APPLIED STATISTICS IN PHYSICAL EDUCATION

Master of Physical Education (M.P.Ed.)
Course Material for Students circulation

Edited by

Dr. S. Glady Kirubakar
Assistant Professor

Y.M.C.A. COLLEGE OF PHYSICAL EDUCATION
A Project of the National Council of YMCAs of India
No.497, Anna Salai, Nandanam, Chennai 600035.
Ct. No. 044-24344816; Email: ymcanandanam1920@gmail.com

MCC202 - APPLIED STATISTICS IN PHYSICAL EDUCATION

Unit I

Introduction Meaning and Definition of Statistics. Function, need and importance of Statistics. Types of Statistics. Meaning of the terms, Population, Sample, Data, types of data. Variables; Discrete, Continuous. Parametric and non-parametric statistics.

Unit II

Data Classification, Tabulation and Measures of Central Tendency Meaning, uses and construction of frequency table. Meaning, Purpose, Calculation, and advantages of Measures of central tendency Mean, median and mode.

Unit III

Measures of Dispersions and Scales Meaning, Purpose, Calculation and advances of Range, Quartile, Deviation, Mean Deviation Standard Deviation, Probable Error. Meaning, Purpose, Calculation and advantages of scoring scales; Sigma scale, Z Scale, Hull scale

Unit IV

Probability Distributions and Graphs Normal Curve. Meaning of probability- Principles of normal curve Properties of normal curve. Divergence form normality Skewness - and Kurtosis. Graphical" Representation; -in-. Statistics: Line diagram, Bar diagram, Histogram, Frequency Polygon, 0 give Curve.

Unit V

Inferential and Comparative Statistics Tests of significance; Independent “t” test, Dependent “t” test chi square test, level of confidence and interpretation of data. Meaning of correlation co-efficient of correlation - calculation of coefficient of correlation by the product moment method and rank difference method. Concept of ANOVA and ANCOVA. Note: It is recommended that the theory topics be accompanied with practical, based on computer software of statistics.

REFERENCE BOOKS:

1. Best J. W (1 971) Research in Education, New Jersey; Prentice Hall, Inc
2. Clark D.H. (1999) Research Problem in Physical Education 2nd edition, Eaglewood Cliffs, Prentice Hall, Inc.
3. Jerry R Thomas and Jack K Nelson (2000) Research Methods in Physical Activities; Illinois; Human Kinetics.
4. Roth stain A (1985) Research Design and Statistics for Physical Education, Englewood Cliffs: Prentice Hall, Inc

5. Sivaramakrishnan. S. (2006) Statistics for Physical Education, Delhi; Friends Publication
6. Thirumalaisamy (1998), Statistics in Physical Education, Karaikudi, Senthilkumar publications.

Unit I INTRODUCTION

Meaning and Definition of Statistics:

- Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data. – Croxton and Cowden.
- Statistics are numerical statement of facts in any department of enquiry placed in relation to each other. – Bowley
- Statistics are aggregate of facts. It means that a single or isolated fact, though numerically stated, cannot be called as statistics. Statistics deals with groups, but not individual items.
- Statistics are affected to a marked extent by a multiplicity of causes.
- Statistics are numerically expressed.
- Statistics should be enumerated or estimated.
- Statistics should be collected with reasonable standard of accuracy.
- Statistics should be placed in relation to each other.
- Statistics should be collected in a systematic manner for a pre-determined purpose.

Functions, Need and importance of statistics: According to Bowley, "A knowledge of statistics is like knowledge of foreign language or of algebra. It may prove of use at any time under any circumstance.

Need & Importance:

1. Statistics is essential for a country. It supplies essential information to run a government.
2. Statistics is an indispensable tool in all the aspects of economic study, in business to makemaximum profits.
3. Statistics is necessary for the formulation of policies to start new courses, consideration offacilities and further research in education.
4. Statistical methods and statistical data are indispensable in research work.
5. In physical education and sports sciences, it helps to collect the data.
6. Helps to analyse the team performance in the tournament based on their previous win or loss.
7. In cricket, football and other sports statistics helps to give the various information to the players, coaches and opponents regarding each players performance and fitness level.

8. Help to prepare budget for the organization of sports meet, tournament, and other sportsrelated activities.
9. Help to draw norms for the performance of players.
10. Help the commentators, coaches and sports journalist to estimate the performance of aplayer or team based on their previous statistics.

Functions:

- Simplifies complexity.
- One of the important functions of statistics is to present statements in a precise and definite form.
- Help to compare the data.
- Enlarges individual experiences.
- Formulates and test hypothesis.
- Tests the laws of sports sciences.
- The statistical technique for extrapolation is highly useful for forecasting future event.
- The extent of relationships between different data can be measured.

Types of statistics:

Descriptive statistics: Descriptive statistics (in the count noun sense) are summary statistics that quantitatively describe or summarize features of a collection of information, while descriptive statistics in the mass noun sense is the process of using and analyzing those statistics. Descriptive statistics is distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.

Inferential statistical: Statistical inference is the process of deducing properties of an underlying probability distribution by analysis of data. Inferential statistical analysis infers properties about a population: this includes testing hypotheses and deriving estimates. The population is assumed to be larger than the observed data set; in other words, the observed data is assumed to be sampled from a larger population. Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data and does not assume that the data came from a larger population.

Comparative statistics: Comparative statistics is the comparison of two different variables outcomes, before and after a change in some underlying exogenous parameter (sports training,

etc.,). As a type of statistical analysis, it compares two different equilibrium states, after the process of adjustment (ifany).

Parametric statistics: Parametric statistics is a branch of statistics which assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters. Most well-known elementary statistical methods are parametric. In the literal meaning of the terms, a parametric statistical test is one that makes assumptions about the parameters (defining properties) of the population distribution(s) from which one's data are drawn, while a non-parametric test is one that makes no such assumptions.

T-test

ANOVA, ANCOVA

Wilcoxon signed rank test.

Whitney-Mann-Wilcoxon (WMW) test.

Kruskal-Wallis (KW) test.

Friedman's test.

Non-parametric statistics: Nonparametric statistics refer to a statistical method wherein the data is not required to fit a normal distribution. Nonparametric statistics uses data that is often ordinal, meaning it does notrely on numbers, but rather a ranking or order of sorts.

1-sample sign test. ...

1-sample Wilcoxon signed rank test. ...

Friedman test. ...

Goodman Kruska's Gamma: a test of association for ranked variables.

Kruskal-Walli’s test. ...

The Mann-Kendall Trend Test looks for trends in time-series data.

Mann-Whitney test. ...

Mood's Median test.

Population and Sample: In statistics and quantitative research methodology, a data sample is a set of data collected and/or selected from a statistical population by a defined procedure. The elements of a sample are known as sample points, sampling units or observations. ... The sample usually represents a subset of manageable size.

- A population includes all the elements from a set of data.
- A sample consists of one or more observations from the population.

Types of Data: The collected data in any statistical investigation are known as raw data. It is useful to distinguish between two broad types of variables: qualitative and quantitative (or numeric). Each

is broken down into two sub-types:

Qualitative data: Categorical variables are also known as **discrete** or qualitative variables. Categoricalvariables can be further categorized as nominal, ordinal or dichotomous.

Nominal, Nominal variables are variables that have two or more categories, but which donot have an intrinsic order. For example, a coach could classify their team players' fitness into distinct categories such as highly fit, normal fit, less fit, not fit. So "level of fitness" is a nominal variable with 4 categories called highly fit, normal fit, less fit, not fit. **Dichotomous** variables are nominal variables which have only two categories or levels. For example, if we were looking at gender, we would most probably categorize somebody as either "male" or "female". This is an example of a dichotomous variable (and also a nominal variable). Another example might be if we asked a person if they owned a mobile phone. Here, we may categorize mobile phone ownership as either "Yes" or "No".

Ordinal, Ordinal variables are variables that have two or more categories just likenominal variables only the categories can also be ordered or ranked. So if you asked physical education teacher if they like the sports policies of the state and they could answer either "Not very much", "They are OK" or "Yes, a lot" then you have an ordinal variable. Why? Because you have 3 categories, namely "Not very much", "They are OK" and "Yes, a lot" and you can rank them from the most positive (Yes, a lot), to the middle response (They are OK), to the least positive (Not very much).

Quantitative variables: Continuous variables are also known as quantitative variables. Continuous variables can be further categorized as either interval or ratio variables.

Interval variables are variables for which their central characteristic is that they can be measured along a continuum, and they have a numerical value (for example, speed is measure in seconds or minutes). So, the difference between 15sec and 30sec is the same as 12sec to 27sec. However, speed is measured in second or minute is NOT a ratio variable.

Ratio variables are interval variables, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable. So, temperature measured in degrees Celsius or Fahrenheit is not a ratio variable because 0C does not mean there is no temperature. However, temperature measured in Kelvin is a ratio variable as 0 Kelvin (often called absolute zero) indicates that there is no temperature whatsoever. Other examples of ratio variables include height, mass, distance and many more. The name "ratio" reflects the fact that you can use the ratio of measurements. So, for example, a distance of ten meters is twice the distance of 5 meters.

	Qualitative	Quantitative
Conceptual	Concerned with understanding human behaviour from the informant's perspective	Concerned with discovering facts about social phenomena
	Assumes a dynamic and negotiated reality	Assumes a fixed and measurable reality
Methodological	Data are collected through participant observation and interviews	Data are collected through measuring things
	Data are analysed by themes from descriptions by informants	Data are analysed through numerical comparisons and statistical inferences
	Data are reported in the language of the informant	Data are reported through statistical analyses

Unit II

Meaning, uses and construction of frequency table: Frequency distribution is the method of classification of raw data. "Frequency distribution is a classification according to the number possessing the same values ofthe variables". - Ericker

Classification is the process of arranging things in groups or classes according to their resemblances and affinities, and giving expression to the unity of attributes that may subsist amongst a diversity of individuals. - R.L. Connor

Classification is the process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts. - Secrist Objectives are,

1. To condense the mass of data.
2. To present the facts in a simple form
3. To bring out clearly the points of similarity and dissimilarity.
4. To facilitate comparison.
5. To bring out the relationship.
6. To prepare data for tabulation.
7. To facilitate the statistical treatment of data.

Frequency distribution is simply a table in which the data are grouped into classes and thenumber of cases which fall in each class is recorded. There are individual observation, discrete frequency distribution and continuous frequency distribution.

Class limit: The class limits are the lowest (smallest) and highest (largest) values in the class. Class limits are also known as class boundaries.

Class interval: The difference between the lower limit and the upper limit of the class in known as the class interval, denoted as alphabet 'I' or 'c'.

Exclusive method (overlapping): In this method, the upper limit of the one class interval is the lower limit of the next class.

Inclusive method (non-overlapping): In this method, the upper limit of one class is included in that class itself.

Class frequency: The number of observations falling within a class interval is called its class frequency or frequency.

Cumulative frequency: Cumulative frequencies (cf) are derived by the cumulation of the frequencies of successive values. Less than cumulative frequency is obtained by adding successively the frequencies of all the previous variables including the class against which it is written. The cumulation is started from the lowest size to the highest size. More than cumulative frequency distribution is obtained by finding the cumulation total of frequencies starting from the highest to the lowest class.

Meaning, merits and demerits of mean:

A measure of central tendency is a typical value of entire group or data.

Mean is a value which is typical or representative of a set of data. – Murry R. Speigal

Mean is an attempt to find one single figure to describe whole of figures. - Clark and SekkadeMean is a single number describing some features of a set of data. Wallis and Roberts

Characteristics of mean:

1. It should be rigidly defined so that there is no confusion with regard to its meaningconnotation.
2. It should be easy to understand.
3. It should be simple to compute.
4. Its definition should be in the form of a mathematical formula.
5. It should be based on all the items in the data.
6. It should not be unduly influenced by any single item or a group of items.
7. It should be capable of further algebraic treatment.
8. It should be capable of being used in further statistical computation.
9. It should have sampling stability.

Merits:

1. It is easy to understand.
2. It is easy to calculate
3. It is used in further calculation
4. It is rigidly defined.
5. It is based on the value of every item in the series.

6. Its formula is rigidly defined. The mean is the same for the series, whoever calculates it.
7. It can be used for further analysis and algebraic treatment.
8. It provides a good basis for comparison.
9. The mean is a more stable measure of central tendency (ideal average).

Demerits (limitations):

1. The mean is unduly affected by the extreme items.
2. It is unrealistic
3. It may lead to a false conclusion
4. It cannot be accurately determined even if one of the values is not known.
5. It is not useful for the study of qualities like intelligence, honesty, and character.
6. It cannot be located by observation or the graphic method.
7. It gives greater importance to bigger items or a series and lesser importance to smalleritems.

Meaning, merits and demerits of median:

Median is the value of item that goes to divide the series into equal parts. It is also known as 'positional average'. The series has to be arranged in ascending or descending order before finding the median.

Median may be defined as the value of that item which divides the series into two equal parts. One half contains values greater than median and the other half containing values less than median. - L.R. Connor

The median, as its name indicates, is the value of the middle item in a series, when items are arranged according to magnitude. - Yau Lun Chou

Merits:

1. It is easy to understand and easy to compute
2. It is quite rigidly defined.
3. It eliminates the effect of extreme items.
4. It is amenable to further algebraic process.
5. Since, it is positional average; median can be computed even if the items at the extremes are unknown.
6. Median can be calculated even from qualitative phenomena i.e.., honesty, character, etc.,
7. Median can sometimes be known by simple inspection.
8. Its value generally lies in the distribution.

Demerits (limitations):

1. Typical representative of the observations cannot be computed if the distribution of item is irregular. For example, runs scored by a batsman in last five ODI matches are, 1, 2, 3,100 and 175, the median is 3.
2. Where the number of items is large, the prerequisite process i.e.., arraying the items is a difficult process.
3. It ignores the extreme items.
4. In case of continuous series, the median is estimated, but not calculated.
5. It is more affected by fluctuations of sampling than in mean.
6. Median is not amenable to further algebraic manipulation.

Meaning, merits and demerits of mode:

Mode is the most common item of a series. Mode is the value which occurs the greatest number of frequencies in a series. The mode of a distribution is the value at the point around which the item tends to be most heavily concentrated. - Croxton and Cowden

Merits:

1. It is easy to understand as well as easy to calculate. In certain cases, it can be found out by inspection.
2. It is usually an actual value as it occurs most frequently in the series.
3. It is not affected by extreme values as in the mean.
4. It is simple and precise.
5. It is the most representative average.
6. The value of mode can be determined by the graphic method.
7. Its value can be determined in an open-end class-interval without ascertaining the classlimit.

Demerits (limitations):

1. It is not suitable for further mathematical treatment.
2. It may not give weight to extreme items.
3. In a bimodal distribution there are two modal classes, and it is difficult to determine thevalue of the mode.
4. It is difficult to compute, when there are both positive and negative items in a series andwhen there one or more items are zero.
5. It is stable only when the sample is large.

6. Mode is influenced by magnitude of the class intervals.
7. It will not give the aggregate value as in average.

Unit III

Range:

The range is the simplest measure of dispersion. It is a rough measure of dispersion. Its measure depends upon the extreme items and not on all the items.

Range = Largest value – Smallest value (R= L – S)Coefficient of range = (L-S) / (L+S)

Merits:

1. It is simple to compute and understand.
2. It gives a rough but quicker answer.

Demerits:

1. It is not reliable, because it is affected by the extreme items.
2. Usually, frequency distribution may be concentrated in the middle of the series. But the range depends on extreme items, it is an unsatisfactory measure.
3. It cannot be applied to open end cases.
4. It is not suitable for mathematical treatment.

Quartile Deviation:

By eliminating the lowest 25% and the highest 25% of items in a series, the central 50% values which are ordinarily free of extreme values is known as quartile deviation. To obtain quartile deviation half of the distance between the first and the third quartiles is calculated, which is also known as 'Semi Inter Quartile Range'.

Quartile deviation (QD) = (Q3-Q1)/2

Coefficient of quartile deviation = (Q3-Q1) / (Q3+Q1)

For symmetrical distribution series, Q3= QD+Median Q1=Median – QD

Merits:

1. It is simple to understand and easy to compute.
2. It is not influenced by the extreme values.
3. It can be found out with open end distribution.
4. It is not affected by the presence of extreme items.

Demerits:

1. It ignores the first 25% of the items and the last 25% of the items.
2. It is positional average, hence not amenable to further mathematical treatment.
3. Its value is affected by sampling fluctuations.
4. It gives only a rough measure.

Mean Deviation:

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency, i.e., the mean, median or mode. All deviations are taken as positive i.e., + or – signs are ignored. Mean deviation is denoted as δ (Delta).

Mean deviation is the average amount of scatter of the items in a distribution from either the mean or the median, ignoring the signs of the deviation. – Clark and Schekade.

Mean deviation = $\sum D / |N|$

Coefficient of mean deviation = Mean deviation / mean or median or mode

Merits:

1. It is simple to understand and easy to compute. Mean deviation is a calculated value.
2. It is not much affected by the fluctuations of sampling.
3. It is based on all items of the series and gives weight according to their size.
4. It is less affected by the extreme items.
5. It is rigidly defined.
6. It is flexible because it can be calculated from any measure of central tendency.
7. It is better measure for comparison.

Demerits:

1. It is non-algebraic treatment.
2. Algebraic positive and negative signs are ignored.
3. It is not a very accurate measure of dispersion.
4. It is not suitable for further mathematical calculation.
5. It is rarely used.

Standard Deviation:

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulas. Standard deviation is also

called 'Root-Mean Square Deviation' or 'Mean Error' or 'Mean Square Error'.

Standard deviation is defined as positive square-root of the arithmetic mean of thesquares of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by σ (Sigma). Standard deviation indicates the spread of the middle 68.26 percent of scores taken from the mean.

Merits:

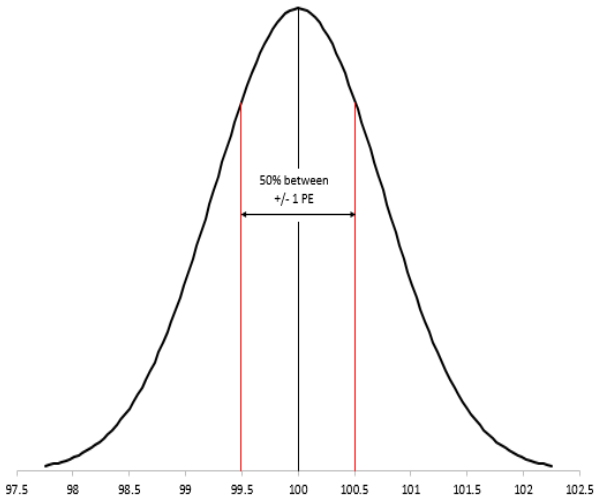
1. It is rigidly defined, and its value is always definite and based on all the observations and the actual signs of deviations are used.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.
4. It is possible for further algebraic treatment.
5. It is less affected by the fluctuations of sampling, and hence stable.
6. Squaring the deviations make all of them positive, there is no need to ignore the signs.
7. It is the basis for measuring the coefficient of correlation, sampling and statisticalinferences.
8. The standard deviation provides the unit of measurement for the normal distribution.
9. It can be used to calculate the combined standard deviation of two or more groups.

Demerits:

1. It is not easy to understand, and it is difficult to calculate.
2. It gives more weight to extreme values because the values are squared up.
3. It is affected by the value of every item in the series.
4. As it is an absolute measure of variability, it cannot be used for the purpose ofcomparison.

Probable error:

In statistics, **probable error** defines the half-range of an interval about a central point for the distribution, such that half of the values from the distribution will lie within the interval and half outside. Probable error is a quantity formerly used as a measure of variability: equal to 0.6745 times the standard deviation. A normally distributed population has half of its elements within one probable error of the mean.



Probable error, PE = 0.6744898σ

Z SCALE:

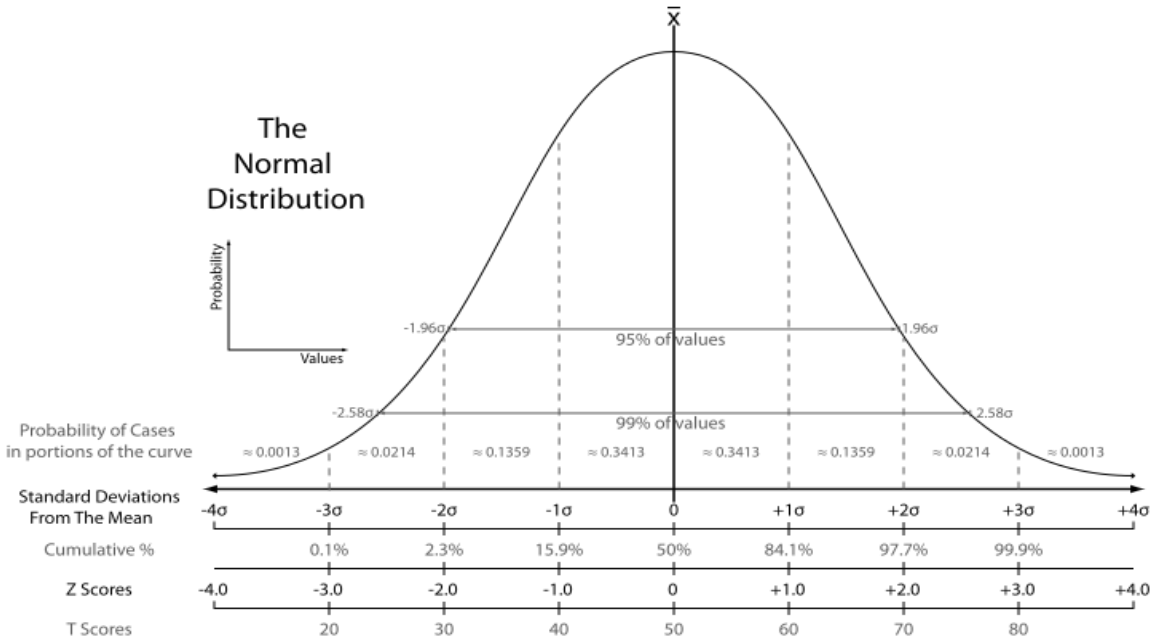
The standard score of a raw score x[1] is.

$Z=(X- \mu)/ \sigma$

μ is the mean of the population.

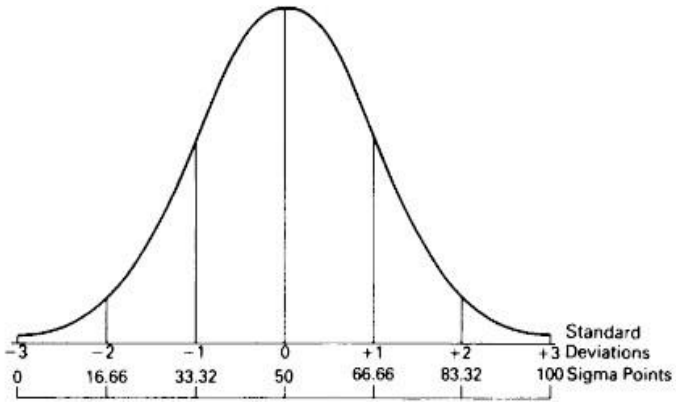
σ is the standard deviation of the population.

The absolute value of z represents the distance between the raw score and the population mean in units of the standard deviation. z is negative when the raw score is below the mean, positive when above.



Sigma scale:

In the construction of a sigma scale the distribution curve is divided into 100 equal parts along its horizontal axis, commencing with the 0 at 3 standard deviations below the mean and finishing with the 100 at 3 standard deviations above the mean. Figure showing Sigma points and standard deviation



To calculate the sigma points for a specific test score, the formula is,Sigma points = 16.66Z+50

Were,

$$Z = \frac{X - M}{S.D.}$$

X = Raw scoreM = Mean

S.D = Standard Deviation

Hull scale:

The Hull scale (named after its originator) is another way of transforming Z scores into a simpler measure of relative position defined in points from 0 to 100.

In the Hull scale again divide the distribution curve into 100 equal parts, but this time the starting point (0) is positioned 3.5 standard deviations below the mean and the finishing point. (100) 3.5 standard deviations above the mean.

Hull points = 14.28Z+50

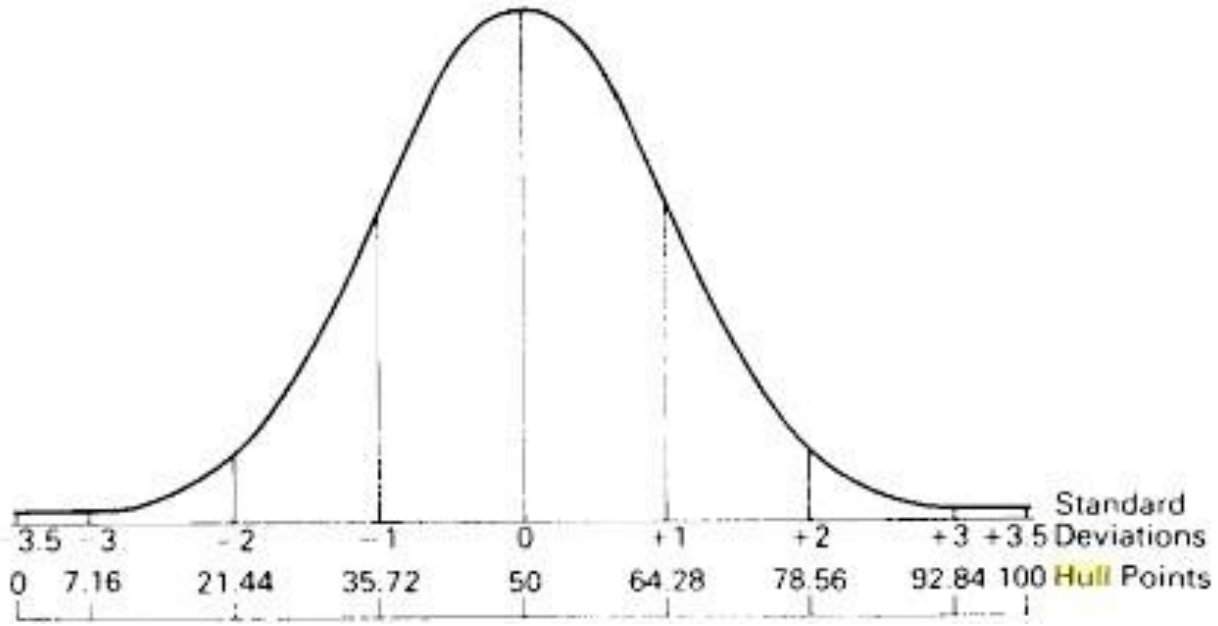


Figure showing Hull points and standard deviations.

T-scale:

Another method of scaling which divides the distribution curve into 100 equal parts obtains an even greater spread of raw scores than in the Hull scale. The starting point (0) is placed 5 standard deviations below the mean and the finishing point (100), 5 standard deviations above the mean.

T-scale points = 10Z+50

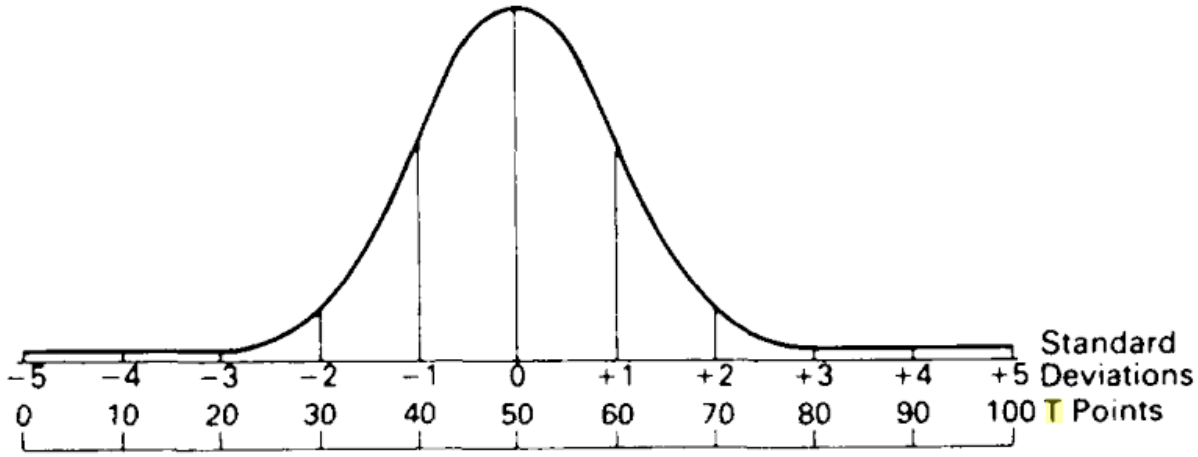


Figure showing T points and standard deviations.

Unit IV

Principles and properties of Normal curve:

The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the "bell curve," although the tonal qualities of such a bell would be less than pleasing. It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss.

Bell-shaped, the mean μ = 0 and the standard deviation σ = 1, The area under the whole normal curve is 100% (or 1, if you use decimals). Many histograms for data are similar in shape to the normal curve, provided they are drawn to an appropriate scale.

1. The normal curve is symmetrical: The Normal Probability Curve (N.P.C.) is symmetrical about the ordinate of the central point of the curve. It implies that the size, shape, and slope of the curve on one side of the curve is identical to that of the other. That is, the normal curve has a bilateral symmetry. If the figure is to be folded along its vertical axis, the two halves would coincide. In other words, the left and right values to the middle central point are mirror images.

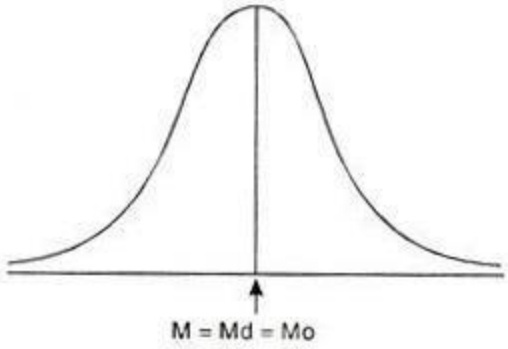
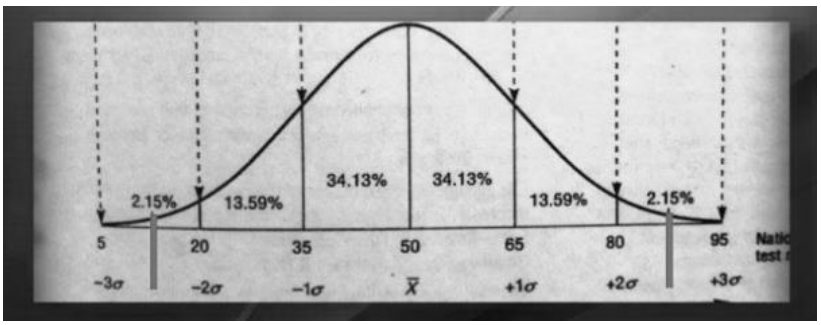


Fig. 6.2 N.P.C., M = Mdn = Mode



2. The normal curve is unimodal: Since there is only one point in the curve which has maximum frequency, the normal probability curve is unimodal, i.e., it has only one mode.
3. Mean, median and mode coincide: The mean, median and mode of the normal distribution are the same and they lie at the center. They are represented by 0 (zero) along the base line. [Mean = Median = Mode]
4. The maximum ordinate occurs at the centre: The maximum height of the ordinate always occurs at the central point of the curve, that is, at the mid-point. The ordinate at the mean is the highest ordinate and it is denoted by Y0. (Y0 is the height of the curve at the mean or mid-point of the base line).

Y_0 is given by $Y_0 = \frac{Ni}{\sigma\sqrt{2\pi}}$ where $\pi = 3.1416 \cdot \sqrt{2\pi} = 2.5066$

5. The normal curve is asymptotic to the X-axis: The Normal Probability Curve approaches the horizontal axis asymptotically i.e., the curve continues to decrease in height on both ends away from the middle point (the maximum ordinate point); but it never touches the horizontal axis. It extends infinitely in both directions i.e., from minus infinity (-∞) to plus infinity (+∞) as shown in Figure below. As the distance from the mean increases the curve approaches to the base line more and more closely.

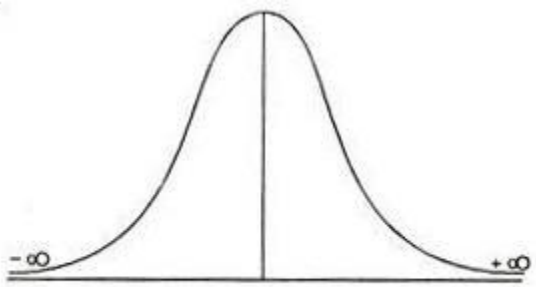


Fig. 6.3 Normal Curve is Asymptotic to the X-axis

6. The height of the curve declines symmetrically: In the normal probability curve the height declines symmetrically in either direction from the maximum point. Hence the ordinates for values of X = μ ± K, where K is a real number, are equal. For example, The heights of the curve or the ordinate at X = μ + σ and X = μ – σ are exactly the same as shown in the following Figure:

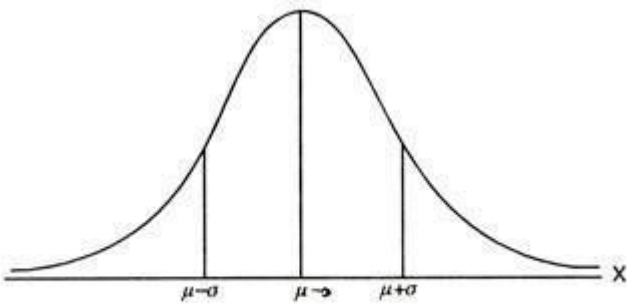


Fig. 6.4 Ordinates of a normal curve

7. The points of Influx occur at point ± 1 Standard Deviation (± 1 a): The normal curve changes its direction from convex to concave at a point recognized as point of influx. If we draw the perpendiculars from these two points of influx of the curve on horizontal axis, these two will touch the axis at a distance one Standard Deviation unit above and below the mean (± 1 σ).
8. The total percentage of area of the normal curve within two points of influxation is fixed: Approximately 68.26% area of the curve falls within the limits of ±1 standard deviation unit from the mean as shown in figure below.

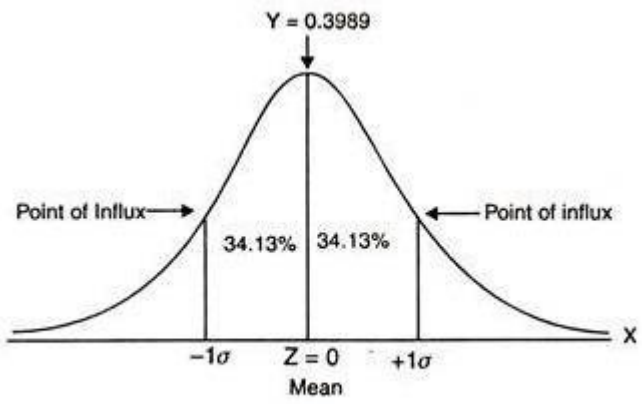


Fig. 6.5 N.P.C., 68.26% area of the curve within the limits of ±1σ

9. Normal curve is a smooth curve: The normal curve is a smooth curve, not a histogram. It is moderately peaked. The kurtosis of the normal curve is 263.
10. The normal curve is bilateral: The 50% area of the curve lies to the left side of the maximum central ordinate and 50% lies to the right side. Hence the curve is bilateral.
11. The normal curve is a mathematical model in behavioural sciences: The curve is used as a measurement scale. The measurement unit of this scale is ± σ (the unit standard deviation).
12. Greater percentage of cases at the middle of the distribution: There is a greater percentage of cases at the middle of the distribution. In between -1σand + 1σ, 68.26% (34.13 + 34.13), nearly 2/3 of eases lie. To the right side of +1σ, 15.87% (13.59 + 2.14 + .14), and to the left of-1σ, 15.87% (13.59 + 2.14 + .14) of cases lie. Beyond +2σ. 2.28% of eases lie and beyond -2σ also 2.28% of cases lie.

Thus, most eases lie at the middle of the distribution and gradually number of cases on either side decreases with certain proportions. Percentage of cases between Mean and different a distance can be read from the figure below:

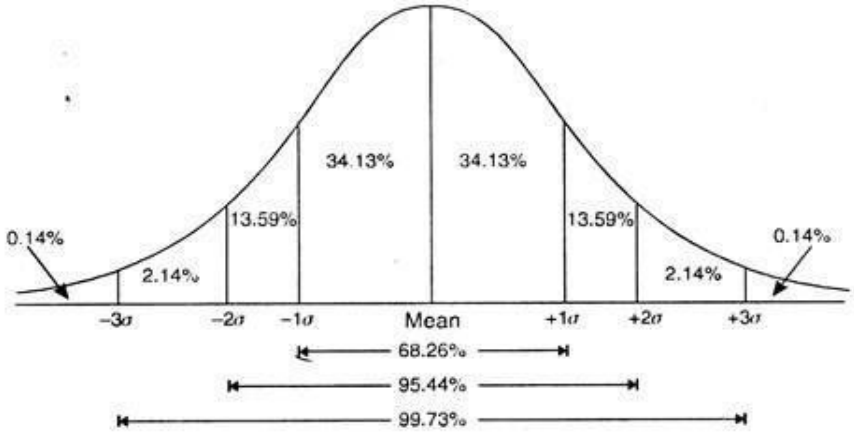


Fig. 6.6 The percentage of the cases falling between Successive Standard Deviations in Normal Distribution.

13. The scale of X-axis in normal curve is generalised by Z deviates

$$Z = \frac{X - M}{\sigma} = \frac{x}{\sigma}$$

14. The equation of the normal probability curve reads (equation of the normal probability curve) in which

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

x = scores (expressed as deviations from the mean) laid off along the base line or X-axis. y = the height of the curve above the X axis, i.e., the frequency of a given x-value.
The other terms in the equation are constants:

N = number of eases

a = standard deviation of the distribution

π = 3.1416 (the ratio of the circumference of a circle to its diameter) e = 2.7183 (base of the Napierian system of logarithms).

15. The normal curve is based on elementary principles of probability and the other name of the normal curve is the 'normal probability curve'.

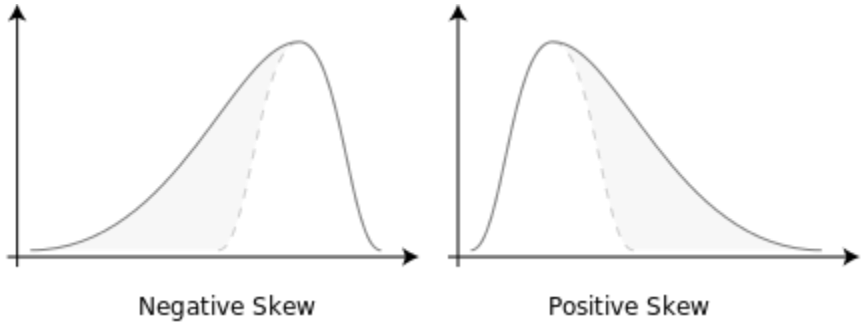
Characteristics of normal curve:

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the centre and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and thestandard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.
7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean

Divergence from normality: Two major types of divergence from normality. The two types are: 1. Skewness 2. Kurtosis.

Skewness: Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.
Skewness essentially measures the relative size of the two tails.

Negative skew: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be left-skewed, left-tailed, or skewed to the left, despite the fact that the curve itself appears to be skewed or leaning to the right; left instead refers to the left tail being drawn out and, often, the mean being skewed to the left of a typical center of the data.A left-skewed distribution usually appears as a right-leaning curve.



Positive skew: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be right-skewed, right-tailed, or skewed to the right, despite the fact that the curve itself appears to be skewed or leaning to the left; right instead refers to the right tail being drawn out and, often, the mean being skewed to the right of a typical center of the data. A right-skewed distribution usually appears as a left-leaning curve.

Skewness in a data series may sometimes be observed not only graphically but by simple inspection of the values. For instance, consider the numeric sequence (49, 50, 51), whose values are evenly distributed around a central value of 50. We can transform this sequence into a negatively skewed distribution by adding a value far below the mean, e.g. (40, 49, 50, 51). Similarly, we can make the sequence positively skewed by adding a value far above the mean, e.g. (49, 50, 51, 60).

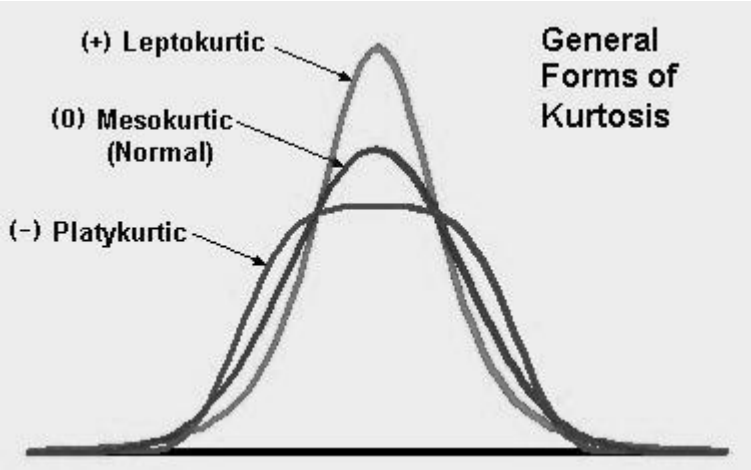
Kurtosis: Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. Kurtosis is a measure of the combined sizes of the two tails. It measures the amount of probability in the tails. The value is often compared to the kurtosis of the normal distribution, which is equal to 3. More specifically, kurtosis refers to the tails or the 2 ends of the curve. Leptokurtic - a “positive” or tall and thin distribution (fatter tails). Platykurtic - a “negative” or flat and wide distribution (thin tail)

Mesokurtic: Distributions with zero excess kurtosis are called mesokurtic, or mesokurtotic. The most prominent example of a mesokurtic distribution is the normal distribution family, regardless of the values of its parameters. A few other well-known distributions can be mesokurtic, depending on

parameter values: for example, the binomial distribution is mesokurtic.

Leptokurtic:A distribution with positive excess kurtosis is called leptokurtic, or leptokurtotic. "Lepto- " means "slender". In terms of shape, a leptokurtic distribution has fatter tails. Examples of leptokurtic distributions include the student's t-distribution, Rayleigh distribution, Laplace distribution, exponential distribution, Poisson distribution and the logistic distribution. Such distributions are sometimes termed super-Gaussian.

Platykurtic:A distribution with negative excess kurtosis is called platykurtic, or platy kurtotic. "Platy- " means "broad". In terms of shape, a platykurtic distribution has thinner tails. Examples of platykurtic distributions include the continuous or discrete uniform distributions, and the raised cosine distribution. The most platykurtic distribution of all is the Bernoulli distribution with p = ½ (for example the number of times one obtains "heads" when flipping a coin once, a coin toss),for which the excess kurtosis is −2. Such distributions are sometimes termed sub-Gaussian.



GRAPHICAL REPRESENTATION IN STATISTICS:

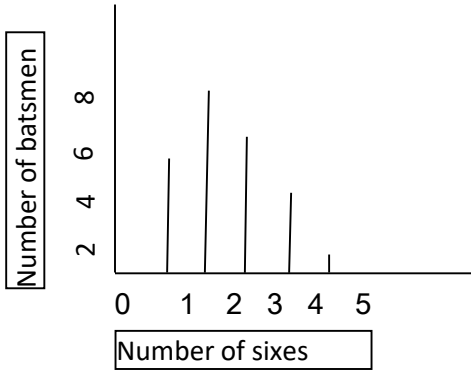
There are five types,

1. Line diagram
2. Bar diagram
3. Histogram
4. Frequency polygon
5. Ogive curve

1. **Line diagram:** This is the simplest of all the diagrams. On the basis of size of the figures, heights of barsor lines are drawn. The distance between lines is kept uniform. It makes comparison easy. This

diagram is not attractive; hence it is less important.

Number of sixes		1	2	3	4	5
Number of batsmen		5	8	6	4	1

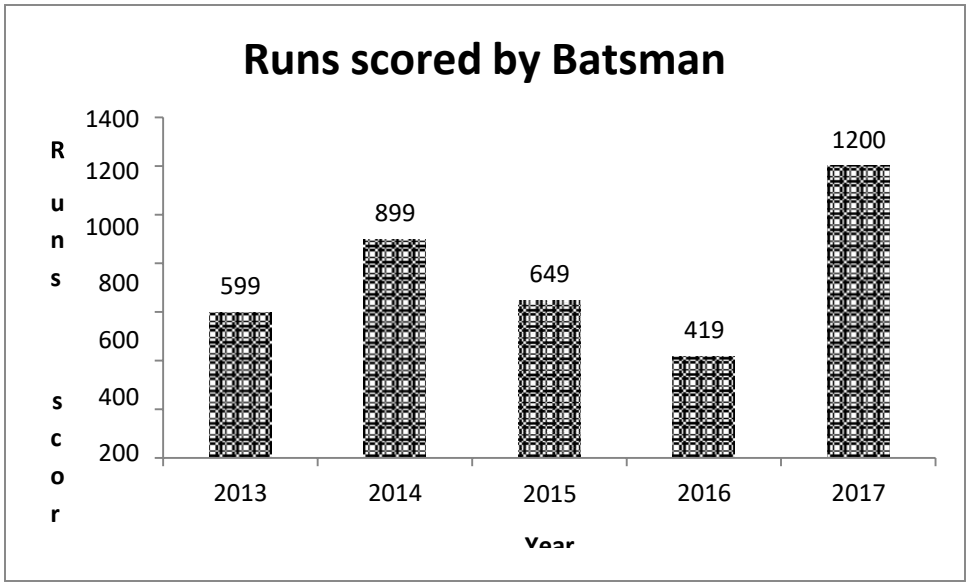


2. **Bar diagram:**

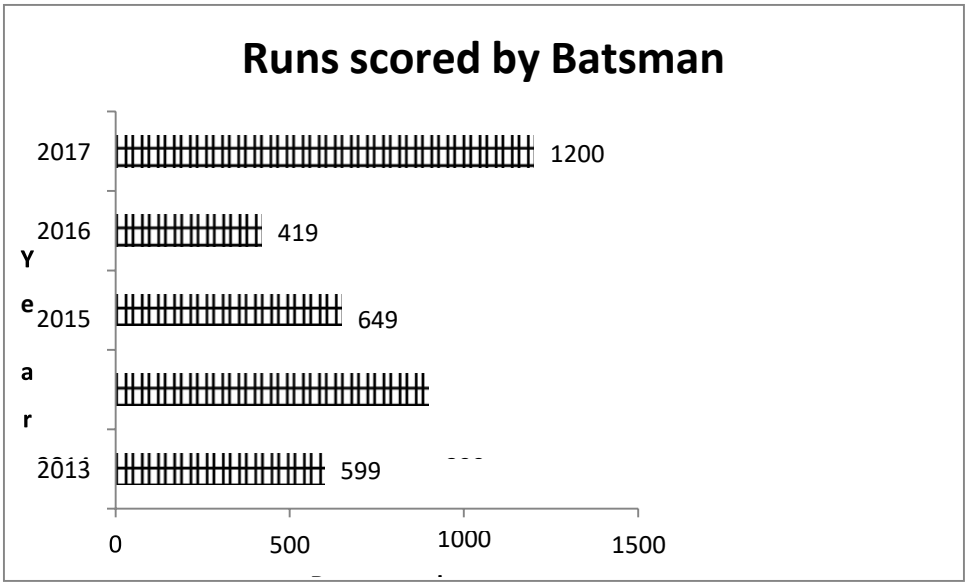
(i) Simple bar diagram vertical bar diagram

A simple bar diagram can be drawn either on horizontal or vertical base. Bars on horizontal base are more common. A bar diagram is simple to draw and easy to understand.

Year (last five years)	2013	2014	2015	2016	2017
Runs scored by Batsman	599	899	649	419	1200

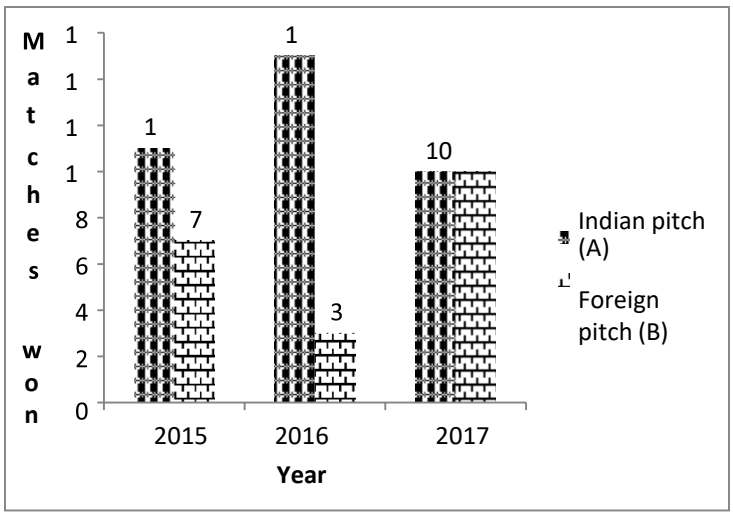


(i) simple bar diagram horizontal bar diagram



(ii) Multiple bar diagram (compound bar diagram)

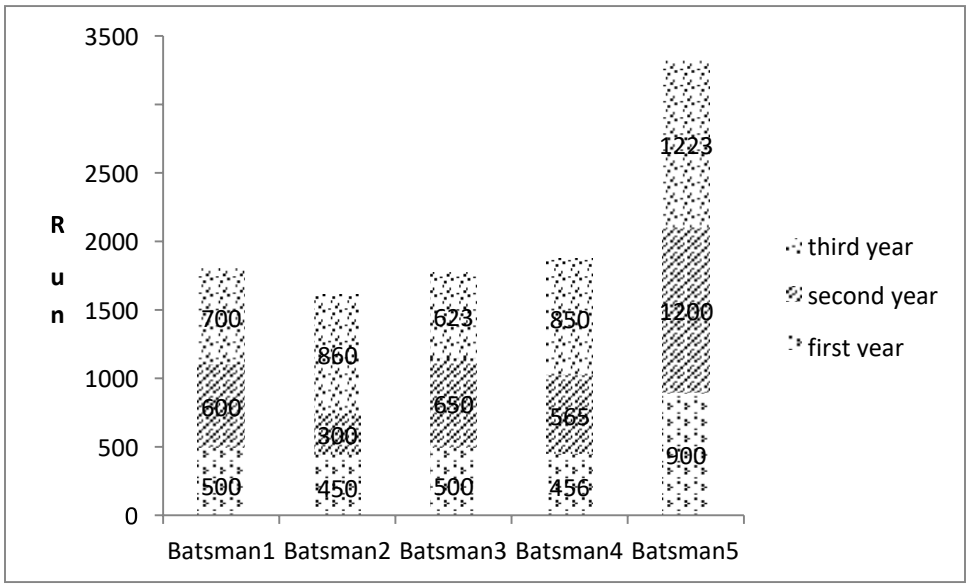
Multiple bar diagrams are used to denote more than one phenomenon. Multiple bars are useful for direct comparison between two values. The bars are drawn side by side. In order to distinguish the bars, the different colours, shades, etc., may be used and a key index to this effect be given to understand the different bars. For example, the data below gives the year wise one day matches win by India in Indian and foreign pitches in last three years.



Year (last three years)	Indian pitch (A)	Foreign pitch (B)
2015	11	7
2016	15	3
2017	10	10

(iii) Subdivided bar diagram (component bar diagram)

The bar is subdivided into various parts in proportion to the values given in the data and may be drawn on absolute figures or percentages. Each component occupies a part of the bar proportional to its share in the total. To distinguish different components from one another, different colours or shades may be given.



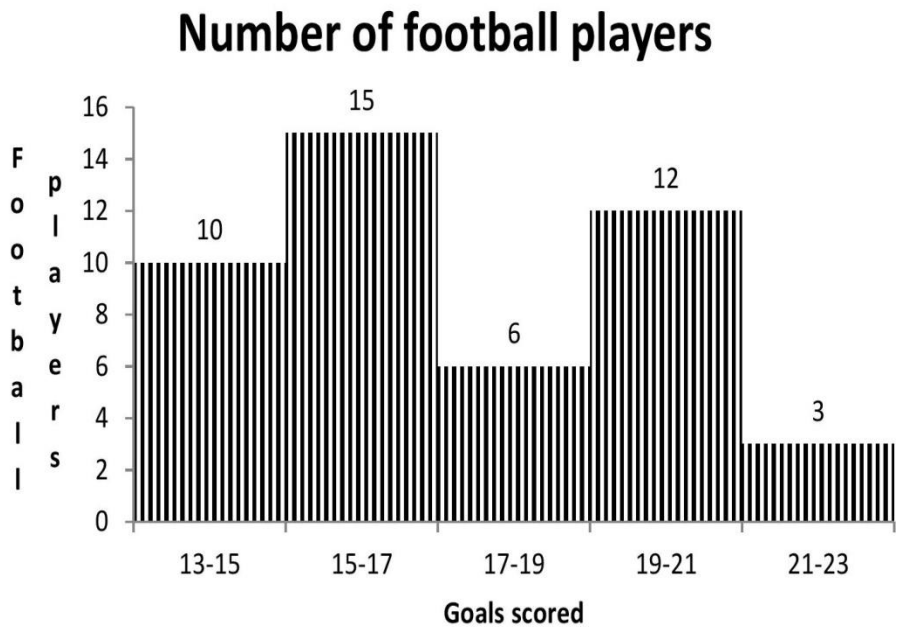
The same model can be drawn by calculating percentage of runs in each year and that bar diagram is known as 'percentage subdivided bar diagram'.

HISTOGRAM:

One of the most important and useful methods of presenting frequency distribution of continuous series in known as histogram.

In this, the magnitude of the class interval is plotted along the horizontal axis and the frequency on the vertical axis. Each class has lower and upper values and this will give two vertical lines representing the frequency. Histogram is also known as 'block diagram' or 'staircase chart'.

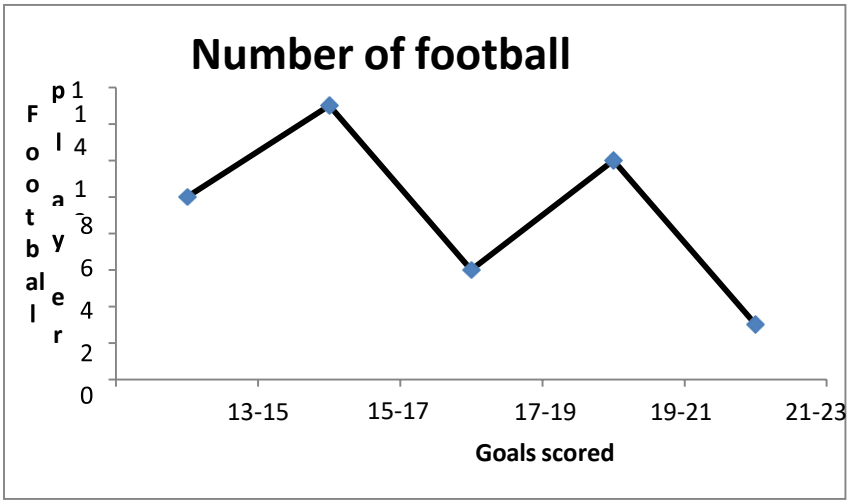
Goals scored	13-15	15-17	17-19	19-21	21-23
Number of football players	10	7	6	5	3



FREQUENCY POLYGON:

A grouped frequency distribution can be represented by a histogram. A simple method of smoothing the histogram is to draw a frequency polygon. This is done by connecting the mid- point of the top of each rectangle with the mid-point of the top of each adjacent rectangle, by straight lines. Mode can easily be found out.

Goals scored	13-15	15-17	17-19	19-21	21-23
Number of footballs players	10	7	6	5	3



OGIVE CURVE:

When cumulative frequencies are plotted on a graph, then the frequency curve obtained is called “Ogive” or “Cumulative frequency curve”.

The class limits are shown along the X-axis and cumulative frequencies along the Y-axis. In drawing an ogive, the cumulative frequency is plotted at the upper limit of the class interval. The successive points are later joined together to get an ogive curve.

There are two methods of constructing ogive,

1. Less than ogive
2. More than ogive

x	f	Less than cumulative frequency (cf)	More than cumulative frequency
0-10	3	3	80
10-20	9	12	77
20-30	15	27	68
30-40	30	57	53
40-50	18	75	23
50-60	5	80	5

x	Less than cumulative frequency (cf)
10	3
20	12
30	27
40	57
50	75
60	80

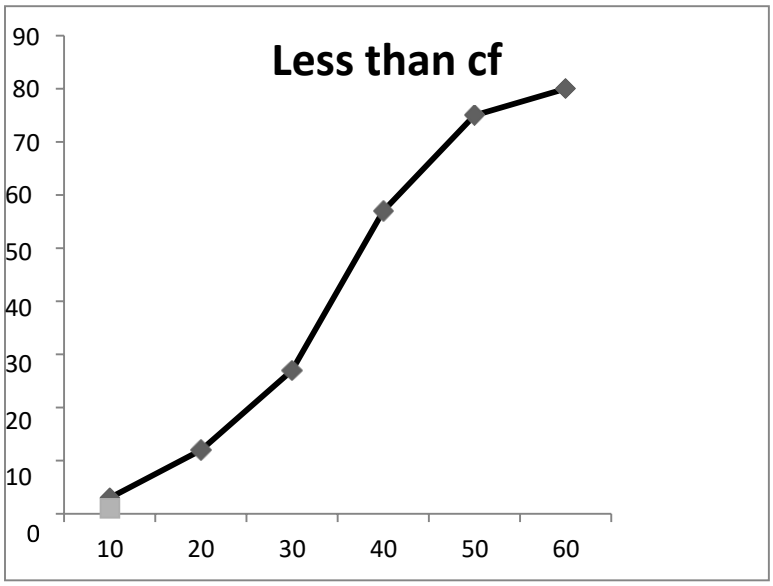
x	More than cumulative frequency (cf)
0	80
10	77
20	68
30	53
40	23
50	5

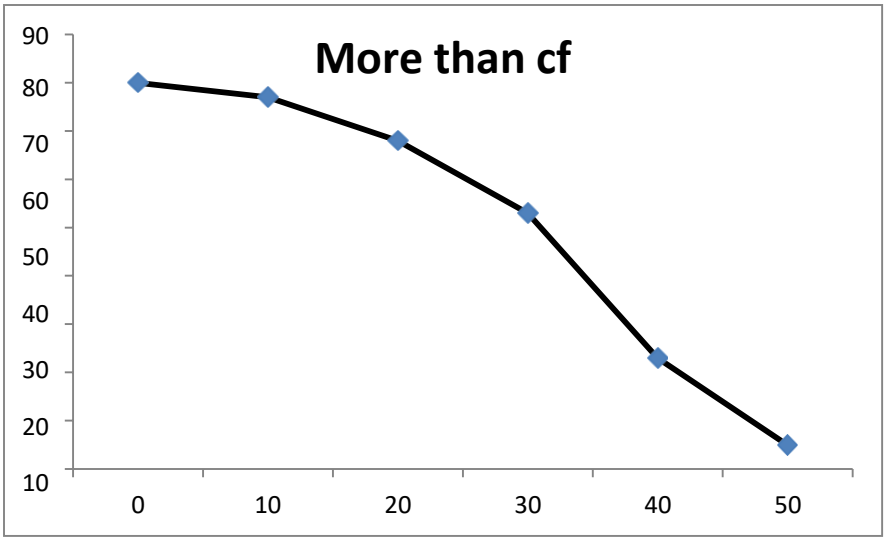
LESS THAN OGIVE:

In less than ogive, the less than cumulative frequencies are plotted against upper class boundaries of the respective classes. Then the point is joined by a smooth free hand curve and has the shape of an elongated S.

MORE THAN OGIVE:

In more than ogive, the more than cumulative frequencies are plotted against the lower-class boundaries of the respective classes. Then the points are joined by a smooth free hand curve and have the appearance of an elongated S, upside down.


Less than cf
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
60


More than cf
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50

UNIT – V

Independent 't' test:

The independent t-test, also called the two-sample t-test, independent-samples t-test or student's t-test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups.

Dependent 't' test:

The dependent t-test (also called the paired t-test or paired-samples t-test) compares the means of two related groups to determine whether there is a statistically significant difference between these means. To calculate dependent t-ratio, it needs one dependent variable that is measured on an interval or ratio scale and needs one categorical variable that has only two related groups.

Chi square test:

Chi square χ^2 test is applied in statistics to test the goodness of fit to verify the distribution of observed data with assumed theoretical distribution. Therefore, it is measured to study the divergence of actual (observed, f_o) and expected frequencies (f_e). If there is no difference between the actual and expected frequencies, chi square χ^2 is zero.

$$x^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

Merits and demerits:

1. Test is based on events or frequencies, whereas in theoretical distribution, the test is based onmean and standard deviation.
2. To draw inferences, this test is applied specially testing the hypothesis but not useful forestimation.
3. The test can be used between the entire set of observed and expected frequencies.
4. For every increase in the number of degrees of freedom, a new chi square distribution isformed.
5. It is general purpose test and as such is highly useful in research.

Meaning of correlation:

Correlation analysis attempts to determine the degree of relationship between variables. Correlation means that between two series or groups of data there exists some casual connection. "The relationship of quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation."

– Croxton and Cowden

Always, correlation lies between **+1** and **-1**.

Types of correlation:

1. Positive and negative correlation
2. Simple and multiple correlation
3. Partial and total correlation
4. Linear and non-linear correlation.

Positive and negative correlation depends upon the direction of change of the variables. If an increase in the value of one variable is accompanied by an increase in the value of the other variable; or decrease in the value of one variable is accompanied by a decrease in the value of the other variable, then the correlation is called as **positive correlation**. For example, height and weight, power and strength, etc.

If two variables tend to move in opposite directions so that an increase in the value of one variable is accompanied by a decrease in the value of the other variable; decrease in the value of one variable is accompanied by a increase in the value of the other variable, then the correlation is called as **negative correlation**.

If we study only two variables, the relationship is described as **simple correlation**.

If we study more than two variables simultaneously, then the relationship is describes as multiple correlation. The study of two variables excluding some other variables is called **partial correlation**. In **total correlation**, all the variables are taken into account.

If the ratio of change between two variables is uniform, then there will be **linear correlation** between them. For example,

Batsman 1	70	80	90
Batsman 2	30	50	70

In a **curvi-linear or non-linear correlation**, the amount of change in one variable does not bear a constant ratio of the amount of change in the other variable(s).

Coefficient of correlation:

Correlation is a statistical technique used for analyzing the behavior of two or more variables. Its analysis deals with the association between two or more variables. Statistical measures of correlation are proof only of co-variation between series, not of functional or causal

relationship. Co-efficient of correlation lies between +1 and -1.

When r = +1, it means that there is perfect positive relationship between the variables.When r = -1, it means that there is perfect negative relationship between the variables.When r = 0, it means that there is no relationship between the variables.

(i) Karl Pearson's product moment coefficient of correlation:

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

where, $\overline{x}$, $\overline{y}$, mean of x, y variable.

(ii) Spearman's rank order coefficient of correlation:

Om 1904, Charles Edward Spearman, a British psychologist found out the method of ascertaining the coefficient of correlation by ranks.

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2-1\right)}$$

Concept of ANOVA and ANCOVA:

The ANOVA. An "Analysis of Variance" (ANOVA) tests three or more groups for mean differences based on a continuous (i.e., scale or interval) response variable (dependent variable).

In ANCOVA the letter "C", stands for 'covariance'. "Analysis of Covariance"(ANCOVA) has a single continuous response variable. ANCOVA compares a response variable by both a factor and a continuous independent variable (e.g., comparing pushups (shoulder strength) score by both circuit training and core training). The term for the continuous independent variable (IV) used in ANCOVA is "covariate".