VALIDATION IN PHARMACEUTICAL INDUSTRY

Concepts, Approaches & Guidelines

2nd Edition

P.P. Sharma
M.Pharm.
Former Dy. Drugs Controller
Govt. of NCT of Delhi

VANDANA PUBLICATIONS
DELHI - 110 034

Published by
Vandana Publications
LU-56, Vishakha Enclave,
Delhi - 110034

First Edition 2007
Second Edition 2013

ISBN 978-81-905957-6-6

Price:
India & Nepal Rs. 1600.00
Other Countries US $ 80.00

Printed at
Rakmo Press Pvt. Ltd.
C-59, Okhla Industrial Area, Phase-I
New Delhi-110020

Dedicated to
the memory of
my parents

Preface to Second Edition

The first edition of this book was well received by the pharmaceutical industry and pharmacy education institutions. Over the years, since this book was published, there have been some changes e.g. acceptance limits in media fill trials, acceptance limits in cleaning validation. Thus, there was a need for review of this book.

In the first edition, process validation, in general, was discussed and then some special processes were discussed. But validation of packaging process was not discussed. In this edition, a chapter has been added on packaging process validation.

Medical devices is an emerging field in India. Several medical devices have been notified as drugs by Government of India. However, processes used in medical devices are much varied than the processes used in the pharmaceutical industry. Therefore, a chapter has been included in this edition, on validation of medical devices.

Before process validation can be undertaken effectively, two important activities are calibration of measuring devices and qualification of equipment/instruments. Unless these two activities have been done correctly, validation will not be reliable. Therefore, a chapter has been included on calibration and qualification of equipment/instruments in this edition.

All the chapters of the first edition of this book have been reviewed and updated.

I hope, this volume will be found useful by technical staff in the pharmaceutical industry, regulatory officers in the regulatory departments and the faculty and students in the pharmacy education institutions. I also look forward to comments from the readers and critics.

I am thankful to all those who have helped me to bring out this book including Mr. Satish Agrawal who has given shape to this edition and Rakmo Press Pvt. Ltd., New Delhi who have printed this book.

January 2013 **P.P. Sharma**

Preface to First Edition

The first text of GMP under Schedule M to the Drugs and Cosmetics Rules did not have validation as an element of GMP. Even the first text of WHO GMP did not have validation as a separate element though validation was obliquely mentioned. But the revised text of GMP under Schedule M, now has validation as a separate element. The WHO GMP text (revised) has an elaborate element namely, qualification and validation. Thus validation is a regulatory requirement.

For regulatory compliance, it is necessary to understand the concepts, approaches, organizing validation and other practical guidelines. There is no Indian publication covering all the aspects of validation which can be used by the pharmaceutical industry. Forgein publications are very costly and also the technical staff, particularly in the small scale and medium scale industry, may not be aware of the sources from where foreign publications can be obtained.

I have made an attempt to fill this void. Understanding some important principles and tools of statistics is necessary to analyze the data generated during validation experimental work. In view of this, a chapter on important principles and tools of statistics which are used in validation and quality control, has been included in the book. I hope, the validation team in pharmaceutical industry, regulatory officers of states and central governments, pharmaceutical consultants and pharmacy faculty in colleges will find this volume useful.

I am thankful to my friends, notably, Sh. Devinder Pal who encouraged me to write book on validation and to those who helped me in collection of information. My thanks to Dr. R.A. Singh of M/s. Arbro Pharmaceuticals Ltd., New Delhi who has reviewed the chapter on Analytical Method Validation and made useful suggestions. My son Sh. Rajat Sharma has reviewed the chapter on Computer System Validation and deserves my thanks. My thanks

to other contributors like Sh. Satish Agrawal who has helped me in giving the shape to the book and Rakmo Press Pvt. Ltd. for printing this book. I look forward for similar support in future also.

This being my first attempt on the subject, I will eagerly await the comments from the readers. Their comments and healthy criticism will help me improve the title in future.

May 2007 **P.P. Sharma**

Contents

List of Tables

List of Figures

Chapter 1

Principles and Tools of Statistics used in Validation and Quality Control of Drugs

1. COMMONLY USED TERMS AND THEIR MEANINGS

We can see applications of statistics in every day life ranging from working out odds associated with gambling to assessing performance in sporting events. Application of statistics in sciences is inevitable. Pharmaceutical sciences are no exception. In case of pharmaceutical sciences, knowledge of statistics is required to interpret data generated from all types of physico-chemical & biological evaluations. Statistical input in sampling and testing for quality control, stability testing, process validation are applications that can be routinely seen in pharmaceutical industry. In view of this, it is relevant to include a chapter in this book on statistical principles and tools which are used in the quality control of drugs & validation of processes.

Statistics has been defined as the science of collecting, analyzing and interpreting data related to an aggregate of individuals by Kendall and Buckland[1]. Statistics can be subdivided into two subcategories, namely, descriptive and inferential.

Descriptive statistics provides general information about statistical properties of data like mean, median, standard deviation etc. Inferential statistics is involved in drawing conclusions based on information derived from experimentation. It is almost impossible to produce two identical units. Weights of tablets derived from the same batch differ. The information described within such variable can be manipulated, presented and statistically analyzed to provide a wholesome idea about the properties that are variable. The presence of such variability is the main reason for necessity of inferential statistics.

1.1 Variables

Sokal and Rohlf[2] described variable as a property with respect to which individuals in a sample differ in some ascertainable way. In a process, if biological and/or chemical measurements are made, replicate measurements of a particular property will exhibit different numerical values. It is necessary to characterize the nature of variable in question before carrying out statistical procedure irrespective of the fact that it is descriptive or inferential. Because this will have direct bearing on the choice of appropriate statistical technique. Usually variables are of the following types:

- measurement variables;
- ranked variables;
- attributes.

(i) Measurement variables

Measurement variables are of two types:

- continuous;
- discrete.

Continuous variable can assume an infinite number of values between the highest & lowest values on a scale of variation. Usually a pharmaceutical formulation is required to have active therapeutic agent between 90% 110% of the nominal amount. The drug content in the pharmaceutical formulation can assume an infinite number of possible masses within these limits. However, depending on the sensitivity of the method of the chemical assay, values will be restricted to the observed values.

Discrete variables are those variables which have a fixed number of values. These variable always have integer values.

(ii) Ranked variables

When ranking scales are used, these represent numerically ordered system. For example, rejected tablets out of batch are analyzed, the results are:

- 50% tablets have broken edges;
- 30% tablets have spots;
- 20% tablets have black particles.

(iii) Attributes (Nominal variables)

Nominal variables are qualitative in nature. These can not be measured. For example, side effects associated with therapeutic agents. When attributes are combined with frequencies, they are referred to as enumeration data.

1.2 Statistical Population and Samples

The total number of observations that constitute a particular group may be defined as the population. Generally samples are relatively smaller group of observations that have been taken from the population. Any particular property associated with a population is known as parameter. Suppose there is a batch of 3,00,000 tablets. All the tablets in the batch constitute a population out of this batch, 100 tablets are removed for weighing. These 100 tablets constitute sample. Weight of tablet is a parameter. On the basis of results of weighing of these 100 tablets (taken out randomly) assumption about the nature of population can be drawn.

1.3 Measurement of Central Tendency and Variation of Data

After data have been collected in an experiment or study or examination, the next thing is to examine the nature of data. Most frequently used two properties are:

- central nature (tendency);
- variability.

Central nature of data can be described by a number of methods and terms like mean, mode, median. The term average is used by general public. Statistician call it mean. Mean is the most commonly employed method to describe tendency. It refers to the centre of distribution of data.

(i) Mean

The mean is obtained by dividing the sum of observations by the number of observations. Algebraically, if $x_1, x_2, x_3, \ldots, x_n$ denote the n individual observations, the mean (denoted by $\bar{X}$) is given by the formula:

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

$$= \frac{1}{n} \sum_{j=1}^{n} x_j$$

Weighted mean of the values x_1, x_2, x_3, . . . , x_n having weights w_1, w_2, w_3, . . . , w_n respectively can be found out with the help of formula:

$$X = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n}{w_1 + w_2 + w_3 + \ldots + w_n}$$

$$= \frac{1}{n} \sum_{j=1}^{n} w_jx_j$$

It is appropriate to use mean for the description of tendency of the data when the data is distributed in Gaussian fashion i.e. distributed equally on either side of the mean.

(ii) Median

Median can be defined as that value which have been arranged in the order of the magnitude so that the number of observations less than median is equal to the number of observations greater than median. Thus it may be noted that if number of observations is odd, the median is the middle most value when observations are arranged either in ascending or descending order. If the number of observations is even, the median is the mean of the two middle most values. The values of weights of eleven tablets from a batch of tablets are given in table 1.1.

Table 1.1: Values of weights of individual tablet

S.No.	Weight of tablets in mg
1	122
2	120
3	120
4	119
5	123
6	121
7	122
8	120
9	120
10	120
11	118

Arrange the data in the order of magnitude:

118, 119, 120, 120, 120, 120, 120, 121, 122, 122, 123

The median is defined as central value i.e. value at position 6. The median in this example is 120 mg.

(iii) Mode

Mode is that value of variable which has the highest frequency. Although measure of tendency is the easier to find out from a frequency distribution, but has fewer applications in quality control. The assay values of active medicament in 10 ampoules are given in table 1.2.

Table 1.2: Assay values of active medicament in individual ampoule

S.No.	Contents in mg/ml
1	200
2	203
3	205
4	198
5	203
6	195
7	205
8	203
9	201
10	206

The most common value in the above mentioned set of data is 203 mg/ml. Thus mode in the example is 203 mg/ml.

1.4 Measurement of the Variation of Data

Mean, median, mode provide no information about the variability of the data from which central measures were obtained. In addition to the tendency it is also important to know measure of variability or dispersion of data. This information provides a measure of relative proximity of data set. Methods by which variation of data may be calculated and presented include range, mean deviation, variance, standard deviation.

1.4.1 Range

The range may be defined as the difference between the smallest and largest values in a given set of measurements. Since the calculation of range involves only two measurements (i.e. lowest and highest points), it truly does not describe the variation of the entire data. The main use of range is to define the variability associated with non-normally distributed data.

1.4.2 Mean Deviation

The mean deviation is a measure of data variation that is calculated as the average deviation from mean. Mathematically it is represented by the formula given below:

$$MD = \frac{\Sigma (X_j - \bar{X})}{N}$$

where,

$X_j - \bar{X}$ = absolute value of the deviation (difference) of the values in data set from the mean

N = number of observations in the data set

The following illustration will make it more clear. Ten tablets taken from a batch weigh as given in table 1.3.

Table 1.3: Values of weights of individual tablet

S.No.	Weight of tablets in mg
1	118
2	121
3	122
4	122
5	122
6	119
7	118
8	118
9	121
10	119

First calculate mean

$$\text{Mean} = \frac{(118+121+122+122+122+119+118+118+121+119)}{10}$$

$$= 120$$

Now calculate mean difference

$$MD = \frac{\Sigma (X_j - \bar{X})}{N}$$

$$= \frac{[(118-120)+(121-120)+(122-120)+(122-120)+(122-120)+(119-120)+(118-120)+(118-120)+(121-120)+(119-120)]}{10}$$

$$= \frac{\Sigma(2)+(1)+(2)+(2)+(2)+(1)+(2)+(2)+(1)+(1)}{10}$$

$$= 1.6$$

Mean deviation is calculated using absolute values of difference between measurement and the mean. Algebraic sign is not taken into consideration

1.4.3 Variance

Before variance is defined, it will be useful to understand sum of squares. In the calculation of the mean deviation, the algebraic sign was ignored to provide a numerical outcome. The addition of the successive squared differences generate a statistical term, known as the sum of squares. Mathematically, it can be represented as under:

$$SS = \Sigma (X_j - \bar{X})^2$$

Variance is the mean sum of squares. Mathematically, it can be written as:

$$\sigma^2 = \frac{\Sigma (X_j - \mu)^2}{N}$$

where,

σ^2 = variance of population
X_j = numerical value of each measurement
μ = population mean value
N = Number of observations

Variance of the sample (denoted by s^2) is given by the formula:

$$s^2 = \frac{\Sigma (X_j - \bar{X})^2}{N - 1}$$

where,

$\bar{X}$ = mean of the sample data
N = number of observations in the sample

It may be understood that σ^2 is population variance and s^2 is sample variance. Since sample variance is considered to be a biased estimate of the population variance, in the denominator, the term (N – 1) is used to remove this bias.

Thus sample variance of groups of sample selected randomly may not be exactly equal to one another & variance of a single sample does not provide a good estimation of the variance of the population. However, a good estimation of the population variance can be made from sample data provided denominator of the equation to calculate variance is modified to N–1. Additionally an average of several sample variances may be calculated.

Now let us consider whether mean deviation or variance provides more information on variability of data. Both of these terms provide similar information regarding the spread of data around the mean. However, the variance can be related to probability. Therefore, variance is considered to be more relevant parameter.

1.4.4 Standard deviation

This measure of dispersion of data is commonly used. Standard deviation is defined as the square root of the variance.

Mathematically, it can be written as follows:

$$\sigma = \sqrt{\frac{\Sigma (X_j - \mu)^2}{N}} \quad \text{(Standard deviation of population)}$$

$$s = \sqrt{\frac{\Sigma (X_j - \bar{X})^2}{N - 1}} \quad \text{(Standard deviation of sample)}$$

As a thumb rule, standard deviation is approximately one fifth to one sixth of numerical value of the range. If such a relationship is not achieved after calculation of these two parameters, calculations should be rechecked.

1.4.4 Standard deviation of the mean

Standard deviation of the mean also known as standard error of the mean (SEM) is commonly used in the statistics. However, the difference between standard deviation and standard deviation of the mean should be clearly understood. Standard deviation describes the variability (dispersion) of a set of data around a central value.

From this an estimate of the variability of the data in a population can be derived. Standard deviation of the mean is the measure of the variability of a set of mean values which are calculated from individual groups of measurement (i.e. samples) drawn from a population.

The standard error of the mean is calculated using the individual mean values of different samples

$$s = \sqrt{\frac{\Sigma X^2 - [(\Sigma X)^2 / N]}{N - 1}}$$

Standard deviation of the mean can be calculated by repeatedly sampling a population. But in practice it becomes difficult to sample a population repeatedly. According to statistical theory, SEM can be found out by dividing the standard deviation of a set of data by the square root of the number of observations in the data set. Thus,

$$SEM = (s/\sqrt{N})$$

where,

s = standard deviation of a set of data
N = Number of observations in the data set

This equation enables to calculate SEM with fewer samples. However, it should be remembered that SEM calculated from standard deviation will not exactly match the true value of standard deviation of the mean values of the different sets of samples from a population.

1.4.5 Coefficient of variation

The term, coefficient of variation describes the variability of a set of data. It is defined as the ratio of the standard deviation to the mean of data. Thus,

$$CV\ (\%) = \frac{s}{\bar{X}} \times 100$$

Coefficient of variation is used to evaluate the variability of the data sets. Magnitude of the coefficient of variation will depend on the nature of data. Generally in pharmaceutical analytical experiments, coefficient of variation is low because variability associated with experiments is usually low. In contrast with this, coefficient of variation of biological experiments may be quite large.

Sometimes it may be as high as 100% because variability of such measurements is usually high.

1.4.6 Accuracy

Accuracy can be defined as the closeness of a measured value to the true value. True value means the value which would be expected in the absence of error. In the pharmaceutical analysis, it is usual to describe accuracy of an analytical method. Some of the methods by which difference between observed and expected values may be described are given below.

(i) *Absolute error:* Mathematically, absolute error can be found out by employing the following formula:

$$error_{abs} = O - E$$

where,

O = observed value or alternatively the observed mean
E = expected value or true value

Example given below will make it more clear. A pharmaceutical preparation containing 100 mg of active medicament has been analyzed by three different methods and the values are 92, 97, 103 mg respectively.

Table 1.4: Assay values of active medicament by three different methods

Method	Observed value	Expected value	Absolute error
1	92	100	–8
2	97	100	–3
3	103	100	3

From these figures it may be seen that absolute error of method two & method three are same but the observed values 97 and 103 are not identical.

(ii) *Relative error:* This term was evolved to overcome the problem encountered in absolute error method. This term describes the error as a proportion of the true or expected value while calculating relative error sign of difference whether positive or negative is ignored. It is represented by the formula.

$$\text{error}_{rel} = \frac{\text{error}_{abs}}{E} = \frac{O - E}{E}$$

When represented as percentage value, formula is as follows:

$$\% \text{ error}_{rel} = \frac{\text{error}_{abs} \times 100}{E}$$

$$= \frac{(O - E) \times 100}{E}$$

Greater numerical values of relative error are indicative of decrease in accuracy. Relative error can be used to compare the accuracies of different measurements.

1.4.7 Precision

Precision describes variability (dispersion) of a set of measurements. It does not provide any indication of the closeness of an observation to the expected value. High precision is associated with low dispersion of values around a central value.

Precision is generally expressed as the standard deviation of a series of measurements obtained from one sample.

2. PROBABILITY AND PROBABILITY DISTRIBUTION

We use probability in our daily lives. When we toss up a coin and when it falls flat on earth it will have two options either "tail" or "head". That means out of two options, it can have one option at a time i.e. 50% chances of occurrence. Mathematically, probability can be defined as that aspect of mathematics which is concerned with calculating the likelihood of the occurrence of an event. In section 1 of this chapter, an introduction to descriptive statistics has been provided. The other sub-discipline of statistics is inferential statistics which provides information about a large population to be estimated on the basis of the statistical analysis of a smaller sample from that population. For example, if there is a batch of 1,00,000 tablets of a formulation and out of it, a sample of 200 tablets is collected and analyzed, such a situation will fall under the scope of inferential statistics. The Indian Pharmacopoeia (IP) prescribes under general notices that fiducial limits of error are stated in biological assays and in all cases, fiducial limits of error are based on a probability of 95% (p = 0.95). There are various other situations where the phenomenon of probability is used in pharmaceutical

industry. Therefore, an overview on probability will be relevant in this chapter.

Now, basic rules of probability will be discussed which include range of values and probability distribution

(i) Range of values

The probability an event occurring must fall between 0 and 1. A probability of 0 Means that an event-will never occur. A probability of 1 means that an event-will always occur. The probability of an event may be calculated the number of times an event occurs by the number of all possible outcomes. The events may be mutually exclusive events or independent events.

(ii) Mutually exclusive events

Probability of the occurrence of two or more mutually exclusive events may be calculated by the addition of the individual probabilities for each event. Suppose there are two events A and B. The probability (P) of any of the events occurring can be described by the equation:

$$P (A \text{ or } B) = P (A) + P (B)$$

The term mutually exclusive means that if one event occurs, then the other event(s) does/do not occur.

(iii) Independent events

Independent events means, unlike mutually exclusive events, events can occur independently. In the above example, if the events A and B can occur independently, the probability of such a situation can be described mathematically by the following equation:

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

Addition law of probability can also be applied to calculate probability of events that are not mutually exclusive, but in the modified form. Mathematically, it can be described as:

$$P (A \text{ and } B) = P (A) + P (B) - P (AB)$$

2.1 Probability Distribution

In inferential statistics, probability theory is used to make assumptions about the properties of populations on the basis of

data recorded from smaller samples taken from a population. A key component of such estimation is the use of probability distribution i.e. relationships between particular variables and their probability of occurrence.

Observations in the samples can be categorized as discrete or continuous. A discrete observation is one of countable finite number e.g. a tablet categorized as within limits or out of limits. A continuous observation is that which can be measured more and more. For example, removal of 20 tablets periodically during compression of a batch of tablets and weighing them.

2.1.1 Binomial Distribution

One of the distributions that is commonly employed in pharmaceutical industry is the binomial distribution. This distribution is used when the outcome of an event is "go" or "no go". Another requirement of binomial trial is that each trial must be independent i.e. occurrence of one event should not influence subsequent events. Sum of the probabilities of all events must be equal to one.

As stated above, binomial is a two parameter distribution, i.e. p the probability of one of the two outcomes, N the number of trials or observations. If out of 100 tablets taken out of a batch, 5 are rejects, the probability (p) of rejection is estimated as 0.05 and N = 100, the probability of passing tablets will be

$$1 - p = q \quad \text{i.e.} \quad 1 - 0.05 = 0.95.$$

If we were to calculate the probability of selecting (i) two defective tablets, (ii) two non defective tablets, and (iii) one defective and one non-defective tablet each in a sample of two tablets, the overall probability will be

$$p^2 + pq + qp + q^2 = 1$$

If this example is extended to consider three samples, the possible outcomes will be

- three defective tablets (p^3)
- two defective tablets and one non-defective tablet (ppq, pqp, qpp) or 3 p^2q
- one defective tablets and two non-defective tablet (pqq, qpq, qqp) or 3 pq^2
- three non-defective tablets (q^3)

So the equation will be

$$p^3 + 3\ p^2q + 3pq^2 + q^2 = 1$$

Thus, expansion of the binomial term, $(p + q)^n$ for defined values of the exponent n will be as given in the table 1.5.

Table 1.5: Expansion of binomial term, $(p+q)^n$

n	Expansion of $(p + q)^n$
1	$p + q$
2	$p^2 + 2\ pq + q^2$
3	$p^3 + 3\ p^2q + 3\ pq^2 + q^3$
4	$p^4 + 4\ p^3q + 6\ p^2q^2 + 4\ pq^3 + q^4$
5	$p^5 + 5\ p^4q + 10\ p^3q^2 + 10\ p^2q^3 + 5\ pq^4 + q^5$
.	.....

If the N_p and N_q are both equal to or greater than 5, cumulative binomial probabilities can be closely approximated by areas under the standard normal curve.

2.1.2 Continuous Probability Distribution

When the variable may adopt an infinite number of outcomes, such a distribution is known as continuous probability distribution or normal distribution. Determining weight variation periodically of a batch of tablets can be considered as a continuous probability distribution. Because of continuous nature of distribution, it is impossible to assign a probability to an exact value of the variable. However, it is possible to calculate the probability of an event occurring within a range. Data of 10 samples each of 20 tablets are tabulated in table 1.6.

Table 1.6 shows the weights of tablets taken at intervals during the compression of tablets. One way to organize such figures is to show their pattern of variation, i.e. to count the number of times each value occurs. The results of count are called a frequency distribution. Sets of observations could be formed in an arrangement which shows the frequency of occurrence of the values of the variable in ordered class. Such an arrangement is called grouped frequency distribution. The interval along the scale of measurement of each ordered class is termed as a cell. Frequency for any cell is the number

Table 1.6: Weights of individual tablet in 10 samples

Sample No.	Weight of Individual Tablet in mg																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	122	120	120	119	123	121	122	120	120	120	118	121	122	122	122	119	118	118	121	119
2	120	122	121	123	121	119	118	117	118	119	122	121	121	119	120	120	123	121	120	121
3	122	117	119	120	120	121	121	121	120	120	121	120	122	118	120	118	118	120	121	120
4	120	120	122	120	120	122	120	120	120	119	121	121	120	118	120	120	121	120	122	120
5	119	123	124	118	119	120	120	122	121	122	122	124	123	119	121	120	120	119	117	118
6	120	120	120	122	121	116	119	120	118	121	121	121	118	122	120	120	121	120	120	122
7	120	122	120	121	120	123	121	124	124	124	120	121	121	120	120	125	122	121	122	120
8	119	123	120	120	124	121	122	120	120	118	119	123	124	121	119	119	122	121	120	119
9	117	118	118	118	120	120	119	122	117	120	121	119	121	121	123	121	120	121	122	120
10	120	121	121	120	120	122	122	120	120	121	120	118	121	120	120	122	120	122	123	120

of observations in that cell. If the frequency of that cell is divided by the total number of observations, it is called relative frequency. Frequency distribution of data in table 1.6 is given in table 1.7.

Table 1.7: Frequency distribution of data of table 1.6

Class intervals	Frequency
115.5-116.5	1
116.5-117.5	5
117.5-118.5	18
118.5-119.5	20
119.5-120.5	68
120.5-121.5	41
121.5-122.5	29
122.5-123.5	10
123.5-124.5	7
124.5-125.5	1

Frequency distribution curve of this data will be as shown Fig 1.1.

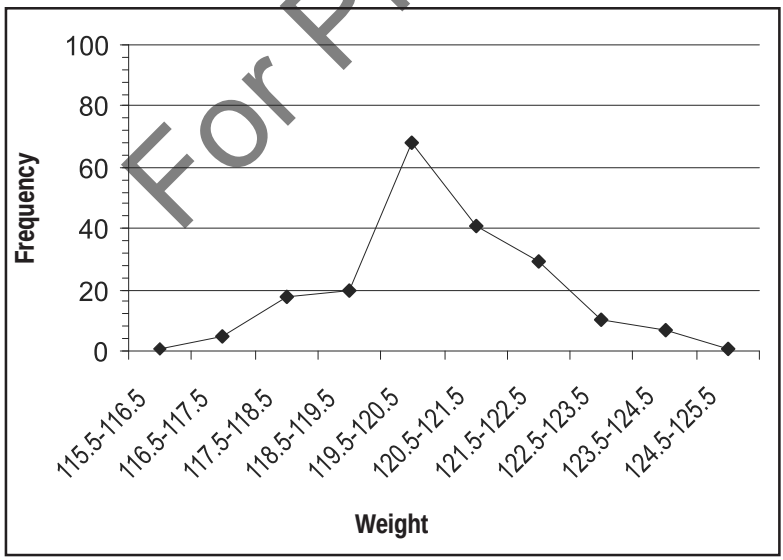

Fig. 1.1 : Frequency curve

From this figure, it may be seen that curve is symmetric about a central value. If the spread of frequency distribution is more, the curve will be like shown in Fig 1.2.

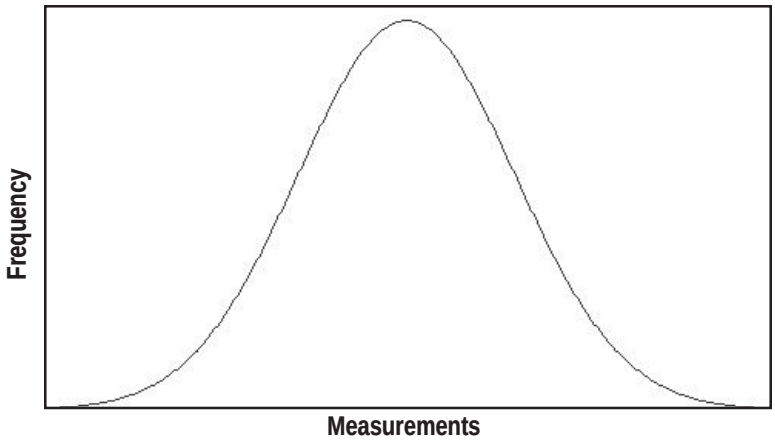

Fig. 1.2: Frequency curve with more spread

2.1.2.1 Normal Distribution

Normal Distribution, also referred to as the Gaussian distribution, is the most important theoretical distribution in statistics and is used in several inferential statistical tests. If a large population is examined with reference to a certain attribute (variable), the resultant distribution would be normal. According to central limit theorem, if a large number of samples are removed from any distribution with a finite variance and mean, the distribution of variable tend to be normal. In other words, if the sample size is large enough, the data will be distributed in normal fashion, independent of nature of distribution from which samples are removed.

The shape of the normal distribution is shown in Fig 1.3.

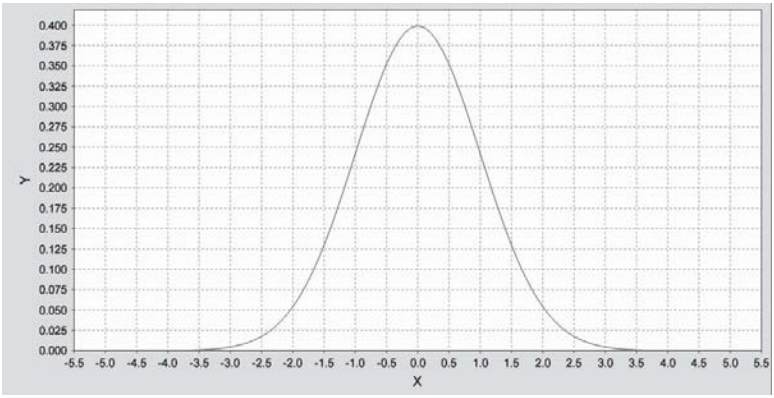

Fig. 1.3: Normal distribution curve

Normal distribution has the following properties:

- it is symmetrical;
- it is bell shaped;
- it extends from – infinity to infinity;
- it has an infinite number of observations;
- the shape of the curve is defined by mean & standard deviation;
- the mean, median and mode are numerically equal.

The central value is designated as μ, the mean. These curves indicate that the most of the values in the distribution are near the mean and as the values are farther from the mean, they are less prevalent. Although theoretically, the data comprising a normal distribution may have values between – infinity and + infinity, but the values sufficiently farther from the mean have little chances of being observed. Normal curves are defined by two parameters, namely, the mean (μ), a measure of location, and the standard deviation (σ), a measure of spread. The population is the totality of data from which sample data is derived. If the batch size of a batch is 1,00,000 tablets, out of which 200 tablets are collected as samples (10 different samples of 20 tablets each), population is 1,00,000 tablets. The sample mean ($\bar{X}$) is an unbiased estimate of the true population mean (μ), although $\bar{X}$ cannot be expected to be equal to μ. If an experiment is repeated several times and if all the $\bar{X}$s are averaged, this grand average would be equal to μ.

Another property of the normal distribution is that area under the normal curve is exactly 1 irrespective of values of μ and σ.

2.1.2.2 Standard normal distribution

As stated in the previous section that each normal distribution is unique by its mean and standard deviation, therefore, calculation of the probability of an event occurring using each unique distribution will require calculation of the probability density function for that variable. It will be a very difficult task. To overcome this difficulty, the standard normal distribution, a generic distribution which possesses a mean value of 0 and a standard deviation of 1 is used. Areas under the curve associated with this distribution have been calculated and the same may be used to estimate the probability of occurrence of an event of which full distribution has not been calculated. The method by which it is done is commonly referred to as "z transformation".

In performing z transformation, two mathematical steps must be performed:

- the mean must be transformed from actual value to 0;
- the standard deviation must be transformed from actual value to 1.

z transformation can be described mathematically as:

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

where,

z = transformed value of the x axis
X = defined value from the original data set
σ = standard deviation of the original data set

To interpret z values, table describing the areas under the standard normal distribution should be consulted. Readers may refer to standard books on statistics for tables.

2.1.2.3 t distribution (Student's t distribution)

t distribution was first described by William Sealy Gossett in 1908. W.S. Gossett used pseudonym of student and therefore, it is also known as Student's t distribution. This distribution is used in statistical analysis when the sample size is small because the distribution of mean after sampling do not correctly conform to the normal distribution. The t distribution of sample size of N may be calculated employing the following equation:

$$t = \frac{\bar{X} - \mu}{s / \sqrt{N}}$$

where,

$\bar{X}$ = mean of the sample
μ = mean of the population
s = sample standard deviation
N = number of observations per sample

If a sample is drawn from the population having normal distribution, we can calculate the mean ($\bar{X}$) and sample standard deviation (s) and use these values as good estimates of the corresponding population parameters. If several samples are drawn

from the population having normal distribution, then a series of mean will be generated whose standard deviation can be calculated. The standard deviation of the mean (standard error) can be calculated with the help of the following equation:

$$SEM = \frac{s}{\sqrt{N}}$$

where

s = sample standard deviation
N = number of observations

When the sample size is large, the points that can be observed are:

- the sample mean ($\bar{X}$) is derived from normal population;
- SEM is reliable estimate of the population standard deviation;
- application of the equation describing t statistic will result in a normal distribution with a mean and standard deviation of 0 and 1 respectively.

In case of small sample sizes, the mean ($\bar{X}$) is also derived from normal distribution but the sample standard deviation will vary from sample to sample. Therefore, SEM is not good estimate of the standard deviation of the distribution.

In view of this, it can be stated that whenever the sample standard deviation is large, the t statistic is small and whenever the sample standard deviation is small, the t statistic is large. The tails of t distribution may be longer. Because of variation from one sample to another, the t distribution is usually used to calculate confidence intervals and to compare mean values for small samples.

Main characteristics of the t distribution include:

- It is symmetrical (as is the case in normal distribution).
- The tails are longer than the standardized normal distribution (z distribution).
- The shape of distribution is dependent on sample size.
- As the sample size increases, the shape of t distribution tends to become similar to z distribution.
- A parameter related to sample size commonly used in statistics about the t distribution is the degree of freedom. In the case of one sample test, the number of degrees of freedom is defined as:

$df = N - 1$,

where N is the sample size.

- Since t distribution is affected by sample size, it is time consuming to report the areas under each t distribution corresponding to different probabilities for each degree of freedom. Therefore, t is normally reported as the t static corresponding to defined probabilities and different degrees of freedom (e.g. one tail test, two tail test & 1,2,3,...). Readers may refer to standard books of statistics or IS:6200 (Part I)[3] for t distribution tables.

Chi (χ^2) distribution

The χ^2 (pronounced as ky-squared) distribution is another important distribution. Mathematically, the χ^2 distribution may be represented as:

$$Y = Y_0 (\chi^2)^{0.5\,(v-2)} e^{-0.5\chi 2}$$
$$= Y_0\, \chi^{v-2} e^{-0.5\chi 2}$$

where,

v = number of degrees of freedom (N – 1)

Y_0 = constant dependent on V

χ^2 = chi-squared static

The total area under the curve is equal to 1. The equation is too complex to explain. But one point that is important to note is that the equation contains one variable parameter (v), the rest is a constant or the value of χ^2 distribution from which corresponding ordinate is being calculated. Therefore, the χ^2 distribution is described by this one parameter (degrees of freedom). The number of degrees of freedom directly influences the shape of χ^2 distribution. Readers may refer to standard books of statistics or IS:6200 (Part II)[4] for critical values of χ^2 distribution.

F distribution

Another important distribution is F distribution. It is derived from the sampling distribution of the ratio of two independent estimation of the variance from normal distribution. The F distribution is used to test the equality of two variances and also for multiple hypothesis testing (ANOVA). Considering two samples of

known sizes which have been drawn from two populations having normal distributions of defined variances may be defined as:

$$F = \frac{N_1 s_1^2 / (N_1 - 1)\ \sigma_1^2}{N_2 s_2^2 / (N_2 - 1)\ \sigma_2^2}$$

where

N_1 & N_2 = sample sizes
s_1^2 & s_2^2 = variances of two samples
σ_1^2 & σ_2^2 = variances of normal distribution from which samples are drawn

Like t distribution, F distribution is also based on the assumption of the normality and independence of the observations. Readers may refer to standard books of statistics or IS: 6200 (Part I) for critical values of F distribution.

3. STATISTICAL HYPOTHESIS TESTING

In statistical hypothesis testing, assumptions are made regarding the likelihood of an event and then, using appropriate methods, the validity of these assumptions is examined. Let us consider an example of manufacture of tablets through a validated process. In such a scenario, it will be assumed that batch will pass all quality control parameters. Now, if for any reason, batch does not pass, then the assumption regarding quality of the batch was incorrect and therefore, an alternative proposal was valid, i.e. batch will not pass all quality control parameters. This illustrates the basis for statistical hypothesis testing i.e. first an assumption is made and then data are collected from which conclusions concerning the validity of initial assumption may be formulated.

As stated in the previous sections, sample mean is representative of population mean. But there may be situations where it is not true. Similarities and differences between sample and population statistics take in into mechanics of statistical hypothesis testing. Consider an experimental batch of tablets. Out of this batch, three samples one in beginning, one in middle and the third at the end of compression of tablets have been collected. If we assume that means of three samples should be representative of population mean, this assumption is commonly known as "null hypothesis". In fact, this is starting position in statistical hypothesis testing.

In the above example, (i) either mean values of all the three samples would be representative of population mean, or (ii) mean values of the three samples would not be representative of population mean. In the former case, null hypothesis will be accepted and in the later case, null hypothesis will not be accepted. The later case raises a problem, how do we interpret a non-acceptance of the null-hypothesis? In statistics, it is done by accepting the alternative hypothesis, i.e. one sample mean is not representative of population mean.

Statistical hypothesis testing is a measure of whether the null hypothesis is accepted or rejected. If the null hypothesis is rejected, the alternative hypothesis is accepted. Thus null hypothesis and alternative hypothesis may be expressed as:

H_0 *(null hypothesis):* there is no difference between sample means and population mean

H_a *(alternative hypothesis):* there is a difference between the sample means and population mean

3.1 Level of significance and critical regions of acceptance and rejection of the null hypothesis

As the collection of data is advanced, it is necessary to state the level of significance because this defines the terms of acceptance and rejection of null hypothesis. It is a convention to use a value of 0.05 to define probability or improbability of an event. Thereby meaning that if a probability value of 0.05 or less is associated with an event, there is sufficient evidence to conclude that the null hypothesis is not acceptable and that the alternative hypothesis is valid.

In statistics, level of significance is written as proportion and is denoted by Greek Letter α (alpha). The choice of α is arbitrary. Usually, a value 0.05 is used in statistical hypothesis testing. As the level of significance is increased, it becomes more difficult to reject null hypothesis. This is important as scientific experiments are generally designed to reject the null hypothesis. After establishing the level of probability, the regions of acceptance and rejections of the null hypothesis associated with defined probability distributions are used.

It is usual in statistical hypothesis testing to define rejection of the null hypothesis before calculating the test statistics. For this,

the possible outcomes of the statistical test must be considered before the data is collected.

3.2 One-Tailed and Two-Tailed Tests (outcomes)

In the process of statistical hypothesis testing, null and alternative hypothesis, level of significance and whether the experimental design (test) is one or two tailed are stated. This concept of one-tailed or two-tailed tests refers to possible outcomes of the study. If there is only one outcome of interest for investigation, the test statistic must be interpolated using one-tailed outcome and if there are two possible statistical outcomes, a two-tailed test must be used.

When statistical analysis is carried out, the sampling distribution and the conventional probability distributions are divided into two regions which facilitate the interpretation of the importance of the calculated test statistic to be performed. The values of test statistic associated with acceptance of the null hypothesis fall within the first region which is called region of non-significance. The second region defined the values of test statistic which are associated with the rejection of null hypothesis and acceptance of alternative hypothesis. This region is called region of significance. The magnitudes (and therefore numerical boundaries) of these regions are dependent on two factors, one, the level of significance (α) and two, whether the test is one-tailed or two-tailed for all probability distributions associated with test statistic. For some distributions, e.g. the F distribution, t distribution χ^2 distribution, the number of degrees of freedom associated with the experimental design also defines the boundaries of the regions of acceptance and rejection.

Assuming α=0.05, the critical region of the z distribution can be represented in three ways as shown in Fig 1.4, 1.5 and 1.6.

In all the three cases, the chosen level of significance is 0.05, as represented by shaded regions. Calculated values of z statistic that fall within this region allow the analyst to reject the null hypothesis. It may be observed that one-tailed and two-tailed tests differ in their distribution of the region of rejection. In the two-tailed test, the rejection region is equally divided into two sections each relating to a single tail at either extreme of the probability distribution. In contrast with this, in one-tailed test, the rejection region is distributed at either tail of the probability distribution.

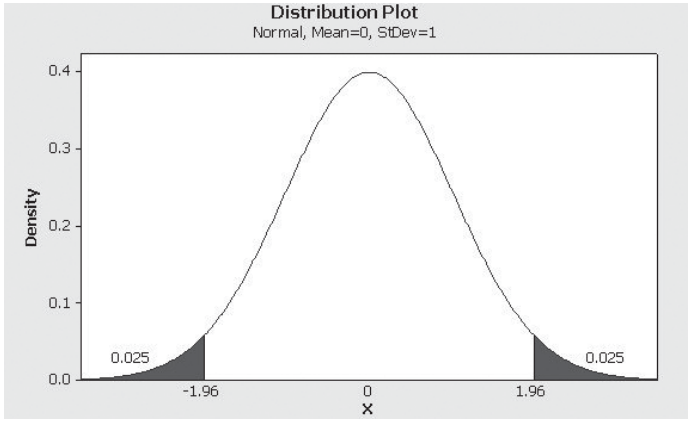

Fig. 1.4: Standardized normal distribution showing critical regions of z statistic for a two tailed-test

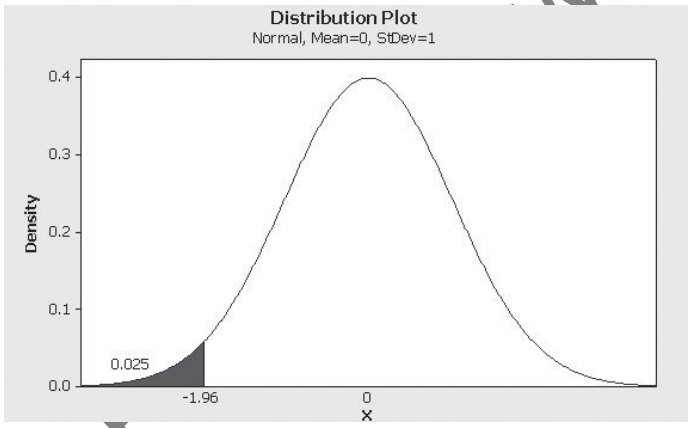

Fig. 1.5: Standardized normal distribution showing critical regions of z statistic for a one-tailed test

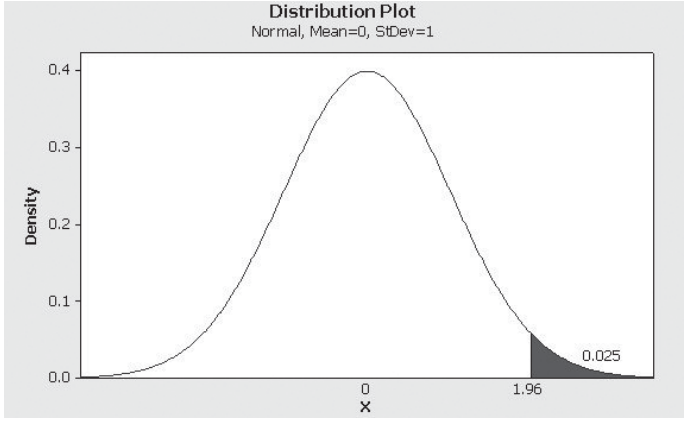

Fig. 1.6: Standardized normal distribution showing critical regions of z statistic for a one-tailed test

3.3 Errors in Decision Making

There are two types of errors which may be made in determining the outcome of the experiment, Type I and Type II. When null hypothesis is rejected when it is true, it is called Type I error. When the null hypothesis is accepted when it is false, it is called Type II error. Therefore, it is common to calculate the probability of making Type II error whenever a null hypothesis has been accepted. It is done using z statistic.

A reciprocal relationship exists between Type I and Type II errors. Thus by attempting to reduce one type of error, the chances of increasing the other type of error increase. Therefore, it is advisable to decide what levels of each type of errors are acceptable. Conventionally the type I (α) error is selected to be 0.05 because this is deemed to be sufficiently small probability of committing this type of error.

3.4 Power of a Statistical Test

For the relationship between statistical decision making and errors, the term power of study is used. The power of a statistical test may be defined as the probability that the null hypothesis is rejected, when in fact, it is false or it can be said that the power is a measure of the ability of the statistical test to validate a research hypothesis when the research hypothesis is, in fact, true. As stated earlier, Type II error may be defined as the probability of not rejecting the null hypothesis when it should be rejected, the power is quoted as $1-\beta$. The power of statistical test increases as the difference between null hypothesis and the alternative hypothesis increases.

In any experimental design, the power of a statistical test is an important consideration. If the power of a statistical test is maximized, it would result in greater confidence about the fidelity of a statistical outcome. Its magnitude is influenced by several factors:

- the selected probability relating to Type I (α) error;
- the magnitude of the difference between the true mean (alternative hypothesis) and the mean associated with null hypothesis;
- the sample size;
- the nature of statistical test

3.5 Choice of Statistical Test

The key stages that have been identified in the process of statistical hypothesis testing include:

- statement of null and alternative hypotheses;
- selection of the level of significance (α) and consideration of the probability of committing a Type II error;
- identification of the nature of the experimental outcome, i.e. whether the result is one or two-tailed;
- identification of the critical statistic which determines the area of rejection of null hypothesis.

One of the salient steps in the process of statistical hypothesis testing involves the choice of statistical test. It is a crucial stage because the outcome of the analysis will determine the fate of research hypothesis (study) and therefore, it is important that statistical test is selected according to the features of experimental design. The selection of the most appropriate statistical test is dependent on:

- desired power of the statistical test;
- nature of population from which the observations are taken;
- nature of measurement of variable.

These factors combinedly are called as the statistical model. Considering all these factors, the correct choice of statistical test may be made. The choice is assumed to be appropriate for the statistical model. It can be stated in other models that the assumptions of the statistical test are satisfied by the conditions of the experimental design.

3.5.1 Parametric and Non-Parametric Analyses

Generally statistical tests may be categorized as:

- parametric analyses;
- non-parametric analyses.

The choice of analysis is made according to the statistical model. Since this information is not readily available to the statistician, assumptions about the statistical model are made in many cases. Therefore, whenever a particular statistical method is recommended to compare two sets of data, the ability of the test to reject the null hypothesis when, in fact, it is false is as function of

the nature of the assumptions of the statistical model. When the assumptions are few, i.e. the features of experimental design are fully known, then the conclusions drawn by statistical test are valid and the output of the analysis is conclusive.

In contrast with this, when several assumptions have been made about the nature of the statistical model, then it is likely the output of the statistical analysis will be general in nature and may not be conclusive. What is important is to ensure that the conditions of the experiment (i.e. statistical model) and subsequent statistical test are matched. This will enhance performance of the analysis.

Essentially parametric and non-parametric analyses differ in the nature of assumptions associated with their use. Parametric tests, for example, t test, F test, z test can only be used when a number of assumptions have been confirmed. If the assumptions are valid, the use of parametric tests is appropriate because this will ensure that the quality of output from the statistical analysis is optimized. The following assumptions should hold good for the selection of parametric statistical method:

- The samples should be drawn from a population having normal distribution.
- The samples should be independent.
- The variance of the population under study should be similar.
- The variable under examination must be measured on an interval or ratio scale in which the values obtained may be conveniently manipulated employing conventional arithmetic.

Usually a small number of replicate samples of a variable are collected for analysis. In these situations, it will be difficult to examine whether the observations were derived from a normal distribution. Therefore, in parametric analysis, where sample size is small, an assumptions is made about the nature of the population from which each data set is derived. With the marked departure from the assumptions, the result obtained from analysis may not be reliable.

One assumption which is well defined is the nature of data. In the parametric analysis, the data (variable) is continuous in nature and can be mathematically manipulated to generate descriptive statistics (e.g. mean, variance, standard deviation). But when the data fall into other categories like nominal and ordinal scales, such data cannot be analyzed using parametric tests. For such data, non-

parametric analyses must be employed. It will be useful to discuss briefly nominal, ordinal and interval and ratio data.

(i) *Nominal Data:* Nominal data is classified into groups which are given name or title. For example, a group of male patients, a group of patients aged more than 50 years. Nominal data is usually expressed in terms of frequencies of observations associated with each category. For the analysis of such data, techniques like χ^2 analysis, binomial based analysis may be used.

(ii) *Ordinal Data:* Nominal and ordinal data are similar but the categories in ordinal data are not independent and they differ in magnitude. For example, description of bitterness – not bitter, slightly bitter, moderately bitter, extremely bitter. Thus, it may be seen that data is categorized but there is relationship between individual categories. But such relationship is not present in nominal data.

Many non-parametric tests are referred to as "ranking tests". These tests can be employed for the analysis of ordinal data. If the non-parametric analysis is to be carried out of an ordinal data, the pre-requisite is that data should possess an underlying continuum, i.e. spread of typical responses within any category. Ordinal data can be described successfully using frequency, relative frequency or percentage. Median and range are used to depict the central tendency and variability respectively.

(iii) *Interval and Ratio Data:* This type of data represents a higher level of organization than nominal and ordinal data. Such data may be characterized by knowledge of the distances between two values on any particular scale, i.e. numerical distance between two values. In an interval scale, there is no true zero. However, in ratio data, there is defined zero point.

A classical example of interval scale is the measurement of temperature in either Centigrade (ºC) or Fahrenheit (ºF). The value, 0º is arbitrary and does not represent an absence of temperature as there as sub-zero temperatures. Because of lack of true 0 value, two or more values cannot be compared directly within a temperature scale.

The interval scale is referred to as quantitative scale and the information held can be manipulated using arithmetic procedures.

As such, interval data can be analyzed using parametric statistical tests. However, it should be ensured that all the assumptions of parametric statistics are valid. If there are doubts about the validity of assumptions, non-parametric tests should be applied.

The ratio scale is also a quantitative scale. But differs from the interval scale in that it has true zero point. Zero refers to an absence of measurable value. Since the data from ratio scale may be manipulated arithmetically, it can be analyzed using parametric methods. But again it should be ensured that all the assumptions for parametric tests are valid. In case of doubt, non-parametric method should be used.

Interval or ratio data, the distribution of which is skewed (non-normal), can be successfully described employing frequency, relative frequency or percentage. Median and range are employed to determine the central tendency and variability respectively. In case of interval or ratio data which is distributed normally, can be described employing frequency, relative frequency, percentage or the z score. Central tendency and variability can be depicted by mean and standard deviation respectively.

For more information on statistical hypothesis testing, readers may refer to standard books of statistics or Pharmaceutical Statistics by David S. Jones[5].

4. STATISTICAL ESTIMATION USING CONFIDENCE INTERVALS

As stated earlier in this chapter, the mean ($\bar{X}$) and the standard deviation (s) of sample data are used to estimate the true (population) mean and true (population) standard deviation. Now the question is how reliable is sample data in representing population data? This question can be considered using confidence intervals. The confidence intervals are quoted as a mean and range. The latter represents the probability of observing true mean. The imposed probability is chosen by the person doing the statistical manipulation of data. The most frequently used confidence intervals are 90%, 95% and 99%.

Pharmacopoeias prescribe fiducial limits of errors which are based on probability. For example, IP prescribes, under general notices, as under:

"fiducial limits of error are stated in biological assays. In all cases, fiducial limits of error are based on a probability of 95%."

4.1 Confidence Intervals for the Population Mean and the Normal Distribution

Normal Distribution is one of the most employed method for calculating confidence intervals. If the data conforms to the normal distribution, two-tailed confidence interval may be calculated with the help of the following equation:

$$P\ \% = \bar{X} \pm \frac{z_{p\%}\sigma}{\sqrt{N}}$$

where,

P % = selected confidence interval
$\bar{X}$ = observed mean
σ = population standard deviation
$z_{p\%}$ = z value corresponding the percentage confidence interval

(*Note:* The term $\sigma / \sqrt{N}$ refers to the standard error of the mean.)

The z value is an important parameter in the calculation of confidence intervals. As there are two possible outcomes, one interval below and one interval above the mean, confidence limits are two tailed events. It is for this reason that the formula has ± (plus-minus) sign. The z value chosen for inclusion in the equation mentioned above is dependent on the selected value of probability. Suppose we want to calculate 95% confidence interval, we choose a z value of 1.96 because this value corresponds to the area under the standard normal distribution that includes 95% of all values. Similarly, the z value of 1.65 and 2.58 are chosen for 90% and 99% confidence intervals.

After calculation of these values, it is necessary to understand the meaning of confidence intervals. After calculating these two values, calculate the mean value. The observed mean value is not likely to be identical to the population mean value. But the calculated confidence interval provides an estimation of the reliability of the measured mean. We can say that we are 95% certain that true mean will lie within the range defined by the confidence interval or we can say that if 100 samples were selected and their means and

confidence intervals were calculated, it is likely that 95 such confidence intervals would contain true mean.

4.2 Confidence Intervals for Differences between Means

As stated earlier, usually confidence intervals are calculated to provide an estimation of the mean of the population. However, at times, it is useful to describe the confidence intervals on the differences between means. In this case also, normal distribution may be used assuming that the means are derived from a normal distribution. The following equation can be used for this purpose:

$$P\ \% = (\bar{X}_1 - \bar{X}_2) \pm z_{p\%} \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

where,

$\bar{X}_1$ and $\bar{X}_2$ = means of sample 1 and 2 respectively

s_1^2 and s_2^2 = variances of sample 1 and 2 respectively

$z_{p\%}$ = the z value relating to chosen level of probability (e.g. 90%, 95%)

N_1 and N_2 = sample sizes in sample 1 and 2 respectively

$\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$ = standard error of the difference between two independent means

4.3 Confidence Intervals for Standard Deviations

Determination of confidence intervals for standard deviation is used to examine the variability of data (whenever there is unexpectedly high variation in a sample). The chi (χ^2) distribution is used to calculate confidence intervals for the population standard deviation. Tables of probability distribution are available in the standard books of statistics and also in IS: 6200 (Part I, Part II and Part III). From these tables, values may be obtained which provide information about the areas under the probability curve. If we were to calculate 95% confidence interval of the population standard deviation, the region under the χ^2 distribution that equates 95% is taken. It is between $\chi^2_{0.025}$ and $\chi^2_{0.975}$. It means that 2.5% of observations lie both below $\chi^2_{0.025}$ and above $\chi^2_{0.975}$. Considering this, the 95% confidence interval is calculated using the following equation:

$$\chi^2_{0.025} < \frac{(N-1)\,s^2}{\sigma^2} < \chi^2_{0.975}$$

where,

N – 1 = number of degrees of freedom (sample size minus 1)
s = sample standard deviation
σ = population standard deviation
$\chi^2_{0.025}$ and $\chi^2_{0.975}$ = χ^2 statistics relating to probabilities of 2.5% and 97.5% for n–1 degrees of freedom

The equation may be rearranged to give the following equation:

$$\frac{s\sqrt{N-1}}{\chi^2_{0.975}} < \sigma < \frac{s\sqrt{N-1}}{\chi^2_{0.025}}$$

4.4 Confidence intervals for proportion

Confidence intervals of proportion may be calculated using approximation to the normal distribution and sample standard deviation. The confidence interval for the population proportion may be calculated using the equation given below:

$$P\ \% = p \pm z\sqrt{\frac{pq}{N}}$$

where,

p = proportion of successes
q = proportion of failures (1–p)
N = sample size
z = the z value relating to a defined probability level

5. CONTROL CHARTS

5.1 Shewart Control Charts

5.1.1 Shewhart Control Chart for Average and Range ($\bar{X}$ & R)

The word, "Control" has a special meaning when we use it in context with SQC. A process is described as in control when a stable (uniform) system of chance causes to be operating.

To explain control charts, I will use the data from in-process control that is usually carried out while tablets are compressed in a pharmaceutical unit. Indian Pharmacopoeia (I.P.) prescribes a

general requirement for all tablets except coated tablets (baring film coated tablets) and tablets that are required to comply with the test for uniformity of content for all active ingredients. This requirement is uniformity of content popularly known as weight variation. This in-process control is done by weighing 20 tablets individually. Average weight is calculated from the weights obtained. Not more than two of the individual weights should deviate from the average weight by more than the %age shown in table 1.8.

Table 1.8: Pharmacopoeial limits of weight variation in tablets

Average weight of tablet	% deviation
80 mg or less	10
80 mg but < 250 mg	7.5
250 mg or > 250 mg	5

To illustrate the control chart for X and R data for weight variation of Nifedipine Tablets IP, 10 mg has been tabulated in Table 1.6. Sometimes, measurement of weight of each tablet are plotted for each sample with upper & lower tolerance limits (*see* Fig. 1.7). Such a chart is not the Shewhart control chart. It can, however indicate whether all tablets weigh or not within the tolerance limits. Averages of samples can also be plotted (Fig 1.8). This type of chart could be useful as it might show trends more clearly than the chart shown in Fig 1.7. However, without the limits provided by the Shewhart technique, it will not indicate whether the process shows lack of control in the statistical sense of the meaning of the word, control.

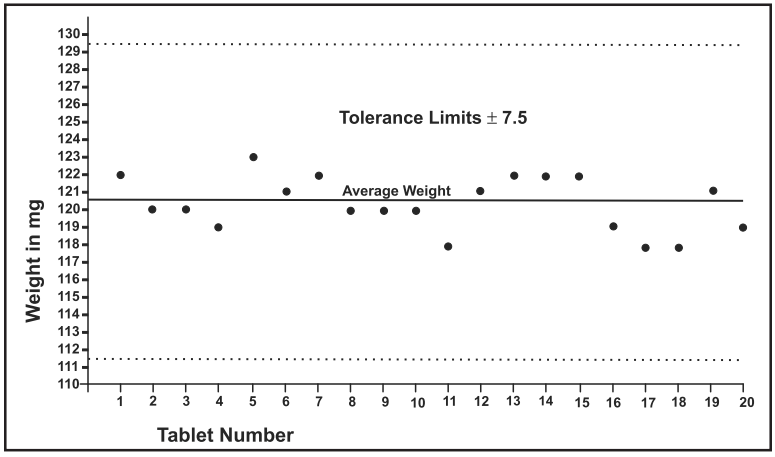

Fig. 1.7: Upper and lower limits of tolerance

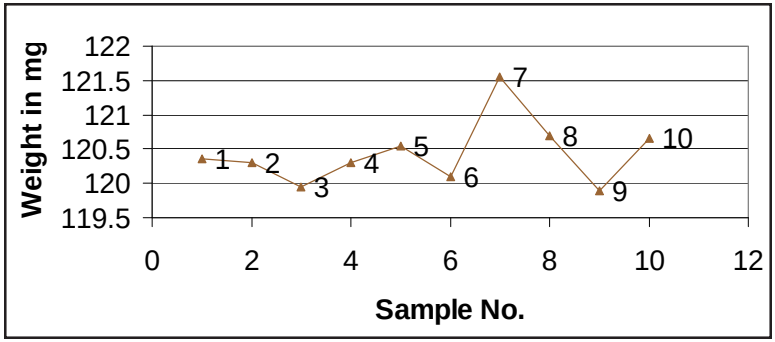

Fig. 1.8: Plot of averages of tablets samples

Note : These are not Shewhart Control Charts

It may be noted that it would have been misleading to indicate tolerance limits on chart. It is the individual tablet that has to meet tolerance limits, not the average sample. Average samples may well be falling within limits and still individual tablets might fall out side tolerance limits.

It will be necessary to understand some basic concepts before preparing control charts. Variation appears inevitable in nature and manufacturing processes are no exception whether one is attempting to control dimension or weight or feel or colour and so on. It follows from this that it is necessary to have some simple methods of describing patterns of variation. Statisticians have developed such methods. One useful method involves a frequency distribution, other involves finding of a measure of central tendency of a distribution (i.e. an average) combined with some measure of the dispersion (or spread) of the distribution. Table 1.9 shows the weights of tablets taken at intervals during the compression of tablets. One way to organize such figures is to show their pattern of variation that is to count the number of times each value occurs. The results of such a count are called a frequency distribution (see fig. 1.9).

Class Intervals	Tally Marks	Frequency
115.5 -116.5	|	1
116.5 -117.5	𝍸	5
117.5 -118.5	𝍸 𝍸 𝍸 |||	18
118.5 -119.5	𝍸 𝍸 𝍸 𝍸	20
119.5 -120.5	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||	68
120.5 -121.5	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |	41
121.5 -122.5	𝍸 𝍸 𝍸 𝍸 𝍸 |||	29
122.5 -123.5	𝍸 𝍸	10
123.5 -124.5	𝍸 ||	7
124.5 -125.5	|	1
	Total	200

Fig. 1.9: Frequency Distribution Chart

Table 1.9: Weight of individual tablet in mg of 10 samples

Sample No.	Weight of Individual Tablet in mg																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	122	120	120	119	123	121	122	120	120	120	118	121	122	122	122	119	118	118	121	119
2	120	122	121	123	121	119	118	117	118	119	122	121	121	119	120	120	123	121	120	121
3	122	117	119	120	120	121	121	121	120	120	121	120	122	118	120	118	118	120	121	120
4	120	120	122	120	120	122	120	120	120	119	121	121	120	118	120	120	121	120	122	120
5	119	123	124	118	119	120	120	122	121	122	122	124	123	119	121	120	120	119	117	118
6	120	120	120	122	121	116	119	120	118	121	121	121	118	122	120	120	121	120	120	122
7	120	122	120	121	120	123	121	124	124	124	120	121	121	120	120	125	122	121	122	120
8	119	123	120	120	124	121	122	120	120	118	119	123	124	121	119	119	122	121	120	119
9	117	118	118	118	120	120	119	122	117	120	121	119	121	121	123	121	120	121	122	120
10	120	121	121	120	120	122	122	120	120	121	120	118	121	120	120	122	120	122	123	120

Sets of observations could be formed in an arrangement which shows the frequency of occurrence of the values of the variable in ordered class. Such an arrangement is called grouped frequency distribution. The interval, along the scale of measurement, of each ordered class is termed a cell. Frequency for any cell is the number of observations in that cell. If the frequency of that cell is divided by the total number of observations; it is called relative frequency for the cell.

Grouped frequency distribution of data in table 1.9 is shown in table 1.10. Frequency distribution can be presented in graphic form (frequency histogram, frequency bar chart & Frequency polygon) also. Reader is advised to refer to IS : 7200 (Part II)-1975 for more information on presentation of statistical data. Presentation of sample data by means of frequency distribution is often bulky & time consuming. But some form of statistical presentation is necessary. This requires at least two numbers, one to measure central tendency of the data and other to measure its spread or dispersion.

The most commonly used measures of central tendency are the median, mode & mean, or more correctly, we can say arithmetic mean. In statistical language it implies average. The arithmetic mean

Table 1.10: Grouped Frequency Distribution

Cell Mid Points	Cell Boundaries	Observed Frequency
	125.5	-
125		1
	124.5	-
124		7
	123.5	-
123		10
	122.5	-
122		29
	121.5	-
121		41
	120.5	-
120		68
	119.5	-
119		20
	118.5	-
118		18
	117.5	-
117		5
	116.5	-
116		1

of a set of n numbers is the sum of the numbers divided by n. It can be expressed algebraically,

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots\ldots X_n}{n} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

Two measures of dispersion which are very useful in statistical quality control are the range (R) and sample standard deviation (s). It will be seen that the range is of special importance in control charts for variables. The range (R) is the difference between the largest & smallest of a set of numbers.

The sample standard deviation of a set of numbers from the arithmetic mean is designated by s. It can be expressed algebraically

$$s = \sqrt{\frac{(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + (X_2-\bar{X})^2 + \ldots\ldots\ldots\ldots\ldots (X_n-\bar{X})^2}{n-1}}$$

Standard deviation can be calculated from the above mentioned formula.

One of the objectives of sampling procedure is to gain knowledge of pattern of variation of the production process from which the sample was drawn. Statisticians use different words & phrases to define the unknown pattern of variation from which the known sample has been drawn. Often, it is called universe, parent distribution or population. True, but unknown, numerical values which describe the universe are called parameters. It is necessary to rely on numerical values devived from samples drawn from that universe to draw conclusions about an unknown universe. Such numerical values are sample mean, median, RMS or standard deviation, range, variance etc. These values summarize the information contained in the sample data. Each of the values is referred to as a static of the sample and may be used to estimate the corresponding parameter of the unknown universe.

Control charts for $\bar{X}$, R & s can supply a basis for judgment on a question of practical importance. The question which can be phrased in different manner is:

"Do the figures indicate a stable pattern of variation?"

or

"Is this variation the results of a constant cause system"

or

"Do these measurements show statistical control"

In more statistical language these question could be phrased as:

"were all the samples drawn from the same bowl" or "is there one universe from that samples seem to come"?

In quality control of manufacturing process, the answer, "No, this is not a constant cause system" makes you to hunt for an assignable cause of variation and also to attempt to remove it, if it is possible. On the contrary, the answer, 'yes, this is a constant cause system", relaxes you leaving the process alone. The rule for establishing control limits that will be basis for deciding answer "yes" or "No" should be practical one. More often, 3-sigma or three standard deviation limits are followed.

If certain points on $\bar{X}$ chart fall outside 3s limits, there is good reason to suspect some factor contributing to quality variation. This factor can be identified & correction may be carried out.

$\bar{X}$ & R charts are usually plotted on rectangular cross-section paper leaving 8 or 10 rulings to an inch. The vertical scale at the left is used for statistical measures, $\bar{X}$ & R. The horizontal scale is used for sub-group number. Other information like dates, hours or lot numbers may also be indicated on the horizontal scale. Each point can be indicated on the chart by a dot or circle or cross. Points on control chart may or may not be connected.

A central line (usually dark) at the value of $\bar{\bar{X}}$ is drawn at center of the paper. Upper control limit (UCL) is the value of $\bar{X}$ + 3-sigma and lower control limit (LCL) is the value of $\bar{X}$ – 3-sigma. Upper control limit is drawn above the central line & lower control limit is drawn below the central line. These lines are either dashed or less intense (*see* fig. 1.10).

Similarly, in case of R charts, a dark horizontal line should be drawn at the value of $\bar{R}$ in the center. UCL_R should be drawn above the central line and LCL_R below the central line (*see* Fig 1.11). If the sub-group size is six or less, the lower control limits for R does not exist.

The actual work of control chart start with first reading of any measurement like weight, volume, diameter etc. It should be kept in mind that information given by control chart is influenced by variation in measurements as well as by variation in the quality characteristic being measured. Though method of measurement will

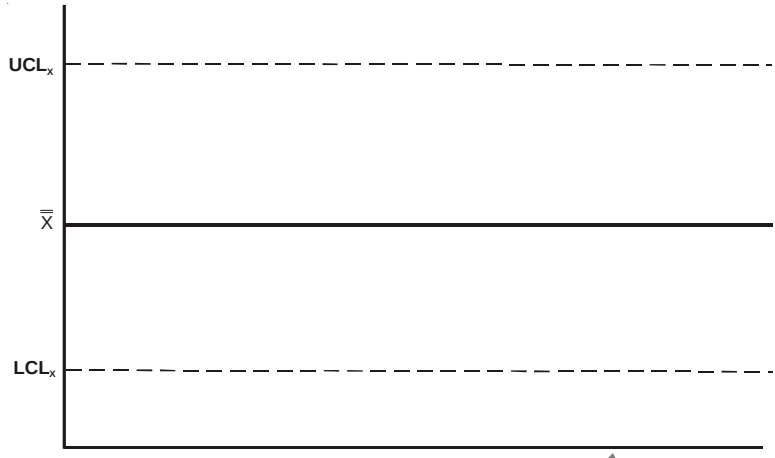

Fig. 1.10: Control chart for averages

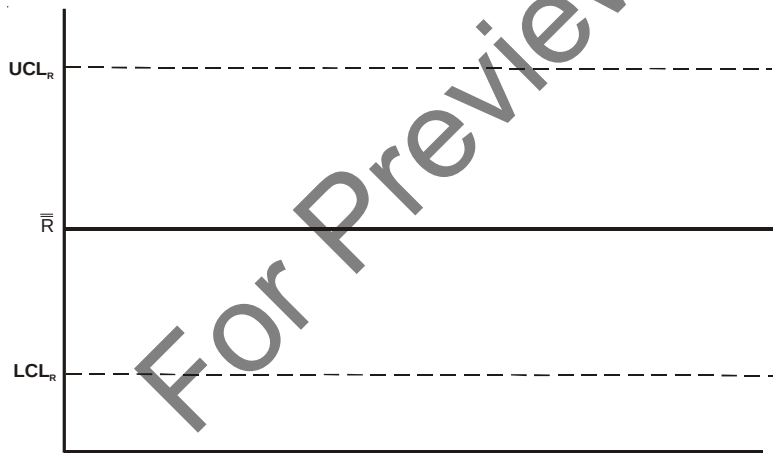

Fig. 1.11: Control chart for ranges

have its own inherent variability but what is important is that this variability should not be increased by mistakes in reading measurements or by making mistakes in recording data. All the measurements are recorded as per laid down procedure. $\overline{X}$ are found for each sample and values of $\overline{X}$ are plotted on the control chart as explained above.

If the product is being manufactured for quite some time. Control limits for the parameter can be found out from the data already available. Grand average of the subgroups, $\overline{\overline{X}}$ and the average of the ranges, $\overline{R}$ can be calculated from the data and control limits can be calculated from the following formulas:

$$UCL_{\bar{x}} = \bar{\bar{X}} + A_2 R$$
$$LCL_{\bar{x}} = \bar{\bar{X}} - A_2 R$$
$$UCL_R = D_4 \bar{R}$$
$$LCL_R = D_3 \bar{R}$$

Factor, A_2, D_3 & D_4 are given in appendix to IS:397, Methods for Statistical Quality Control during Production. These control limits may require modification before these are extended to future production. Grand average, $\bar{\bar{X}}$ can be found out by totaling $\bar{X}$ and by dividing number of samples. From the data given in table 1.9, $\bar{\bar{X}}$ is 120.33 and $\bar{R}$ is 5.5. From the formulas uppr control limit and lower control limit can be found out as illustrated below:

$$UCL_{\bar{x}} = \bar{\bar{X}} + A_2 R$$
$$LCL_{\bar{x}} = \bar{\bar{X}} - A_2 R$$
$$UCL_{\bar{x}} = 120.33 + 0.18 \times 5.5 = 121.32 \text{ and}$$
$$LCL_{\bar{x}} = 120.33 - 0.18 \times 5.5 = 119.34$$

The $\bar{X}$ (average) for each sample from the data given in Table 1.9 can be found out. The averages of these ten samples can be plotted on a control chart as shown in Fig 1.12.

Similarly, control limits for R can be calculated from formulas:

$$UCL_R = D_4 \bar{R}$$
$$LCL_R = D_3 \bar{R} \text{ i.e.}$$
$$UCL_R = 1.59 \times 5.5 = 8.74$$
$$LCL_R = 0.41 \times 5.5 = 2.25$$

Note: For values of factors A_2, D_3 & D_4 see appendix to IS:397.

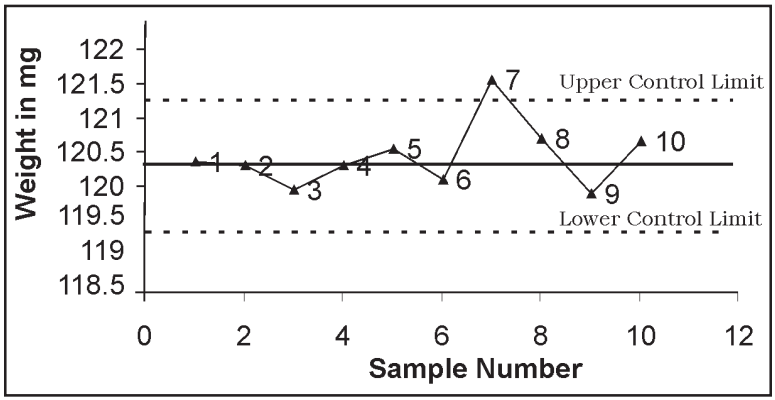

Fig. 1.12: Control chart for averages of 10 samples

Similarly range (R) for each sample from the data given in Table 1.9 can be found out and the values of range can be plotted on Y axis of control chart. Upper control limit (UCL_R) and lower control limit (LCL_R) can be drawn as per values calculated. The values of range of the samples shown in table 1.9 have been plotted in Fig 1.13.

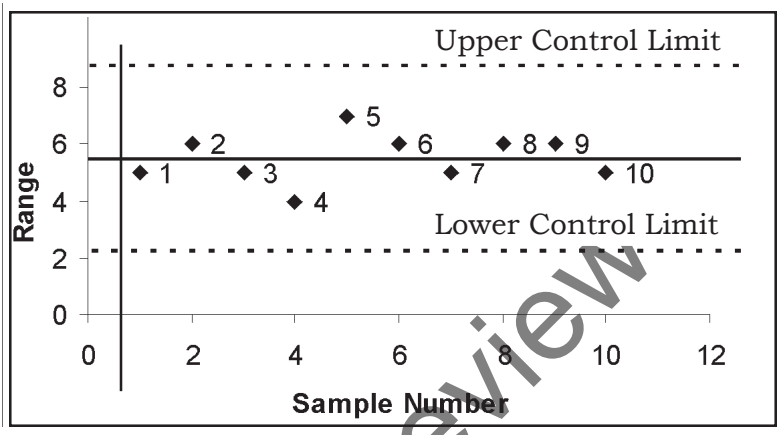

Fig. 1.13: Control chart for range of 10 samples

In the control chart for $\bar{X}$ there is only one point out side the upper control limit while in the control chart for R, all the points are within the control limits

If the points fall outside the control limits on either the $\bar{X}$ or R charts, lack of control is indicated. The points falling outside the control limits can be represented by a special symbol to make them obvious. Even though all the points may fall within control limits seven or more points falling in succession on the same side of the central line indicate lack of control. On the contrary, one or two points may fall outside the control limits. Even in the best manufacturing processes, occasional errors occur which constitute assignable causes of variation but that may not constitute a basis for action. Therefore, there is a need for practical working rule on the relationship between satisfactory control and the number of points falling outside limits. One such rule stipulates to consider not more than 1 out of 35 or 2 out of 100 points outside control limits as evidence of control* The control chart data gives estimates of the following:

* Control Chart Method of Controlling Quality during Production, ANSI Standard Z1.2-1958 (Reaffirmed 1975), American National Standard Institution, NY.

(i) the centering of the process (μ may be estimated as $\overline{\overline{X}}$),
(ii) the dispersion of the process (σ may be estimated as R/d_2)

Estimates of μ & σ are subject to sampling errors. Therefore, any conclusion obtained from short period (such as 25 points on control chart) must be regarded as tentative. These conclusions may be confirmed or changed by continuing the control chart. Practically all (all but 0.27 %) of normal distribution falls within limits of $\mu \pm 3\sigma$, or we can say, for practical purposes the spread of distribution may be thought as approx. 6σ.

When a control process must meet two specification limits, upper & lower, all possible situations may be grouped as under:

- the spread of the process (6σ) is a appreciably less than the difference between upper & lower limit;
- the spread of the process (6σ) is approximately equal to the difference between the upper & lower limit.
- the spread of the process (6σ) is appreciably greater than the difference between the upper & lower limit.

In situation one, practically all the products manufactured will meet the specification as long as the process is in control. In such a situation, there may be hardly any need for machine adjustments. In situation two also, if the process is in control, as shown by the control chart, there may be hardly any need for frequent machine adjustment. But in situation three, some points may fall above the upper control limit and some points may fall below the lower control limit and in turn some items of the product may be outside the upper & lower specification limits. In the last situation machine adjustments may be frequent. In such situation, even the increase of tolerance limits may be necessary, if the tolerance limits are too tight.

In case of a specification which has single limit, the key to the classification of situation involving minimum limit is the position of low value process distribution ($\mu - 3\sigma$) with respect to lower limit again there may be three situations:

- the low value of the process distribution ($\mu - 3\sigma$) is appreciably above L;
- the low value of the process distribution ($\mu - 3\sigma$) is approximately at L;

– the low value of the process distribution ($\mu - 3\sigma$) is appreciably below L;

In first situation, there is a margin of safety. In second situation, the product will just barely meet the specification as long as the process is in control. In the third situation nonconforming products are inevitable unless a fundamental change is made in the process. The fundamental change may be either a decrease in process dispersion or an increase in the process average. Similar reasoning will apply, if there was an upper limit. If control chart shows lack of control, obvious action is to hunt for the assignable causes of variation and try to correct them.

For recording data, record sheets like the one given below can be used.

Name of company:

Address:

RECORD SHEET FOR $\bar{X}$ & R CHART

Name of Product

B.No.

Characterstic measured

Recorded by

Date

Time	Item Inspected	Average	Range
	1,2,3,4,5,6,7,8,9,10....	$\bar{X}$	R

$UCL_{\bar{x}}$ =

$LCL_{\bar{x}}$ =

UCL_R =

LCL_R =

5.1.2 Shewhart Control Chart for Fraction Rejected

Though $\bar{X}$ & R charts are powerful instruments for the diagnosis of quality problems but they have some limitations also. One limitation is that they are charts for variables (those quality characteristics which can be measured or expressed in numbers). In a product, there are also quality characteristics which can be observed only (i.e. attributes) when the items of a product are observed for its attributes, these may be classified into one of two classes, either conforming or nonconforming to the specification. Moreover, indiscriminate use of $\bar{X}$ & R charts will be impracticable & uneconomical. For attributes, there are different types of control charts. These are:

- the p chart, the chart for fraction rejected as non-conforming to specification;
- the n p chart, the control chart for number of non-conforming items;
- the u chart, the control chart for number of non-conformities per unit.

Out of the attributes control charts, the p chart is most verstile and most widely used. This is the chart for fraction rejected as nonconforming to specification. It can be applied to quality characteristic that can be observed only as attributes. For example, visual inspection of tablets on inspection belt or visual examination of clear liquid containers. Fraction rejected may be defined as the ratio of the number of nonconforming articles found in any inspection or series of inspections in the total number of articles actually inspected. Percent rejected is 100p i.e. 100 times the fraction rejected.

Shewhart control chart model with 3σ limits may be expressed algebraically as:

Central line $y = E(y)$

$UCL_Y = E(y) + 3\sigma_y$

$LCL_Y = E(y) - 3\sigma_y$

Where y is the random variable or control static. E(y) is the expected value of statistical variable and σ_y is the standard deviation.

The mean or expected value of the binomial is p and its standard deviation is $\sqrt{p\ (1-p)}\ /\sqrt{n}$

Therefore, 3σ limits for p chart are:

$$UCL_p = p+3\sqrt{p\ (1-p)}\ /\sqrt{n_i}$$
$$LCL_p = p-3\sqrt{p\ (1-p)}\ /\sqrt{n_i}$$

As applied to 100% inspection, a control chart for fraction rejected may have one, several or all of the following purposes.

- To discover the average proportion of non-conforming units submitted for inspection over a period of time.
- To bring to the attention of the management any changes in this average quality level.
- To discover those out of control high spots that call for action to identify and correct causes for quality defect.
- To discover those out of control low spots that indicate either relaxed inspection standards or erratic causes of quality improvement which might be converted into causes of consistent quality improvement.
- To suggest places for the use of $\bar{X}$ & R charts to detect quality problems.

Often, p control chart is applied to inspection station where many different quality characteristics are to be checked. For example, inspection of tablets before packing into strips, inspection of stripped tablets. Now the question is whether to have one control chart or several? A single control chart is the most common answer with the idea that causes of rejection may be looked into from the supporting data on record sheets. However, it may be useful, at times, to use separate control charts for certain selected nonconformities. Some nonconformities may be corrected by rework or reprocessing. For example tablets with broken edges, strips with empty pockets etc. In some cases, it may be useful to have one control chart for spoilage and another for reprocessing.

Like in the case of Shewhart control chart for variables, in the control chart for fraction rejected, the most natural basis for selection of subgroups is the order in which production takes place. In all control charts, subgroups should be so selected as to minimize the chance for variation within any subgroup. An assignable cause of variation in any inspection operation is difference in inspectors. This is more true in inspection by attributes. In visual inspection, where judgement play an important role, There is great chance for difference among inspectors.

Steps in starting control chart in brief are:

- Record the data for each subgroup on number of articles inspected and rejected. Any occurrences that might give indication to an explanation of points out of control or to changes in the quality level should be noted on the data sheet as supplementary remarks.
- Compute p for each subgroup using the following formula:

$$p_1 = \frac{\text{Number of rejects in subgroup } r_i}{\text{Number inspected in subgroup } n_i}$$

- Compute p, the average fraction rejected using the following formula:

$$p = \frac{\text{Total number of rejects during period } \Sigma r_i}{\text{Total number inspected during period } \Sigma n_i}$$

However, it is desirable to have data for at least 25 subgroups before computing p and establishing trial control limits, if it is possible.

- Compute trial control limits for each subgroup based on the observed average fraction rejected (p)
- Plot each point as obtained. Plot trial control limits as soon as calculated. From this it can be noted whether the process appears to be in control.

A prototype of proforma printed below can be used as record sheet for p chart. when more than one nonconformities are to be recorded a proforma printed at the end of this chapter can be used.

RECORD SHEET FOR p CHART

Name (of company)

Site

Name of Product

Batch / Lot No.

Inspection station

Recorded by

Date	Time	Number inspected	Number of rejects	% rejected	remarks

$UCL_p = p+3\sqrt{p\ (1-p)}\ /\sqrt{n}$

$LCL_p = p-3\sqrt{p\ (1-p)}\ /\sqrt{n}$

Checked by
Supervisor QC

It is not uncommon that when the decision is made to use p chart for any manufacturing operation, data are available for the period immediately past. This data can be used for preparing p charts thus avoiding the period for which no control limits are currently available. When a period is over and trial limits have been computed on the basis of p, the control chart may show any condition from an excellent control to absence of any control i.e. (all points falling within control limits to a very few points falling within limits). If the control chart indicates control, p_o should be assumed to be p. It is generally desirable eventhough p is considered a too high fraction rejected to be satisfactory in the long run. There is no use to set a standard which is not attainable. Even if the chart shows hopeless absence of control, it is generally better to continue the p chart for sometime without any control limits and without any standard value of fraction rejected (p_o) until the situation could be improved somewhat. If the control limits are to be respected, it has to be demonstrated that it is possible to stay within control limits most of the times. Unless such an evidence exists, there is hardly any use to draw control limits on p chart. In most cases, the control chart for preliminary period will show a condition between two extremes of statistical control & complete absence of statistical control. There will be few points outside control limits. In such situation, the best procedure is to eliminate points above control limits & then to recompute p. Then, judgement may be applied to the revised p when establishing the standard fraction rejected (p_o) to be used in future.

Once p_o is established 3-sigma (3σ) limits are computed based on this value. Then, as soon as data are obtained, points and limits should be plotted promptly on the control chart. In case of $\bar{X}$ & R charts, generally it is not desirable to draw lines connecting points. But contrary is true for p charts. A line connecting points is helpful in the interpretation of the chart.

Suppose a pharmaceutical company is manufacturing co-trimoxazole Tablets I.P. and has manufactured several batches & the data for fraction rejected for 10 batches is found as under:

Table 1.11: Values of rejects in 10 samples of tablets

Batch	Number rejected	Total number inspected
1	500	98000
2	800	98000
3	600	98000
4	1000	98000
5	750	98000
6	650	98000
7	900	98000
8	550	98000
9	850	98000
10	700	98000

Note: Total number of tablets inspected has been taken same for each batch for simplification. Fraction rejected (p) and 100p for this data is given in table 1.11 and table 1.12.

Table 1.12: Values of fraction rejected

	Fraction rejected(p)	Percent fraction rejected (100 p)
1.	500/98000 = 0.0051	0.51
2.	800/98000 = 0.0081	0.81
3.	600/98000 = 0.61	0.61
4.	1000/98000 = 0.0010	0.76
5.	750/98000 = 0.0076	0.66
6.	650/98000 = 0.0066	0.91
7.	900/98000 = 0.0091	0.91
8.	550/98000 = 0.56	0.56
9.	850/98000 = 0.86	0.86
10.	700/98000 = 0.0071	0.71

It is convenient to plot 100 p rather than p. Average of 100 p in the above example is 0.73. UCL and LCL can be found out as follows:

$$\text{UCL} = 0.73 + 3\sqrt{0.73(1-0.73)/10}$$
$$= 1.15$$
$$\text{LCL} = 0.73 - 3\sqrt{0.73(1-0.73)/10}$$
$$= 0.31$$

To plot these values, a central line should be drawn at the value of 0.73. Upper Control Limit at the value of 1.15 and Lower Control Limit at the value of 0.31 should then be drawn. Values of 100p of 10 batches mentioned above should thereafter be plotted. These values have been plotted in Fig. 1.14.

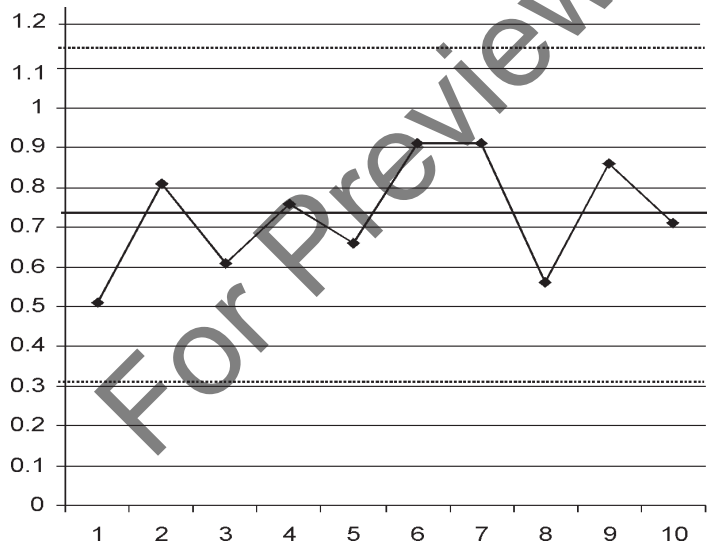

Fig. 1.14: Control Chart for Fraction Rejected

The standard fraction rejected should be reviewed periodically. However, review might be done at irregular intervals when there is enough evidence to justify a change where many charts are being used, it is a good practice to fix a review period like a week, a month etc. where a subgroup consists of a days production, review period could be even 2 months. Where subgroup consist of production lots or lots submitted for inspection, frequency of subgroups can be used for deciding review period. This period may be decided as once every 20 subgroups, once every 40 subgroups etc.

Whenever sustained decrease in average percent rejected is observed, it indicates real quality improvement. In such situations, it will be good idea to revise p_o downward on the contrary, whenever there is sustained increase in the average percent rejected, it indicates poor quality level. In such situations p_o should not be revised upward unless it is demonstrated that it is due to certain changes made and is inevitable because of changes.

It is matter of experience & it has been shown that mere introduction of p chart often causes some quality improvement. Many a times, it may be because of focussing of attention of production personnel to the quality level and may not have any relation with actual use of control limits. In the long run, however, quality improvement attained through p chart may be attributed to the concentration of personnel on assignable causes of troubles indicated by the points falling above the upper control limit. Such out of control points are known as high spots. These are reported to production supervisors and management. Detection and correction of assignable causes of poor quality are technical jobs. In such cases, it may do no good merely to bring pressure on supervisor on the basis of high spot reports. The supervisor may already be knowing the trouble. What is needed is technical help in finding out its causes & correction.

Low spots on the control chart (i.e. points falling below the lower control limits), on the contrary call or different kind of attention. These spots, sometimes, indicate faulty inspection. This may necessitate providing better inspection standards or providing better inspection training. In other cases, low spots may be worth examining to discover the reasons why quality for one subgroup was better than the standard. A knowledge of these reasons may help to bring about quality improvement.

5.1.3 Control Chart for Nonconformities

We have discussed $\bar{X}$ & R control charts & p control charts. The $\bar{X}$ & R control charts can be applied to any quality characteristic that can be measured. The control chart for p can be applied to the inspection data that either rejects or accept the article inspected. Thus these control charts are broadly useful in any statistical quality control programme. The control chart for nonconformities, usually called as c chart has a restricted field of usefulness. However, there are certain manufacturing & inspection situations in which c chart

may be needed. It is first necessary to determine whether its use is appropriate from the viewpoint of statistical theory before it can be used in any individual case. A nonconforming article is an article, which in some way fails to, conform to one or more given specifications. Each instance of lack of conformity to the specifications in an article is a nonconformity. Every nonconforming article contains one or several nonconformities. Where it will be appropriate to make a total count of the number of non-conformities in each article or in each group of an equal number of similar articles. For example, if the tablets of a product were to be examined for chipping, broken edges, black spots/particles a record is to be made for all these nonconformities.

In case of control charts for nonconformities, the underlying distribution of number of nonconformities per item is of poisson type wherein the opportunities for nonconformities are numerous even though the chances of nonconformities occurring at any one spot are very small. Therefore, it will be appropriate to use a control chart technique based on the Poisson distribution.

Poisson Distribution

If the n is large and many terms are involved, calculations involving use of binomials are often cumbersome. A simple approximation may be obtained to any term of binomial. This approximation is called Poisson's Exponential Binomial Limit. In statistical literature this is referred to as the Poisson Law or Poisson distribution or simply as Poisson. The larger the value of n and the smaller the value of p, the closer is the Poisson approximation.

It may be remembered that $\mu_{n\,p} = n\,p$ is the average value of the expected number of occurrences. For ordinary use in the discussion of the Poisson, $\mu_c = c$. The average (expected value and standard deviation of the Poisson distribution are:

$$\mu_c = \mu_{np} \text{ and } \sqrt{\mu_c} = \sqrt{\mu_{np}}$$

The Poisson is, therefore, a distribution for which the standard deviation $\sqrt{\mu_c}$ is always the square root of the average μ_c

In different kinds of manufactured articles, the opportunities for nonconformities are numerous, even though the chances of a nonconformity occurring in any one spot are small. The limits on control chart for c are based on this assumption.

As mentioned earlier standard deviation of Poisson is μ_c Thus 3-sigma limits on c chart are as follows:

$$UCL = \mu_c + 3\sqrt{\mu_c}$$
$$LCL = \mu_c - 3\sqrt{\mu_c}$$

When standard value of average number of nonconformities per unit, c_o, is not used, μ_c may be taken equal to the observed average c. This is always done in the calculation of trial control limits and control limits are fixed as:

$$UCL = c + 3\sqrt{c}$$
$$LCL = c - 3\sqrt{c}$$

Since the Poisson is not a symmetrical distribution, the upper and lower 3-sigma (3σ) limits do not correspond to equal probabilities of a point on the control chart falling outside limits even though there has been no change in the universe. This fact at times has been advanced as reason for the use of probability limits on c chart. The use of 0.995 and0.005 probability limits have been favoured (American Standard Z1.3-1958).

As stated earlier, theoretical, condition for the applicability of the Poisson distribution require that the count of number of occurrences of an event should have an infinite number of opportunities to occur and a very small constant probability of occurrence at each opportunity, For practical purposes, this infinite may be considered as very large. Similarly in practical situations, unknown probability of occurrence of a nonconformity is not quite constant. As long as, there are minor failures to meet the exact condition of applicability, the results obtained may be good enough for practical purposes.

Slight departures of the actual distribution from the true Poisson are likely to cause the standard deviation to be slightly greater than $\sqrt{c}$. Limits based on $3\sqrt{c}$ may really be at little less than 3-sigma. But this fact in itself does not justify to discard $\sqrt{c}$ or $\sqrt{c_O}$ as basis for calculating limits.

The c chart has been used to advantage in four types of situations. These are:

- It has been applied to a count of nonconformities, all of which must be eliminated by 100% inspection. For example, visual

inspection of tablets on inspection belt. In this use, the c chart is primarily an instrument to reduce cost of reprocess of control points. Sometimes the c chart calls attention to the lack of definite inspections standards or to irregularities in the application of inspection standards.

- In situation where a certain number of non-conformities per unit are tolerable even though it is desirable to hold their number to a minimum, the c chart may be applied to periodic samples of production. The objective of application of the c chart here is the improvement of quality of outgoing product with the idea of fewer, rejection by customer's inspection and better consumer acceptance of the product. Similar to the application of 100% inspection, the use of c chart gives management up to date information on the quality level and helps to increase uniformity of product by putting pressure on out of control points.
- It can be applied for special short studies of the variation of a particular product or manufacturing operation.
- It can be applied to sampling acceptance procedures based on defects (nonconformities) per unit.

When there is an evident change in the area of opportunity for occurrence of a nonconformity from subgroup to subgroup, the conventional c chart is not applicable, For example, if a number of units constitute a subgroup of size n, where n varies from subgroup to subgroup, nonconformities per unit (c/n) may be more appropriate control statistic. In such cases, if total nonconformities observed in each subgroup were plotted, the central line on the chart as well as control limits would have to change from one subgroup to the other. this would make chart confusing & difficult to interpret.

The symbol u is used to represent nonconformities per unit (c/n). The central line on u chart will be μ_o and control limits will be 3-sigma limits:

$$UCL = \mu_u + 3\sqrt{\mu_u}/\sqrt{n_i}$$

$$LCL = \mu_u - 3\sqrt{\mu_u}/\sqrt{n_i}$$

When a standard value for u is used, u_o is substituted for μ_u in above equations when the average value is from a series of subgroups is to be used for trial control limits to test for a constant

chance cause system and estimate μ_u, u can be found from the following equation:

$$u = \frac{\text{Total nonconformities found}}{\text{Total units inspected}} = \frac{\sum c_i}{\sum n_i}$$

and trial control limits are set as follows:

$$UCL = u + 3\sqrt{u}/\sqrt{n_i}$$
$$LCL = u - 3\sqrt{u}/\sqrt{n_i}$$

It may be mentioned here that statistic u does not follow the Poisson distribution. But the statistic nu does follow the Poisson distribution.

Data may be recorded in data record sheet. A prototype of Proforma which can be used to record nonconformities is given below.

RECORD SHEET FOR C CHART

Name of the company

Address

Name of the Product

Batch No.

Name of the Inspector

Date	Time	Number of items inspected	Number of items nonconforming					Total	Remarks
			1	2	3	4	5		

Name & Signature of Supervisor

A list of nonconformities may be made assigning each nonconformity a number, for example, if tablets were being examined, nonconformities could be listed like this:

1. Capping
2. Chipping
3. Sticking
4. Rough edges
5. Broken tablets
6. Black particles/Black spots

A similar list can be prepared for other dosage froms if the items are being examined for more than one nonconformities.

6. LINEAR REGRESSION AND CORRELATION

Many a time, the terms, regression and correlation are used by some persons interchangeably. But it should be understood that both the terms are not the same. In simple regression where two variables are involved, it is assumed that one variable is dependent and the other is independent. In contrast with this, in correlation, of the two variables both of them are assumed to be dependent variables. A classified example of regression is the relationship between concentration of a therapeutic agent in solution and the associated analytical response (e.g. ultra-violet absorption, peak area in HPLC). In this example, the response of one variable, analytical response is dependent on the magnitude of the second independent variable, i.e. concentration of therapeutic agent is not dependent on analytical response.

The simplest relationship between two variables is a linear or straight-line relationship. Linear regression analysis is a statistical tool/technique which mathematically defines the nature of linear relationship between an independent and a dependent variable. Some of the applications of linear regression analysis in pharmaceutical sciences include:

- the construction of calibration curves in pharmaceutical analysis;
- the accuracy of an analytical method for the determination of an API within defined matrix;
- identification of a linear relationship between two physico-chemical properties.

Mathematically, description of straight line is

$$y = a + bx$$

where,

y and x = dependent and independent variables
b = slope of the line
a = intercept of the line

From this equation, it is observed that the relationship between two variables may be described from two parameters, i.e. the slope and the intercept. Therefore, these two fundamental parameters are calculated in the linear regression.

In the linear regression, relationship between two variables is described in terms of a line of best fit. As mentioned above, the properties of straight line are defined in terms of slope (sometimes also referred to as the regression coefficient) and the intercept of the line on the y axis (i.e. when the corresponding value of the x axis is 0). The slope (b) is the gradient of the line. It is calculated with the help of the following equation:

$$b = \frac{N(\Sigma Xy) - (\Sigma X)(\Sigma y)}{N(\Sigma X^2) - (\Sigma X)^2}$$

where

N = number of pairs of data
ΣX = sum of all X values
Σy = sum of all y values
ΣX^2 = sum of squares of all X values
$(\Sigma X)^2$ = square of the sum of all X values
ΣXy = sum of the product of each pair of values

Another important point about he mathematical description of linear relationship is the need for simultaneous expression of the slope and intercept to understand the properties of relationship. Accordingly, in linear regression analysis, both the slope (b) and the y intercept (a) should be calculated. Sometimes, the slope has negative value. A negative slope refers to an inverse relationship between two variables.

The application of linear regression analysis to data derived experimentally involves more than the defining of the slope and the intercept of the line of best fit. Typically an user of this technique

will like to examine hypothesis concerning the regression. In some cases, he would like to compute the confidence intervals associated with regression parameters. But before this could be done, several assumptions about the regression process must be satisfied. These assumptions include:

- For each value of x, the corresponding y value has been sampled from a population having normal distribution.
- Each of the population from which y values have been sampled must have a similar variance.
- Sampling process is unbiased and the y values have been randomly sampled from their populations.
- No error is associated with the measurement of each x value.
- A linear relationship exists between the independent and dependent variables.

7. ANALYSIS OF VARIANCE (ANOVA)

Analysis of variance is a parametric statistical technique. It has wide application in scientific research. This technique can be used to analyze both paired and independent data and may also be used to simultaneously compare several variables. However, the mechanics of the technique are more complex than other parametric methods.

Since it is parametric statistical technique, its use is valid only when the conditions of experimental design conform to the assumptions which are similar to those for other parametric statistical methods and include the following:

- The data must be measured on an interval or ratio scale.
- Populations from which samples have been drawn must possess similar variances.
- Populations from which samples have been drawn must have normal distribution.
- Observations must be independent of each other.

In ANOVA, the variances of each set of data are compared. From this comparison, conclusions about similarity or dissimilarity are drawn. In ANOVA, the two variances known as "mean square error" and the "mean square between groups" are calculated and compared. Alternative names of these terms are "mean square within groups" and "mean square treatment". It is important to understand

the methods by which these two variances are calculated. The mean square within groups is effectively the mean of variability of the groups. Suppose there are three sets of data, A; B; C, mean square error can be calculated from the formula given below:

$$MS_{error} = \frac{\sigma_A^2 + \sigma_B^2 + \sigma_C^2}{3}$$

Since one of the assumption of ANOVA is the homogeneity of variances, the above formula is one measure of the population variance. If the sample sizes are not equal, then the mean square error is calculated by pooling the variances, weighting each variance as a function of the number of degrees of freedom.

Mean square between groups is the measure of the intergroup variability. According to central limit theorem, the variance of means sampled from a common population is equal to the population divided by the sample size ($s^2 = \sigma^2 / N$). This equation may also be written as $\sigma^2 = s^2 \times N$. Therefore, it may be observed that mean square error and mean square between groups provide independent estimates of the variance of the population from which individual sets of data have been sampled.

The fundamental basis of the ANOVA is the mathematical comparison (i.e. ratio) of these two estimates. If these two estimates are in agreement, the null hypothesis is accepted and if two estimates differ, the null hypothesis is rejected in favour of alternative hypothesis.

Significance or lack of significance of differences between sample means is determined by calculating the ratio of mean square treatment to the mean square error and interpreting the significance of this ratio by reference to the F distribution.

$$F_{calculated} = \frac{\text{Mean square treatment}}{\text{Mean square error}}$$

$$= \frac{\sigma^2_{treatment}}{\sigma^2_{error}}$$

If it is assumed that the sets of the data have been sampled from the same population, then the two estimates of the population variance will be equal. If that is so, according to the above equation, the F ratio will be 1. In effect, it is unlikely. This reflects the variability

of the magnitudes of means derived from any single population. But the degree of this deviation from normality is small. As the magnitude of the calculated F ratio departs from 1; it becomes more probable that null hypothesis will be rejected. As is the case in all statistics, a critical value of F is defined before the collection of the data. If the calculated value of F is equal to or exceeds (≥) the critical value, the null hypothesis is rejected in favour of alternative hypothesis and if the calculated value of F is less than the critical value, the null hypothesis is accepted.

The steps involved in performing ANOVA include the following:

- State the null hypothesis.
- State the alternative hypothesis.
- State the level of significance.
- State the number of tails associated with experimental design.
- Select the most appropriate statistical test.
- Perform the statistical test.

7.1 A priori and post-hoc comparisons of treatments

Earlier it has been stated that in ANOVA, either similarity between the means of individual treatment groups is defined or in alternative hypothesis, difference in the means is defined. One of the problem of ANOVA is the nature of rejection of null hypothesis. There may be several reasons for this outcome. In this case, three means may be different significantly or alternatively two means differ significantly. This cannot be clearly understood by ANOVA. Therefore, further statistical analyses are required. These statistical tests are referred to as:

– a priori (planned) comparison;
– post-hoc (unplanned) comparison.

These terms refer to the nature of the decision to evaluate the individual treatments for their similarity or dissimilarity. In a priori, no information is available about the differences in the means of treatment groups as it is chosen before the collection of the data. But in contrast with this, post-hoc comparisons are made after computation of the means of each treatment group and the completion of ANOVA.

7.1.1 A priori (planned) comparisons

This comparison is called planned comparison as an analyst has predetermined comparison in mind before the experiments are carried out. This may involve a comparison between an established product and a newly developed product in a comparison design or between a control and a defined treatment. In this type of tests, the chances of making a type I error are lower than the post-hoc comparisons.

Suppose there are four treatments in a multiple comparisons test. Usually the null hypothesis will state that there is no difference in the four means ($\bar{X}_1 = \bar{X}_2 = \bar{X}_3 = \bar{X}_4$). Also consider that after completion of ANOVA, the null hypothesis has been accepted. However, even there it is possible that a difference exists between one set of means, but this difference has been marked in the outcome of ANOVA because of general similarities of other treatments. This difference could have been between any of the following pairs:

- $\bar{X}_1$ and $\bar{X}_2$;
- $\bar{X}_1$ and $\bar{X}_3$;
- $\bar{X}_1$ and $\bar{X}_4$;
- $\bar{X}_2$ and $\bar{X}_3$;
- $\bar{X}_2$ and $\bar{X}_4$;
- $\bar{X}_3$ and $\bar{X}_4$;

If the comparison had been planned before the commencement of the experiment, the analyst would have 1 in 6 probability of selecting the specific difference between two treatments, that is the difference associated with type I error.

One of the methods that may be employed to identify a priori differences between treatments within a multiple comparisons experimental design is a two independent samples t test.

7.1.2 Post-hoc (unplanned) comparisons

As stated above, post-hoc comparisons are performed after statistical examination of data, these are different than a priori comparisons. There are several types of post-hoc comparisons. A detailed discussions of these is outside the scope of this chapter. The most popular post-hoc comparisons include:

– Fisher's LSD test;
– Tukey's HSD test;
– Dunnett's test.

These tests, in brief, are given below:

(i) Fisher's Least Significant Difference (LSD) test

This test is similar to the two independent samples t test. The basic formula for calculation of LSD is rearrangement of the equation which describe t statistic. Both the tests, Fisher's LSD test and t test use mean square error as the estimate of the pooled variance. However, one important difference is that LSD test may be used only when the F statistic associated with ANOVA is significant whereas this is not the case with two-independent samples t test.

Fisher's LSD (FLSD) is calculated from the following equation:

$$FLSD = t_{critical} \sqrt{s^2 \left(\frac{1}{N_A} + \frac{1}{N_B} \right)}$$

where,

N_A and N_B= sizes of two independent treatment group
s^2 = pooled variance
$t_{critical}$ = critical t statistic associated with experimental design

(ii) Tukey's Honestly Significant Difference (HSD) test

Tukey's HSD test is used to compare treatments with identical sample sizes. As in the case of Fisher's LSD test, an ANOVA is performed first and then this test is used to identify differences between the individual treatments. But unlike Fisher's LSD test, Tukey's HSD test may be executed when a non-significant F statistic has been recorded in the ANOVA.

With the help of the Tukey's HSD test, a critical difference between means of treatment is calculated. If the actual difference between two groups is equal to or exceeds (>=) this critical value, then the null hypothesis is rejected and it may be concluded that the difference between two treatments is significant. The equation that is used in the calculation of HSD statistic is another variant of the equation used for the calculation of two independent samples t test. But the major difference is that Tukey's HSD does not use the critical

value of t associated with the experimental parameters. Tukey's HSD uses a different statistic, that is, the "studentized range statistic" (q), thus

$$HSD = q\sqrt{\frac{s^2}{N}}$$

where,

q = studentized range statistic
N = number of observations in either group
s^2 = an estimate of pooled variance derived from mean square error in the ANOVA

The studentized range statistic is obtained by consulting the table of critical values of this statistic. For the table of this statistic, readers may refer to the standard books of statistics. The selection of this statistic for inclusion in the above equation requires knowledge of the following parameters:

- the number of degrees of freedom associated with mean square error in the ANOVA;
- the total number of means in the experimental design.

(iii) Dunnett's test

Dunnett's test is used when the analyst wants to compare the mean of a control group to the means of other groups. This test is considered one of the most powerful test. This test does not increase the chances of type I error.

In this test, critical statistic is calculated with the help of the following formula:

$$\text{Critical statistic} = t_d\sqrt{\frac{2 \times s^2}{N}}$$

where,

t_d = Dunnett's t statistic
s^2 = pooled variance obtained from ANOVA
N = number of observations in either the treatment or control group which are identical

Dunnett's t statistic (t_d) may be obtained from the table of critical values. Information about two experimental parameters will

be required to enable selection of the t_d statistic which relate to the conditions of experiments. These include:

- the number of treatments including control;
- the number of degrees of freedom associated with pooled variance.

As is the case in other statistics, acceptance or rejection of null hypothesis depends on whether the difference between the means of treatment and control group is equal to or greater than the critical statistic or alternatively is less than the critical value. In the former case, the null hypothesis is rejected and in the latter case, null hypothesis is accepted.

7.2 Two-way Analysis of Variance (Randomized Blocks)

The two-way model is an extension of two-independent samples t test in which more than two groups or treatments are compared. As in the case of two-independent samples t test, each individual (often referred to as a "block") is subjected to every treatment.

Generally, the order in which treatments are assigned to individuals is randomized unless special design like cross-over is used. A table of random numbers can be used to randomize the order of treatments to be tested on each individual or experimental unit.

Analysis of two way design is similar to the one-way ANOVA. However, the difference is that an additional source of variation is present. By simultaneous assessment of the effects of the two factors on dependent variable, the two-way ANOVA can provide information on the interdependency of the two factors.

The computations in the two-way ANOVA are identical to those for one-way ANOVA. But as two factors are under consideration, there are more calculations and the interpretation becomes slightly more complex. In practice, the computations in the two-way ANOVA involve the compression of two factors to form two one-way analysis of variance.

If there are three independent factors, it will be three-way ANOVA. So when the method used to examine the effects of several independent factors on the dependent variable is employed, it is referred to as multifactor ANOVA. The design of the experiment

based on multifactor ANOVA is referred to as a factorial design. A detailed discussion is beyond the scope of this chapter. Readers may refer to standard books of statistics for more information on statistical designs and techniques. One book which can be very useful is Pharmaceutical Statistics by David S. Jones, Pharmaceutical Press, London.

8. PRESENTATION OF STATISTICAL DATA

From the discussion in preceding sections, it is evident that a lot of data will be generated when experiments are carried out. A proforma should be devised to record the items of information along with ancillary information required for identification and verification. Records should show units of measurement, data, time of observation, observers name, instrument(s) used.

Discrepancies and errors may arise at various stages i.e. in the course of recording, summarization, calculation. As far as possible built-in checks should be provided. The persons handling the data should be suitably trained. An extreme observation should be examined carefully before acceptance or rejection. IS:8900 (1978) – criteria for rejection of outlying observation will be an useful guide.

Graphical data presentation is a powerful tool. The construction of graphs is simple. However, certain rules should be kept in mind. All graphs should be considered as complete units of information. A title should be given to the graph. The title should be concise, information & relevant to the information. For graphs which contain two or more plots, a key which identifies the symbols of each plot should be provided.

The axes (x axis, y axis) are important in the construction of graphs because they define pictorial basis of the presentation of data. Usually graphs are composed of sets of data which describe the relationship between a fixed variable and random variable. Choice of the range of numerical values of each axis is important to ensure optimum graphical presentation.

8.1 Types of graphs and plots

There are several types of graphs/plots. These can be broadly divided into two categories:

- graphs/plots which are used to describe relationships between a fixed (independent) variable and a dependent variable;
- graphs which are used to describe distribution of data pictorially.

Some examples of independent and dependent variables are:

- Analytical response like UV absorbance is dependent variable and concentration of an analyte is an independent variable.
- Hardness of a tablet is a dependent variable and force of compression is an independent variable.

An example of graph representing linear relationship between concentration of an analyte and UV response is shown in Fig. 1.15.

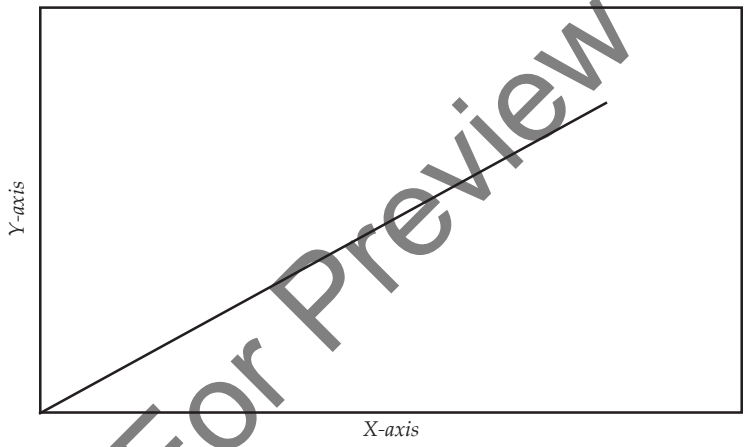

Fig. 1.15: Example of Linear Graph

Some types of graphs are used specifically for the presentation of certain types of data. For example, scatter plots are used to display correlation between data sets. This relationship may be linear or non-linear. A scatter plot is shown in Fig. 1.16. Another example is pie charts. These are commonly used to present data in the form of percentages. They are circular in design. The total area represent 100% i.e. the total frequency. The chart is divided into sections representing the data set according to the proportions. Suppose a pharmaceutical company obtained the following sales figures for different type of drugs:

Antibiotics	40%
Vitamins	30%
Analgesics	10%
Cardio vascular drugs (CV drugs)	5%
Rest	15%

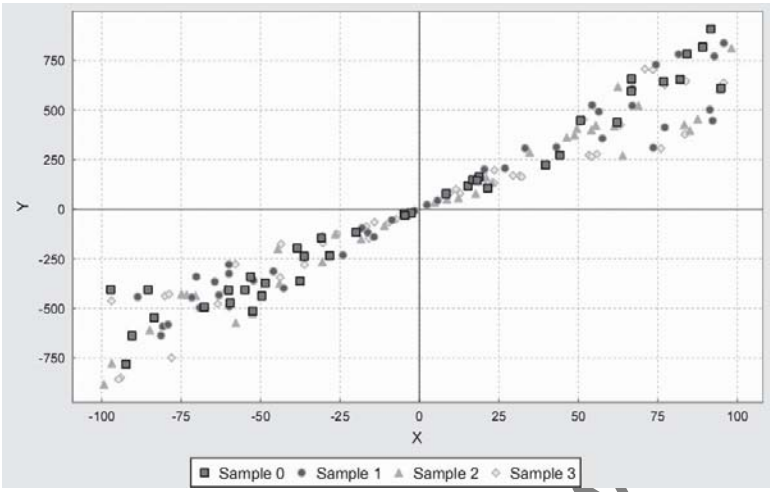

Fig. 1.16: Example of Scatter Plot

This data set can be represented in a pie chart as shown in Fig. 1.17. Graphs which are most commonly used to describe distributions of data pictorially include frequency and cumulative frequency distributions, histograms and stem and leaf displays.

One simple and useful way of summarizing numerical data is to prepare frequency table. In case of attribute type data summarization is done by noting the number of items in two categories i.e. "go" and "no go". The number of items in these two categories when added should be equal to the total number of items inspected.

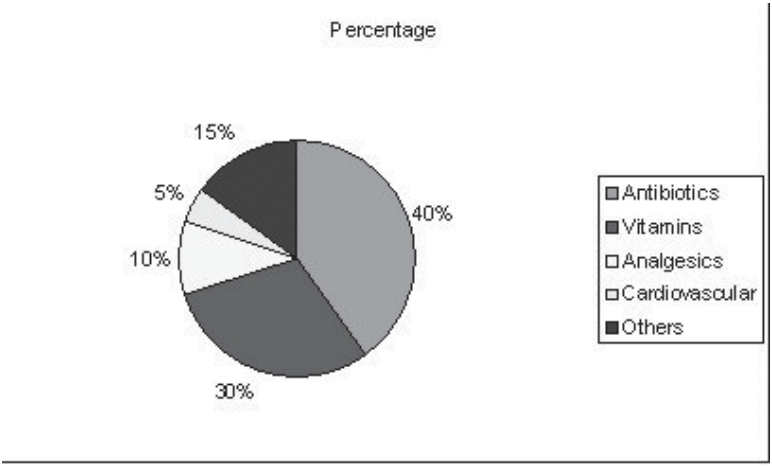

Fig. 1.17: Example of Pie Chart

In case of discrete or continuous variables, the summarization is done by forming classes within which the individual observation differs a little from one another. The total data set then can be replaced with a table which shows the number of observations falling within each identified class. There are no rigid rules for the formation of class intervals.

However, in the IS:7200 (Part I) – Presentation of Statistical Data, some broad guidelines have been mentioned. These guidelines include the following:

- The total number of observations should not be less than 100 for a frequency to show up any definite pattern.
- Practically the intervals should be of equal width for better graphical representation and for easier calculation.
- From the maximum and minimum values, range (R) can be obtained and divided into suitable number of smaller class intervals of equal width. The number of class intervals (k) may vary from 7-15 depending on the number of observations. Class width may be obtained from the following relationship.

  Number of class interval k = Integral part of [Range/Class Width (c)] +1

- Class interval should be defined in such a manner that each observation belongs to one and only one class.
- It is desirable to choose the class interval in such a manner that the mid point of the interval is convenient figure for making a plot or for calculation.
- Observations that are to be grouped should not be rounded off before grouping as grouping is a form of rounding and the successive rounding is likely to lead t_o systematic errors.

Frequency distribution of data of weights of tablets of a product is shown in section 5.1.1 of this chapter (*see* fig. 1.9).

Histograms

Frequency distribution of a data set can also be presented using a histogram. Histograms appear similar to bar charts. Both of them consist of a set of rectangles, both have their bases at x axis with centres at the class marks. Also in both cases, the areas are

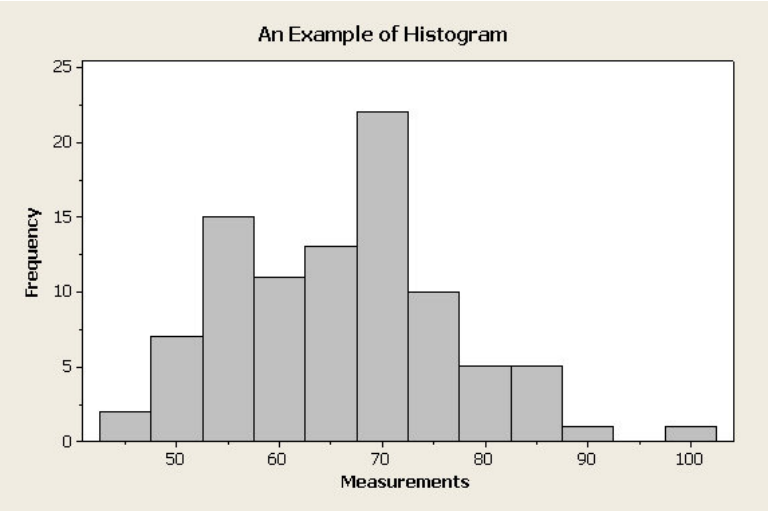

Fig. 1.18: Example of Histogram

proportional to class frequency. Histograms are used primarily to represent graphically the frequency distribution. In addition, the individual bars are joined to one another to form a continuous data display.

Cumulative frequency distribution

Another method of graphical representation may be cumulative frequency distribution. In this representation, the data are presented in the terms of total frequency of all observations which are less than the upper class boundary of class interval or alternatively the data may be presented in the terms of total frequency of observations that are greater than or equal to the lower class boundary of particular class boundary. The former is termed as "less than" cumulative frequency distribution and the latter is termed as "more than" cumulative frequency distribution.

The cumulative frequency distribution may be used to estimate the number of observations occurring within defined class intervals.

There are other graphical representation also but the ones mentioned in this section are used most commonly. The readers may refer to IS:7200 (Part I, Part II & Part III) for more information on presentation of statistical data. Other useful Indian standards relating to statistical principles & techniques are listed below:

IS:4905 Methods for Random sampling.

IS:397 (Part I, Part II) Methods for Statistical Quality Control During Production

IS:6200 (Part I, Part II, Part III) Statistical Tests of Significance

IS:7200 (Part I, Part II, Part III) Presentation of statistical data

IS:8900 Criteria for the rejection of outlying observations.

An attempt has been made in this chapter to give information in brief on principles and techniques which can be used in the quality control and validation activities in the pharmaceutical industry.

References

1. Kendall M.G., Buckland WR, A dictionary of statistical terms, 4th ed. Longman, London, 1982
2. Sokal R.R., Rohlf F.J., Biometry, 2nd ed., W.H. Freeman, New York, 1981
3. IS:6200 (Part I) statistical tests of significance, Bureau of Indian Standards, New Delhi.
4. IS:6200 (Part II) Statistical tests of significance, Bureau of Indian Standards, New Delhi
5. David S. Jones, Pharmaceutical Statistics, Pharmaceutical Press, London.