DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

GEE6201 – RESEARCH METHODOLOGY AND IPR

MODULE II

DATA COLLECTION, ANALYSIS AND INTERPRETATION OF DATA

2.1 Application of Computer in Research

The most emerging tool in the research process is computer. Computer is an essential tool for research, whether for academic purpose or for commercial purpose. Computers play a major role today in every field of scientific research from genetic engineering to astrophysics research. It led the way to a globalized information portal that is the World Wide Web. Using WWW, researcher can conduct research on massive scale. Various programs and applications have eased the way into computing of research process.

There are many reasons why computers are so important in scientific research and here are some of the main reasons:

SPEED: computer can process numbers and information in a very short time. So researcher can process and analyze data quickly. By saving time researcher can conduct further research. A calculation that may take a person several hours to process will take computer mere minutes, if not seconds.

ACCURACY: Computer is incredibly accurate. Accuracy is very much important in scientific research. Wrong calculation could result an entire research or project being filled with incorrect information.

ORGANIZATION: We can store millions of pages of information by using simple folders, word processors & computer programs. Computer is more productive & safer than using a paper filing system in which anything can be easily misplaced.

CONSISTENCY: computer cannot make mistakes through “tiredness” or lack of concentration like human being. This characteristic makes it exceptionally important in scientific research.

2.1.1 COMPUTER IN THE RESEARCH PROCESS

Research process consists of series of actions or steps necessary to effectively carry out research and the desired sequencing of the following steps.

1. Role of Computer in Conceptual Phase

The conceptual phase consists of formulation of research problem, extensive literature survey, theoretical frame work and developing the hypothesis. Use of computers in extensive literature review: computers help for searching the literatures (for review of literature) and bibliographic reference stored in the electronic database of the World Wide Web's. It can thus be used for storing relevant published articles to the retrieved whenever needed. This has the advantage over searching the literatures in the form of books, journals and other newsletters at the libraries which consume considerable amount of time and effort.

2. Role of Computers in Design and Planning Phase

This phase consists of research design preparation and determining sample design. Design and planning phase also consists of population, research variables, sampling plan, reviewing research plan and pilot study. Role of Computers for Sample Size Calculation: Several software's are available to calculate the sample size required for a proposed study. The standard deviation of the data from the pilot study is required for the sample size calculation.

3. Role of Computers in Data collection phase

This Empirical phase consists of collecting and preparing the data for analysis: In research studies, the preparation and inputting data is the most labor-intensive and time consuming aspect of the work. Typically the data will be initially recorded on a questionnaire or record for suitable for its acceptance by the computer. To do this the researcher in conjunction with the statistician and the programmer, will convert the data into Microsoft word file or excel spreadsheet or any statistical software data file. These data can be directly opened with statistical software's for analysis.

Data collection and Storage: The data obtained from the subjects are stored in computes are word files or excel spread sheets or any statistical software data file. This has the advantage of making necessary corrections or editing the whole layout of the tables if needed, which is impossible or time consuming incase of writing in papers. Thus, computers help in data entry, data editing, data management including follow up actions etc. computers also allow for greater

flexibility in recording the data while they are collected as well as greater ease during the analysis of these data. Examples of editors are WordPad, SPSS data editor, word processors, others like ultraedit etc.

Data exposition: Most researchers are anxious about seeing the data: what they look like; how they are distributed etc. you can also examine different dimension of variables or plot them in various charts using a statistical application.

4. Role of Computers in Data Analysis

This phase consist of the analysis of data, hypothesis testing and generalizations and interpretation. Data analysis phase mainly consist of statistical analysis of the data and interpretation of results. Data analysis: many software's are now available to perform the mathematical part of the research process i.e. the calculations using various statistical methods. Software's like SPSS and spreadsheets are the widely used. They can be like calculating the sample size for a proposed study, hypothesis testing and calculating the power of the study. Familiarity with any one package will suffice to carry out the most intricate statistical analysis. Computers are useful not only for statistical analysis, but also to monitor the accuracy and completeness of the data as they are collected. Software's also display the results in graphical chart or in graph form.

5. Role of Computer in Research Publication

This phase consists of preparation of the report or presentation of the results, i.e., formal write-up of conclusions reached. This is the research publication phase. The research article, research paper, research thesis or research dissertation is typed in word processing software and converted to portable data format (PDF) and stored and/or published in the world wide web. Online sites are available through we can convert our word file into any format like html, pdf etc. Various online applications are also available for this purpose. Even we can prepare our document using online word processing software and can store/edit/access it from anywhere using internet.

2.1.2 TOOLS AND APPLICATIONS USED IN THE RESEARCH PROCESS

Statistical Analysis Tool: (SPSS): SPSS is the most popular tool for statisticians. SPSS stands for Statistical Package for Social Sciences. It provides all facilities like data analysis like following and many more.

- Provides Data view & variable view

- Measures of central tendency & dispersion
- Statistical inference
- Correlation & Regression analysis
- Analysis of variance
- Non parametric test
- Hypothesis tests: T-test, chi-square, z-test, ANOVA, Bipartite variable….
- Multivariate data analysis
- Frequency distribution
- Data exposition by using various graphs like line, scatter, bar, histogram, pie chart…

2.2 SPREADSHEET TOOL

Data Analysis Tool: SPREADSHEET PACKAGES: A spreadsheet is a computer application that simulates a paper worksheet. It displays multiple cells that together make up a grid consisting of rows and columns, each cell containing either alphanumeric text or numeric values. Microsoft Excel is popular spreadsheet software. Others spreadsheet packages are Lotus 1-2-3 Quattro Pro, Javeline Plus, Multiplan, VisiCalc, Supercalc, Plan Perfect etc.

OTHER STATISTICAL TOOLS: SAS, S-Plus, LISREL, Eviews etc.

WORD PROCESSOR PACKAGES: A word processor (more formally known as document preparation system) is a computer application used for the production (including composition, editing, formatting, and possibly printing) of any sort of printable material. The word processing packages are Microsoft Word, WordStar, Word perfect, Software, Akshar (Gujarati), AmiPro etc.

PRESENTATION TOOL

PRESENTATION SOFTWARE: A presentation program is a computer software package used to display information, normally in the form of a slide show. It typically includes three major functions: an editor that allows text to be inserted and formatted, a method for inserting and manipulating graphic images and a slideshow system to display the content. The presentation packages are Microsoft PowerPoint, Lotus Freelance Graphics, Corel Presentations, Apple keynote etc.

DATABASE MANAGEMENT PACKAGES (DBMS): Database is an organized collection of information. A DBMS is a software designed to manage a database.

Various Desktop Databases are Microsoft Access, Paradox, Dbase or DbaseIII+, FoxBase, Foxpro/ Visual Foxpro, FileMaker Pro Commercial Database Servers that supports multiuser are Oracle, Ms-SQL Server, Sybase, Ingres, Informix, DB2 UDB (IBM), Unify, Integral, etc. Open source Database packages are MySQL, PostgreSQL, Firebird etc.

BROWSERS: A web browser is a software application which enables a user to display and interact with text, images, videos, music, games and other information typically located on a Web page at a website on the World Wide Web or a local area network. Examples are Microsoft Internet explorer, Mozilla firefox, Opera, Netscape navigator, Chrome (google browser), Safari.

TOOLS THROUGH INTERNET: Search Engines (to search the information) Google (popular search engine); Yahoo! ; Webcrawler; Excite; Altavista;

Online Data/Documentation Management: (to manage your documents online); Dropbox; Google Drive; Google Docs; MS Sky Drive (free); Microsoft 365 (paid version);

Online Data Collection: (To collect data online from different users); Online forms; Online questionnaires; Online surveys;

Collaboration tools: Skype : Voice and video conferencing Google Hangouts :Voice and video conferencing Modern Research tools Zotero Evernote Modern electronic research tools, like Zotero and Evernote, make the collection of research data, and collaboration between colleagues possible, which that in the past would have been difficult, expensive, or even impossible. They also save large amounts of time citing and creating bibliographies. Evernote allows the user to capture digital content, including web pages, PDF files or snippets of web pages, organize them, annotate them, share them, publish them and search them.

2.3 BASIC PRINCIPLES OF STATISTICAL COMPUTATION

Some of the most important principles/applications used in scientific research are data storage, data analysis, scientific simulations, instrumentation control and knowledge sharing.

Data Storage Experimentation is the basis of scientific research. Every experiment in any of the natural sciences generates a lot of data that needs to be stored and analyzed to derive important conclusions, to validate or disprove

hypotheses. Computers attached with experimental apparatuses, directly record data as it's generated and subject it to analysis through specially designed software. Data storage is possible in SPSS data file, lotus spreadsheet, excel spreadsheet, ASCII/DOS text file etc.

Data Analysis: Analyzing tons of statistical data is made possible using specially designed algorithms that are implemented by computers. This makes the extremely time-consuming job of data analysis to be a matter of a few minutes. In genetic engineering, computers have made the sequencing of the entire human genome possible. Data from different sources can be stored and accessed via computer networks set up in research labs, which makes collaboration simpler.

Scientific Simulations: One of the prime uses of computers in pure science and engineering projects is the running of simulations. A simulation is a mathematical modeling of a problem and a virtual study of its possible solutions. Problems which do not yield themselves to experimentation can be studied through simulations carried out on computers. For example, astrophysicists carry out structure formation simulations, which are aimed at studying how large-scale structures like galaxies are formed. Space missions to the Moon, satellite launches and interplanetary missions are first simulated on computers to determine the best path that can be taken by the launch vehicle and spacecraft to reach its destination safely.

Instrumentation Control: Most advanced scientific instruments come with their own on-board computer, which can be programmed to execute various functions. For example, the Hubble Space Craft has its own onboard computer system which is remotely programmed to probe the deep space. Instrumentation control is one of the most important applications of computers.

Knowledge Sharing Through Internet: Through internet, computers have provided an entirely new way to share knowledge. Today, anyone can access the latest research papers that are made available for free on websites. Sharing of knowledge and collaboration through the Internet has made international cooperation on scientific projects possible. Through various kinds of analytical software programs, computers are contributing to scientific research in every discipline, ranging from biology to astrophysics, discovering new patterns and providing novel insights. When the work in neural network based artificial

intelligence advances and computers are granted with the ability to learn and think for them, future advances in technology and research will be even more rapid.

Use of computer in research in science is so extensive that it is difficult to conceive today a scientific research project without computer. Many research studies cannot be carried out without use of computer particularly those involving complex computations, data analysis and modeling. Computer in scientific research is used at all stages of study-from proposal/budget stage to submission/presentation of findings.

2.4 IMPORTANCE OF STATISTICS IN RESEARCH

Statistics in Engineering

Engineers are finding increasing practical usefulness for some of the basic concepts of the branch of mathematics called probability and statistics. While this unit is devoted to engineering experimentation involved in research, the statistical concepts and methods discussed are also applicable to other areas. In order to give some practically useful statistical tools, the approach adopted here is to concentrate on general concepts and specific calculation method for commonly occurring problems without emphasis on proofs.

Suppose the researcher buy a carbon steel rod, the manufacturer says that modulus of elasticity (E) of the steel is 206.8 GPa, while actual testing of specimen of several rods, the E value is found varying between 205.7 GPa to 207.9 GPa. Hence, the statistics may be helpful in deciding the probability of the E value falling below a chosen value. Similarly when a specimen is prepared using machining with a value of 25mm, diameter, in actual practice when several specimens are prepared, the variation between specimens could be from 24.5mm to 25.5mm. This uncertainty will produce an associated uncertainty in lts.

In the engineering experimental analysis, one can find that statistical methods have much important application such as specifying the uncertainty of measured data, planning of experiments to get the maximum amount of significant data with the least effort and expense. Moreover the hypothesis could be tested rationally to decide the probability of their being true or false.

Every measurement we make have some inaccuracy and this can be broken down into so called systematic and random or chance errors. Systematic error repeats themselves if the measurement is repeated and could be removed by calibrating the instrument and the experimental set up. Random errors do not behave in this way rather they exhibit scatter. The best that we can do is to statistically estimate what their largest values might reasonably be. We can then quote the measurement as, say 25.20 + .06 to indicate that we are, say 95% sure that the value is somewhere between 25.14 and 25.26.

Many of the statistical methods were originally developed for use in biological and social sciences, where the interrelationship between variables are not clear-cut as they often are in engineering. In many circumstances the engineers and physicists have to deal with poorly defined situations or forced to use data that have low precision. Proper application of statistical analysis can greatly help them in such situations.

Scope of Statistical Analysis

Often physicists and engineers run experiments but their variables are subject to much more precise control than those in biological and social sciences. Statistics have been found very useful, indeed necessary, in sorting out the subtle effects of various interacting variables in situations of this sort. As engineering is becoming more involved with biological and social system, the need for engineers for some statistical background increases. Even in pure technical problems, statistics has been found very helpful in giving a rational basis for reaching judgments when the interaction of variables is subtle.

There are several uses of statistical tools. Some of the important applications of statistical tools in engineering include,

- At the most basic level, we are interested in the analysis of experimental data so that the results can be unequivocally described by appropriate statistical parameters.
- In statistical inference, we use statistics for making reliable decisions utilizing tests of hypothesis and confidence limits. Hypothesis tests are used to determine whether there is a significant difference between the

characteristics of an observed set of data and a proposed mathematical model of the data.

- Confidence limits allow us to determine the range in which the true characteristic of a population is likely to lie.
- Analysis of variance is a test for the equality of means and / or variances of group of observations, such as the means resulting from different experimental procedures
- The statistical design of experiments helps the researcher to collect data in a more efficient and economic way that gives value of information per experiment optimally.
- Regression analysis is another tool used in experimentation to determine whether there is a relation between two or more variables and this can also be used for prediction purposes.

2.5 CONCEPT OF PROBABILITY

Probability Definition

Probability is a number between 0 and 1 related to a given event. If the event is absolutely certain, its probability is 1 and if the event is impossible, its probability is zero. Three different definitions of probability are in common use, each one is appropriate for certain types of applications.

Classical definition

If an event can occur in ‘N’ equally likely and mutually exclusive ways, and if ‘n’ of these ways have an attribute ‘A’, than the probability P(A) of the occurrence of ‘A’ is defined to be n/N. This definition is applicable to study of games of chance. For example, if there are 4 aces in a deck of 52 cards, the probability of drawing on ace in one try is 4/52 = 0.777.

Frequency definition

If an experiment is carried out ‘N’ times and an event ‘A’ occurs ‘n’ times, then if we let N approach infinity, the limit of n/N is defined to be the probability P(A) of an event ‘A’. This is the most popular definition and allows treatment of many practical problems in which classical definition could not be applied.

For example, it is neither desirable nor economical to test all the finished components in a continuous production system. Theoretically, it is not possible to study an infinite population; hence a sample size large enough to be reliable but small enough to be economical is chosen to make decision about the quality of components produced.

Subjective definition

In this the probability P (A) of a proposition "A" is a measure of the "degree of belief" one holds in the proposition. This is perhaps the broadest definition and a necessary one. Suppose we are trying to choose between two alternative strategies, each of which is likely to produce certain results, and we cannot experimentally try out each case, then one has to rely on 'expert judgment' to assign numerical probability in such cases. (eg: betting in horse race, dropping a bomb or not dropping on enemy target). Here again, the classical and frequency interpretation are useless and a subjective judgment is necessary.

Irrespective of the interpretation one puts on probability, once a number has been chosen, any further calculations are independent of the definition.

Probability laws

A special and precise system of language and notation is used in probability theory. Two events A and B are said to be 'independent' if the occurrence of one event (say A) has no effect on the probability of occurrence of the other (say B). The two events A and B are called 'mutually exclusive' if one of them happens, the other cannot, i.e. they have no elements in common.

Multiplicative low:

1. If A, B, C …….. are independent events then the probability that all events occur, is called joint probability, is the product of their respective probabilities.

   P (ABC…….) = P(A) P(B) P(C) ……..
2. If A and B are mutually exclusive than P (AB) = 0. (or) P (ABC…..) = 0.
3. If two events A and B are not independent, then

   [P(A and B) = P (AB) = P(A). P (B/A) = P (B) P (A/B). Where,]

P (B/A): conditional probability of B with respect to A. that is, probability of B occurring, assuming that A has occurred.

P (A/B: Conditional probability of A with respect B. that is, probability of A occurring, assuming that B has occurred.

Whether events are independent, or not are not always obvious or clear cut in practical applications. By critically examining the situation, one has to decide whether the events are dependent or independent. When the correct interpretation is not obvious after brief consideration, more theoretical and/or experimental studies may be helpful in reaching a decision.

Additive law:

The probability that event 'A' or 'B' occur is given by the relation,

P (A and / or B) = P (A) + P (B) – P(AB)

The events A and B need not be independent, so long as we know their joint probability P (AB) i.e., not mutually exclusive events.

If A and B are mutually exclusive (means that if one of them happens, the other cannot), then P (AB) = 0. In that case,

P (A+B) = P (A and/or B) = P (A) + P (B) (or) P (A+B) = P (A) +P (B)

This can be generalized to any number of events by a process of continued replication,

P (A and/or B and/or C) = P (A+B+C) = P(A)+P(B)+P(C) - P(AC)-P(BC)-P(ABC).

Example:

A machine produces parts that are either good (90%), slightly defective (2%), or obviously defective (8%). Produced parts get passed through an automatic inspection machine, which is able to detect any part that is obviously defective and discard it. What is the quality of the parts that make it through the inspection machine and get shipped?

Solution:

Let G (resp., SD, OD) be the event that a randomly chosen shipped part is good (resp., slightly defective, obviously defective). We are told that $P(G) = .90$, $P(SD) = 0.02$, and $P(OD) = 0.08$.

We want to compute the probability that a part is good *given* that it passed the inspection machine (i.e., it is *not* obviously defective), which is

$$P(G|OD^c) = \frac{P(G \cap OD^c)}{P(OD^c)} = \frac{P(G)}{1 - P(OD)} = \frac{.90}{1 - .08} = \frac{90}{92} = .978$$

Baye's rule:

Baye's method allows us to modify a probability estimate as additional information becomes available. If B is an event, which may be possible by k mutually exclusive and exhaustive ways of Ai, i = 1,2…..k, then

$$P\left(Ai/B\right) = \frac{p(Ai)p\left(B/Ai\right)}{\sum p(Ai)p\left(B/Ai\right)}$$

This is known as posterior probability

Example:

A factory production line is manufacturing the CCD using three machines, A, B and C. Of the total output, machine A is responsible for 25%, machine B for 35% and machine C for the rest. It is known from previous experience with the machines that 5% of the output from machine A is defective, 4% from machine B and 2% from machine C. A CCD is chosen at random from the production line and found to be defective. What is the probability that it came from

(a) Machine A (b) machine B (c) machine C?

Solution: Let D = {CCD is defective}, A = {CCD is from machine A},

B = {CCD is from machine B}, C = {CCD is from machine C}.

Through definition, that P(A) = 0.25, P(B) = 0.35 and P(C) = 0.4. Also P(D|A) = 0.05, P(D|B) = 0.04, P(D|C) = 0.02.

A statement of Baye's theorem for three events A,B and C is

$$\begin{aligned} P(A|D) &= \frac{P(D|A)P(A)}{P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{0.05 \times 0.25}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4} \\ &= 0.362 \end{aligned}$$

Similarly

$$\begin{aligned} P(B|D) &= \frac{0.04 \times 0.35}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4} \\ &= 0.406 \end{aligned}$$

$$\begin{aligned} P(C|D) &= \frac{0.02 \times 0.4}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4} \\ &= 0.232 \end{aligned}$$

Example: At a certain circuit test, 4% of Narrow band amplifiers are with the gain factor greater than 6 and 1% of Wide band amplifiers are with the gain factor greater than 6. The total no. of amplifier circuits are divided in the ratio 3:2 in favor of Wide band amplifiers. If an amplifier is selected at random from all those with the gain factor greater than 6, what is the probability that the amplifier is made of Wide band?

Solution:

Let F={Amplifier with Narrow band }, C={Amplifier with Wide band }, (note that F and C partition the sample space of the amplifiers), T= {Amplifier with the gain factor greater than 6}.

- We know that P(F) = 2/5, P(C) = 3/5,

  P(T|F) = 4/100 and P(T|C) = 1/100.
- We require P(C|T). Using Bayes' Theorem :
- P(C|T) = [P(T|C)P(C)]/[P(T|C)P(C) + P(T|F)P(F)]

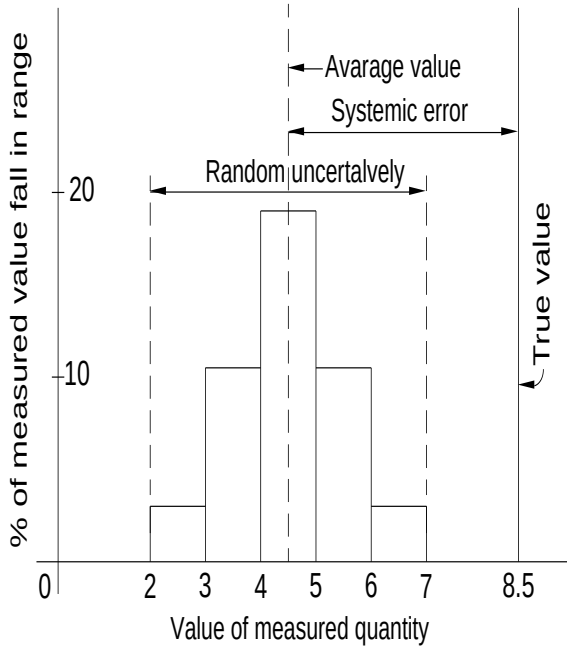
  = {[1/100] × [3/5]} /{[1/100]× [3/5] + [4/100]× [2/5]}

  = 3/11

2.5.1 PROBABILITY DISTRIBUTIONS

Errors and Samples

The act of making any type of experimental observations involves two types of errors: systematic errors (which exert a non random bias) and experimental or random errors. Systematic errors arise because of faulty control of the environment. For instance in the case of an experimental study by properly calibrating the measuring instruments and improving the experimental setup, repetition of this type of error could be minimized.



Random errors are due to many un-assignable causes and exhibit scatter as already stated. They have to be estimated. Infect the main objective of statistical analysis is to deal quantitatively with random or experimental error.

The data collected from an experiment represents a sample of the population from which it was drawn. The population in this case is collection of all possible specimens. As it is impossible to experiment with all specimens of the population, one of the main purposes of statistical analysis is to determine the best estimate of the population parameter from a randomly selected sample. While the population parameters are fixed and invariant, the parameter calculated from a sample contain random errors. Therefore, the sample provides only an estimate of the population parameter.

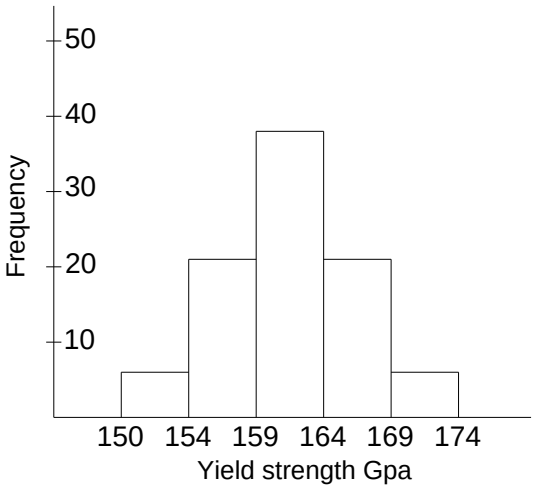
The statistical method, therefore leads to conclusions having a given probability being correct.

Frequency distributions

When large numbers of observations are made from a random sample, a method is needed to characterize the data. The most common method is to arrange the observations into a number of equal class intervals and then determine the frequency of the observations falling within each class interval. The table below shows the yield strength of a particular material after testing 100 specimens in GPa. These sets of 100 data are arranged in the form of frequency tabulation. For instance out of 100 specimens, 5 specimens have value of yield strength in between 150 to 154 GPa.

Class Interval of Y. strength	Class mid points (xi)	Frequency (fi)	(xi) (fi)	Relative frequency	Cum Relative frequency	% of cum Relative frequency
150-154	152	5	760	0.05	5	5
155-159	157	18	2826	0.18	23	23
160-164	162	42	6804	0.42	65	65
165-169	167	27	4509	0.27	92	92
170-174	172	8	1376	0.08	100	100

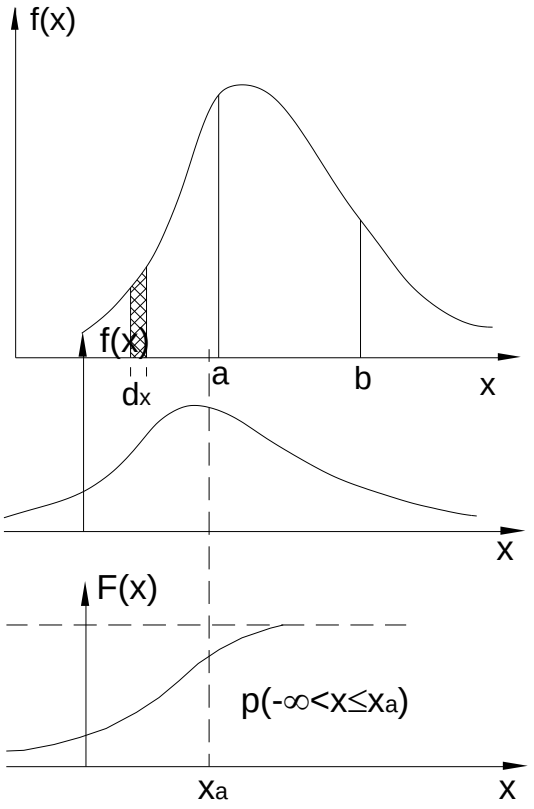
An estimate of the frequency distribution of observations can be obtained by plotting the frequency of observations against class-interval of yields strength as a bar chart, shown in the figure. This type of bar chart is known as “histogram”. As the number of observations increases, the size of class interval can be reduced until we obtain a limiting curve which represents the frequency distribution of the sample. This smooth curve is a function of the random variable yield strength (X). This f(X) is called the probability density function. The sub range width now becomes the infinitesimal dx, but the bar area f(x) d x still has the meaning the probability of finding x in d x. Thus $P(a < x < b) = \int_a^b f(x)\,dx$. If we know the probability density f (x) for some random variable, many useful calculations can be made.



If the frequency of observations in each class interval is expressed as a percentage of total number of observations, the area under the curve, which is plotted on this basis, is equal to unity. If the curve range is from -∞ to +∞, then,

$P(-\infty < x < \infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1.0$

Before inferring probabilities from the frequency distribution, we need to fit the experimental results to standard statistical distributions such as normal exponential lognormal etc.,



Another way to present the experimental data is to arrange in a cumulative manner. The presentation of data as a cumulative distribution is preferred sometimes, because it is much less sensitive than the frequency distribution to the choice of class intervals. Thus for each probability density

function f (x), there is an associated function F(X), called the cumulative distribution function denoted by.

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

This is just the probability that x is less than some chosen value x_a

$$P(-\infty < x \leq x_a) = \int_{-\infty}^{x_a} f(x)\,dx$$

Measures of central tendency and dispersion

Before applying any more sophisticated statistical techniques, one always compute two simple parameters viz mean and variance.

Mean:

The arithmetic mean (AM) or average is the most common and important measure of the central value of any array of data. It is the best indicator of central tendency. If the observations are devoted by x_1, x_2….x_n, than the mean $\bar{x}$ is given by,

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{n}$$ where n is the number of observation

Variance: The most important measure of dispersion of sample data is given by variance (s^2). It is calculated as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$ where $(x_i - \bar{x})$ is the deviation of each observation from the mean $\bar{x}$ of n observations, It is a measure of the scatter. The quantity (n-1) is called the number of degrees of freedom and is equal to the number of observations minus the number of linear relations between the observations. Since the mean $\bar{x}$ represents one such relation used, the number of degrees of freedom for the variance about the mean is (n-1). The other formulas for calculating variance are

$$s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$$ (For ungrouped data)

$$s^2 = \frac{\sum_{j=1}^{r} x_j^2 f_j - \frac{\left(\sum_{j=1}^{r} f_j x_j\right)^2}{n}}{n(n-1)}$$ (For grouped data)

x_j = midpoint of group

f_j = frequency of group

n = total no. of observations

For our example of analyzing yield strength of 100 specimens, the average or mean value is,

$$\bar{x} = \frac{\sum_{j=1}^{n} x_i f_i}{\sum_{j=1}^{n} f_j} = \frac{16275}{100} = 162.75 GP_a$$

$$s^2 = \frac{\sum_{j=1}^{5} x_j^2 f_j - \frac{\left(\sum_{j=1}^{5} f_j x_j\right)^2}{100}}{100 - 1} = 48.67$$

It is usual practice to work with standard deviation (s) which is defined as the positive square root of variance. (absolute dispersion)

$$s = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right]^{1/2}$$

Sometimes, it is desirable to describe the variability relative to the average. If the absolute dispersion is the standard deviation (s) and the mean is (x), the relative dispersion is called the "coefficient of variation" of dispersion C and is given by $C_v = \left[\frac{s}{\bar{x}}\right]$ Suppose if there are two sets of data each having s = 100, but one has x = 200 and the other the x = 2000, the values for cvi = 100/200 = .5 and cv2 = 100/2000 = .05 obviously, the second data set clustered much more closely around the mean than the first one.

Range: The range is another measure of dispersion. It is simply the difference between the largest and the smallest observation $R = (X_u - X_l)$

There are two other common measures of central tendency other than average viz mode and median.

Mode: The mode is that value of observation which occur most frequently (M_o)

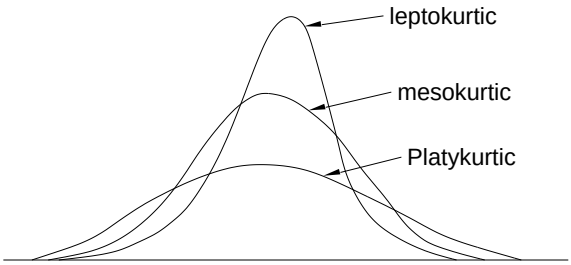
Median: The median is the middle value of a group of observations. For a frequency distribution, the median is the value that divides the area under the curve into two equal parts (M_d).

There are certain other measures depending upon the shape of the distribution which may be useful for analysis.

Skewness: Skewness indicates lack of symmetry. A distribution is said to be skewed if mean (M), median (M_d) and mode (M_o) fell at different points i.e. (M)≠ (M_o) ≠(M_d).

Karl Pearson's coefficient (Sk) is one measure used to for skewness $Sk=3(M-M_o)/\sigma$ where σ = standard deviation of the distribution.

Kurtosis: Kurtosis enables us to have an idea about the flatness or peakedness of the curve. In is also known as "convexity of curve". It is measured by coefficient β_2 or its derivative δ_2. The shapes are defined as mesokurtic (neither flat nor peaked), platy kurtic (flatter than normal) and leptokurtic (more peaked)



Example:

Consider the training data set of experimentation conducted using different test bench to assure the level of quality and it is represented as a target variable 'QoS' (suggesting possibilities of satisfying the QoS). Now, define the possibility of satisfying the QoS condition with different test bench by proper justification.

Test Bench	1	2	3	4	1	1	2	4	2	3	4	2	4	3	2	1	4
QoS	Yes	No	Yes	No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes

Solution:

Step I: Convert the data set into a frequency table

Frequency table

Test bench	Yes	No
1	3	1
2	3	2
3	1	2
4	3	2
Total	10	7

Step II: Create Likelihood table

Test bench	Yes	No	Probability	
1	3	1	= 4/17	0.24
2	3	2	= 5/17	0.29
3	1	2	= 3/17	0.18
4	3	2	=5/17	0.29
Total	10	7		
Probability	= 10/17	= 7/17		
	0.59	0.41		

Now, the result signifies Testbench2 and Testbench4 has higher probability

2.6 POPULAR PROBABILITY DISTRIBUTIONS

We have defined the probability density functions in terms of a histogram based on real world data. In order to design a meaningful experiment or proper analysis of a set of data, it is necessary to make a realistic choice regarding some standard distributions to be used to fit the data gathered. The reason for using them includes the generalization, the added speed and convenience of routine calculations. There are several standard distributions. Basically they are grouped under i) continuous probability distributions and ii) discrete probability distributions and iii) sampling distributions.

It is not possible to discuss all distributions and their scope of applications. To give an exposure, under the category of continuous distributions i) Normal or Gaussian ii) Weibull and iii) Exponential will be covered. Under the discrete distributions, i) Binomial, ii) Poisson will be covered. In sampling, t-distribution, chi-square and f-distribution will be covered.

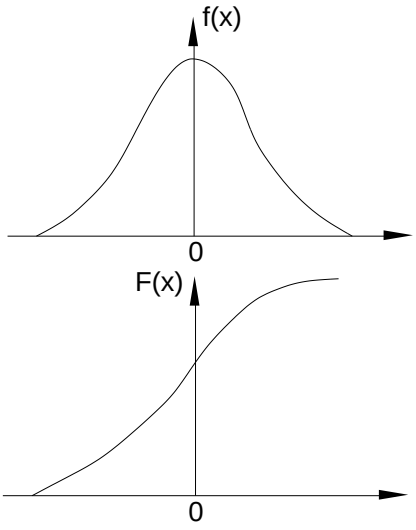
Gaussian or Normal distribution:

Gaussian or normal distribution is the most important and commonly used distribution to fit real world data. In addition to its goodness of fit for practical data, there are some theoretical reasons for its importance. Repeated measurements of lengths, diameter etc., distribution of yield strength, tensile strength etc. have been found to follow this curve. The shape of the curve is bell

shaped and symmetrical. By definition, the probability density for the Gaussian distribution is,

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$; -∞<x<∞ where σ : standard deviation and μ : population average value (mean).

We can choose any positive value for σ and real (+, -, o) value for μ, giving sufficient adjustability to fit many sets of practical data. For any allowed values of σ and μ the curve is symmetrical about μ and has a total area of precisely 1.0. The value of σ decide the spread of the curve while μ .locates the centre. The cumulative distribution F (x) must be found by numerical integration since it cannot be integrated analytically.



In order to place all normal distributions on a common basis, the Gaussian distribution is frequently expressed in terms of the standard normal variable 'z' given by

$$z = \frac{x-\mu}{\sigma} \; giving \; f(x) = \frac{1}{\sqrt{2\pi}} exp\left(\frac{-z^2}{2}\right)$$

For this standard normal curve, μ-o and σ =1. Again the total area under the standardized curve is 1.0. Cumulative distributions $F(x) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}z^2\right)$ values are available in the form of table for various values of 'Z' falling between -∞ and a specified value of 'Z'.

To use normal table for any calculations, we need to have numerical values μ and σ, the true mean and standard deviation of the "total population" (infinite size sample)

In practice, these are never available, and we must instead use their estimates $\bar{x}$ and s, calculated from the available sample. These estimates are more reliable for larger samples and we can quantify the reliability using confidence intervals.

For any variable that follows a Gaussian distribution it may be noticed that,

68.3% of valves fall within ± 1 σ of μ

95.4% of valves fall within ± 2 σ of μ

99.7% of valves fall within ± 3 σ of μ

Example : It is established that the average response of a particular phenomena is 1.250 with standard deviation as 0.002. One experimenter conducted several repeated experiments and found the average valve varies between 1.245 and 1.255. Find what % of results non confirm.

Solution:

We may assume the population mean (μ) = 1.250

and standard deviation of population (σ) = .002

using standard Gaussian variable, z, $z = \frac{x-\mu}{\sigma}$

$z_u = \frac{1.255\text{-}1.250}{.002} = 2.5$

$z_L = \frac{1.245\text{-}1.250}{.002} = \text{-}2.5$

$P(1.245 \le x \le 1.250) = P(-2.5 \le z \le 2.5)$

$= 2(1.0000 - 0.9938)$

$= 2 \times .0062 = .0124$

$= 1.24\%$

A random sample of 10 components from a bin is considered with the weightage as: 1, 2, 2, 3, 4, 5, 6, 6, 7, 7. Assume that, its life times are roughly normally distributed.

a. Calculate the sample mean.

b. Calculate the sample standard deviation.

c. Calculate the standard error of the mean.

d. Calculate the 99% confidence interval.

(Consider the Degree of freedom is 9, $\alpha_{0.005} = 3.25$)

Solution:

a. Sample mean $\bar{x} = \Sigma[(x_i/n)] = 43/10 = 4.3$

b. sample standard deviation = 2.2136

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

where n = Number of data points

$\overline{x}$ = The mean of the x_i

c . standard error of the mean = $\sigma_M = \frac{\sigma}{\sqrt{N}}$ = 0.7

d. 99% confidence interval = $\{x \pm \alpha x \sigma_m\}$ = {4.3 ± [3.25 x (0.7)]}

= 4.3 ± 2.275

The 99% confidence interval about its mean = (6.575, 2.025)

Example:

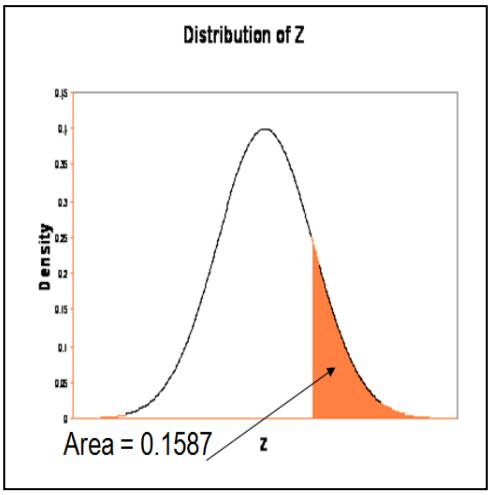
Consider the packet delivery of a communication network is normally distributed with μ = 80 and σ = 10. Find what percentage that a communication network have the packet delivery greater than 90?

Solution:

Z score transformation = Z = (90 - 80) /10 = 1

The percentage greater than 90 is equivalent to the area under the Standard Normal curve greater then Z = 1.

From tables of the Standard Normal distribution, the area to the right of Z=1 is 0.1587 (or 15.87%)



Weibull distribution

The normal or Gaussian distribution is an unbounded symmetrical distribution with long tails extending from −∞ to +∞. Many engineering random variables follow a bounded non – symmetrical distribution. Weibull distribution is one such distribution. It is used in many engineering problems because of its versatility. Originally used to describe the fatigue life of components, it is now widely used to describe the life of parts / components like ball bearings, gears and electronic components.

One of the reasons for the popularity of this distribution is that the data can be conveniently plotted on the Weibull paper and the conformance of the data to Weibull distribution can be evaluated by the linearity of the cumulative distribution function in the same way as the normal distribution.

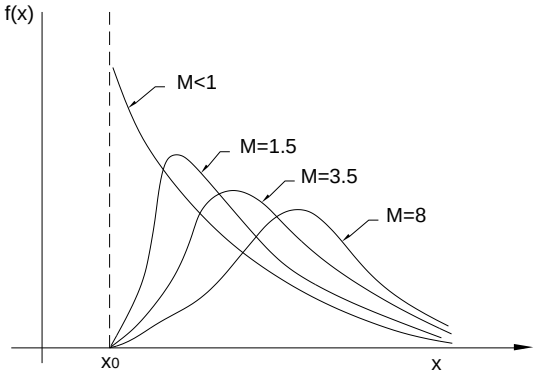
The Weibull distribution density function is given by

$$f(x) = \frac{m}{\theta}\left[\frac{x-x_o}{\theta}\right]^{m-1} exp\left[-\left(\frac{x-x_o}{\theta}\right)^{m}\right]$$ where

θ : characteristic or scale parameter f(x)

m : shape or slope parameter

x_0 :expected minimum value of x and is often referred as the threshold parameter



The cumulative distribution function

$$f(x) = 1 - exp\left[-\left(\frac{x-x_0}{\theta}\right)^{m}\right]$$

In the case of two parameter, x_0 = 0, then

$$f(x)=\frac{m}{\theta}\left(\frac{x}{\theta}\right)^{m-1}\exp\left[-\left(\frac{x}{\theta}\right)^{m}\right] \qquad f(x) = 1 - exp\left[-\left(\frac{x}{\theta}\right)^{m}\right]$$

The shape of the curve varies with change in shape parameter valves, i.e. m values as shown in the figure.

Example:

Let X = the ultimate band width of the test circuit at 200 degrees K. Suppose X has a Weibull distribution with parameters m = 20, and θ = 100. Find:

(a) P(X ≤ 105); (b) P(98 ≤ X ≤ 102); (c) the value of x such that P(X ≤ x) = 0.10

Solution:

(a) P(X ≤ 105) = F(105; 20, 100) $= 1 - e^{-(105/100)^{20}} = 1 - 0.070 = 0.930$

(b) P(98 ≤ X ≤ 102) = F(102; 20, 100) - F(98; 20, 100)

$$= e^{-(0.98)^{20}} - e^{-(1.02)^{20}}$$

= 0.513 – 0.226 = 0.287

(c) P(X ≤ x) = 0.10

$$P(X \le x) = 1 - e^{-(x/100)^{20}} = 0.10$$

Then, $e^{-(x/100)^{20}} = 0.90$

$$(x/100)^{20} = -\ln(0.90)$$

$$\Rightarrow \quad x/100 = [-\ln(0.90)]^{1/20}$$

$$\Rightarrow \quad x = 100[-\ln(0.90)]^{1/20}$$

x = 89.36

Example:

The random variable X can modeled by a Weibull distribution with m = ½ and θ = 1000. The spec time limit is set at x = 4000. What is the proportion of items not meeting spec?

Solution

The fraction of items not meeting spec is

$$\begin{aligned} P(X > 4000) &= 1 - P(X \le 4000) \\ &= 1 - F(4000) \\ &= e^{-\left(\frac{4000}{1000}\right)^{1/2}} \\ &= e^{-2} \\ &= 0.1353 \end{aligned}$$

The above calculation concludes that, about 13.53% of the items will not meet spec.

Exponential distribution

The exponential distribution is a special case of Weibull distribution when m = 1 and $x_0 = 0$.

For Weibull distribution, the probability density function is

$$f(x) = \frac{m}{\theta}\left[\frac{x-x_0}{\theta}\right]^{m-1} exp\left[-\left(\frac{x-x_0}{\theta}\right)^{m}\right]$$

Putting m = 1 and $x_0 = 0$

$$f(x) = \frac{1}{\theta} exp\left(-\frac{x}{\theta}\right) = \frac{1}{2} e^{-x/\theta}$$

Putting $\lambda = \frac{1}{\theta}$ the exponential distribution becomes, $f(x) = \lambda\, e^{-\lambda x}$; for x>0 and $\lambda > 0$

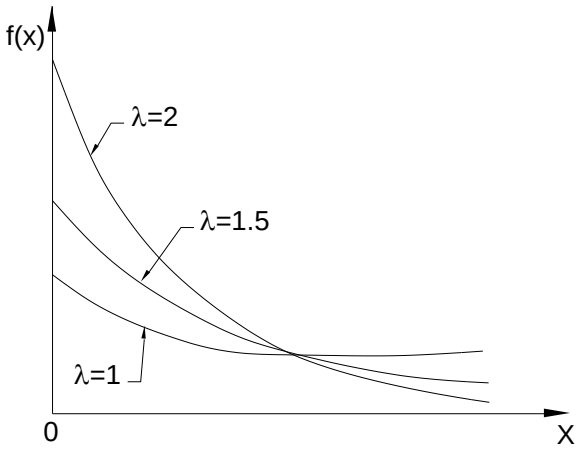
The mean of exponential distribution $(\bar{x}) = \frac{1}{\lambda}$ and variance $(s^2) = \left(\frac{1}{\lambda}\right)^2 \therefore s = \frac{1}{\lambda}$.

In this distribution both mean and standard deviation are equal i.e., X = s

The cumulative distribution function of exponential distribution is,

$$F(x) = 1 - e^{-\lambda x}; x > 0$$

The exponential distribution plays a key role in the theory of queues and in reliability analysis. This is used to represent activity where most of the events take place in a relatively short time, while there are a few which take very long time. The service times in queuing systems, inter-arrival time of vehicles in a highway, the life of some electronic components are some of the examples where exponential distribution could be used. To model failure of complete system this distribution is better suited.



Example:

Suppose the response time X of certain CMOS device (the elapsed time between the output end of a device and at the input) has an exponential distribution with expected response time equal to 5 sec.

(a) What is the probability that the response time is at most 10 seconds?

(b) What is the probability that the response time is between 5 and 10 seconds?

(c) What is the value of x for which the probability of exceeding that value is 1%?

Solution:

The E(X) = 5=θ, defines λ = 0.2.

(a) The probability that the response time is at most 10 sec is:

$$\begin{aligned} P(X \le 10) &= F(10,0.2) \\ &= 1 - e^{-(.2)(10)} \\ &= 1 - 0.135 \\ &= 0.865 \end{aligned}$$

or P (X>10) = 0.135

(b) The probability that the response time is between 5 and 10 sec is:

$$\begin{aligned} P(5 \le X \le 10) &= F(10;0.2) - F(5;0.2) \\ &= (1-e^{-2}) - (1-e^{-1}) \\ &= 0.233 \end{aligned}$$

(c) The value of x for which the probability of exceeding the value as 1%?

$$P(X \le x) = 1 - e^{-\lambda x} = 0.99$$

$$e^{-\lambda x} = 0.01$$

$$-\lambda x = \ln(0.01)$$

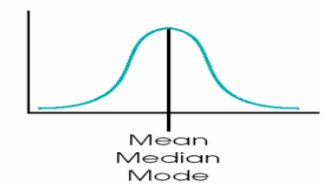
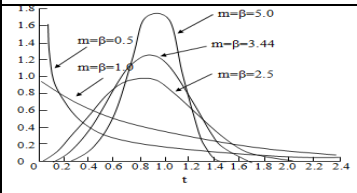
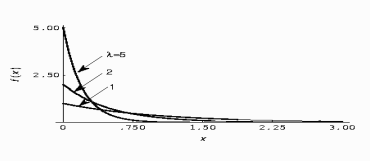
$$x = \frac{4.605}{0.2}$$

$$x = 23.025 \sec$$

$$F(10) = 0.99$$

$$= 1 - e^{-\lambda_N (10)}$$

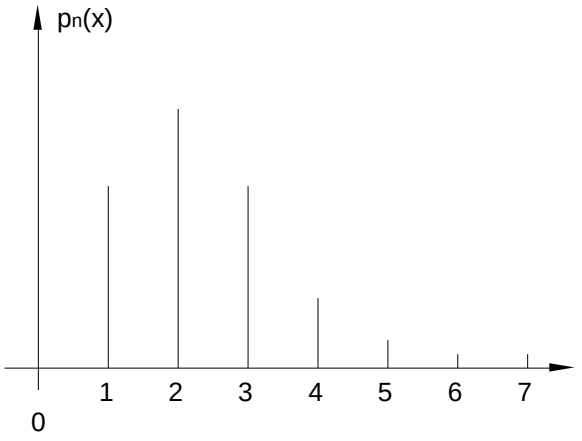
Conclusion on Continuous distribution

Statistical tool	Normal or Gaussian distribution	Weibull distribution	Exponential distribution
Mean (Population average value)	μ (True mean) (Its estimate is represented as $\overline{X}$		$\overline{X} = 1/\lambda$
Variance	σ^2		$S^2 = (1/\lambda)^2$
Standard deviation	σ		$S = 1/\lambda$
Range of X	$-\infty < x < \infty$		
Probability density function (pdf)	f(x) = $P_G(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$	F(X)= $(m/\Theta)[(X-X_0)/\Theta]^{m-1} \exp[(X-X_0)/\Theta]^m$	$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$
	If $z = \lvert x - \mu \rvert/\sigma$ then, F(X)= $\frac{1}{\sqrt{2\pi}} \int_{-z}^{z} \exp[-z^2/2]$		
Cumulative distributions	F(X)= $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp[-z^2/2]$	$F(x) = P(X \le x) = 1 - e^{-\left(\frac{x}{\theta}\right)^m}$	$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } 0 \le x < \infty \\ 0, & \text{otherwise} \end{cases}$
Characteristics	**mean** = median = mode symmetrical Unbounded symmetrical distribution Many complexly-determined	Θ = Characteristics or scale parameter f(x); M=shape or slope parameter; X_0=Expected minimum value of X =Threshold parameter	Mean and standard deviation are equal A non-negative random variable It exhibits Memoryless (Markov) property.
Pattern of distribution			
Application	Repeated measurements of lengths, diameter, yield strength, tensile strength, RF signal strength Distribution to fit real world data	Life of components, Parts To solve bounded non symmetrical distribution. Eg. Engineering random variables, Fatigue test	commonly in reliability & queuing theory Inter arrival time between two IP packets (or voice calls) Time to failure, time to repair etc.

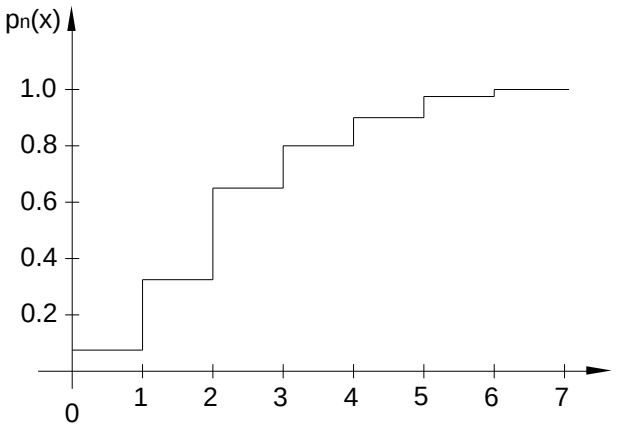
Binomial distribution

In continuous distribution like Gaussian the probability of the random variable taking on any specific value is zero. For those applications where such probabilities are non-zero, we use discrete distributions. Binomial and Poisson are most commonly used discrete distributions.

The binomial distribution applies to situations where events are judged on a yes or no basis. In quality control after inspection, a part is declared as defective or not defective. We assume the parts come from a population which has fixed percentage of good and defective items, and this percentage remains the same as we withdraw samples to be tested.



Let the occurrence of selecting a good part is 'success' and its non occurrence as 'failure'. Let 'p' denote the probability of success and 'q' denote the probability of failure such that p + q = 1. The number of successes in n trails may be 0, 1, 2, 3, r……n and obviously a chance variate. In our case, n being the number of parts in a batch being tested.



The probability of 'x' successes in 'n' trails is given by $p_n(x) = p^x q^{n-x} \frac{n!}{x!(n-x)!}$

$P_n(x)$ is the probability of getting x good parts in a batch of n. It is also written as $p(x) = nC_x p^x q^{n-x}$. This is the probability mass function of Binomial distribution. The mean of the distribution $(\bar{x}) = n.p$ and variance $(\sigma^2) = n.p.q$. The probability of getting 'x' or less good parts in a batch of n is given by $C_n(x) = \sum_{x=0}^{x} p_n(x)$

In practical applications, we are often in the position of not knowing the true valve of ‘p’ and we try to estimate it from data on a finite size sample.

Example:

Four different make USBs are available. Give the selections to select any two at a time.

Solution:

- Consider as USB A, B, C and D are the Four different make, then the possible choices are: AB AC AD BC BD CD, i.e. there are 6 possible combinations.
- Confirming this with the Binomial distribution, defines

$$\frac{N!}{r!(N-r)!} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{(4)(3)(2)(1)}{(2)(1)(2)(1)} = \frac{12}{2} = 6$$

Example:

3components to be selected from 100 components, find the number of possible combinations with the justification of method to be adopted.

Solution:

The no.of possible combinations are through binomial distribution

$$\frac{N!}{r!(N-r)!} = \binom{100}{3} = \frac{100!}{3!97!} = \frac{100*99*98}{3*2} = \frac{970,200}{6} = 161,700$$

Example:

Consider 4 different high speed Networks (NWK) to be tested for its speed by 4 trials and the NWK_1 have a 60% chance of high speed data transformation. Assuming that the speeds of NWKs are independent of each, what is the probability that:

a. The NWK_1 will give highest speed at 0 trial, 1st trial, 2nd trial, 3rd trial, or all 4 trials?

b. The NWK_1 will give highest speed at least in 1st trial

c. The NWK_1 will give a majority of the trials

Solution:

Let ‘X’ is the equal numbers of trials have NWK_1 to test the communication speed

a. Using the definition of binomial distribution,

$$\binom{4}{0} p^0 q^{4-0} = \frac{4!}{0!(4-0)!} \cdot .60^0 \cdot .40^4 = .40^4 = .0256,$$

$$\binom{4}{1} p^1 q^{4-1} = \frac{4!}{1!(4-1)!} \cdot .60^1 \cdot .40^3 = 4 * .60 * .40^3 = .1536,$$

$$\binom{4}{2} p^2 q^{4-2} = \frac{4!}{2!(4-2)!} \cdot .60^2 \cdot .40^2 = 6 * .60^2 \cdot .40^2 = .3456,$$

$$\binom{4}{3} p^3 q^{4-3} = \frac{4!}{3!(4-3)!} \cdot .60^3 \cdot .40^1 = 4 * .60^3 \cdot .40^1 = .3456,$$

$$\binom{4}{4} p^4 q^{4-4} = \frac{4!}{4!(4-4)!} \cdot .60^4 \cdot .40^0 = .60^4 = .1296$$

b. P(at least 1) = P(X ≥ 1)

= 1 - P(none)

= 1 - P(0)

= 0.9744. Or, P(1) + P(2) + P(3) +P(4)

= 0.9744.

c. P(NWK$_1$ have a high speed on majority trials)

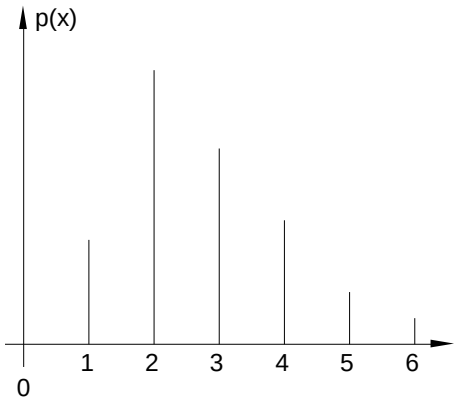
= P(X ≥ 3)

= P(3) + P(4)

= 0.3456 + 0.1296 = 0.4752.

Poisson distribution

Poisson distribution is the discrete version of the exponential distribution. Poisson distribution is used to model the number of independent events that occur in a fixed amount of time or space. This distribution finds application in a wide variety of situations like number of typographical errors in a page, number of defects or flaws in a square meter of cloth, number of alpha particles emitted by a radio active substance etc.



If λ represents the average number of occurrence of an event, the probability of occurrence of x events is given by $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$; for x=0, 1, 2,

3……2τ>0. The mean and variance of Poisson distribution are equal mean $(\bar{x}) = \lambda$ and σ^2. The cumulative probability distribution is given by

$$C(x) = \sum_{i=0}^{x} e^{-\lambda}.\lambda^{i}$$

Example: The surface defects noticed in each 10m length of cold drawn bar stock in number is given below. What is the probability of being or beyond the specification limit of 3 defects per 10m length?

Defects / 10m	Nos.	Defects / 10m	Nos.
0	40	4	1
1	9	5	1
2	4	6	1
3	2		

The average number of defects is calculated as given below

Defects	No.	Total
0	40	(0) (40) = 0
1	9	(1) (9) = 9
2	4	(2) (4) = 8
3	2	(3) (2) = 6
4	1	(4) (1) = 4
5	1	(5) (1) = 5
6	1	(6) (1) = 6
	58	**38**

The average number of defect

$\lambda = \frac{38}{58} = 0.655$ $\quad e^{-\lambda} = e^{-.655} = 0.519$

$$P(0) = \frac{(0.519)(0.655)^0}{0!} = 0.519$$

$$P(1) = \frac{(0.519)(0.655)^1}{1!} = 0.339$$

$$P(2)=\frac{(0.519)(0.655)^2}{2!}=0.111$$

$$P(3)=\frac{(0.519)(0.655)^3}{3!}=0.024$$

∴.Probability of being at specification limit of 3 defects / length p (3)=0.024

Probability of being beyond the specification limit = p(x>3)

ie p(x>3) = 1-p(x≤3)

= 1 – [p(0)+p(1)+p(2)+p(3)]

= 1 – [0.519+0.339+0.111+0.024]

= 1 – 0.993

∴ P(x>3) = 0.007

Example:

(a) If calls to your cell phone are a Poisson process with a constant rate λ=2 calls per hour, what's the probability that, if you forget to turn your phone off in a 1.5 hour (Seminar), your phone rings during that time?

(b) How many phone calls do you expect to get during the Seminar?

Solution:

(a) X ~ Poisson (λ=2 calls/hour); P(X≥1)=1 – P(X=0)

⇨ $P(X=0)=\frac{(2*1.5)^0 e^{-2(1.5)}}{0!}\frac{(3)^0 e^{-3}}{0!}=e^{-3}=.05$

⇨ ∴P(X≥1)=1 – .05 = 95% chance

(b) Phone calls expected to get during the Seminar is

E(X) = λt = 2(1.5) = 3

Sampling distribution:

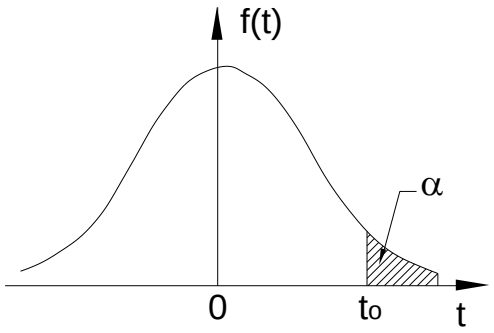
The major problem in statistics is relating the population and the samples that are drawn from it. There are two main aspects in this viz : i) What the population tell us about the behavior of samples drawn from it, ii) What does the sample or series of samples tell us about the population from which the samples came? We discuss the first question.

Suppose we have population of 300 items, we take all possible samples of n = 10 and determine the mean of each sample $\bar{x}$. This would give population means of $\bar{x}$'s, one for each sample of 10. We call this as sample statistic.

Similarly, we can have sampling distribution for s^2 or for any other sample statistic. Such distributions are called as "sampling distributions".

Distribution of sample means.

The mean of a sample provides an unbiased estimate of the mean of the population from which the sample was drawn. The sample values of the mean will be normally distributed about the population mean μ. The mean of the distribution of $\bar{x}, \mu\bar{x}$ is equal to the mean of the population of x's if the population is very large. (i.e. n > 30) The standard deviation of the sampling distribution is called the standard error of the mean and it is equal to $\sigma_{\bar{x}} = {}^{\sigma}/_{\sqrt{n}}$. We can use this to establish the standard variable 'z', which is normally distributed with mean 'o' and standard deviation 1.



$$z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

Student's 't' distribution.

In many situations we may not know the standard deviation (σ.) of the population. In such cases, we cannot substitute's' for 'σ' in the equation of the previous para and assume that the statistic is normally distributed unless 'n' is greater than 30. However, if x is approximately normally distributed and n < 30, the sampling distribution for the mean is the student's 't' distribution. The statistic 't' is given by $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ for ϑ = n-1 degree of freedom.

The 't' – distribution is symmetrical about the line t = 0 and maximum ordinate is at t = 0. There is a different curve for each value ϑ. As ϑ increases, the distribution of 't' approaches normal distribution. The number of d.f of a statistic denoted by ϑ is defined as 'n' (independent observations) in the sample minus the number of population parameters which must be estimated from the samples. In the present case's' is the only statistic and hence ϑ =(n-1). The't' distribution is available in tables where for a given ϑ and probability α, the t_0 value can be found. This gives p ($t > t_o = \alpha$).

Chi-square distribution.

The distribution of sample variances s^2 from a normal population with variance σ^2 is given by chi-square (χ^2) distribution.

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

The chi-square statistic, χ^2 can also be computed for a given set of data by computing the observed frequency of data in each class interval and the expected frequency for the same class interval, predicted by the theoretical distribution as to $\chi^2 = \sum_i^k \frac{(f_o - f_e)}{f_o}$ where f_o : observed frequency and f_e : expected frequency, k : number of class intervals. If $\chi^2 = 0$, then the observed and theoretical frequencies agree exactly, whereas if $\chi^2 > 0$ they do not. The larger the value of χ^2 the greater is the discrepancy. The Chi-square values are tabulated and available as published data. The χ^2 statistic is tabulated by d.f Vs (1-α) or the significance level. This is extensively used to establish goodness of fit test in data analysis.

F – distribution

Let $N_1(\mu_1, \sigma_1)$ and $N_2(\mu_2, \sigma_2)$ are two normally distributed populations with mean μ_1 and μ_2 and variances σ_1 and σ_2 respectively. If we take two independent samples of sizes n_1 and n_2 from the two populations respectively with variances of samples as $s^2_1$and s^2_2 The F-distribution tell us whether the two samples come from populations having equal variances. The larger two sample variance must be in the numerator. For instance S, is larger, then

$F_{\vartheta_1 \vartheta_2} = {s_1^2}/{s_2^2}$ where ϑ_1=(n-1) and ϑ_2=(n_2-1) are the degrees of freedom.

Simple calculation of s_1 and s_2 and comparison of s_1 and s_2 can be misleading. The values of F-distribution is available in the form of table for different values of ϑ_1 and ϑ_2 and chosen value of 'α', the level or significance.

2.7 SAMPLE DESIGN

Sampling

All items of interest to investigate such as individuals or of their attributes or results of operations, whether can be numerically specified are termed as

'population' or 'universe' in statistics. A part or a small portion selected from the population is called sample and the process of such selection is called "sampling". Sampling is resorted to when either it is impossible to enumerate the whole population or when it is too costly to enumerate in terms of money and time. To serve a useful purpose sampling should be unbiased or representative.

The aim of sampling is to get as much information as possible ideally the whole information about the population from which the sample is drawn. In particular, given the form of parent population, we would like to estimate the parameters of the population or specify the limits within which the population parameters are expected to lie with a specified degree of confidence. It is, however, to be clearly understood that the logic of the theory of sampling is the logic of induction. That is, we pass from particular (sample) to general (population) and hence results will have to express in terms of probability.

Sample design.

A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample especially the size of the sample. Sample design is determined before data are collected. There are many sample designs from which a researcher can choose. Some are relatively more precise and easy to use than others. A researcher must select / prepare a same design which should be reliable and appropriate for his research study.

While developing a sampling design, the researcher must pay attention to the following aspects.

- the type of universe, whether finite or infinite and also clearly define the set of objects which constitute the universe.

- sampling unit should be decided before selecting sample. For example the unit may be state, district or a zone etc. or specific items or individuals

- source list : known as sampling frame from which sample is to be drawn. If source list not available the researcher has to prepare it.

- size of sample : refers to the number of items to be selected from the universe to constitute a sample. While deciding the size of sample,

researcher must determine the desired precision as also an acceptable confidence level for the estimate.

- Parameters of interest: In deciding sample design one must consider the question of specific population parameters which are of interest such as average or other measures.

- Budgetory constraint : Cost considerations from practical point of view have a major impact relating to not only the size but also the type.

- Sampling procedure : The researcher should decide the type of sample and the technique to be used in selecting the items, which gives smaller sampling error for given sample size and cost.

Sampling errors are the random variations in the sample estimates around the true population parameters. They occur randomly and equally likely to be on either direction, their nature happens to be compensatory type and the expected value of such errors happens to be equal to zero. Sampling error is the sum total of frame error, chance error and response error. Sampling error decreases with increase in sample size and it happens to be a smaller magnitude in the case of homogeneous population. Sampling error can be measured for a given sample design and size. The measurement of sampling error is usually called the "precision of the sampling plan". While selecting a sampling procedure, researcher must ensure that the procedure couses relatively small sampling error and helps to control the systematic bias in a better way.

Types of sample Designs

There are different types of sample designs based on two factors viz., the representation basis and the element selection technique. On the representations basis, the sample may be probability sampling or it may be non–probability sampling. Probability sampling is based on the concept of random selection whereas non–probability sampling is non-random' sampling. On element selection basis, the sample may be either unrestricted or restricted. When a sample element is drawn from a large population, then the sample so drawn is known as **"**unrestricted sample"' whereas all other forms of sampling are covered

under the term “restricted sampling”. Thus, the sample designs are basically two types viz., non-probability and probability sampling.

Representation basis / Element basis	Probability sampling (1)	Non – probability sampling (2)
1) Unrestricted sampling	Simple random sampling	Haphazard sampling or convenience sampling
2) Restricted sampling	Complex random sampling such as cluster, stratified, systematic etc.	Purposive sampling such as quota, judgment sampling.

Non-probability sampling

Non – probability sampling does not afford any basis for estimating the probability that each item in the population has a chance of being included in the sample. This sampling is also known as purposive, deliberate and judgment sampling. In this, items for the sample are selected deliberately by the researcher using his choice. In such a design, personal element has a great chance of entering into selection of the sample. The investigator may select a sample in such a way that it yield results favorable to his point of view and if that happens, the entire inquiry may get vitiated. Thus, there is a danger of bias entering into this type of sampling technique.

The experience of the investigator, his impartiality and capability of taking sound judgment may result in selecting a sample which is tolerably reliable. Sampling error in this type of sampling cannot be estimated and the element of bias, great or small, is always there. This type of sampling technique is usually adopted in small inquiries and researches by individuals due to the relative advantage of time and money inherent in this type. Quota sampling is an example of non-probability ones. Here, the interviewers are simply given quota and left their discretion. Quota samples are essentially judgment samples and inferences drawn on their basis are not amenable to statistical treatment in a formal way.

Probability Sampling

Probability sampling is also known as 'random sampling' or 'chance sampling'. Under this, every item of the universe has equal chance of being included in the sample. The results obtained from random sampling can be assured in terms of probability. That is, we can measure the errors of estimation or significance of results obtained from a random sample. This method is much superior over the deliberate sampling design. In brief, the implications of a random sampling (or simple random sample) are:

- each element in the population has an equal probability of being selected into the sample and all choices are independent of one another.
- each possible sample combination has an equal probability of being chosen.

Suppose a population has a finite size of 6 numbers and we want to select sample size of 3 from it, then there are 6C3 = 20 possible distinct sample of the required size. If we choose any one of the samples in such a way that each has the probability of 1/20 of being chosen, we will then call this a random sample.

The method of obtaining a random sample can be simplified in actual practice by the use of random numbers table. Suppose we are interested in taking a sample size of 10 units from a population of 5000 units. We can number each item from 3001 to 8000. we can then select from the random number table random numbers which are not less than 3001 and not greater than 8000. A portion of that is given below.

2052	6641	3992	9792	7979	5911
3170	5624	4167	9525	1545	1396
7203	5356	1300	2693	2370	7483
3408	2769	3563	6107	6913	7691
0560	5246	1112	9025	6008	8126

If we decide to read the table numbers randomly from left to right, starting from the first row itself, we obtain the following numbers: 6641, 3992, 7979, 5911, 3170, 5624, 4167, 7203, 5356, and 7483. Which are >3000 and <8000. The units bearing the above serial numbers form the required random sample.

Complex random sample design

Probability sampling under restricted sampling techniques may result in complex random sample designs. They may also be called as “mixed sampling designs” as many such designs represent a combination of probability and non-probability procedures in selecting a sample. Some of the popular sampling designs are briefly discussed below.

Systematic Sampling

If in a sampling, we select every i th item on a list, then this constitutes systematic sampling. An element of randomness is introduced in this by using random numbers to pick up the starting unit. For instance, if a 4% sample is desired, the first item would be selected randomly from the first 25 units, and thereafter every 25^{th} unit would be automatically included in the sample. Though this is not a random sample in strict sense, it is often treated as a random sample.

Stratified Sampling

If a population from which a sample is to be drawn does not constitute a homogeneous group, stratified sampling technique is generally applied in order to obtain a representative sample. In this method, the population is divided into several sub-populations that are individually more homogeneous than that of the total population, and then we select items from each stratum to constitute a sample. If each strata is of different size, then proportional allocation of number of items to be selected from each strata is made. Within each strata simple random sampling approach is adopted.

Cluster Sampling

In cluster sampling the total population is divided into a number of relatively small sub-divisions which are themselves clusters of still smaller units. Then some of these clusters are randomly selected for inclusion in the overall sample. Suppose we want to estimate the proportion of defective parts in a population of 20000 units, packed in 400 cases, each combining 50 parts, we

would consider the 400 cases as clusters and randomly select 'n' cases and examine all the parts in each randomly selected cases. If clusters happen to be some geographic subdivisions, then the cluster sampling is known as "Area sampling".

Multistage Sampling

This sampling is a further development of cluster sampling and is applied in big inquiries extending to a large geographical area, say the entire country. The first stage is to select a large primary sampling unit such as states in a country. Then we may select certain districts and do the survey. This represents a two stage sampling design with ultimate sampling units being cluster of districts.

Sample size determination in Simulation Experiments

When we use simulation to study a stochastic system, we represent one or more of the variables in the model by probability distributions from which we draw samples. Since these samples are randomly drawn we have some degree of imprecision in the result, which is highly influenced by the choice of sample size. As we use the information furnished by the simulation experiment as the basis for decision regarding the operation of the real system, we want this information to be as accurate and precise as possible or at least we want to know the degree of imprecision present. It is therefore essential that a statistical analysis be conducted to determine the required sample size.

The sample size may be determined in either of two ways: 1) prior to and independently of the operation of the model; ii) during the operation of the model and based upon the results generated by the model.

Prior determination approach

It is frequently possible to justify the use of certain forms of prior analysis based upon the knowledge of the model. Many forms of analysis are based upon the assumption that the responses of the model are independent and normally distributed. These assumptions are justified because of the central limit theorem from probability theory. It is usually sufficient that the response is the additive sum of a large number of contributing effects. The significance of control limit

thereon lies in the fact that it permits us to use sample statistics to make inferences about population parameters without knowing the nature of population, other then we get from sample.

In the most straight forward case, where we can invoke the central limit theorem and assume no auto correlation, we can take a confidence limit approach to determine the sample size required for estimating the parameters to a specified level of precision. These parameters are population mean, standard deviation etc..

Suppose we wish to determine an estimate of $\bar{x}$ of the true population μ, such that, P [(μ-d) ≤ x ≤ (μ+d)] = (1 - ∝) where ∝ is the sample mean, μ is the population mean and (1 - ∝) is the probability that the interval μ ± d contains $\bar{x}$, The problem is to determine the sample size such that the above equation holds. If our assumptions of normality are valid, we can show that, $n=\left[\frac{\sigma z^{\alpha}/_{2}}{d^2}\right]^2$ where $Z\frac{\propto}{2}$ is the two tailed standardized normal statistic for the probability we seek. In many cases, we must either guess at the value of σ or run a short pilot experiment. If we have some idea of what the highest and lowest possible responses of the system might be (feasible range) we can assume the range is approximately equal to 4 σ to estimate 'σ'.

If the variance is unknown, then do a pilot run or take a preliminary sample and obtain an estimate of the variance (s^2), then compute the total number of observations necessary. Thus, when σ is unknown, take a sample and estimate using 't' distribution, $n = \frac{t^2 S^2}{d^2}$ where t = tabulated t value for the desired confidence level and d. f of initial sample, d = half width of desired confidence interval and s^2 is the estimated variance from sample.

Use of automatic stopping rule

In this approach the confidence intervals for output values are determined during a simulation run and then terminate the execution, when a predetermined confidence level has been reached. Thus, we can avoid the inefficiencies of runs that are either too long or too short. One of the approaches to incorporate automatic stopping rule is to run the simulation in two stages. First run a sample

size 'n' use the results to estimate n* by one of the methods previously described. If n* <n, the run is over, otherwise, extend the run by (n* - n).

2.8 HYPOTHESIS TESTING

Hypothesis: What it is?

A hypothesis may be defined as a proposition or a set of propositions set forth as an explanation for the occurrence of some specified group of phenomena. They are either asserted merely as a provisional conjecture to guide some investigation or accepted as probable in the light of established facts. Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods that relates an independent variable to some dependent variable.

Very often we make decisions about populations based upon sample information. Such decisions are known as "statistical decisions". While attempting to make such decisions, it is usual to make assumptions or guesses about the population involved. Such assumptions, which may or may not be true, are called "statistical hypotheses", and in general they are statements about the probability distribution of the population.

Experimentation provides a method of hypothesis testing. The researcher define a problem and then propose a hypothesis. Test results either confirm or disconfirm the hypothesis. The confirmation or rejection is always stated in terms of probability rather than certainty. The ultimate purpose of experimentation is to generalize the variable relationship so that they may be applied outside the laboratory to a wider population of interest.

Null hypothesis and Alternative hypothesis

This is used in the context of statistical analysis. If we are to compare method A with method B about its superiority and if we proceed on the assumption that both the methods are equally good, then this assumption is termed as the "null hypothesis". As against this, we may think that the method A is superior or the method B is inferior, we are then stating what is termed as "alternative hypothesis". The null hypothesis is generally symbolized as Ho and the alternative hypothesis as H2. Suppose we want to test the hypothesis that

the minimum yield strength for accepting a component made of steel is μ_o, then we would say that the null hypothesis symbolically expressed as,

$H_0 = \mu_{H0} = \mu_0$ (μ_{H0} hypothesis mean) or it could be simply written as

$H_0 : \mu = \mu_0$

If our sample data do not support the null hypothesis, we conclude that something else is true. What we conclude rejecting the null hypothesis is known as alternative hypothesis. In other words a set of alternatives to the null hypothesis is known "alternative hypothesis". We may consider three possible alternatives as

i) $H_1: \mu \neq \mu_0$ ii) $H_1 : \mu > \mu_0$ iii) $H_1 : \mu < \mu_0$

Alternative hypothesis is usually the one which we wishes to prove and the null hypothesis is one, which we wishes to disprove or reject.

Level of significance

This is a very important concept in the context of hypothesis testing. It is always some percentage (usually 5% or 1%), which should be chosen with great care, thought and reason. If we choose a 5% significance level, then this implies that Ho will be rejected when the sampling result (observed evidence) has a less than 0.05 probability of occurring if Ho is true. In other words, the researcher is willing to take as much as 5% risk of rejecting the null hypothesis when it (H_0) happens to be true. Thus, the level of significance is the maximum value of the probability of rejecting Ho, when it is true and is usually determined in advance before testing the hypothesis.

Type I and Type II errors:

In the context of testing of hypothesis, there are two types of errors one can make. We reject Ho, when Ho is true and we may accept Ho when in fact Ho is false or not true. Rejecting Ho, when it is true is known as "Type-I error" and accepting Ho, when it is not true is called as "Type II error". In other words Type – I error means rejecting of a hypothesis which should have been accepted and Type – II error means accepting the hypothesis which should have been rejected. Type – I error is devoted by '$\propto$' (alpha) and known as $\propto$- error, also called the

"level of significance of the test". The type – II error is denoted by 'β' (Beta) known as β-error. These errors are shown in the decision table.

State of nature	Decision	
	Accept H_o	**Reject H_o**
H_o (True)	Correct decision	Type – I Error (∝)
H_o (False)	Type – II error (β)	Correct decision

If we fix Type – I error at 5% (assuming level of significance as 5%), it means that there are about 5 chances in 100 that we will reject Ho, when Ho is true. We can control Type – I error by fixing it at a lower level. But with a fixed sample size 'n' when we try to reduce Type – I error, the probability of committing Type – II error increases. Both cannot be controlled simultaneously. A trade-off between the two types of errors has to be made taking into account, the cost or penalties attached to both type of errors.

In the example quoted earlier, if we conduct test with 10 specimens to get the mean value of yield strength for steel to be used in the construction of building (μo), Type –I error to reject Ho when infact is true would mean that steel would be shipped back to the supplier with resulting financial loss and penalty to the builder due to delay in construction. On the other hand, a Type – II error, accepting Ho, when it is false, would mean that steel of insufficient strength would be used to construct the building resulting in poor quality and failure of the building leading to loss. Thus, a Type – II error should be minimized in engineering situations.

Two tailed and one tailed tests

In the context of hypothesis testing, these two terms are quite important. The two tailed test is appropriate when the null hypothesis, Ho is some specified value and the alternative hypothesis, H1 is a value not equal to the specified value, which means, it may be more or less than the specified. Symbolically, the two tailed test is appropriate when we have

H_0 : μ = μo and H1 : μ ≠μO which may be

μ > μo (or) μ < μo

Thus, in two tailed test there are two rejection regions, one on each tail of the curve as shown in the figure.

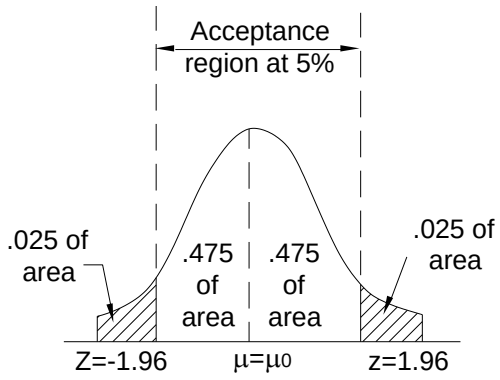


Fig. 1. Reject H0 if the sample mean $\bar{x}$ falls in either of the region. (Two tailed test at 5%)

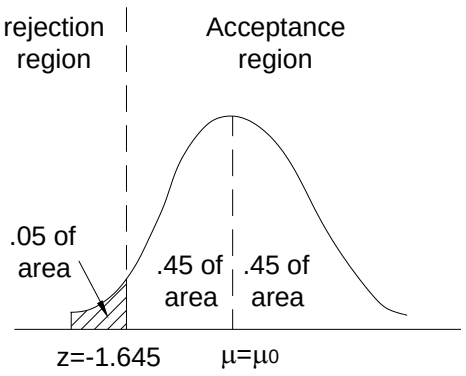


Fig. 2. Reject if the sample mean $\bar{x}$ falls in either of the region. (Two tailed test at 5%)

If the significance level is 5% and two tailed test is to be applied, the probability of the rejection area will be 0.05 (equally split on both tails) and the acceptance region will be 0.95 of the area as shown in fig.1. If our sample mean deviate from, μo in either direction, then we shall reject the null hypothesis, but if the sample does not deviate significantly from μo, we shall accept the Ho.

One tailed test is used when the population mean is either lower or higher than some hypothesized value. For instance, if our Ho: μ=μo and H1: μ<μo, then we are interested in what is known as left tailed test (wherein there is one rejection region on the left tail) as shown in fig.2. In case our Ho : μ= μo and Ho: μ > μo, we are then interested in what is known as right – tailed test and the rejection region will be on the right tail of the curve.

It should be remembered that accepting Ho on the basis of sample information does not constitute the proof that Ho is true. We only mean that there is no statistical evidence to reject it, but we do not certainly say that Ho is true.

Steps Involved in Testing of Hypotheses.

The various steps involved in testing hypothesis include,

i. Making a formal statement of the null hypothesis Ho and alternative hypothesis H1.
ii. Selection of significance level, Generally in practice either 5 % or 1 % level is adopted for the purpose.
iii. Determination of appropriate sampling distribution
iv. Selecting a random sample and compute the sample characteristic
v. Determine the critical region for rejecting the null hypothesis, Ho.

vi. Compare the probability. If the calculated probability is $\leq \propto$ value in the case of one at ailed test and $\propto/2$ in the case of two tailed tests then reject the null hypothesis.

standard text books on statistics do give details of type of hypothesis with conditions and what test statistic to be used with the formula and criteria for refection of H_o. One can refer to them to get more information on this.

The tests of hypothesis (to test significance) can be classified as i) Parametric tests or standard tests of hypothesis; and ii) Non-parametric test or distribution free tests. The parametric tests usually assume certain properties of the parent population from which we draw samples. Assumptions like observations come from normal population, sample size is large, and assumptions about population parameters such as mean, variance etc. must hold good before parametric tests can be used. In case, the researchers do not want to make such assumptions, they can use non-parametric methods to test hypothesis.

The important parametric tests are i) Z-test ii) t-test, iii) X^2-test and iv) F-test. All these tests are based on the assumption of normality ie., the source of data is considered to be normally distributed. In some cases it may not be normally distributed, but still we use it on account of the fact that we mostly deal with samples and that the sampling distributions closely approach normal distribution.

The Z - test is used for judging the significance of several statistical measures, particularly the mean. The relevant statistic used 'Z' is to be calculated and compared with table value (to be read from table showing area under normal curve) at a specified level of significance $\propto$. This is more frequently used by researchers.

The t-test is based on t-distribution and is appropriate to compare the significance of difference between the means of two samples in the case of small samples (ie $n < 30$). The X^2 test is based on X^2 distribution and is a parametric test used for comparing a sample variance to a theoretical population variance. F-test, which is based on F-distribution, is used to compare the variance of two independent samples. This test is also used in the context of Analysis of variance (ANOVA) for judging significance of more than two samples.

Example: A manufacturer of switch assures a mean life of 2000 Hrs. with a standard deviation of 200 Hrs based on their testing for a large number of samples over a period of time. Later on the company made a small change in the spring. To determine whether this has changed the life of the switch, a sample of 100 switches tested to give a sample mean as 1960 Hrs and standard deviation to 180. Is there a change in the product?

Let us assume the assured life of 2000 Hrs represents the population mean μ, 200 Hrs represent the population standard deviation.

The null hypothesis, H_o : $\mu_o = 2000$

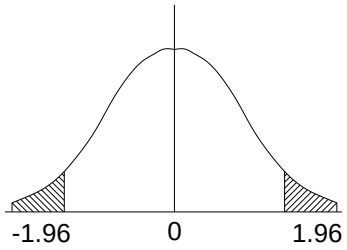
Alternate hypothesis, H_1 : $\mu \neq \mu_o$

Significance level is not given, we assume…..= 5% or .05

The sample size (n)=100.

As the sample size >30, we assume normal distribution σ is also known ie. σ = 200

As the alternative hypothesis H_1 is $\mu \neq \mu_o$, it may be more than or less than μ o and hence use two tailed test.

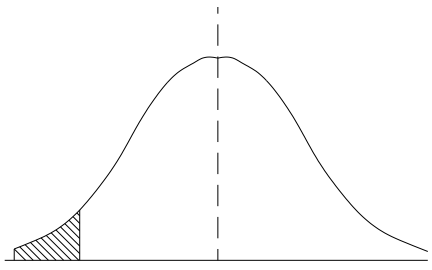


The critical region is $z < z_{\propto/2} and > z_{\propto/2}$ from standard normal table find values corresponding to $-Z_{.025}$ and $Z_{.025}$ The value are – 1.96 and 1.96.

The calculated of value of $Z = \bar{x} - \mu = \frac{1960-2000}{180/\sqrt{100}} = \frac{-40}{18} = -2.22$

As the calculated value falls in the critical region, the null hypothesis is rejected and we conclude that the change in spring has significantly altered the life of the switch. There are 5 chances in 100 that we have rejected Ho, when it is true, we can minimize this error by selecting ∝ = 1% = 01. Now the critical region is $z < -z_{.005}$ and $> Z_{.005}$. The corresponding table values are - 2.58 and 2.58. Since the calculated Z of - 2.22 falls outside of the critical region, we accept Ho at1% significance level.

In reality we are concerned only with, whether the change in spring reduced the life of the switch; it is not a serious consequence if the life was increased. Suppose, we form alternative hypothesis, $H_1 = \mu < \mu_o$. that is $H_1 = \mu < \mu_o = 2000$. In that case we have to deal with one tailed test and the critical region for rejecting H_o is $Z<Z_{-.01}$ or Z <-2.326. Since Z=-2.22, we do not reject the null hypothesis, Ho.



Example: ('t' test) (single group specimen) (no.of samples less than 30)

A random sample of branded smart phone user survey carried at selected 15 locality are as follows; 60, 85, 72, 59, 62, 73, 82, 67, 56, 69, 75, 58, 80, 61, 67. Suggested population mean is 68. Discuss is this supports the theory or not.

Solution:

Null hypothesis Ho: The population mean is 68.

Ho : m = 68

Sample mean $\overline{x} = \Sigma[(x_i/n)] = (1026/15) = 68.4$

Sample Varience $S^2 = \Sigma[(x_i - \overline{x})^2)/(n-1)]$

$= (1/14)\ [(60-68.4)^2+(85-68.4)^2+(72-68.4)^2+(59-68.4)^2 + (62-68.4)^2 +(73-68.4)^2 +(82-68.4)^2 +(67-68.4)^2 +(56-68.4)^2 +(69-68.4)^2 +(75-68.4)^2 +(58-68.4)^2 +(80-68.4)^2 +(61-68.4)^2 +(67-68.4)^2]$

$= (1/14)[1193.6] = 85.26$

Now $t_{calc} = |[(\overline{x} - m)/(s/\sqrt{n})]| = |[(68.4 - 68)/(9.23/3.74)]| = 1.78$

For $t_{14,\ 0.05} = 2.145$

⇨ $t_{calc} < t_{14,\ 0.05}$ represents the population mean supports the theory and it is a reasonable one.

Example:

A random sample of 20 WiFi system has the power gain in dB as follows; 20, 12, 15, 18, 10, 11, 17, 14, 16, 13, 9, 12, 16, 19, 21, 20, 14, 19,16, 14. Discuss about the suggested population mean as 18.

Null hypothesis Ho: The population mean is 18.

Ho : m = 18

Sample mean $\overline{x} = \Sigma[(x_i/n)] = (306/20) = 15.3$

Sample Varience $S^2 = \Sigma[(xi - \overline{x})^2)/(n-1)]$

$= (1/19)\ [(15-18)^2+(18-18)^2+(10-18)^2 +(11-18)^2 +(17-18)^2 +(14-18)^2 +(16-18)^2 +(13-18)^2 +(9-18)^2 +(12-18)^2 +(16-18)^2 +(19-18)^2 +(21-18)^2 +(20-18)^2 +(14-18)^2+(19-18)^2+(16-18)^2+(14-18)^2]$

$= (1/19)[380] = 20$

Now $t_{calc} = |[(\overline{x} - m)/(s/\sqrt{n})]| = |[(15.3 - 18)/(4.47/4.47)]| = 2.7$

For $t_{19,\ 0.05} = 2.093$

=> $t_{calc} > t_{19,\ 0.05}$ represents the population mean not supporting the theory and it is not a reasonable one.

Hence **Null hypothesis Ho to be rejected**

Example: ('t' test) (Two group specimen) (no.of samples less than 30)

Two groups of specimens are collected from experimentation follows normal population is given below;

specimens 1	66	73	68	78	81	58	73	57	68	67	70
specimens 2	84	61	80	82	76	63	59	69	--	--	--

Test whether the specimen variance differs significantly @5% level or not.

Solution:

Null hypothesis Ho: The specimen variance differs significantly @5% level.

Sample mean $\overline{x} = \Sigma[(x_i/n_1)] = (759/11) = 69$; $\overline{y} = \Sigma[(y_i/n_2)] = (574/8) = 72$

Sample Varience $S^2 = [\{\Sigma(x_i - \overline{x})^2 + \Sigma(y_i - \overline{y})^2\}/(n_1+n_2-2)]$

$= (538 + 704)/(11+8-2) = 1242/17 = 73$

$S = 8.5$

Now $t_{calc} = |[(\overline{x} - \overline{y})/(s(\sqrt{(1/n_1)} + \sqrt{(1/n_2)})]| = |[(69 - 72)/(8.5)\ (0.654)]| = 0.54$

From table for $[(n_1+n_2-2) = 17df]$

$t_{17,\ 0.05} = 2.11$

=> $t_{calc} < t_{14,\ 0.05}$ represents the Ho is accepted.

Example:

Theory predicts that the preposition of USB usage of five different manufacturers should be 6:3:5:1:2. During the examination of 2000 storage devices, utility of five USBs are observed as 720, 370, 535, 125 and 250. Justify is the experimental research supports the theory or not. ($x^2_{0.05} = 9.488$).

$O_1 = 720$; $E_1 = [(6/17) * 2000] = 706$; $O_2 = 370$; $E_2 = [(3/17) * 2000] = 353$

$O_3 = 535$; $E_3 = [(5/17) * 2000] = 588$; $O_4 = 125$; $E_4 = [(1/17) * 2000] = 118$

$O_5 = 250$; $E_5 = [(2/17) * 2000] = 235$;

$x^2_{calcu} = \Sigma[(O_i - E_i)^2/E_i]$ (Where 'O' s the observed and 'E' is the estimated)

$=\Sigma[\{(720-706)^2/706\}+\{(370-353)^2/353\}+\{(535-588)^2/588\}$

$+\{(125-118)^2/118\}+\{(250-235)^2/235\}]$

$= 0.28+0.82+4.8+0.42+0.96 = 7.28$

⇨ $x^2_{calcu} = 7.28$; $x^2_{4, 0.05} = 9.488$

⇨ $x^2_{calcu} < x^2_{4, 0.05}$, hence Null Hypothesis H_0 to be accepted.

That is the experimental result supports the theory.

Example:

The following information is obtained concerning an investigation of 50 consumers, about the LAN and WLAN usage by service providers A and B

	small scale industries		Total
service provider	LAN	WLAN	
A	6	16	22
B	16	12	28
Total	22	28	50

Can it be inferred that service providers B is relatively more utilized with LAN than in WLAN? Use x^2 test (x^2 value for one degree of freedom at 5% level of significance is 3.841)

Solution

Null hypothesis, Ho = Infers that service providers B is relatively more utilized with LAN than in WLAN

To define the expectation of each cell, multiply the marginal row and column totals for that cell and divide by the overall total.

⇨ For each cell estimation is: E= [(row total×column total)/sample size]

$E(6)_c = [(22 * 22)/50] = 9.68$; $E(16)_c = [(22 * 28)/50] = 12.32$;

$E(16)_c = [(28 * 22)/50] = 12.32$; $E(12)_c = [(28 * 28)/50] = 15.68$

$x^2_{calcu} = \Sigma[(O_i - E_i)^2/E_i]$

$= \Sigma[\{(6-9.68)^2/9.68\}+\{(16-12.32)^2/12.32\}+\{(16-12.32)^2/12.32\}+\{(12-15.68)^2/15.68\}]$

$= 1.4 +1.1 + 1.11 + 0.86 = 4.47$

From tables for n = (R-1) (C-1) = (2-1) (2-1) = 1 degree of freedom

⇨ ($x^2_{0.05} = 3.841$)

⇨ $x^2_{calcu} > x^2_{0.05}$;

Hence, null hypothesis Ho to be rejected.

It concludes that service providers B is not utilized the LAN more than the utilization in WLAN.

Limitations of Tests of hypothesis

There are several limitations that a researcher should keep in mind while using tests of hypothesis. They include.

- It should not be used in a mechanical fashion. The test of hypothesis is not decision making itself; but they are only aids to decision making. Therefore proper interpretation of statistical evidence is important to intelligent decision making.
- The test results simply indicate whether the difference is due to fluctuations of sampling or because of other reasons.
- When the test result shows that a difference is statistically significant it simply suggests that the difference is probably not due to chance.
- Statistical inferences based on the significance tests cannot be said to be entirely correct evidences concerning the truth of the hypothesis. The error in inferences is higher with small samples compared to larger samples.

2.9 ANALYSIS OF VARIANCE

What is ANOVA?

Analysis of variance (ANOVA) is a statistical technique used when multiple sample cases are involved. The significance of difference between the means of two samples can be judged either through Z-test or t-test. But, when more than two samples are involved, the ANOVA technique enables us to perform the significance test simultaneously. It is an important tool in the analyst hand. Professor R.A. Fisher developed a very elaborate theory of ANOVA and its use in practical applications. Later on Prof. Snedocor has greatly contributed to its further development.

The essence of ANOVA is that the total amount of variation in a set of data is broken down into two types, that amount which can be contributed to chance and the amount which can be contributed to specific causes. There may be variation between samples and also within sample items. ANOVA consists in splitting the variance for analytical purposes. Using ANOVA technique, one can investigate any number of factors, which are said to influence the dependent

variable. One may as well investigate the differences among various categories of within each of these factors which may have large number of possible values.

For example, the output of a process may be classified by machines and operators. From this cross classification, it could be determined whether the mean qualities of output of various operation differed significantly. Similarly, it could be determined whether the mean qualities of outputs of various machines differ significantly. This study may help the researcher as how to improve the uniformity in quality of output whether by improving the operations or standardization of machines.

We make two estimates of population variance viz., one based on between sample (treatments) variance and the other within sample (error variance) variance. The two estimates of population variance are compared using F-test computing F-ratio as

$$F = \frac{\text{Estimate of population variance based on between sample variance}}{\text{Estimate of population variance based on within sample variance}}$$

This value of F is than compared with the F-limit for given degree of freedom (from tables). If the calculated F value is equal to or exceeds F – limit value, it is concluded that there are significant difference between the samples or treatment means.

One – way analysis of variance

In one–way analysis, a single factor is considered at different levels (treatments) and each treatment having certain number of observations. Consider a single factor at 'm' levels or treatments, each treatment containing 'n' observations. Let the j^{th} observation of the i^{th} treatment is denoted by x_{ij}.

Treatment	Observations	Totals	Mean
1	x_{11} x_{12} …. X_{1j} ….x_{1n}	x_1	$\bar{x}_1$
2	X_{21} x_{22} …. X_{2j} ….x_{2n}	X_2	$\bar{x}_2$
i	X_{i1} x_{i2} …. X_{ij} ….x_{in}	X_i	$\bar{x}_i$
m	X_{m1} x_{m2} …. X_{mj} ….x_{mn}	X_m	$\bar{x}_m$

$x_i = \sum_{j=1}^{n} x_{ij}$; $\bar{x_i} = {}^{x_i}/_{n}$ for i = 1,2 ….m.

The "dot" subscript notation implies summation over the subscript it replaces. Therefore the grand mean $\bar{x}$.

Is given by $\bar{x}.. = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}}{m.n}$

The total variability of the observations described by the total sum of squares (SST) is given by

SST = $\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}..\right)^2$

The total sum of squares can be partitioned into treatment sum of squares or sum of squares between samples (SSTr) and on error sum of squares (SSE) or within sum of squares. That is,

SST = SSTr + SSE, where

SSTr = $n\sum_{i=1}^{m}(\overline{x_{i.}} - \bar{x}..)^2$ and $SSE = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}..\right)^2$

The SST r is based on the difference between the treatment means and the grand mean, while SSE is based on the difference between an observation within a treatment and the treatment mean. To get the mean square (MS) of SST, SSTr and SSE, divide each quantity by its appropriate degrees of freedom. For example, the d.f for total variance is equal to the total number of items minus 1 ie (m.n – 1). The d.f for SSTr is (m -1) and for SSE it is m (n-1). The one way ANOVA table will be as given below

Source of variation	Sum of squares	Degrees of freedom	Mean square (Ms)	F-ratio
Between treatments (m)	$n\sum_{i=1}^{m}(\bar{x}_{i.} - \bar{x}_{..})^2$	m-1	$\frac{SST_r}{(m-1)}$	$\frac{Ms\ Between}{Ms\ With\ in}$
Samples (SST_r) within the treatments (SSE)	$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}i.\right)^2$	m(n-1)	$\frac{SSE}{m(n-1)}$	
Total sum of squares (SST)	$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}..\right)^2$	(mn-1)		

To test the null hypothesis that the means of 'm' treatments are equal, we compare the variance based on the variation between treatment means ie MS for

SSTr with the variance based on variation within this means ie MS for SSE by computing F-ratio as shown in the last column of the above table. If the F-ratio calculated is larger than the F statistic for F.., m-1, m (n-1) taken from table for level of significance 'x' we reject the null hypothesis and conclude the variable x is influenced by the level of treatments.

Example: Suppose a factor is given three levels of treatment and for each treatment, there are four observations as given in the table, find whether the factor levels are significant.

Levels	observations	Total	mean
A	6 7 3 8	24	6
B	5 5 3 7	20	5
C	5 4 3 4	16	4

$x.. = 60, \bar{x}_{..} = \frac{60}{12} = 5$

Total sum of squares (SST) = $[(1^2 + 2^2 + 3^2 + 5^2) + (0^2 + 0^2 + (-2)^2 + 2^2 + (0^2 + 1^2 + (-2)^2 + (-1^2)] = 32$

Sum of squares within Treatments (SSE) = $[(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2] + [(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2] + [(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2] = 24$

Sum of squares between Treatments (SSTr) = $4[(6-5)^2 + (5-5)^2 + (4-5)^2] = 8$

Source of variation	**SS**	**D f**	**M.S**	**F - ratio**	**$\propto$ = 5% F - limit**
Between treatments	8	(3 -1) = 2	8/2 = 4.00	4/2.67 = 1.5	F .05, 2,9 = 4.26 (table)
Within treatments	24	3(4-1) = 9	24/9 = 2.67		
Total	32	(3X4)-(1) = 11			

The calculated value of F is 1.5, which is less than the table value of 4.26 at 5% level of significance with d f ϑ_1, = 2, ϑ_2, = 9. Hence, the null hypothesis of no difference between means could not be rejected. The difference may be due to chance.

Example: The number of products manufactured through three different shifts over a 2months period is shown in the table below.

Products	Shift A	Shift B	Shift C
Pen drive	60	79	85
Hard Disk	80	78	75
Power pack	43	54	61

Use the 5% level of significance to test for independence of products towards various shifts.

Solution:

Null hypothesis: Ho = $\mu_1 = \mu_2 = \mu_3$

⇨ No significant difference during the test of independence of products towards various shifts.

Products	Shift A **(x_1)**	Shift B **(x_2)**	Shift C **(x_3)**	x_1^2	x_2^2	x_3^2
Pen drive	60	79	85	3600	6241	7225
Hard Disk	80	78	75	6400	6084	5625
Power pack	43	54	61	1849	2916	3721
Σ	183	211	221	11849	15241	16571

N = Number of observations = 4 rows x 3 columns = 12

Total observation = $T = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 = 615$

Correction factor = $(T^2)/N = (100)^2/12 = 31519$

Sum of Square Total = SST = $\{\Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 - [(T^2)/N]\}$

= 43661-31519 = 12142

Column Sum of Squares = SSC = $\{(\Sigma x_1^2/N_1) + (\Sigma x_2^2/N_2) + \{(\Sigma x_3^2/N_3) - [(T^2)/N]\}$

= 8372 + 11130 + 12210 – 31519 (where $N_1=N_2=N_3=4$)

= 193

Error Sum of Square = SSE = SST – SSC = 11949

F-Table ANOVA Table

Source of variation	Sum of Square	Degree of freedom (D.f)	Mean sum of squares	Fratio(Variation ratio)
Between columns	SSC= 193	C-1=3-1=2	SSC/D.f=(193/2)=96.5	F_{calc} = (96.5)/(1328) = 0.07
Error	SSE= 11949	N-C=12-3=9	MSE = SSE/D.f =(11949/9)=1328	

$F_{calc} = 0.07$

From table (C-1, N-C) (2,9) $F_{(2,9,0.05)} = 4.26$

Defines $F_{calc} < F_{(2,9,0.05)}$; Hence Ho is accepted.

Two – way Analysis of variance

Two – way ANOVA technique is used when the data are classified on the basis of two factors. For instance in the case of process output, two factors considered could be operators and machines. Such a two way design may have repeated measurements of each factor or may not have repeated values. The technique is little different in the case of repeated measurements where we also compute interaction variations.

i) When repeated values are not there.

As we do not have repeated values, we cannot directly compute the sum of squares within treatments (SSE) as we had done in the case of one-way analysis. We have to calculate this residual or error variation by subtraction, once we have calculated the sum of squares for total variance (SST) and for variance between treatments (SSTr). ie, SSE = (SST – SSTr – SSB) where SSB is the sum of square between columns (blocks). The total sum of squares (SST) is given by

$SST = SST_r + SSB + SSE$

$SST = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}..\right)^2$ d.f: (m.n–1)

$SST_r = n\sum_{i=1}^{m}(x_i - \bar{x}..)^2$ d.f: (m-1)

$SSB = m\sum_{j=1}^{n}\left(x_{.j} - \bar{x}..\right)^2$ df: (n-1)

$SSE = \sum_{i=1}^{m}\sum_{j=0}^{n}\left[\left(x_{ij} - \bar{x_i}\right) - \left(x_{.j} - \bar{x}..\right)\right]^2$ df: (n-1)(m-1)

where m = number of treatments (rows) and n = numbers of columns (blocks). The AVOVA table can be set up as below:

Source of variation	Sum of squares	d f	Mean square (MS)	F - ratio
Between block (columns) SSB	$m\sum_{j=1}^{n}\left(x_{.j} - \bar{x}..\right)^2$	(n – 1)	$\frac{SSB}{(n-1)}$	$\frac{MS\ of\ SSB}{Ms\ of\ SSE}$
Between treatment (rows) SSTr	$n\sum_{i=1}^{m}(x_{i.} - \bar{x}..)^2$	(m – 1)	$\frac{SSTr}{(m-1)}$	$\frac{MS\ of\ SSTr}{Ms\ of\ SSE}$
Residual error SSE	SST-SSTr-SSB	(n -1) (m -1)	$\frac{SSE}{(n-1)(m-1)}$	
Total SST	$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}..\right)^2$	(m. n – 1)		

The MS residual or residual variance provides the basis for the F – ratios. The MS residual is always due to the fluctuations in sampling and hence serves as the basis for significance test. Both the F – ratios are compared with their corresponding table values for a given d f and at a specified level of significance. As usual if it is found that the calculated F – ratio concerning variation between columns (blocks) is equal to or greater than its value, then the difference between block means is considered significant. Similarly, the F – ratio concerning variation between treatment means (rows) can be interpreted.

ii) When repeated value are there

In the case of two – way design with replication we can obtain a separate independent measure of inherent or smallest variations. By introducing replication, we are able to separate out the variation due to interaction effects from that due to random error. An interaction effect is present if the observed values for different levels of one factor are altered by the presence of other factor. The sum of squares due to error within each cell (sample for each treatment) is determined by first determining the subtotal sum of squares (SSS) and then sum of square within cells computed by subtracting SSS from SST. Ie, SSWC = (SST – SSS). Finally the interaction sum of squares is given by,

SSI = [SSS – (SSTr + SSB)]

Suppose, the number of columns (blocks) is n, and the number of treatments ‘m’, and within each treatment the number of replications is ‘r’, then,

$SST = \sum x_{ij}^2 - \frac{T_{..}^2}{m.n.r}$; where T is the cell subtotal and T.. is the grand total

Treatment sum of squares (between rows) is,

$SST_r = \sum \frac{Ti_.^2}{n.r} - \frac{T_{..}^2}{m.n.r}$; where $T_{.i}$ is the sum of column total of cell totals

The block sum of squares (between columns) is,

$SSB = \sum \frac{T_j^2}{m.r} - \frac{T_{..}^2}{m.n.r}$;Where $T_{.j}$ is the sum of each column

$SSB = \sum \frac{T^2}{r} - \frac{T_{..}^2}{m.n.r}$

Sum of squares within cells,

SSWC = [SST – SSS]

The interaction sum of squares is given by

SSI = SSS – SSTr – SSB.

Once the computations are made, the results could be in the ANOVA table as given below.

Source of variation	S S	D f	M. S.	F - ratio	5% F - limit
Between blocks (columns)	SSB	(n – 1)	$\frac{SSB}{(n-1)}$	$\frac{MS\ of\ SSB}{MS\ of\ SSWe}$	
Between treatments	SSTr	(m – 1)	$\frac{SSTr}{(m-1)}$	$\frac{MS\ of\ SSTr}{MS\ of\ SSWe}$	
Within cells (error)	SSWc	m. n (r – 1)	$\frac{SSWe}{m.n(r-1)}$	–	
Interaction	SSI	(m-1) (n-1)	$\frac{SSI}{(m-1)(n-1)}$	$\frac{MS\ of\ SSI}{MS\ of\ SSWe}$	
Total		(m. n. r – 1)	–		

To check whether there is a significant interaction the two factors, the F – ratio for interaction is checked against the tabulated value.

2.10 DESIGN OF EXPERIMENTS

Importance of Experimental Design

The experimental studies generally provide a logical and systematic method for answering questions “what will happen if this is done, when certain variables are carefully controlled or manipulated?” Infact deliberate manipulation is a part of the experimental method. In an experiment, the researcher measures the effect of an experiment intentionally. Experimentation not only provide a method for hypothesis testing but also permit drawing of inferences about causality. As a result of experiments, relationship between data and the unknown in the universe is established.

The greatest benefit can be gained from statistical analysis, when experiments are planned in advance so that data are taken in a way that will provide the most unbiased and precise results commensurate with the desired expenditure and time spent. The major benefit from statistically designed experiment are,

- more information per experiment will be obtained than with unplanned experimentation.
- provides an organized approach to the collection and analysis of information

- credibility of the conclusions of experimental programme since the variability and sources of experimental error are made clear by statistical analysis.
- Ability to discover interactions between experimental variables.

The experimental design approaches are assuming greater significance due to the strong emphasis shown to improve quality. The technique developed in Japan by Genichi Taguchi employ experimental design methods in a unique way to produce robust designs.

Randomization

If an experimental programme involve large number of tests, it is necessary to randomize the order in which the specimens or combination of factors are selected for testing. Thus, by randomization any one of the many specimen involved in the experiment has an equal chance of being selected for a given test. By this approach, bias due to uncontrolled second order variables is minimized. For example in any extended testing programme errors can arise over a period of time owing to subtle changes in the characteristics of testing equipments or in the proficiency of the operator of the test. Similarly, when test specimens are prepared from a large forgings or ingots, the possibility of variation of properties in the forgings must be considered.

Variables and factors

Any experimental investigation starts with preliminary familiarization of experiments. This is more aimed at defining the problem. Once the goals of investigation are well defined, the large number of variables which are possible should be reduced a few important ones. The statistical design will be very useful. As the experiment moves to optimization stage, the number of variables should be reduced further. Hence, again statistical design provides effective means.

The output of an experiment run is the data termed as response variables. Response variables can be classified as quantitative, qualitative or Quantal. A quantitative response, which is measured by a continuous scale, is the most common and easiest to work with in statistical analysis. Qualitative responses

like luster, odor etc. can be ranked on an ordinal scale, eg: worst (0) to best (1). Quantal or binary response produces one or two values, go or no-go ie. Pass or fail.

Factors are experimental variable that are supposed to be controlled by the experimenter. An important part of planning of experimental program is identifying the significant variables that affect the response and deciding how to exploit them in the experiment. Factors may be independent in the sense that the level of one effect factor is independent of the level of other factors. However, two or more factors may interest with one another ie. The effect on the response of one variable depends upon the levels of other variables.

In addition to primary variables that are under the control of experimenter, there are other variables which may not be strictly under experimenter's control. The use of randomization of test runs removes this unconscious bias. Also, the use of experimental blocks produces relatively homogeneous test conditions and minimizes this. When an experiment is run with blocking, the effect of background variables is removed from the experimental error.

Planning of experiments

We should realize that not all primary factors may be capable of variation with equal facility. Complete randomization may also be impractical often the final experiment is a compromise between information that can be obtained and the cost involved in that. In developing the experimental design, the physical reality gets precedence over strict adherence to statistics.

It is important to make some initial estimate of the overall repeatability before embarking on major experimental test programme. The experience of previous experiments may also be useful in this. In case the pilot experiments indicate large variability in response then primary variables may not have been properly identified. It is not necessary to conduct statistically designed experiments all the way through completion. Infact there are advantages in conducting experiments in stages, taking into account the advantages of information gathered in early results.

The statistically designed experiments may be classified as under.

i) Blocking Designs

The Blocking designs use blocking techniques to remove the effect of background variables from the experimental error. The common type of blocking designs include,

Randomized block and Balanced incomplete block	: Remove the effect of singly extraneous variables
Latin square and Youden square	: Remove the effect of two extraneous variables
Gracco - Latin square and Hyper Latin square	: Remove the effect of three or more extraneous variables.

ii) Factorial designs

Factorial designs are experiments in which the levels of each factor is combined with the levels of all other factors to conduct experiment simultaneously. This is an important class of experiment. In factorial we have a) Full factorial designs b) Fractional factorial designs and c) Use of orthogonal arrays.

iii) Response Surface designs

The response surface designs are used to develop empirical functional relation between the factors (independent variables) and the response (output variables).

2.11 Factorial Designs

Factorial designs are experiments in which the level of each factor is combined with the levels of all other factors to conduct experiment simultaneously. This is an important class of experiment. In factorial we have a) Full factorial designs b) Fractional factorial designs and c) Use of orthogonal arrays.

2.11.1 Full factorial Designs

A full factorial design is one, in which we control several factors and investigate their effects at each of two or more levels. The experimental design consist of making an observation at each of all possible combinations that can be formed for the different levels of factors. Each different combination is called a "treatment combination". The approach in a factorial experiment is much different than in traditional experiment, in which all factors but one are held constant. The simplest and most common type of factorial design is one that uses two levels, i.e., 2^n factorial design.

To illustrate the point, let us consider three factors A, B, C each at two levels, one at low level indicated by a (-) sign and the other at higher level by a(+) sign. We are interested in the result of the experiment due to change in factor 'A' alone, B-alone and C-alone. These three potential output changes are called "main effects".

There are also first order effects due to joint change in factors A and B, A and C and B and C. These interactions are denoted by AB, AC and BC respectively. A second order interaction effect due to simultaneous changes in A, B and C is denoted by ABC. This experiment is a three factor 2^3 design.

The main effect of a factor is the change in response produced by a change in the level of the factor.

$$\text{Main effect of } x_i = \frac{\sum responses\ at\ high\ x_i - \sum responses\ at\ low\ x_i}{Half\ the\ number\ of\ runs\ in\ experiment}$$

Thus, the main effect of factor X_B is,

$$\text{Main effect of } x_B = \frac{(Y_b + Y_{ab} + Y_{bc} + Y_{abc}) - (Y_1 + Y_a + Y_c + Y_{ac})}{4}$$

Similarly, other main effects X_A, X_c, can be calculated.

The effect for the interaction of B & C is given by,

$$\text{BC interaction} = \frac{(Y_1 + Y_a + Y_{bc} + Y_{abc}) - (Y_b + Y_{ab} + Y_c + Y_{bc})}{4}$$

Using Yate's notation, the three factor 2^3 factorial design table will be as given below.

Yate's std order	Run No.	Level of Factors						
		X_A	X_B	X_C	$X_A X_B$	$X_A X_C$	$X_B X_C$	$X_A B_C$
1	1	–	–	–	+	+	+	–
d	2	+	–	–	–	–	+	+
b	3	–	+	–	–	+	–	+
ab*	4	+	+	–	+	–	–	–
c	5	–	–	+	+	–	–	+
ac*	6	+	–	+	–	+	–	–
bc*	7	–	+	+	–	–	+	–
abc@	8	+	+	+	+	+	+	+

*: First order interaction @: Second order interaction.

Example: A heat treatment involves three processes A, B, C, each treatment set at two levels (a high and low). The tensile strength is measured in KP_a (is, response 'Y') as indicated in the table below.

Treatment	Response	X_A	X_B	X_C	X_{AB}	X_{AC}	X_{BC}	X_{ABC}
(1)	1041	-	-	-	+	+	+	-
a	1014	+	-	-	-	-	+	+
b	1014	-	+	-	-	+	-	+
ab	696	+	+	-	+	-	-	-
c	1255	-	-	+	+	-	-	+
ac	1096	+	-	+	-	+	-	-
bc	1193	-	+	+	-	-	+	-
abc	1089	+	+	+	+	+	+	+

The main effect of

$$A = \tfrac{1}{4}[(1014 + 696 + 1096 + 1089) - (1041 + 1014 + 1255 + 1193)]$$

$$= \tfrac{1}{4}[3895 - 4501] = -606/4 = -152.0.$$

The minus sign indicates that the lower level of factor 'A' resulted in higher values of the response.

To test the significance of the effect of factor 'A', we can perform a paired t-test, by comparing the mean value of factor A at its high level within mean at low level.

$\bar{X}_L = 1126$

$\bar{X}_H = 974$

$\bar{d} = \bar{X}_L - \bar{X}_H = 1126 - 975 = 152$

$Also\ \bar{d} = \frac{\sum d}{4} = \frac{608}{4} = 152$

$\bar{X}_L = 1126, \bar{X}_H = 974, \sum d = 608$

Low A	High A	Difference d	Deviation From d
1041	1014	27	-125
1014	696	318	166
1255	1096	159	7
1193	1089	104	-48

$$S^2 = \frac{\sum(deviation)^2}{(n-1)} = \frac{(-125)^2 + (166)^2 + (7)^2 + (-48)^2}{(4-1)}$$

$S^2 = 15178$; $\therefore S = 15178 = 123.2$

we assume Null hypothesis as, there is no difference in mean due to factor – A.

ie, Ho :μ = o and H1 : μ ≠ o

The t – statistic can be obtained from $t = \frac{\bar{d}}{s\sqrt{n}} = \frac{152}{123.2/\sqrt{4}} = 2.47$

$\therefore$ Calculated t = 2.47; $\vartheta = 4 - 1 = 3$

For a two tailed test, $t_{0.10} = 2.35$. Therefore, the null hypothesis is rejected at a 10% significance level. Similarly the pained t – test can be used to check the significance of interaction effects. The precision of statistical analysis can be improved by replication.

2.11.2 Fractional factorial design

Because of cost and time limitations, we may wish to investigate an experimental model with less than full factorial design. One type of partial layout is called fractional factorial design. This type is usually used when four or more factors are considered. The number of treatment combinations in a factorial design increases with increase in number of factors. If n = 4, with 2 levels, the number of combinations will be, $2^4 = 16$, but if n = 8, $2^8 = 256$ and so on. If the number of main effects will be 'n', the first order interaction will be $n(n-1)/2$,

second order will be $\frac{n(n-1)(n-2)}{(2)(3)}$

Considering a 2^6 factorial experiment, there are six main effects: since n = 6;
First order interaction: n (n-1) / 2 = 15,
Second order interaction: n (n-1) (n-2) / (2)(3) = 20;
Third order interaction: n (n-1) (n-2) (n-3) / (2)(3)(4)(5) = 15:
Fourth order interaction: n (n-1) (n-2) (n-3) (n-4) / (2)(3)(4)(5) = 6;
Fifth order interaction: = 1 and sixth order interaction = 1.
If we neglect higher order interaction effects, we can design an experiment that is some fraction of the total factorial design. For example, if we consider one half replicate of 2^3 factorial design, then N = r (2^n) – ½ (2^3) = 4. We could split the treatment contributions into two blocks of the previous example to study the main effects of A, B, & C.

2.12 Orthogonal arrays

Experiments using orthogonal arrays

In full factorial experiments, we conduct experiments under all combinations of factor levels. For example, if an experiment consists of three factors at four levels each, we need to conduct 4^3 = 64 experiments to try all combinations. Certainly, the interaction effects could also be studied in this. Fortunately, in most practical situations, the interaction effect is significantly less or absent, and additive models provide excellent approximation. In this context, use of the concept of orthogonal arrays provides a powerful tool for frugally exploring relations that involve many factors.

Designed experiments are matrix experiments, which consist of a set of experiments where we change the settings of various input parameters (factors) from one experiment to another. Orthogonal arrays are special matrices which allow the effect of several parameters to be determined efficiently. Dr. Genichi Taguchi introduced the concepts of Robust design to find optimum values of design parameters or process variables wherein he made the use of orthogonal arrays extensively.

As the name suggests, the columns of the matrix of the array are mutually orthogonal. Here, the orthogonality implies that for any pair of columns, all combination of factor levels occur and that they occur equal number of times. This is called the balancing property and its means orthogonality. In order to

preserve the benefit of using an orthogonal array, it is important that all experiments in the matrix be performed. If experiments corresponding to one or more rows are not conducted, or if their data are missing or erroneous the balancing property and, hence the orthogonality is lost. In some situations, incomplete matrix experiments can give useful results, but analysis of such experiments is complicated.

Taguchi has tabulated 18 basic orthogonal arrays, which are called **"Standard orthogonal arrays",** which are available in many standard texts. The process of fitting an orthogonal array to a specific project has been made easy by the standard orthogonal arrays and the graphical tool called '**linear graphs'** developed by Taguchi to represent interactions between pairs of columns in the array. Before constructing an orthogonal array, one must define the following. i) Number of factors to be studied ii) Number of levels of each factor iii) Specific 2 factor interactions to be estimated iv) Specific difficulties if any in the experiment.

The first step in constructing an orthogonal array to fit a specific case study is to count the total degrees of freedom that tells the minimum number of experiments that should be performed to study the main effects of all control factors and chosen interactions. In many problems one of the 18 standard tabulated orthogonal arrays can be directly used to plan the matrix experiment. Orthogonality is not lost by keeping empty one or more columns of the array. In orthogonal array experimentation, one may choose not to estimate any interactions among control factors, but in situations where it is needed linear graphs makes it easy to plan.

The first four standard orthogonal arrays and the L9 (3^4) Orthogonal array of Taguchi is given below for reference.

Standard Orthogonal Arrays (First four)

Orthogonal array	Number of rows	Maximum No. of factors	Maximum No. of columns of these levels			
			2	3	4	5
L4	4	3	3	-	-	-
L8	8	7	7	-	-	-
L9	9	4	-	4	-	-
L12	12	11	11	-	-	-

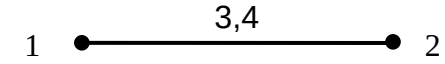
L_9 (3^4) orthogonal array

Expt. No.	Columns			
	1	2	3	4
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

The array L_9 (3^4) has 9 rows and four 3 level columns. Similarly array L_{18} (2^1 3^7) has 18 rows, one 2 level Column and seven 3 level columns. The Number of rows of an array represents the number of experiments.

Linear graphs represent the interaction graphically. It make it easy to assign factors and array. The columns of an array represented by 'data' and the connecting line, the interaction of two columns represented by dots is contained in (confounded with) the column represented by the line.

Linear graphs for L_9



The first step in constructing an orthogonal array for a problem is to decide the total degrees of freedom, which tells the minimum number of experiments to be performed to study all chosen factors. Suppose, in a problem we have four factors, each at three levels, the total number of d.f is 9 (that is, for overall mean = 1, for 4 factors 4 (3-1) = 8. ∴ total is 8 + 1 = 9). A minimum 9 experiments required. The use of matrix method with orthogonal array is explained using a case example. In this example the interaction effects of factors is not taken into account. Ex: Consider a project where we are interested in determining the effect of four process parameters: temperature (A), pressure (B), settling time (C), and cleaning method (D) on the formation of certain surface defects in a chemical vapor deposition on silicon wafers. Three parameter settings are chosen to cover the range of interest. The factors and the chosen levels are given in table 1. The starting values are indicated by an underscore. Our goal is to determine the best settings for each parameter so that the surface defect formation is minimized.

Table - 1

Factors	Levels		
	1	2	3
A. Temperature o^c	T_o - 25	$\underline{T_o}$	T_o + 25
B. Pressure	P_o - 200	$\underline{P_o}$	P_o + 200
C. Settling time (method)	$\underline{t_o}$	t_o + 8	t_o + 16
D. Cleaning method	$\underline{None}$	CM_2	CM_3

There are four factors at three levels each. Therefore the total degree of freedom for this case is 9 including the grand mean. i.e. 4(3 - 1) = 8+1 = 9. Therefore L9 (3^4) array of Taguchi is chosen.

The matrix experiment selected for this project is given in table. 2. It consists of 9 individual experiments corresponding to the 9 rows. The four columns of the matrix represent the four factors as indicated in the table. The entries in the matrix represents the levels of the factors. Thus, the experiment 1 is to be conducted with each factor at the first level. This implies the experiment 1 has factor levels as (T_o – 25) temperature, (P_o – 200) for pressure, (t_o) minutes settling time and no cleaning, similarly for other experiments also. For experiment 4, temperature is at level 2, pressure at level 1, settling time at level 2 and cleaning at level 3(CM3). The experimental settings for experiment 4 can also be referred concisely as A_2, B_1, C_2, D_3.

The matrix experiment of table 2 is the standard orthogonal array L_a of Taguchi. As the name suggests, the columns of this array are mutually orthogonal. For any pair of columns, all combination of factor levels occur and they occur equal number of times. In this matrix for each pair of columns there exists 3x3 = 9 combination of factor levels and each combination occurs precisely once. For columns 1 and 2, the 9 possible combination of factor levels are (1,1); (1,2); (1,3); (2,1); (2,2); (2,3); (3,1); (3,2); and (3,3) occur in experiments (or rows) 1,2,3,4,5,6,7,8 and 9 respectively.

(Table – 2)

Expt. No.	Column no. & factor assigned				Observation η (dB)
	1 (A)	2 (B)	3 (C)	4 (D)	
1	1	1	1	1	η_1 = - 20
2	1	2	2	2	η_2 = - 10
3	1	3	3	3	η_3 = - 30
4	2	1	2	3	η_4 = - 25
5	2	2	3	1	η_5 = - 45
6	2	3	1	2	η_6 = - 65
7	3	1	3	2	η_7 = - 45
8	3	2	1	3	η_8 = - 65
9	3	3	2	1	η_9 = - 70

Suppose for each experiment we observe the surface defect count per unit area at three locations on 3 silicon wafers, so that there are nine observations per experiment. Let us define the following formula for summary statistic, η_i for i th experiment.

η_i = -10 $\log_{10}$ (mean square defect count for except i) where the mean square refers to the average of the squares of the 9 observations in the experiment i. We refer to the η_i calculated using the above formula as the observed η_i. The observed η_i for the 9 experiments is shown in the last column of table 2. The objective of minimizing surface defects is equivalent to maximizing ŋ. The summary statistic ŋ.is called signal – to – noise (s/n) ratio.

In order to estimate the effect of the four process parameters from nine experiments, first the overall mean values of ŋ for the experiment is obtained as,

$$m = \frac{1}{9}\sum \eta_i = \frac{1}{9}[\eta_1 + \eta_2 + \ldots\ldots + \eta_4]$$

The effect of a factor level is defined as the deviation it causes from the overall mean. Let us examine how the experimental data can be used to evaluate the effect of temperature at level A_3. Temperature was level A3 for experiments 7.8, and 9. The average S/N ratio for these experiments, devoted by m_A is, $m_A = \frac{1}{3}[\eta_7 + \eta_8 + \eta_9]$

The effect of temperature at level A_3 is given by (m_{A3} – m). From the table 2, it may be noticed that for experiments 7, 8, 9, the pressure level takes values 1,2 & 3 respectively. Similarly for these three experiments, the levels of settling time and cleaning method also take values 1, 2 & 3. So the quantity m_{A3} represents an average ŋ when the temperature is at level A3, where the averaging is done in a balanced manner over all levels of each of the other three factors.

The average S/N ratio for levels A_1 and A_2 of temperature, as well as those for various levels of other factors can be obtained in a similar way. Thus for example $m_{A2} = \frac{1}{3}[\eta_4 + \eta_5 + \eta_6]$ represents the average S/N ratio for temperature at level 2, and $m_{B3} = \frac{1}{3}[\eta_2 + \eta_5 + \eta_8]$ is the average S/N ratio for pressure at level B_2. Because the matrix experiment is based on orthogonal array, all the level averages possess the same balancing property described for M_{A3}.

By taking the numerical values of η listed in Table 2, the average η for each level of the four Factors can be obtained as listed in table 3. For Example, the average value for B_2 is 1/3[-10 -45 -65] = -40

Factor	Level		
	1	2	3
A-Temperature	-20	-45	-60
B- Pressure	-30	-40	-55
C-Settling time	-50	-35	-40
D-Cleaning method	-45	-40	-40

(Table – 3)

As already stated the primary goal is to determine the best or optimum level of each factor. The optimum level for each factor is the level that gives the highest value for ŋ in the experimental region. In this case, our goal is to minimize the surface count. Since -log is a monotone decreasing function, it implies that we should maximize ŋ. Note that ŋ = -20 is preferable to ŋ = -45 because -20 is greater than -45. From table -3, we can determine the optimum levels for each factor as the highest value of ŋ. The best temp setting is A_1, best pressure is B_1 and best settling time is C_2 and the best cleaning method is either D_2 or D_3. We can, therefore conclude that settings A_1 B_1 C_2 D_2 or A_1 B_1 C_2 D_3 would give the highest ŋ or the lowest surface defect count.

REVIEW QUESTIONS

Part A

1. Is computer takes the important role in research? Justify.
2. List any eight significant applications of computer in research
3. Define the application of spread sheet tool in research.
4. State the principles of statistical computation with example.
5. What is the significance of statistical analysis in research?
6. When do we apply cluster analysis in research?
7. In a production industry, test was carried out 7 times. The probability of non defective component in any test is 0.6 Let X denotes the number of non defective that come up, then find: P(X = 5)
8. State exponential distribution function and estimate for x= 9 and λ= 7.
9. Population of an experimentation has m= 4 and σ = 2. Find the standard score of a randomly selected value being greater than 6. State its significance.
10. State Gaussian distribution function and estimate for x = 8, σ = 1.6 and μ = 4.

11. Draw the histogram for the experimentation observations 7, 3, 6, 4, 8, 5, 7.
12. F(x) is a normally distributed function with mean 11 and variance 5. Calculate F(x) for x = 4.
13. In a R&D division average failure rate is modeled by a Poisson distribution as 35%. Find the probability that exactly 3 will fail out of 7 developed systems.
14. How the sample size of experiments were decided.
15. List the classification and importance of sampling design.
16. Why probability sampling is generally preferred in comparison to non probability Sampling?
17. State the significance of hypothesis testing.
18. State the procedures of hypothesis testing
19. What are the types of errors expected in hypothesis test?
20. Define the condition to accept the alternative hypothesis during the testing of any research with an example.
21. Sate the condition to reject the null hypothesis during the testing of any research with example.
22. State the significance of Type I error and type II error.
23. What are type II errors?
24. Determine the number of second order interaction of 2^6 factorial experiment
25. In Fractional factorial design, If the number of main effects will be 'n' then determine the first order interaction.
26. Brief the significance of orthogonality.
27. Match the following

(A) measure of ANOVA	(i) repeatability of a measurement
(B) Orthogonal array	(ii) Estimate the uncertainties in the results
(C) data analysis	(iii) Less no. of tests cases but high functional coverages
(D) accuracy	(iv) disagreement between true or accepted value
(E) Precision	(v) agreement between a measured and true value
(F) Error	(vi) Two sources of variation in the data

Part B

1. Illustrate the utilization of computers in research
2. Is statistical tools are essential to converge the research data? Explain with suitable example.
3. Describe the processes involved in sample design through example.

4. In research, the experimental set up is built for a specific purpose. Justify and validate through suitable example.
5. Discuss the classifications of important experimental design and its inferences with example.
6. Performance of the hybrid algorithm is tested 12 times by a software jig. The test response is defined either as pass or as fail. Probability of pass on any test is 0.2. Let X denotes the number of test that comes up. Find: (i) P(X = 2) (ii) P (1< X < 4) (iii) P (1 < X < 8). Comment the significance of the findings.
7. "Describe the basic principles of experimental designs with their significance, interms of The principle of replication; The principle of randomization; The principle of local control through suitable example.
8. Random sample of component to be selected from each of three bin to be examined and the number of components selected from each bin is as follows:

   Bin1 : 7 4 3 5; Bin2 : 6 2 5 4; Bin3 : 4 6 3 5. Determine is the selection defines any difference in the mean between bins 1, 2 and 3 at 5% level of significance. (Where, level of significance at 5% for F (2, 8) is 4.46}.
9. Theory predicts that the preposition of mobile phone usage of Six different manufacturers should be 4:2:3:5:1:2. During the examination of 2000 various mobile devices, utility of six specific manufacturer's mobile phones are observed as 490, 230, 370, 535, 125 and 250. Justify is the experimental research supports the theory or not. ($x^2_{0.05}$ = 9.488).
10. A quality assurance department performed a test on "Durability", "Flexibility" and 'User Friendly". Observations of the test shown below are in terms of its grading. Using ANOVA find its significant factor at 0.5% of level of confidence. (F = 3.88, from table).

Test	Observations				
Durability	8	6	5	7	6
Flexibility	9	5	4	8	7
User Friendly	5	8	4	7	6

   Prepare the report to state whether the factor differs significantly @ 5% level or not.
11. In a production industry, 3 systems are used for 5 days in a week. Their hourly usage per day is tabulated below. Carry out all possible ANOVA and make a report.

System	Mon	Tue	Wed	Thu	Fri
A	10	8	8	7	7
B	6	7	4	8	5
C	8	6	6	9	6

12. Quality of a networking system is to be justified with three different jig systems like A, B &

C with two level parameters as low level and high level. The response of a system is measured in its corresponding units as specified below

Expt.No.	Factor levels			Responses
	A	B	C	
1	-	-	-	16
2	+	-	-	20
3	-	+	-	19
4	+	+	-	22
5	-	-	+	21
6	+	-	+	18
7	-	+	+	26
8	+	+	+	24

Using factorial design,

i) Find the main effects and interaction effects of factors A, B & C

ii) Through the statistical test determine the significance between effects of levels of factors.

13. Quality of a networking system is to be justified with three different jig systems like A, B & C with two level parameters as low level and high level. The response observed is as follows 18, 15, 17, 19, 16, 14, 21 and 13. Using factorial design, find its main effects and interaction effects. Summarize its significance in a report form and state your recommendations
14. QoS of three different algorithms are represented by the two level parameters as low level and high level. The response of a system is measured in its corresponding units are as follows 9, 7, 14, 16, 6, 10, 13, 11. Find the main effects and interaction effects of above said algorithms and suggest the quality aspects.
15. Explain the difference between factorial design and fractional factorial design. Use an appropriate example to explain.
16. Define and discuss the balancing property of orthogonal arrays with its merits and issues.