$$

Plot the value of p_i in the chart and in found that the process is in control or not. If any point falls outside the control limits, remove that point and construct revised control chart for the remaining observations.

(b) Sample size is fixed

Consider k samples of equal size n inspect each and every sample and the defective units are denoted as $d_1, d_2, \ldots, d_k$ the fraction defective of i^{th} sample defined as the ratio of the number of defective units in the i^{th} sample (d_i) to the sample size.

$$p_i = \frac{d_i}{n}, \; i = 1,2,\ldots,k$$

The average of the fraction defectives is computed by using the relation

$$\bar{p} = \frac{\sum d_i}{k(n)}$$

$\bar{p}$ indicates the central line. Then the control limits are,

$$\bar{p} \pm 3\sqrt{\bar{p}(1-\bar{p})/n}$$

After drawing the control limits and central line, we have to plot the values of p_i in the chart and draw the conclusion whether the process is in control or not. Suppose any point falls outside the control limits, we remove that points and construct the revised control chart.

2.3. Control chart for number of defectives (np or d – chart)

In a production process, the number of non-conformities are collected after expecting the sampling units. Let d be the number of defective units. It is noted that the proportion of defectives is given by the relation p = d/n, this implies that,

$$E(p) = P$$

The mean and variance of non-conformities units for obtained in terms of population proportion

$$\text{Mean} = E(d) = nP$$

$$\text{Variance} = v(d) = nP(1-P)$$

$$\Rightarrow SE(d) = \sqrt{nP(1-P)}$$

Hence, central line = nP.

Control limits are

$$E(d) \pm 3SE(d)$$

$$= nP \pm 3\sqrt{nP(1-P)}$$

$$\Rightarrow UCL = nP + 3\sqrt{nP(1-P)}$$

$$LCL = nP - 3\sqrt{nP(1-P)}$$

After drawing central line and control limits, we have to plot the values of number of defectives (d_i). Suppose any point falls outside the limits, we have to find out the reasons and rectified the causes.

2.4. Control chart for number of defects per unit (c – chart)

In any production process inspections is carried out for supporting standard items and bad items (defective items) by considering defectives, p chart and np chart are appropriate charts.

Suppose one may try to study about the defects per unit, another control chart, namely c chart is used. According to probability law, number of defects per unit follows Poisson distribution. Generally the population average for number of defects is denoted as C.

$$X \sim P(C)$$

$$P(x) = e^{-C}\frac{C^x}{x!}$$

Here, we known that X is the number of non-conformities and C is the parameter.

Case (i): Standards known

Let C be the average of number of defects in the population. C is a Poisson parameter and the control limits are defined as,

$$C \pm 3\sqrt{C}$$

C indicates the central line.

Case (ii): Standards unknown

From industrial products, collect defective pieces. Inspect each and every defective pieces and noted the number of defects per product. Let $c_1, c_2, \ldots, c_n$ be the number of defects and its average is

$$\overline{C} = \frac{\sum_{i=1}^{n} C_i}{n}$$

This $\overline{C}$ is also follows Poisson distribution and it represents the central line of the chart. We required control limits are

$$\overline{C} \pm 3\sqrt{\overline{C}}$$

After drawing central line and control limits, plot the values of $c_1, c_2, \ldots, c_n$ and conclude that the process is in control or not.

2.5. Control chart for number of defects in variable sample size (u – chart)

This control chart is different from c chart. In c - chart, each and every unit is taken as a sample. Suppose we have many number of defective units. There is no possibility to construct the c-chart, we apply u chart, which is otherwise called c chart for variable size.

Let $n_1, n_2, \ldots, n_k$ be sizes of different samples (or) the sampling units are considered as defective units. Observe the number of defective in each unit and count the total defects in each sample. Let $c_1, c_2, \ldots, c_k$ be the number of defects of the above samples $n_1, n_2, \ldots, n_k$ respectively. The ratio of the number of defects (C_i) to the number of defective units (n_i) is taken as,

$$u_i = \frac{c_i}{n_i}, \quad i = 1,2,\ldots,k$$

Now compute the average of $u_i^{'}s$

$$\overline{u} = \frac{\sum_{i=1}^{k} u_i}{k}$$

It is noted that $\overline{u}$ follows Poisson distribution. According to a standard error of sample mean, we have

$$SE(\overline{u}) = \sqrt{\frac{\overline{u}}{n_i}}$$

The control limits are defined as

$$E(u_i) \pm 3SE(u_i)$$

$$\overline{u} \pm 3\sqrt{\frac{\overline{u}}{n_i}}$$

Here, $\bar{u}$ indicates the central line. After drawing control limits, plot the values of $u_i^{'}s$ corresponding to sample numbers.

Comparison of $\bar{X}$ and R chart with p chart

1. p chart is attribute control chart, i.e. for quality characteristic that can be classified as either conforming or nonconforming to the specifications. For example, dimensions checked by Go-No-Go gauges. Whereas, $\bar{X}$ and R chart is used for quality characteristic that can be measured and expressed in numbers.
2. The cost of collecting the data for p chart is less than the cost of collecting the data for $\bar{X}$ and R chart. For example, 10 shafts might be inspected with "go-no-go" gauge in the time required to measure a single shaft diameter with a micrometer. Secondly, p chart uses data already collected for other purpose.
3. The cost of computing and changing may also be less since p chart can be applied to any number of quality characteristics observed on one article. But separate $\bar{X}$ and R chart is required for each measured quality characteristic, which may be impracticable and uneconomical.
4. p chart is best suited in cases where inspection is carried out with a view to classifying an article as accepted or rejected. $\bar{X}$ and R charts are best suited for critical dimensions.
5. p chart though discloses the presence of assignable causes of variations, it is not as sensitive as $\bar{X}$ and R chart. For actual diagnosis of causes of troubles, $\bar{X}$ and R charts are best, still p chart can be used effectively in the improvement of quality.
6. The sample size is generally larger for p chart than for $\bar{X}$ and R chart. The variations in the sample size influence the control limits much more in $\bar{X}$ and R charts than in p chart.
7. The control chart for fractions defective provides management with a useful record of quality history.

Purpose of the p- chart

Because of the lower inspection and maintenance costs of p charts, they usually have a greater area of economical applications than do the control charts for variables. A control chart for fraction defective may have any one or all of the following purposes:

1. To discover the average proportion of defective articles submitted for inspection, over a period of time.
2. To bring to the attention of the management, any changes in average quality level.
3. To discover, identify and correct causes of bad quality.
4. To discover, identify and correct the erratic causes of quality improvement.
5. To suggest where it is necessary to use $\bar{X}$ and R charts to diagnose quality problems.
6. In a sampling inspection of large lots of purchased articles.

Basis of control limits on c- chart

Control limits on c chart are based on Poisson distribution. Therefore, two conditions must be satisfied.

- The first condition specifies that the area of opportunity for occurrence of defects should be fairly constant from period. The expression may be in terms of defects per unit being employed. For example, while inspecting the imperfections of a cloth it is necessary to take some units area say 100 square meters and count the number of imperfections per unit (i.e. per 100% square meters). Another example, may be number of point's imperfections per square area of painted surface. However, c chart need not be restricted to a single type of defect but may be applicable for the total of many different kinds of defects observed on any unit.
- Second condition specifies that opportunities for defects are large, while the changes of a defect occurring in anyone spot are small. For example, consider a case in which the product is large unit, say a ratio, which can have defects at number of points although any one point has only few defects.

Comparison between attribute charts and variable charts

Choosing a particular type of chart is a question of balancing the cost of collecting and analysing the type of data required to plot the chart against usefulness of the conclusions that can be drawn from the chart.

	Variable Charts	Attribute Charts
1.	**Example** $\overline{X}$, R, σ charts.	p, np, c, u charts.
2.	**Type of Data Required** Variables data (Measured values of characteristics).	Attribute data (using Go-No-Go gauges).
3.	**Filed of Application** Control of individual characteristics.	Control of proportion of defectives or number of defects or number of defects per unit.
4.	**Advantages** ➢ Provides maximum utilisation of information available from data. ➢ Provides detailed information on process average and variation for control of individual dimensions.	➢ Data required are often already available from inspection records. ➢ Easily understood by all persons. Since, it is more simple as compared to $\overline{X}$ and R chart. ➢ It provides overall picture of quality history.
5.	**Disadvantages** ➢ They are not easily understood unless training is provided. ➢ Can be confusion between control limits and specification limits.	➢ They do not provide detailed information for control of individual characteristic. ➢ They do not recognise different degree of defectiveness.