Introduction to Statistical Methods in Data Science
Basics of Data Sciences

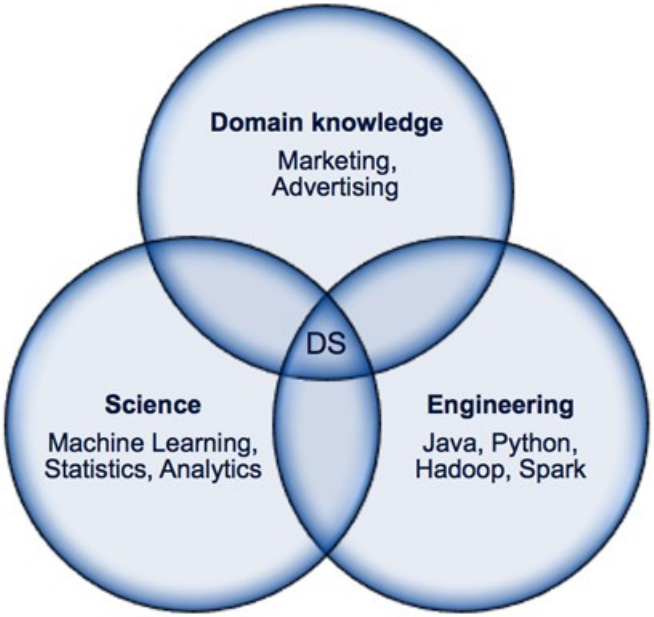
Presented By

Dr. Arvind Kumar Sinha
Associate Professor
Department of Mathematics
NIT Raipur

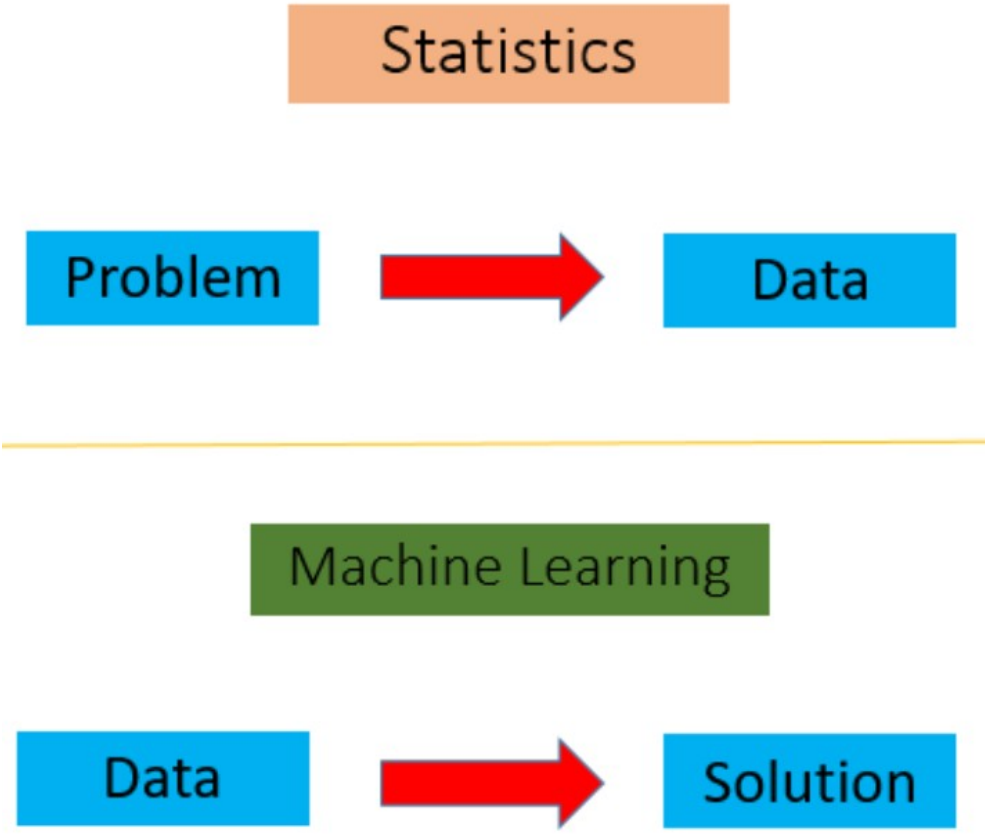
Data Sciences

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to mainly data mining, machine learning and big data.

Data Sciences



Role of Statistics and Machine Learning

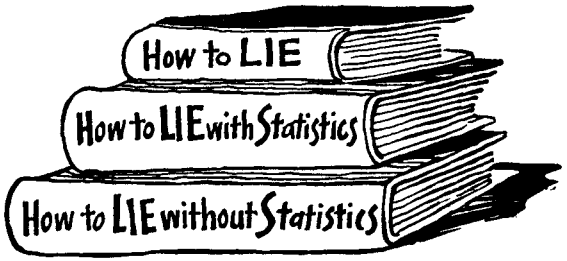


Statistical Methods

- Statistical method is used to search and summarize historical data in order to identify patterns or meaning.
- Descriptive analytics, which use Data Aggregation and Data Mining.
- So Statistics is the basic or fundamental concept of data sciences.

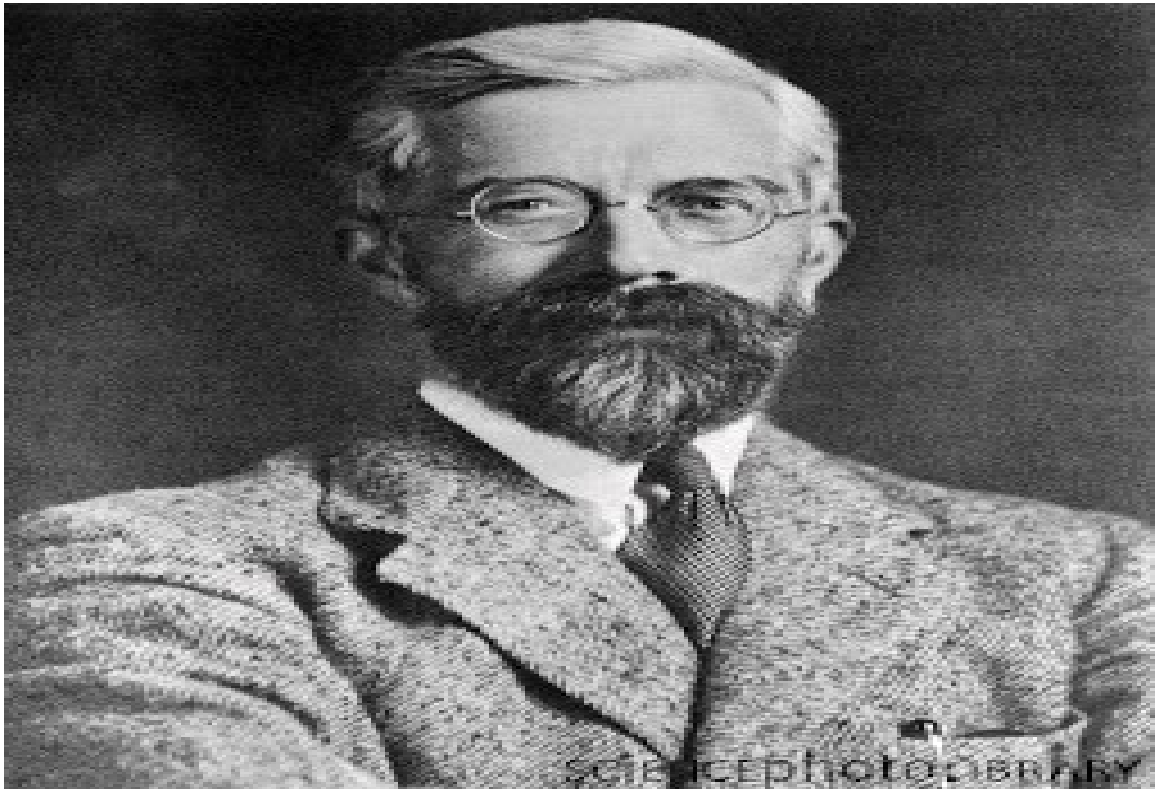
Statistics

"There are three kinds of lies: lies, damned lies, and statistics" (B.Disraeli - twice served as Prime Minister of the United Kingdom)



Great and Famous Statisticians & their contributions

R. A. Fisher



R. A. Fisher

- Sir Ronald Aylmer Fisher, FRS (1890 - 1962) was an **English statistician**, evolutionary biologist, and geneticist. Richard Dawkins described him as "The greatest of Darwin's successors", and the historian of statistics Anders Hald said "Fisher was a genius who almost single-handedly created the foundations for modern statistical science. His contributions to **experimental design, analysis of variance**, and likelihood based methods have led some to call him "**The Father of Statistics**".

Karl Pearson

Karl Pearson

- Karl Pearson (1857 - 1936) was a major contributor to the early development of statistics, and founder of the world's first university statistics department at University College **London i**n 1911. He was also an ardent and controversial proponent of eugenics. His most famous contribution is the **Pearson's chi-square test**. He was an influential English mathematician who has been credited for establishing the discipline of mathematical statistics.

G. Cox

G. Cox

- Gertrude Mary Cox (1900 - 1978) was an influential **American statistician** and **founder of the department of Experimental Statistics** at North Carolina State University. She was later appointed director of both the Institute of Statistics of the Consolidated University of North Carolina and the Statistics Research Division of North Carolina State University. Her most important and influential research dealt with experimental design; she wrote an important book on the subject with W. G. Cochran. In 1949 Cox became the first female elected into the International Statistical Institute and in 1956 she was president of the American Statistical Association. From 1931 to 1933 Cox undertook graduate studies in statistics at the University of California at Berkeley, then returned to Iowa State College as assistant in the Statistical Laboratory. Here she worked on the **design of experiments**

F. Yates

F. Yates

- Frank Yates (1902 - 1994) was one of the pioneers of 20th century statistics from **United Kingdom.** He worked on the design of experiments, including contributions to the theory of analysis of variance and originating Yates' algorithm and the balanced incomplete block design. At Rothamsted he worked on the design of experiments, including contributions to the theory of **analysis of variance** and originating **Yates's algorithm** and the balanced incomplete block design.

Thomas Bayes

Thomas Bayes

- Thomas Bayes was the son of **London** Presbyterian minister Joshua Bayes,[5] and was possibly born in Hertfordshire.[6] He came from a prominent nonconformist family from Sheffield. In 1719, he enrolled at the University of Edinburgh to study logic and theology. On his return around 1722, he assisted his father at the latter's chapel in London before moving to Tunbridge Wells, Kent, around 1734. There he was minister of the Mount Sion chapel, until 1752.

Florence Nightingale
The Lady with the Lamp

Florence Nightingale

- **Florence Nightingale**, OM, RRC, DStJ (/ˈnaɪtɪŋɡeɪl/; 12 May 1820 – 13 August 1910) was an **English social reformer** and statistician, and the **founder of modern nursing**.
- Florence Nightingale exhibited a gift for mathematics from an early age and excelled in the subject under the tutelage of her father.[55] Later, Nightingale became a **pioneer in the visual presentation of information and statistical graphics**.[56] She used methods such as the pie chart, which had first been developed by William Playfair in 1801. While taken for granted now, it was at the time a relatively **novel method of presenting data.**

George W. Snedecor

George W. Snedecor

- **George Waddel Snedecor** (October 20, 1881 – February 15, 1974) was an **American** **mathematician** **and** **statistician**. He contributed to the foundations of analysis of variance, data analysis, experimental design, and statistical methodology. Snedecor's *F*-distribution and the George W. Snedecor Award of the American Statistical Association are named after him.
- The "*F*" of Snedecor's *F* distribution is named in honor of Sir Ronald Fisher.

and many more……

Why study statistics?

1. Data are everywhere
2. Statistical techniques are used to make many decisions that affect our lives
3. No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions efectively

Applications of statistical concepts in the real life world

- Finance – correlation and regression, index numbers, time series analysis
- Marketing – hypothesis testing, chi-square tests, nonparametric statistics
- Personel – hypothesis testing, chi-square tests, nonparametric tests
- Operating management – hypothesis testing, estimation, analysis of variance, time series analysis

What is Statistics?

Statistics: The science of collectiong, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

Statistical analysis is used to manipulate, summarize, and investigate data, so useful decision-making information (results) can be done.

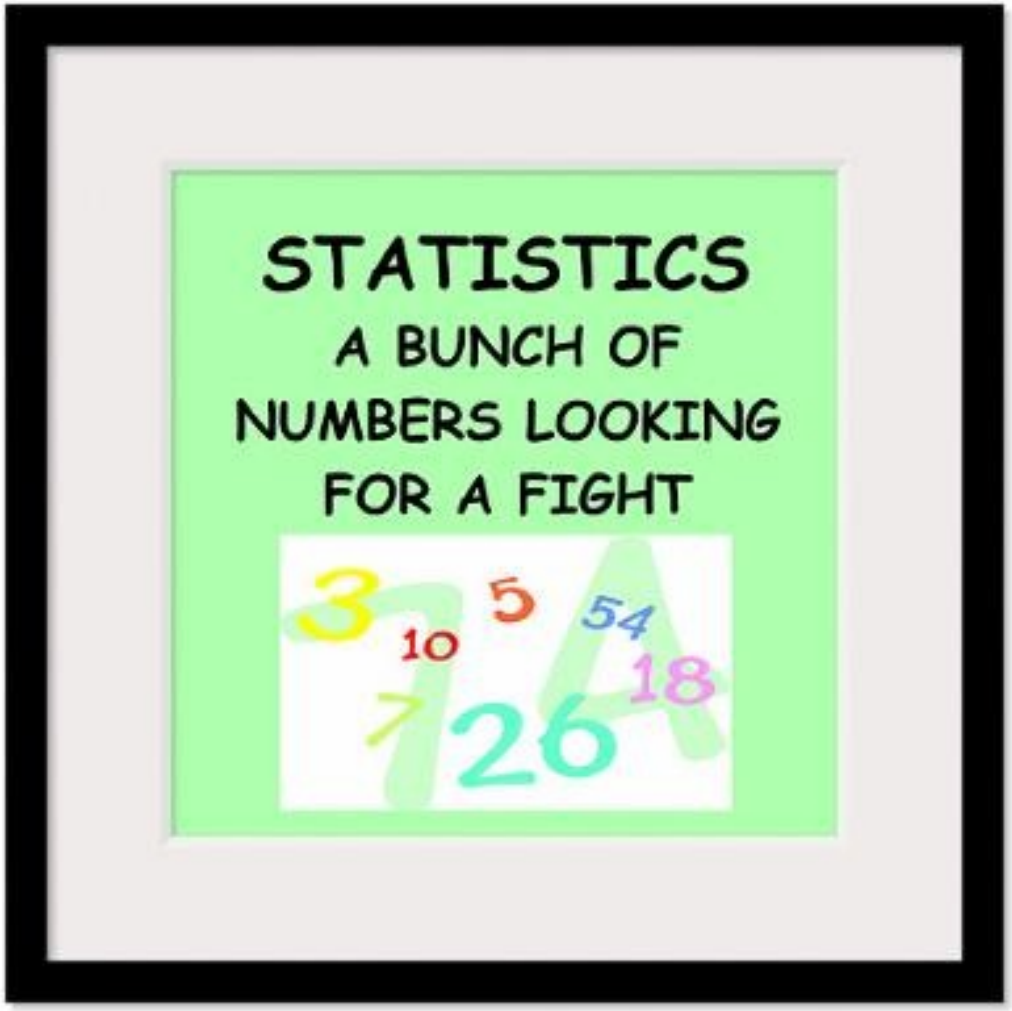
Statistics

Two areas of statistics:

Descriptive Statistics: collection, presentation, and description of sample data.

Inferential Statistics: making decisions and drawing conclusions about populations.

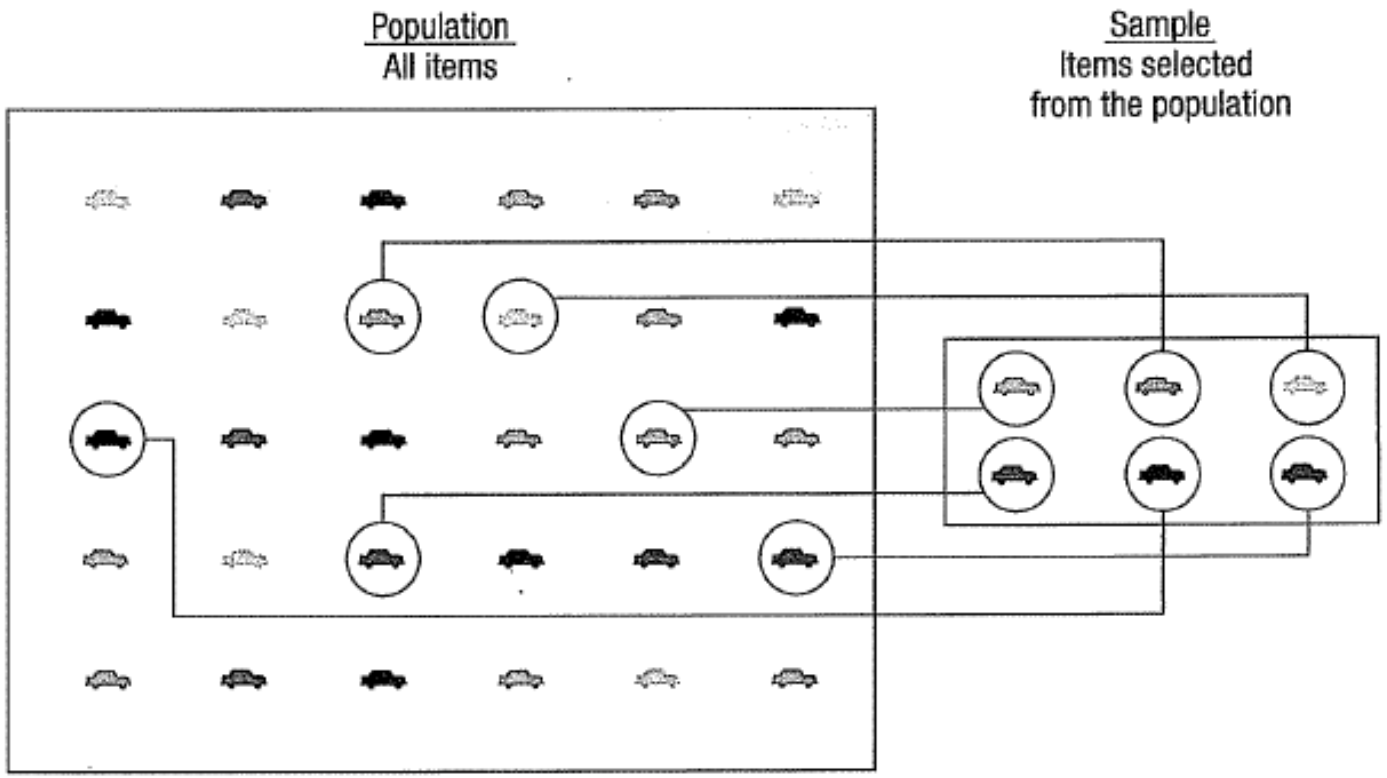
?


STATISTICS
A BUNCH OF
NUMBERS LOOKING
FOR A FIGHT
3
10
5
54
18
7
26

Types of statistics

- **Descriptive statistics** – Methods of organizing, summarizing, and presenting data in an informative way
- **Inferential statistics** – The methods used to determine something about a population on the basis of a sample
  - Population – The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
  - Sample – A portion, or part, of the population of interest

Population and Sample



Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kg

Inference is the process of drawing conclusions or making decisions about a population based on sample results

Sampling

A sample should have the same characteristics as the population (it is representing).

Sampling can be:

- **With replacement**: a member of the population may be chosen more than once (**picking the candy from the bowl**)
- **Without replacement**: a member of the population may be chosen only once (**lottery ticket**)

Sampling methods

Sampling methods can be:

- **Random** (each member of the population has an equal chance of being selected**)**
- **Non-random**

Factors related to actual process of sampling causes **sampling errors**. For example, the sample may **not be large enough** or representative of the whole population.

Factors not related to the sampling process cause **non-sampling errors**. A **defective counting device** can cause a non-sampling error.

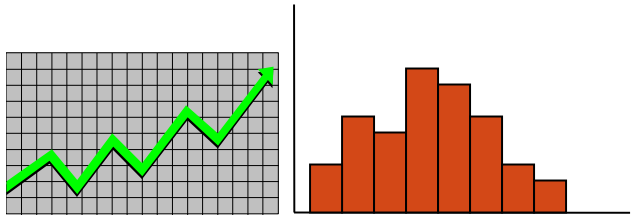
Random sampling methods

- **Simple random sample** (each sample of the same size has an equal chance of being selected)
- **Stratified sample** (divide the population into groups called strata and then take a sample from each stratum)
- **Cluster sample** (divide the population into strata and then randomly select some of the strata. All the members from these strata are in the cluster sample.)
- **Systematic sample** (randomly select a starting point and take every n-th piece of data from a listing of the population)

Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs



- Summarize data
  - e.g., Sample mean = $\frac{\sum X_i}{n}$

Statistical data

- The collection of data that are relevant to the problem being studied is commonly the most difficult, expensive, and time-consuming part of the **entire research project**.
- Statistical data are usually obtained by counting or measuring items.
  - **Primary data** are collected specifically for the analysis desired
  - **Secondary data** have already been compiled and are available for statistical analysis
- A **variable** is an item of interest that can take on many different numerical values.
- A **constant** has a fixed numerical value.

Data

Statistical data are usually obtained by counting or measuring items. Most data can be put into the following categories:

- **Qualitative -** data are measurements that each fail into one of several categories. (**hair color, ethnic groups and other attributes of the population**)
- **Quantitative** - data are observations that are measured on a numerical scale (distance traveled to college, **number of children in a family,** etc.)

Qualitative data

Qualitative data are generally described by words or letters. They are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data. For example, it does not make sense to find an average hair color or blood type.

Qualitative data can be separated into two subgroups:

- **dichotomic** (if it takes the form of a word with two options (**gender** - male or female)
- **polynomic** (if it takes the form of a word with more than two options (**education** - primary school, secondary school and university).

Quantitative data

Quantitative data are always numbers and are the

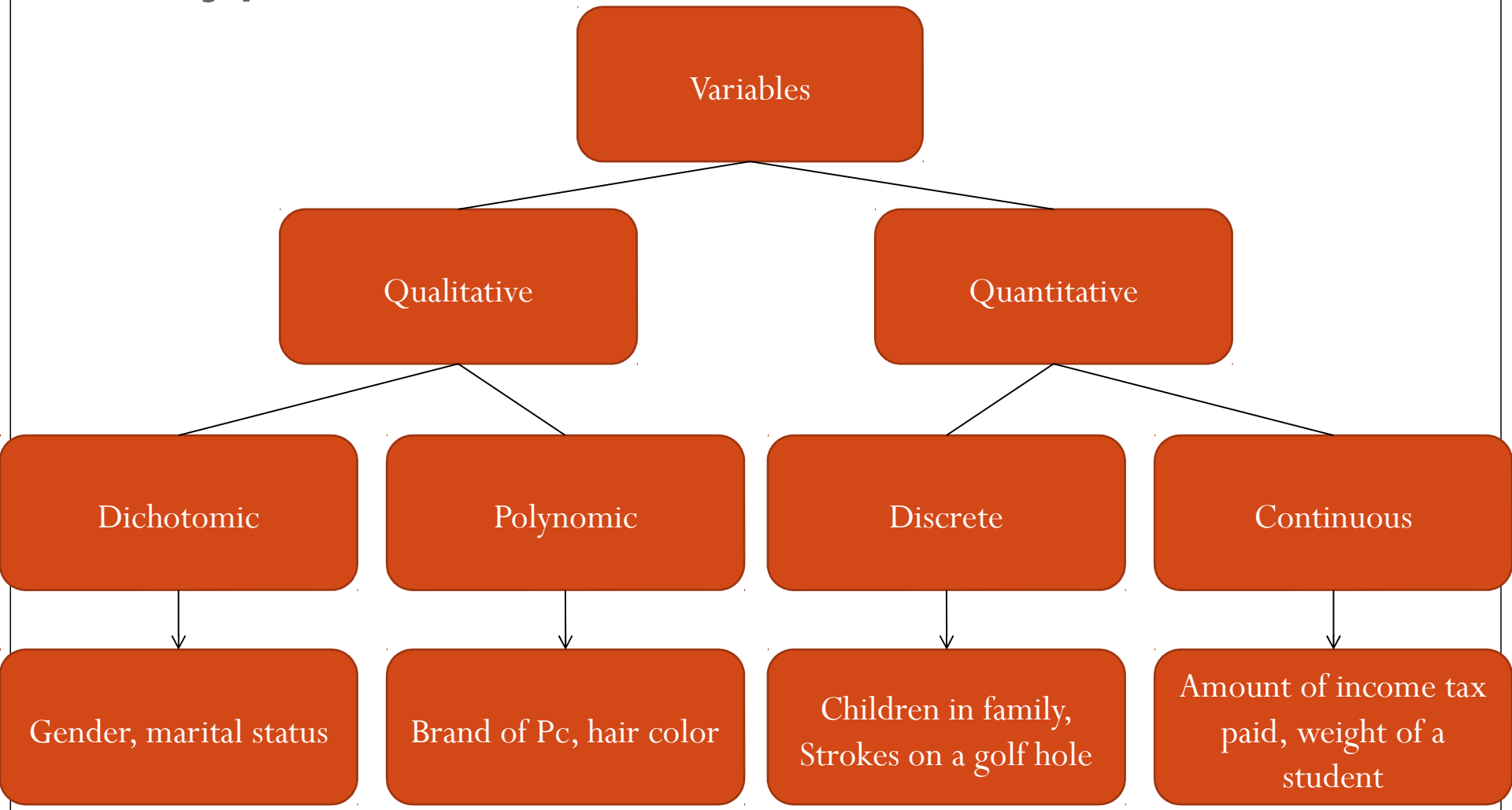
result of counting or measuring attributes of a population.

Quantitative data can be separated into two

subgroups:

- **discrete** (if it is the result of *counting* (the number of students of a given ethnic group in a class, **the number of books on a shelf**, …)
- **continuous** (if it is the result of *measuring* (**distance traveled**, weight of luggage, …)

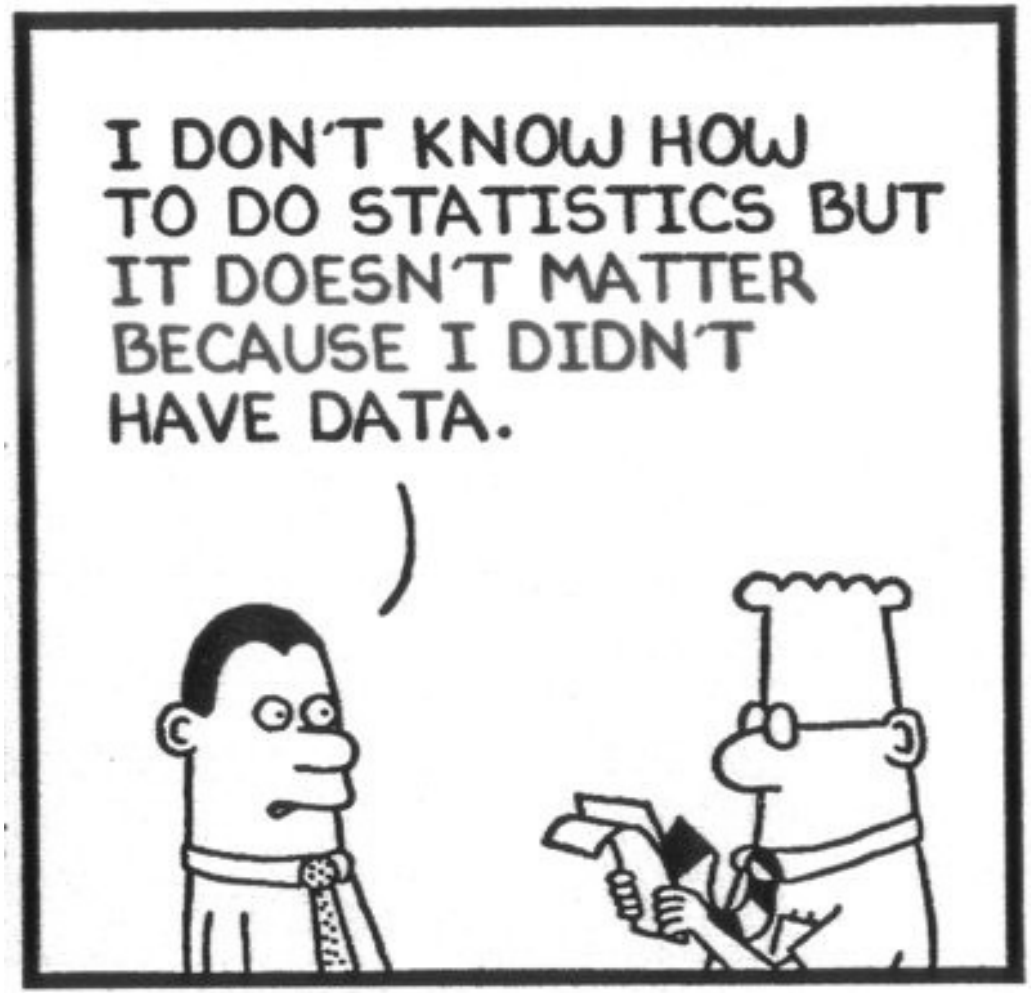
Types of variables



Numerical scale of measurement:

- **Nominal** – consist of categories in each of which the number of respective observations is recorded. The categories are in no logical order and have no particular relationship. The categories are said to be ***mutually exclusive*** since an individual, **object,** or measurement can be included in only one of them.
- **Ordinal** – **contain more information**. Consists of distinct categories in which order is implied. Values in one category are larger or smaller than values in other categories (e.g. **rating-excelent, good, fair, poor**)
- **Interval** – is a set of numerical measurements in which the distance between numbers is of a known, constant size.
- **Ratio** – consists of numerical measurements where the distance between numbers is of a known, constant size, in addition, there is a nonarbitrary zero point.

?


I DON'T KNOW HOW TO DO STATISTICS BUT IT DOESN'T MATTER BECAUSE I DIDN'T HAVE DATA.

What to teach, How to teach -Amir Khan:

https://www.youtube.com/watch?v=mPdf4tNPT6s

Data presentation

Frequency distributions – numerical presentation of quantitative data

- **Frequency distribution** – shows the frequency, or number of occurences, in each of several categories. Frequency distributions are used to summarize large volumes of data values.
- When the raw data are measured on a qunatitative scale, either interval or ration, categories or classes must be designed for the data values before a frequency distribution can be formulated.

Frequency table

- **Absolute frequency "n_i"** (Data Tab→Data Analysis→Histogram)

- **Relative frequency "f_i"**

- **Cumulative frequency "N_i" - Cumulative frequency distribution** shows the total number of occurrences that lie above or below certain key values.

Charts and graphs

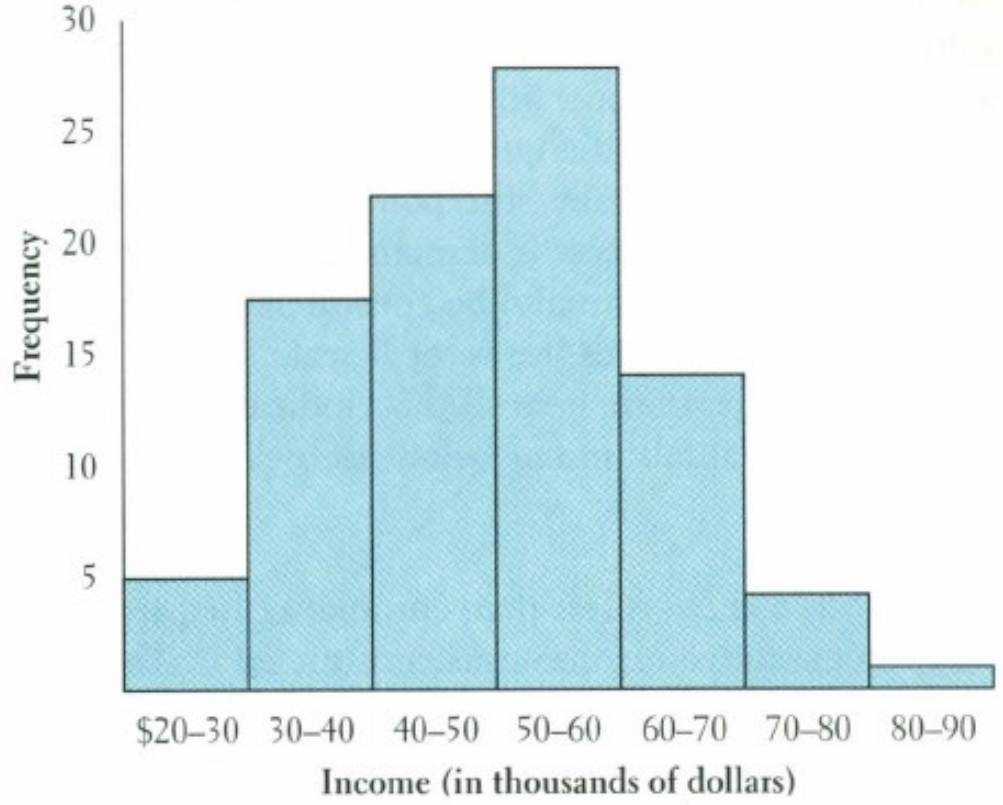
- Frequency distributions are good ways to present the essential aspects of data collections in concise and understable terms

- Pictures are always more effective in displaying large data collections

Histogram

- Frequently used to graphically present interval and ratio data
- Is often used for interval and ratio data
- The adjacent bars indicate that a numerical range is being summarized by indicating the frequencies in arbitrarily chosen classes

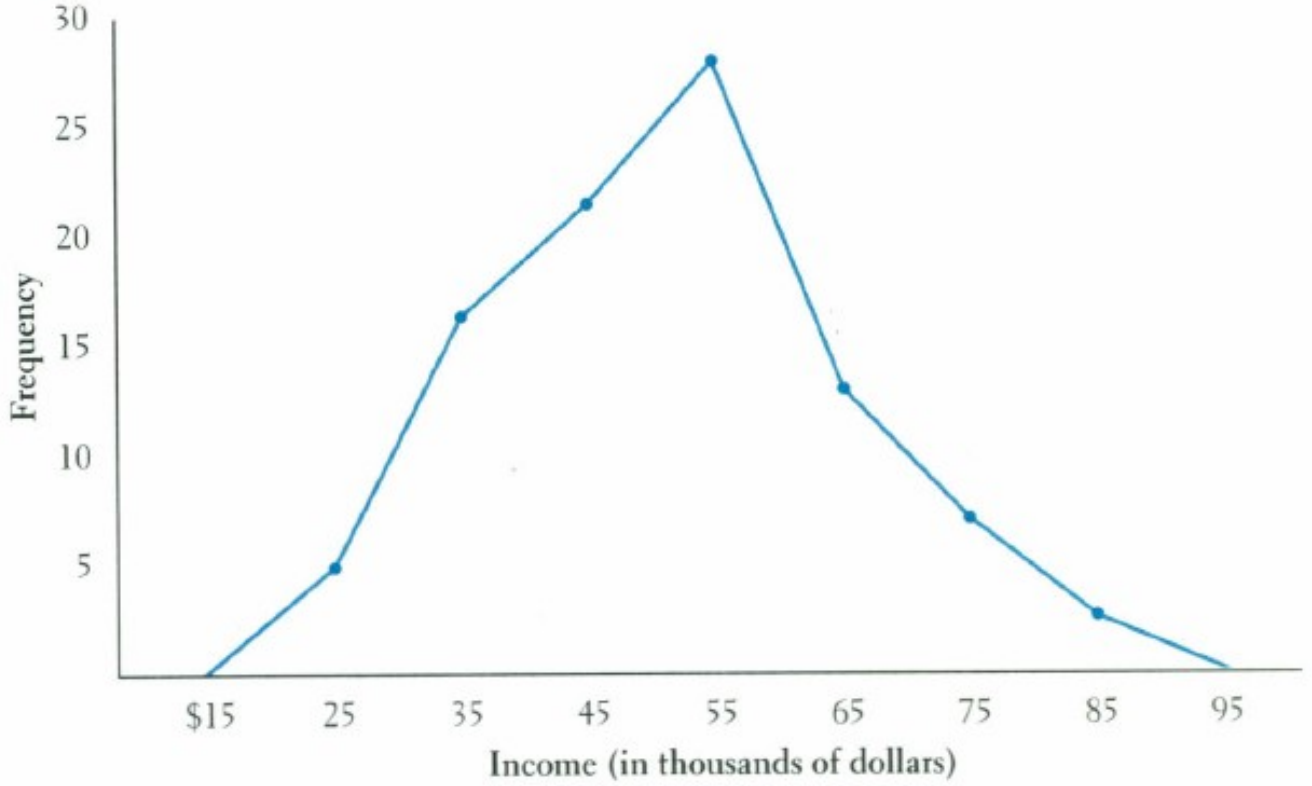
FIGURE 3.7 Histogram—Executive Incomes for the Sunrunner Corporation



Frequency polygon

- Another common method for graphically presenting interval and ratio data
- To construct a frequency polygon mark the frequencies on the vertical axis and the values of the variable being measured on the horizontal axis, as with the histogram.
- If the purpose of presenting is comparation with other distributions, the frequency polygon provides a good summary of the data

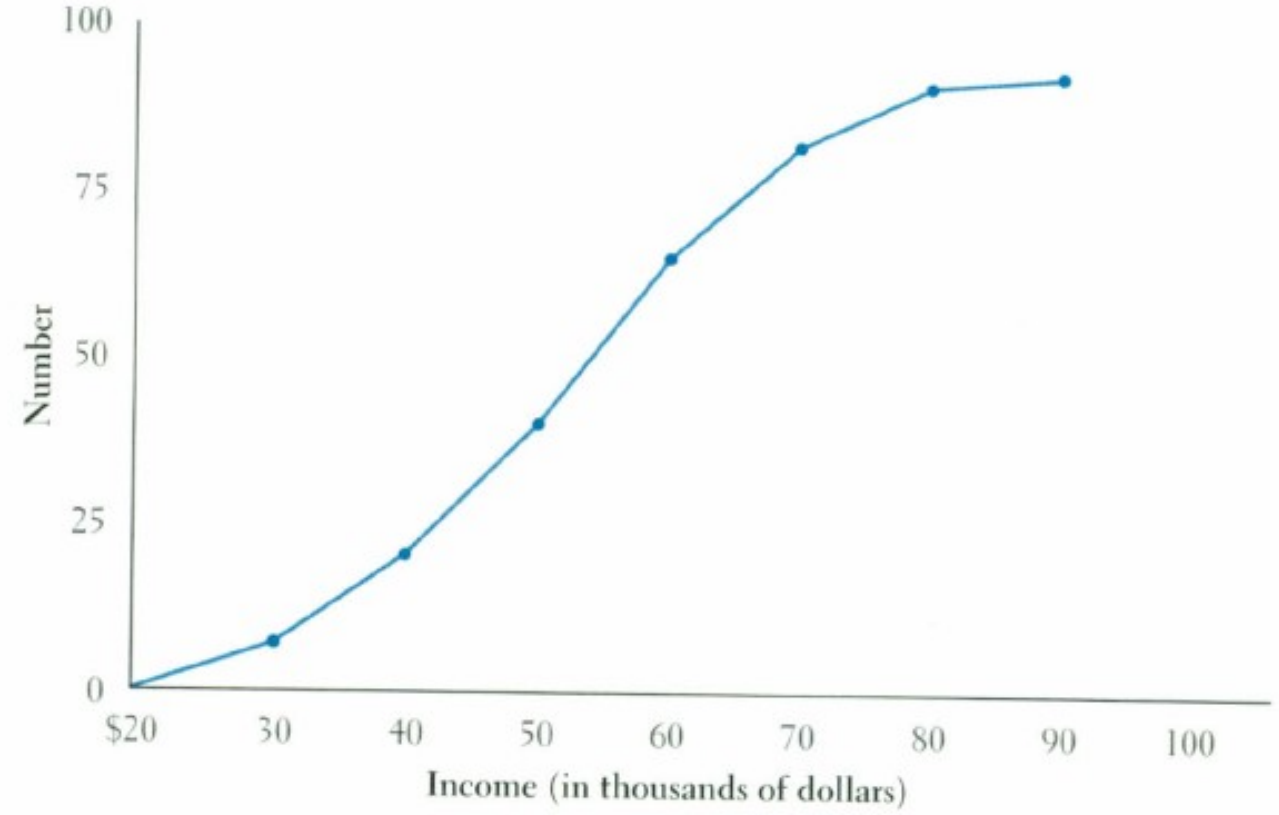
FIGURE 3.8 Frequency Polygon—Executive Incomes



Ogive

- A graph of a cumulative frequency distribution
- Ogive is used when one wants to determine how many observations lie above or below a certain value in a distribution.
- First **cumulative frequency distribution is constructed**
- Cumulative frequencies are plotted at the upper class limit of each category

FIGURE 3.9 Ogive—Executive Incomes (frequencies)



Histogram

.11. Histogram

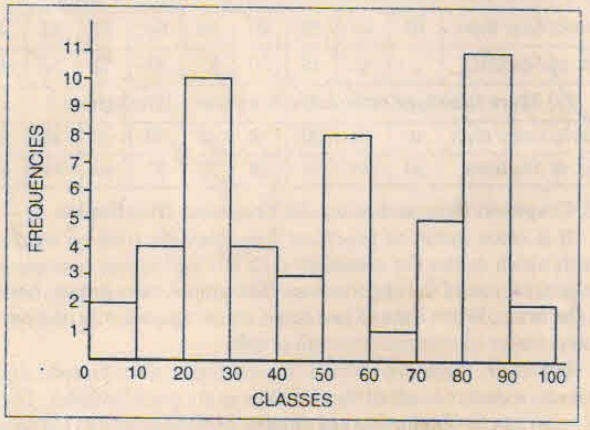
In drawing the histogram of a given grouped frequency distribution ıark off along a horizontal base line all the class-intervals on a suitable :ale; one centimeter to class-interval would be suitable. With the class-in-ervals as bases, we draw rectangles with the areas proportional to the requencies of the respective class-intervals. For equal class-intervals, the eights of the rectangles will be proportional to the frequencies. If the lass-intervals are not equal, the heights of the rectangles will be propor-onal to the ratios of the frequencies to the width of the corresponding lasses. A diagram with all these rectangles is a **Histogram.**

Example 1. *Draw a histogram for the following distribution—*

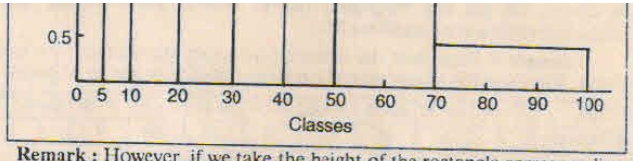
Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	2	4	10	4	3	8	1	5	11	2

Solution : Taking class-intervals on X-axis and frequencies on Y-axis, we have ıe required histogram.

Scale : 1 cm = 10 marks; 1 cm = 1 frequency



Frequency Polygon



Remark : However, if we take the height of the rectangle corresponding the class 0 — 5 as 5, then the height of other class-intervals may be taken as 10, $\frac{15}{2}$, $\frac{20}{2}$, $\frac{25}{2}$, $\frac{40}{2}$, $\frac{30}{4}$, $\frac{15}{6}$ respectively.

2.12. Frequency Polygon

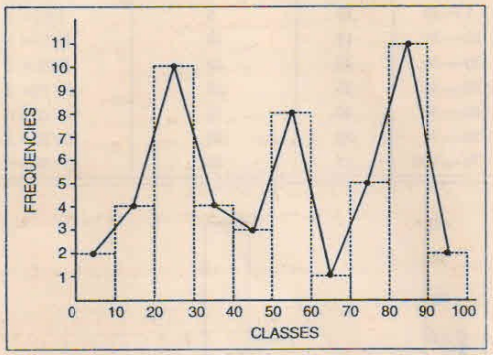
If the various points are obtained by plotting the central values of the class-intervals as x-coordinates and the respective frequencies as the y-coordinates, and these points are joined by straight lines taken in order, they form a polygon called **frequency polygon.**

In a frequency polygon the variables or individuals of each class are assumed to be concentrated at the mid-point of the class-interval. ●

Example 3. *Draw a frequency polygon for the data given below—*

Class	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100
Frequency	2	4	10	4	3	8	1	5	11	2

Frequency Polygon & Frequency Curve



Here in this diagram dotted is the **histogram** and a polygon with lines as sides is the **frequency polygon.**

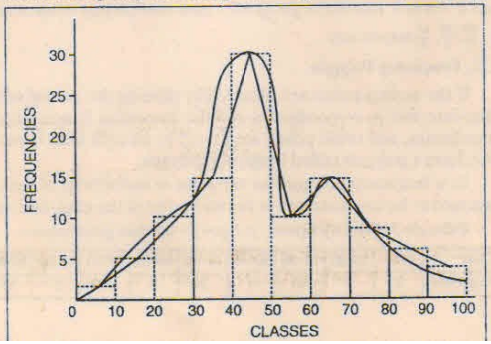
2.13. Frequency Curve

If through the vertices of a frequency polygon a smooth freehand curve is drawn we get the **frequency curve.** This is done usually when the class-intervals are of small widths.

Example 4. Represent the following frequency distribution by a frequency curve. Show also the histogram and frequency polygon on the same graph—

Class	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100
Frequency	2	6	10	17	30	10	14	7	4	1

Solution : Frequency curve along with histogram and frequency polygon



Cumulative Frequency Curve or Ogive

2.14. Cumulative Frequency Curve or the Ogive

If from a cumulative frequency table the upper limits of the class taken as x-coordinates and the cumulative frequencies as the y-coordinates and the points are plotted, then these points when joined by straight lines, we obtain **less than type cumulative frequency polygon.**

If more than cumulative frequency is plotted against the corresponding lower limits of each class and the points plotted are joined by straight lines, we obtain **more than type cumulative frequency polygon.**

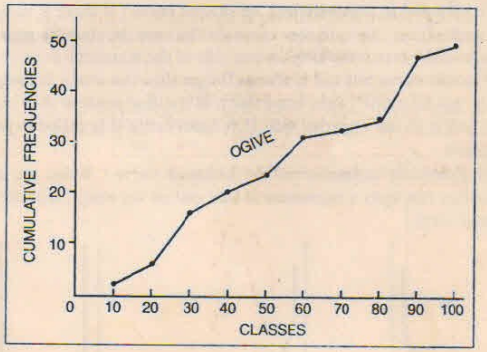
However, when the points plotted are joined by a free hand smooth curve, we obtain cumulative frequency curve.

The cumulative frequency polygon (or curve) is often called an **Ogive.**

Example 5. *Draw a less than type cumulative frequency polygon for the data given below—*

Class	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100
Frequency	2	4	10	4	3	8	1	5	11	2

Solution : Taking upper limits on the X-axis and less than type cumulative frequencies on the Y-axis, the required cumulative frequency polygon is as follows—



2.15. Types of Frequency Curves

The following are some important types of frequency curves which are generally obtained in the graphical representations of frequency distributions –

Types of Frequency Curves

(1) *Symmetrical curve* or *bell shaped curve.*

(2) *Moderately asymmetrical* or *skewed curve.*

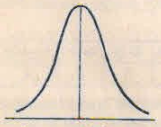
(3) *Extremely asymmetrical* or *J-shaped curve* or *reverse J-shaped.*

(4) *U-shaped curve.*

(5) *A bimodal frequency curve.*

(6) *A multimodal frequency curve.*

(1) Symmetrical curve or bell shaped curve : If a curve can be folded symmetrically along a vertical line, it is called a symmetrical curve. In this type the class frequencies decrease to zero symmetrically on either side of a central maximum *i.e.* the observations equidistant from the central maximum have the same frequency.



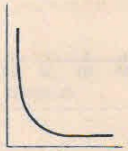
Symmetrical curve

Moderately asymmetrical curve

(2) Moderately asymmetrical or skewed curve : If there is no symmetry in the curve, it is said *skew curve.* In this case the class frequencies decrease with greater rapidity on one side of the maximum than on the other. In this curve one tail is always longer than the other. If long tail is to the positive side (right hand side), it is called *positive skew curve*, if long tail is to the negative side (left hand side), it is called *negative skew curve.*

(3) Extremely asymmetrical or J-shaped curve : When the class frequencies run upto a maximum at one end of the range, they form a J-shaped curve.

J-shaped curve



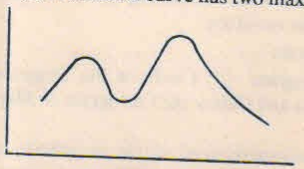
Reversed J-shaped curve

U-shaped curve

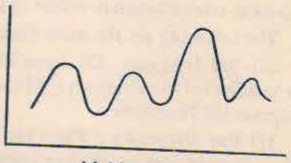
(4) U-shaped curve : In this type of curve the maximum frequency is at the ends of the range and a maximum towards the centre.

Types of Frequency curve & Characteristics of Frequency Distribution

(5) A bimodal curve has two maxima.



Bimodal curve



Multimodal curve

(6) A multimodal curve has more than two maxima.

2.16. Characteristics of Frequency Distribution

To compare the distributions of the same type we study their four main characteristics which describe their nature in a general way. These characteristics are as follows –

(1) Central Tendency : In most of the frequency distributions there is found a point around which largest number of observations tend to cluster. This point is called the *central point* of the frequency distribution. The tendency of the observations to concentrate around a central point is known as *central tendency*.

(2) Dispersion or Variability : The word dispersion used in statistics for variability is defined in two senses –

(i) Dispersion implies that within a group the observations are not uniform in their magnitude *i.e.* they differ in magnitude. Such scatteredness of the values is called *dispersion* or *variation* or *variability*.

(ii) The scatteredness of the deviations (deviation means difference between an observation and any fixed value) is referred to as *dispersion*.

(3) Skewness : Skewness means the lack of symmetry. It is of two types, positive skewness and negative skewness. It determines the nature and extent of the concentration of the observations towards the higher or lower values of the variable.

(4) Kurtosis : The kurtosis of a frequency distribution refers to the shape of the top of its curve *i.e.* Kurtosis is the degree of peakedness in a frequency distribution or the relative flatness of the top of the frequency curve. There are three forms of distributions based on kurtosis –

(i) Mesokurtic (ii) Leptokurtic (iii) Platykurtic

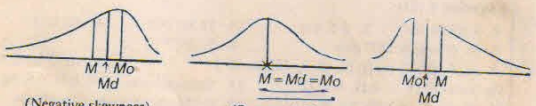
Skewness

5.1. Skewness

Skewness denotes the opposite of symmetry. It is lack of symmetry. As applied to frequency distribution it indicates that the distribution of items on it is not symmetrical. A frequency distribution is said to be symmetrical when the values equidistant from the mean have equal frequencies.

In a symmetrical series the mode, the median, and the arithmetic average are identical. Therefore, skewness or lack of symmetry in a series is shown when these three averages do not coincide. It is further shown when the sum of the positive deviations from the median in a series is not equal to the negative deviations.

However, several measures have been devised to measure the departure of a frequency distribution from symmetry. These measures show which distributions have been pulled away from the ideal symmetrical curve.



(Negative skewness) (Symmetrical) (Positive Skewness)

Skewness can be positive as well as negative. If the mean is greater than the mode or the median, the skewness is positive. If it is less, skewness is negative. In other words it is positive if the large tail of the distribution lies towards the higher values of the variable (right) and negative in the contrary.

In other words, if Mode < Median < Mean, then skewness is positive and if Mean < Median < Mode, the skewness is negative.

5.2. Measures of Skewness

The first coefficient of skewness is defined by Bowley as –

$$\text{Coeff. of Skewness} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$ (Agra 1983)

Skewness

Coeff. Of Skewness = (Mean – Mode) / SD

Kurtosis

frequency distribution).

5.14. Kurtosis

The characteristics related with the nature of the concentration of the items in the central part of a frequency distribution is called a *Kurtosis*.

In other words,

Kurtosis is the degree of peakedness (or flatness) in a curve of the frequency distribution. In fact Kurtosis is an indication for the peakedness of a single humped frequency curve. β_2, γ_2 measures of Kurtosis indicate the degree to which a curve of the frequency distribution is peaked or flat topped.

Karl Pearson in 1905 introduced the three terms –

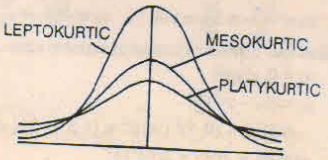
(i) Mesokurtic, (ii) Leptokurtic, (iii) Platykurtic.

(i) A frequency curve which is not very peaked or very flat topped is called **Mesokurtic** or **normal curve**. For such types of curves, $\beta_2 = 3$ and $\gamma_2 = 0$.

(ii) A frequency curve which is more peaked than the mesokurtic is called **Leptokurtic.** For such types of curves $\beta_2 > 3$ and $\gamma_2 > 0$.

(iii) A frequency curve for which flatness of the top is more than the mesokurtic is called **Platykurtic,** for such types of curves, $\beta_2 < 3$ and $\gamma_2 < 0$.

A comparative picture of these three types is as follows –

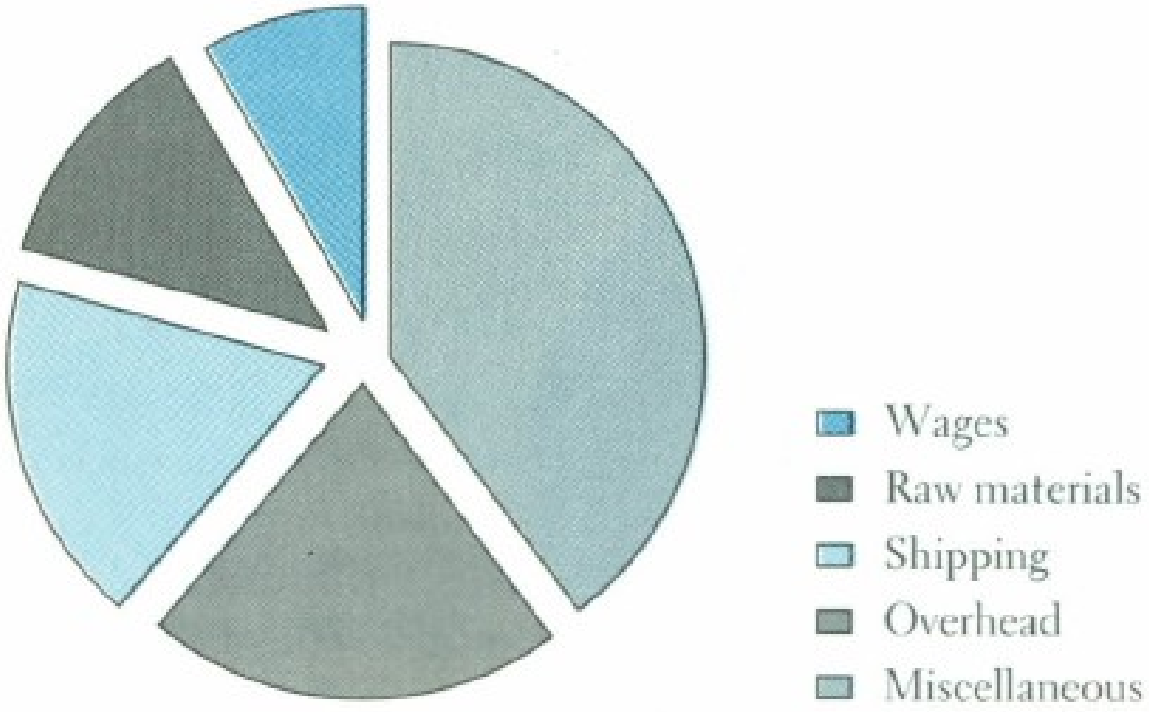


Pie Chart

- The pie chart is an effective way of displaying the percentage breakdown of data by category.
- Useful if the relative sizes of the data components are to be emphasized
- Pie charts also provide an effective way of presenting ratio- or interval-scaled data after they have been organized into categories

Pie Chart

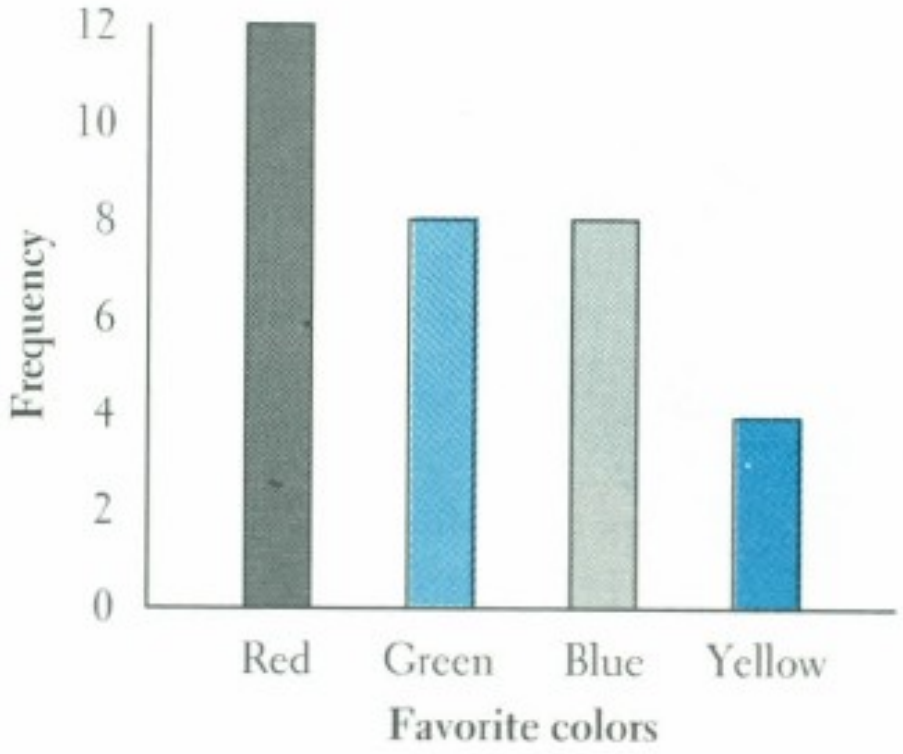
FIGURE 3.3 Pie Chart—Expenditures of Funds for Itrex Company



Bar chart

- Another common method for graphically presenting nominal and ordinal scaled data
- One bar is used to represent the frequency for each category
- The bars are usually positioned vertically with their bases located on the horizontal axis of the graph
- The bars are separated, and this is why such a graph is frequently used for nominal and ordinal data – the separation emphasize the plotting of frequencies for distinct categories

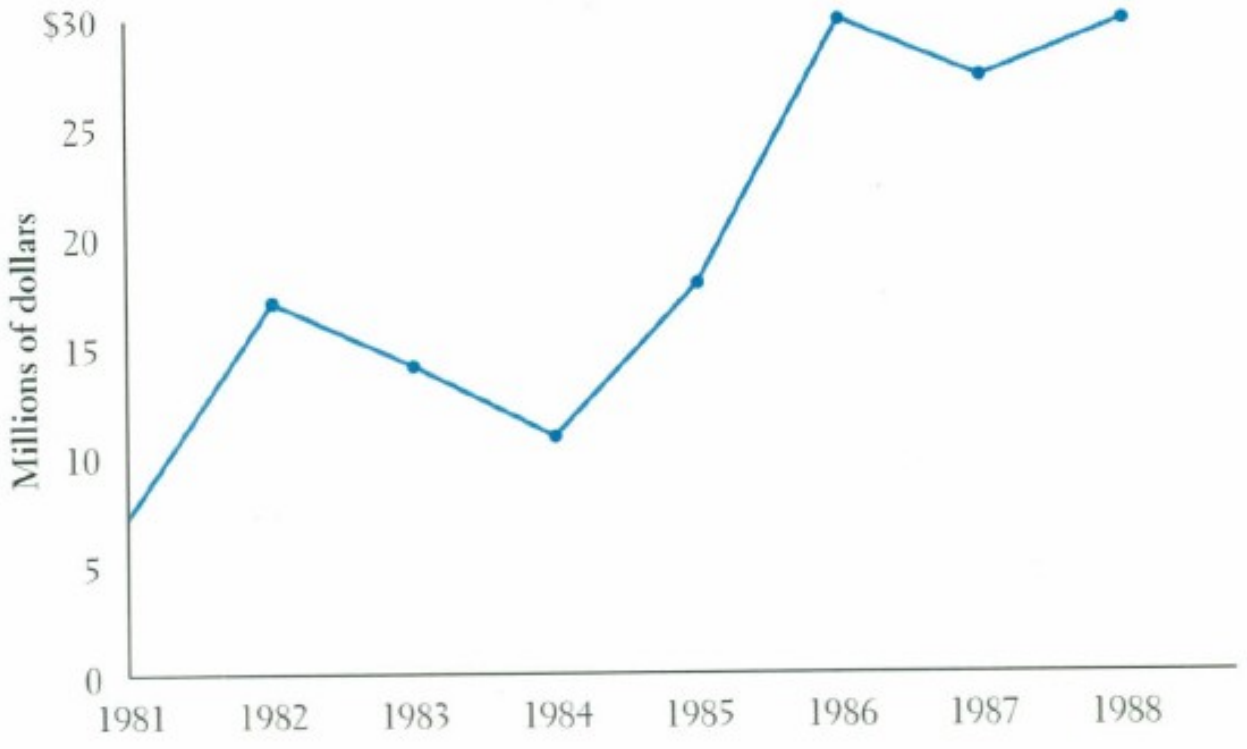
FIGURE 3.4 Bar Chart—Favorite Colors of 32 People



Time Series Graph

- The time series graph is a graph of data that have been measured over time.
- The **horizontal axis** of this graph represents **time periods** and the **vertical** axis shows the **numerical values** corresponding to these time periods

FIGURE 3.13 Time Series Graph—Corporate Revenue, Flightcraft Corp.



Basic Terms

Population: A collection, or set, of individuals or objects or events whose properties are to be analyzed.

Two kinds of populations: *finite* or *infinite*.

Sample: A subset of the population.

Variable: A characteristic about each individual element of a population or sample.

Basic Terms: continued

Data (singular): The value of the variable associated with one element of a population or sample. This value may be **a number**, a word, or **a symbol**.

Data (plural): The **set of values collected** for the variable from each of the elements belonging to the sample.

Experiment: A planned activity whose results yield a set of data.

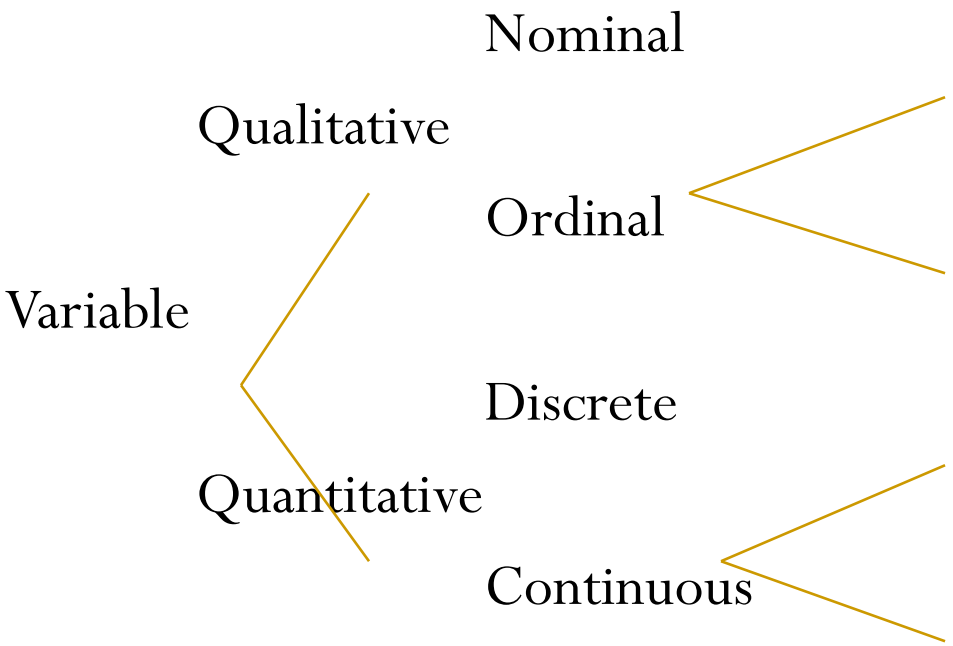
Parameter: A numerical value summarizing all the data of an entire **population**.

Statistic: A numerical value summarizing the **sample data**.

Example: Identify each of the following examples as **attribute** (**qualitative**) or **numerical** (**quantitative)** variables.

1. The good residence room for each student in a statistics class. (Attribute)
2. The amount of gasoline pumped by the next 10 customers at the local Unimart. (Numerical)
3. The amount of radon in the basement of each of 25 homes in a new development. (Numerical)
4. The color of the baseball cap worn by each of the student. (Attribute)
5. The length of time to complete a mathematics homework assignment. (Numerical)
6. The state in which each truck is registered when stopped and inspected at a weigh station. (Attribute)

Qualitative and quantitative variables may be further subdivided:



Nominal Variable: A qualitative variable that **categorizes** (or describes, or **names)** an element of a population.

Ordinal Variable: A qualitative variable that incorporates an **ordered position, or ranking.**

Discrete Variable: A quantitative variable that can assume a countable number of values. Intuitively, a discrete variable can assume values corresponding to isolated points along a line interval. That is, there is a gap between any two values.

Continuous Variable: A quantitative variable that can assume an uncountable number of values. Intuitively, a continuous variable can assume any value along a line interval, including every possible value between any two values.

Note:

1. In many cases, a discrete and continuous variable may be distinguished by determining whether the variables are related **to a count or a measurement**.
2. Discrete variables are usually associated with counting. If the variable cannot be further subdivided, it is a clue that you are probably dealing with a discrete variable.
3. Continuous variables are usually associated with measurements. The values of discrete variables are only limited by your ability to measure them.

Statistics:

https://www.youtube.com/watch?v=MXaJ7sa7q-8

Measure and Variability

- No matter what the response variable: there will always be **variability** in the data.
- One of the primary objectives of statistics: measuring and characterizing variability.
- Controlling (or reducing) variability in a manufacturing process: **statistical process control**.

Example: A supplier fills cans of soda marked 12 ounces. How much soda does each can really contain?

- It is very *unlikely* any one can contains exactly 12 ounces of soda.
- There is variability in any process.
- Some cans contain a little more than 12 ounces, and some cans contain a little less.
- On the average, there are 12 ounces in each can.
- The supplier hopes there is little variability in the process, that most cans contain *close* to 12 ounces of soda.

Data Collection

- First problem a statistician faces: how to obtain the data.
- It is important to obtain *good*, or *representative*, data.
- Inferences are made based on statistics obtained from the data.
- Inferences can only be as good as the data.

Biased Sampling Method: A sampling method that produces data which systematically differs from the sampled population. An **unbiased sampling method** is one that is not biased.

Sampling methods that often result in **biased samples**:

1. **Convenience sample**: sample selected from elements of a population that are easily accessible.

2. **Volunteer sample**: sample collected from those elements of the population which chose to contribute the needed information on their own initiative.

Process of data collection:

1. Define the objectives of the **survey or experiment**.

   Example: Estimate the average life of an electronic component.

2. Define the variable and population of interest.

   Example: Length of time for anesthesia to wear off after surgery.

3. Defining the data-collection and data-measuring schemes. This includes sampling procedures, sample size, and the data-measuring device (questionnaire, scale, ruler, etc.).

4. Determine the appropriate **descriptive** or **inferential data-analysis techniques.**

Methods used to collect data:

Experiment: The investigator controls or modifies the environment and observes the effect on the variable under study.

Survey: Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.

Census: A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.

Sampling Frame: A list of the elements belonging to the population from which the sample will be drawn.

Note: It is important that the sampling frame be representative of the population.

Sample Design: The process of selecting sample elements from the sampling frame.

Note: There are many different types of sample designs. Usually they all fit into two categories: judgment samples and probability samples.

Judgment Samples: Samples that are selected on the basis of being "typical."

Items are selected that are representative of the population. The validity of the results from a judgment sample reflects the soundness of the collector's judgment.

Probability Samples: Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.

Random Samples: A sample selected in such a way that every element in the population has a equal probability of being chosen. Equivalently, all samples of size *n* have an equal chance of being selected. Random samples are obtained either by sampling with replacement from a finite population or by sampling without replacement from an infinite population.

Example of random sample: An employer is interested in the time it takes each employee to commute to work each morning. A random sample of 35 employees will be selected and their commuting time will be recorded. There are 2712 employees.

Each employee is numbered: 0001, 0002, 0003, etc. up to 2712.

Using four-digit random numbers, a sample is identified: 1315, 0987, 1125, etc.

Systematic Sample: A sample in which **every *k*th item of the sampling frame is selected,** starting from the first element which is randomly selected from the first *k* elements.

Note: The systematic technique is easy to execute. However, it has some inherent dangers when the sampling frame is repetitive or cyclical in nature. In these situations the results may not approximate a simple random sample.

Stratified Random Sample: A sample obtained by stratifying the sampling frame and then selecting **a fixed number of items from each of the strata** by means of a simple random sampling technique.

Proportional Sample (or Quota Sample): A sample obtained by stratifying the sampling frame and then selecting a number of items in proportion to the size of the strata (or by quota) from each strata by means of a simple random sampling technique.

Cluster Sample: A sample obtained by stratifying the sampling frame and then selecting some or all of the items from some of, but not all, the strata.

Probability and Statistics

Probability: Properties of the population are assumed known. Answer questions about the sample based on these properties.

Statistics: Use information in the sample to draw a conclusion about the population.

Example: A jar of Chunggam contains 100 candy pieces, 15 are red. A handful of 10 is selected.

Probability question: What is **the probability** that 3 of the 10 selected are red?

Example: A handful of 10 Chunggams is selected from a jar containing 100 candy pieces. 3 Chunggams in the handful are red.

Statistics question: What is **the proportion** of red Chunggams in the entire jar?

Statistics and the Technology

- The electronic technology has had a tremendous effect on the field of statistics.
- Many statistical techniques are repetitive in nature: computers and calculators are good at this.
- Lots of statistical software packages: R - Programming, MINITAB, SYSTAT, STATA, SAS, Statgraphics, SPSS, and calculators.

Responsible use of statistical tools is very important.

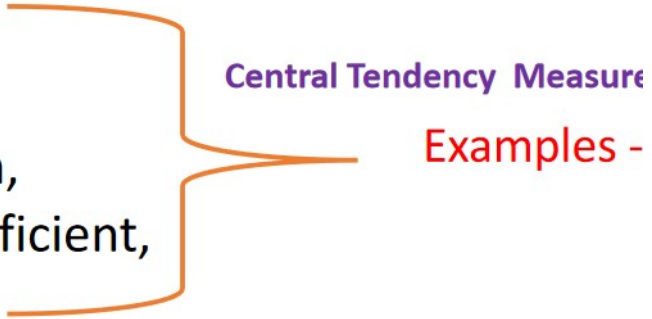
The burden is on the user to ensure that the appropriate methods are correctly applied and that accurate conclusions are drawn and communicated to others.


What I do when a teacher says "this cannot be done the night before"
Adhere the warning and start early
Take it as a personal challenge
GraphJam.com

So in statistical methods some important tools are...

Statistics Methods

- Mean, Median, Mode
- Range, Quartiles, IQR
- Variance, Standard Deviation,
- Covariance, Correlation Coefficient,
- Skewed and Kurtosis
- Regression
- Probability Distributions
- Hypothesis Test
- Z-Test, T-Test
- Chi-Square Test



Basics of Statistic

- Mean The 'average' of all the scores
- Median The central score
- Mode The most common score

Empirical relationship

$$\text{Mean - Mode} = 3(\text{Mean} - \text{Median})$$

Accuracy of the Mean

- Variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{N - 1}$$

- Standard Deviation:

$$s = \sqrt{s^2}$$

- Standard Error:

$$s_e = \frac{s}{\sqrt{n}}$$

Geometric Mean

3.21. Geometric Mean

If $x_1, x_2, x_3, \ldots, x_n$ be the (non-negative) n values of the variate x, none of them being zero, then the geometric mean, G, is defined by

$$G = (x_1, x_2, x_3, \ldots, x_n)^{1/n}.$$

If $x_1, x_2, x_3, \ldots x_n$ occur $f_1, f_2, f_3, \ldots, f_n$ times respectively and N is the total frequency, *i.e.* $N = f_1 + f_2 + \ldots + f_n$, then

$$G = \{x_1^{f_1} x_2^{f_2} \ldots x_n^{f_n}\}^{1/N}$$

or
$$\log G = \frac{1}{N} \sum_{i=1}^{n} f_i \log x_i$$

Thus the logarithm of the Geometric Mean is weighted mean of the different values of $\log x_i$ *whose weights are the frequencies* f_i.

In case of grouped frequency distribution the mid-values are considered as $x_1, x_2, \ldots, x_n$.

3.22. Computation of Geometric Mean

Example 1. *Find geometric mean of* 4, 8, 16.

Solution : $\text{G.M.} = (4 \times 8 \times 16)^{1/3}$

Harmonic Mean

3.25. Harmonic Mean

The harmonic mean of a series of (positive) values is defined as the reciprocal of the arithmetic mean of their reciprocals.

(i) For individual series : Let $x_1, x_2, \ldots, x_n$ be the n values of the variable x. Then harmonic mean denoted by H is given by

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

(ii) For discrete series. If $x_1, x_2, \ldots, x_n$ (none of them being zero) have the frequencies $f_1, f_2, \ldots, f_n$ respectively, then harmonic mean is given by

$$\frac{1}{H} = \frac{1}{N} \sum_1^n \left(\frac{f}{x}\right), \text{ where } N = \Sigma f$$

or

$$H = \frac{N}{\Sigma(f/x)}$$

(iii) For grouped series : In case of a grouped series the mid-values are taken as $x_1, x_2, \ldots, x_n$ with the corresponding class frequencies $f_1, f_2, \ldots, f_n$, then harmonic mean is given by

$$H = \frac{N}{\Sigma \left(\frac{f}{x}\right)}$$

Application of AM

Example 5. *Calculate the average speed of a car running at the rate of* 15 *kilometers per hour during the first* 30 *kilometers ; at* 20 *kilometers per hour during the second* 30 *kilometers and at* 25 *kilometers per hour during the third* 30 *kilometers.* **(Agra 1983)**

Solution : In going first 30 km he takes $\frac{30}{15}$ hr, next 30 km he takes $\frac{30}{20}$ hr and next 30 km he takes $\frac{30}{25}$ hr. Thus a total distance 90 km is covered in $\frac{30}{15} + \frac{30}{20} + \frac{30}{25}$ hours.

Hence average speed $= \dfrac{30 + 30 + 30}{\frac{30}{15} + \frac{30}{20} + \frac{30}{25}} = \dfrac{3}{\frac{1}{15} + \frac{1}{20} + \frac{1}{25}}$

$= \dfrac{3}{0{\cdot}06657 + 0{\cdot}05000 + 0{\cdot}04000} = \dfrac{3}{0{\cdot}15667}$

$= 19{\cdot}15.$

Application of GM

Example 4. Find the average rate of increase in population which in the first decade has increased 15 percent, in the next 22 percent and in the third 44 percent.

Solution : In such problems geometric mean is most suitable average. Therefore, here we shall find the geometric mean G of 15, 22 and 44.

$$\log G = \frac{\log 15 + \log 22 + \log 44}{3}$$

$$= \frac{1 \cdot 1761 + 1 \cdot 3424 + 1 \cdot 6434}{3}$$

$$= \frac{4 \cdot 1619}{3} = 1 \cdot 3873$$

$$G = \text{Antilog } 1 \cdot 3873 = 24 \cdot 40\%.$$

Application of HM

Example 3. *Calculate the average speed of a car running at the rate of* 20 *kilometers per hour during the first* 30 *kilometers; at* 25 *kilometers per hour during the second* 30 *kilometers and at* 30 *kilometers per hour during the third* 30 *kilometers.*

Solution : In the questions related to the average speed, harmonic mean is most suitable. Therefore, here we shall compute the harmonic mean of 20, 25, 30.

$$\therefore \quad \frac{1}{H} = \frac{\frac{1}{20} + \frac{1}{25} + \frac{1}{30}}{3}$$

$$= \frac{0 \cdot 05000 + 0 \cdot 04000 + 0 \cdot 03333}{3}$$

$$= \frac{\cdot 12333}{3} = \cdot 04111$$

$$\therefore \quad H = \frac{1}{\cdot 04111} = 24 \cdot 32 \text{ kilometers per hour.}$$

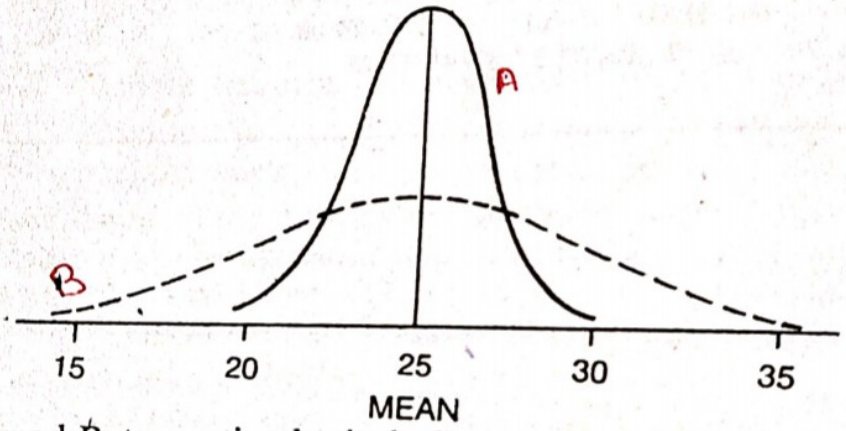
Measures of Dispersion

4.1. Dispersion

To explain the term 'dispersion' let us consider the case of two typists who typed the following number of pages in 6 working days of a week –

	Mon.	*Tues.*	*Wed.*	*Thurs.*	*Fri.*	*Sat.*
I Typist	20	20	25	25	30	30
II Typist	15	20	25	25	30	35

Each of the typists typed 150 pages in 6 days and therefore the mean in both the cases is 25. Thus there is no difference as far as the average



is concerned. But we notice that in the first case the number of pages varies from 20 to 30 and in the second case this number varies from 15 to 35, *i.e.* the greatest deviation from the mean in the first case is 5 and in second case it is 10. Clearly this indicates a difference in the two series. Such variation is called **dispersion.**

Measures of Dispersion

The terms like, *dispersion*, *variation*, *spread*, *scatter*, *deviation* and *variability* give the idea of homogeneity or heterogeneity of the data under study.

4.2. Desiderata for a Satisfactory Measure of Dispersion **(Agra 1983)**

The following are the essential requisites for satisfactory dispersion—

(1) It should be rigidly defined.
(2) It should be based on all observations.
(3) It should be readily comprehensive.
(4) It should be easily calculated.
(5) It should be amenable to algebraic treatment.
(6) It should be affected as little as possible by fluctuation of sampling.

4.3. Measures of Dispersion

The following are the measures of dispersion which are in common use—

(1) *Range.*
(2) *Quartile deviation or semi-interquartile range.*
(3) *Average deviation* or Mean Deviation
(4) *Standard deviation.*

Range

4.4. Range

The simplest possible measure of dispersion is the **range** which is the difference between the greatest and the least values of the variable. Thus in the first case above the range is the difference between 30 and 20, *i.e.* 10 and in the second case the range is the difference between 35 and 15, *i.e.* 20.

There are certain drawbacks as the range is considered. The range is subject to fluctuations of considerable magnitude from sample to sample. It takes into account only the extreme items. The occurrence of the height of a giant or dwarf will increase the range considerably. The range takes no account of the form of the distribution within the range.

Quartile Deviation

4.5. Quartile Deviation or Semi-interquartile Range

The difference between the upper and lower quartiles, *i.e.* $Q_3 - Q_1$ is known as the **interquartile range** and 50% of the total frequency lies in this range. Half of the interquartile range *i.e.*, half of the difference $Q_3 - Q_1$ is called the **semi-interquartile range or the quartile deviation. (Agra 1983)**

Thus quartile deviation, denoted by Q.D., is given by

$$\text{Q.D.} = \frac{1}{2}(Q_3 - Q_1)$$

The quartile deviation is easily computed and is a better measure of dispersion than the range. However, it does not take into account all the items.

The difference between the ninth and first decile is called as **inter-decile range** and 80% of total frequency lies in this range.

Mean Deviation

Average Deviation or Mean Deviation

4.6. Average Deviation or Mean Deviation

Arithmetic average of the group measured from an average (median, mode or mean) or any other arbitrary point taking all deviations as positive, is known as **Mean Deviation**. In other words mean deviation is the sum of the deviations (taken positive) from any point divided by the number of items. Thus

Mean deviation about any point A is given by

$$\delta_A = \frac{1}{N} \Sigma \mid x - A \mid$$

In this formula, A is to be taken mean or median or mode according as the mean deviation from the mean or median or mode is required.

In case of discrete series mean deviation from A, denoted by δ_A is given by

$$\delta_A = \frac{1}{N} \Sigma f(x - A).$$

In case of grouped frequency distribution the mid-values are taken as x.

SD and RMSD

4.8. Standard Deviation and Root Mean Square Deviation

The *mean square deviation*, denoted by S^2, is defined as the mean of the squares of the deviations from an arbitrary point A. Thus

$$S^2 = \frac{1}{N}\Sigma f(x - A)^2$$

The arithmetic mean of the squares of the deviations from arithmetic mean, M of a series is called *variance* and is denoted by σ^2. Thus

$$\sigma^2 = \frac{1}{n}\Sigma(x - M)^2, \quad \text{for individual series}$$

$$\sigma^2 = \frac{1}{\Sigma f}\Sigma f(x - M)^2, \quad \text{for discrete series}$$

$$\sigma^2 = \frac{1}{\Sigma f}\Sigma f(x - M)^2, \quad \text{for grouped series, where } x \text{ is the mid-value of the class.}$$

The root-mean square deviation denoted by S is defined as the positive square root of the mean of the square of the deviations from an arbitrary point A.

SD

Thus $$S = + \sqrt{\frac{1}{N} \Sigma f(x - A)^2}$$

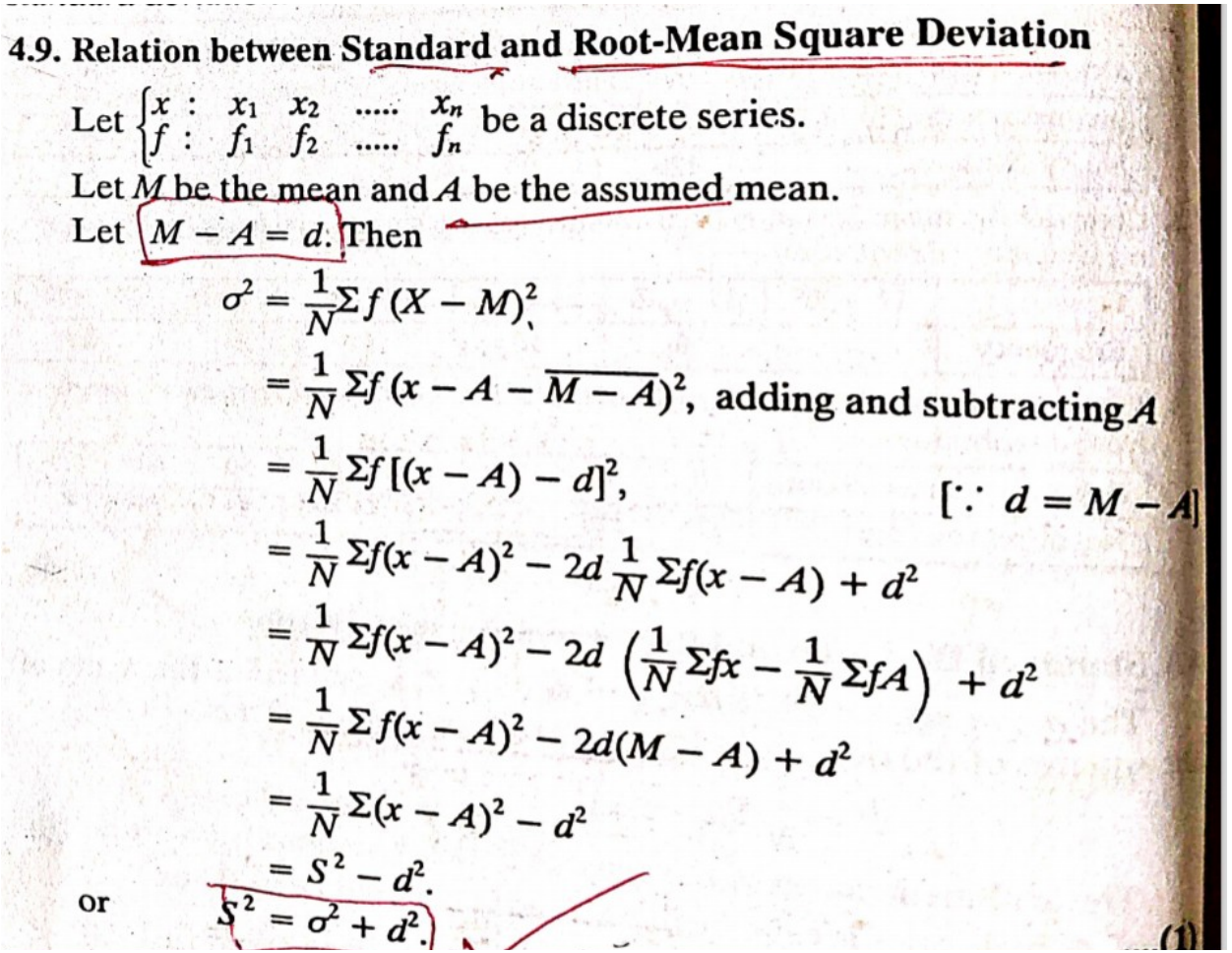
When the deviations are taken from the mean M, the root-mean square deviation is called the *standard deviation* and is denoted by σ. Thus

$$\sigma = + \sqrt{\frac{1}{N} \Sigma f(x - M)^2}$$

In other words,

The positive square root of the arithmetic mean of the squares of the deviations of observations in a series from its arithmetic mean is called *standard deviation.*

Relation between SD & RMSD



4.9. Relation between Standard and Root-Mean Square Deviation

Let $\begin{cases} x : & x_1 & x_2 & ..... & x_n \\ f : & f_1 & f_2 & ..... & f_n \end{cases}$ be a discrete series.

Let M be the mean and A be the assumed mean.

Let $M - A = d$. Then

$$\sigma^2 = \frac{1}{N}\Sigma f (X - M)^2,$$

$$= \frac{1}{N}\Sigma f (x - A - \overline{M - A})^2, \text{ adding and subtracting } A$$

$$= \frac{1}{N}\Sigma f [(x - A) - d]^2, \qquad [\because d = M - A]$$

$$= \frac{1}{N}\Sigma f(x - A)^2 - 2d\frac{1}{N}\Sigma f(x - A) + d^2$$

$$= \frac{1}{N}\Sigma f(x - A)^2 - 2d\left(\frac{1}{N}\Sigma fx - \frac{1}{N}\Sigma fA\right) + d^2$$

$$= \frac{1}{N}\Sigma f(x - A)^2 - 2d(M - A) + d^2$$

$$= \frac{1}{N}\Sigma(x - A)^2 - d^2$$

$$= S^2 - d^2.$$

or $S^2 = \sigma^2 + d^2$. ...(1)

Coefficient of Variation

(4) The nature of standard deviation and arithmetic mean is called *standard coefficient of dispersion* or *coefficient of variation*. Thus, coefficient of variation,

$$\text{C.V. or } V = \frac{\sigma}{m} \times 100$$

Example

Example 2. *The details of runs gained by two batsmen A and B in different innings are as follows —*

A	24	79	31	114	14	02	68	01	110	07
B	05	18	42	53	09	47	52	17	81	56

(i) *Which of the two batsmen is better run scorer ?*

(ii) *Which of the two batsmen has more consistency in the number of runs ?*

(Jiwaji 1994)

Solution

Solution :

(i) To compute the coefficient of variation of runs of the batsman A :

X	$X - M_A$	$(X - M_A)^2$
24	– 21	441
79	34	1156
31	– 14	196
114	69	4761
14	– 31	961
02	– 43	1849
68	23	529
01	– 44	1936
110	65	4225
07	– 38	1444
450	—	17498

Solution conti..

$\Sigma X = 450, n = 10.$

$$M_A = \frac{\Sigma X}{n} = \frac{450}{10} = 45 \text{ runs.}$$

$$\sigma_A = \sqrt{\frac{1}{n}\Sigma (X - M_A)^2} = \sqrt{\frac{17498}{10}} = 41{\cdot}83 \text{ runs.}$$

∴ Coefficient of Variation,

$$V_A = \frac{\sigma_A}{M_A} \times 100\ \% = \frac{41{\cdot}83}{45} \times 100\% = 92{\cdot}96\%$$

(ii) To compute the coefficient of variation of runs of the batsman B :

X	$X - M_B$	$(X - M_B)^2$
05	− 33	1089
18	− 20	400
42	4	16
53	15	225
09	− 29	841
47	9	81
52	14	196
17	− 21	441
81	43	1849
56	18	324
380	—	5462

Solution cont...

$$\Sigma X = 380, n = 10 \ \therefore \ M_B = \frac{\Sigma X}{n} = 38 \text{ runs.}$$

$$\sigma_B = \sqrt{\frac{1}{n}\Sigma (X - M_B)^2} = \sqrt{\frac{5462}{10}} = 23 \cdot 37 \text{ runs.}$$

$\therefore$ Coefficient of Variation,

$$V_B = \frac{\sigma_B}{M_B} \times 100\% = \frac{23 \cdot 37}{38} \times 100\% = 61 \cdot 5\%$$

(i) Since, $M_A > M_B$, batsman A is better run scorer.

(ii) Since, $V_B < V_A$, the batsman B has more consistency in the number of ns.

Empirical Relations

4.24. Empirical Relations between Measures of Dispersion

In case of frequency distributions with bell-shaped frequency curves, we have

(i) Mean Deviation about Mean $= \frac{4}{5}$ Standard Deviation

(ii) Quartile Deviation $= \frac{2}{3}$ Standard Deviation

(iii) Quartile Deviation $= \frac{5}{6}$ Mean Deviation.

Thus $\quad$ S.D. > M.D. > Q.D.

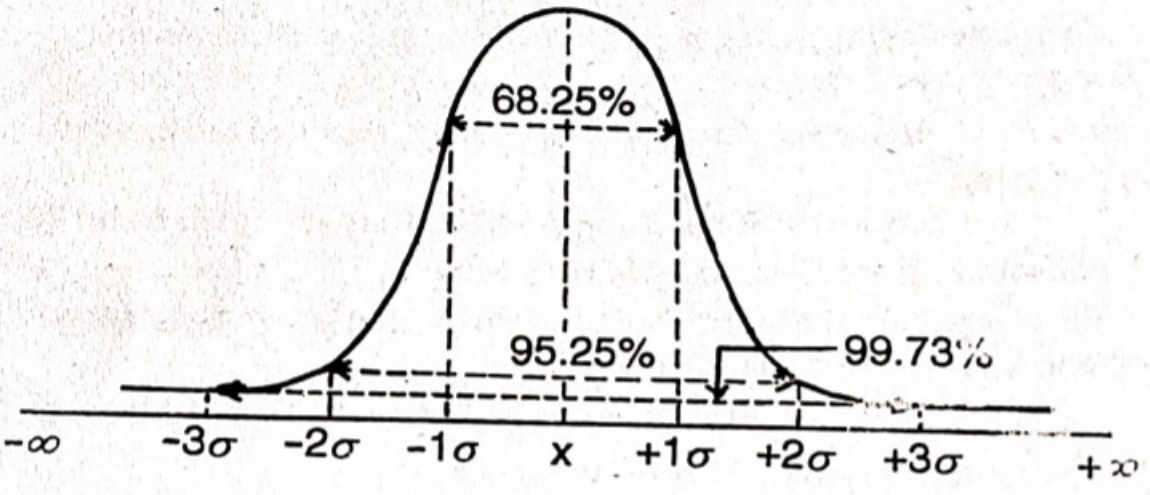
Mean, med, mode, sd,

https://www.youtube.com/watch?v=mk8tOD0t8M0

Bell-shaped Symmetrical Curve

4.25. Percentage of Observations within Mean ± $k\sigma$

If the curve of the frequency distribution is bell-shaped symmetrical then about 68·27% observations lie within $\bar{x} \pm \sigma$, about 95·25% of the observations lie within $\bar{x} \pm 2\sigma$ and about 99·73% of the observations lie within $\bar{x} \pm 3\sigma$. In graphical form we have—



Example:

Example : *Determine whether the frequency distribution given below is bell-shaped symmetrical—*

x	15	16	17	18	19	20	21	22	23	24
f	5	8	12	20	15	8	3	1	2	1

Solution :

x	f	$\xi = x - 18$	$f\xi$	$f\xi^2$
15	5	− 3	− 15	45
16	8	− 2	− 16	32
17	12	− 1	− 12	12
18	20	0	0	0
19	15	1	15	15
20	8	2	16	32
21	3	3	9	27
22	1	4	4	16
23	2	5	10	50
24	1	6	6	36
Total	75		17	265

Solution

Mean, $\bar{x} = a + \frac{\Sigma f\xi}{N} = 18 + \frac{17}{75} = 18 \cdot 23$

$$\text{S.D.}, \sigma = \sqrt{\frac{1}{N}\Sigma f\xi^2 - \left(\frac{\Sigma f\xi}{N}\right)^2}$$

$$= \sqrt{\frac{265}{75} - \left(\frac{17}{75}\right)^2}$$

$$= \sqrt{3 \cdot 5333 - \cdot 0529} = \sqrt{3 \cdot 4804} = 1 \cdot 86.$$

Now,

Solution cont...

(a) $\bar{x} - \sigma = 18 \cdot 23 - 1 \cdot 86 = 16 \cdot 37$

$\bar{x} + \sigma = 18 \cdot 23 + 1 \cdot 86 = 20 \cdot 09$

Number of terms between $16 \cdot 37$ and $20 \cdot 09$ is

$$12 + 20 + 15 + 8 = 55$$

Thus percentage between

$$\bar{x} \pm \sigma = \frac{55}{75} \times 100 = 73 \cdot 33 \quad \text{....(i)}$$

Solution cont...

(b) $\bar{x} - 2\sigma = 18{\cdot}23 - 2 \times 1{\cdot}86 = 18{\cdot}23 - 3{\cdot}72 = 14{\cdot}51$

$\bar{x} + 2\sigma = 18{\cdot}23 + 2 \times 1{\cdot}86 = 18{\cdot}23 + 3{\cdot}72 = 21{\cdot}95$

Number of terms between $14{\cdot}51$ and $21{\cdot}95$ is

$5 + 8 + 12 + 20 + 15 + 8 + 3 = 71$

Thus percentage between $\bar{x} \pm 2\sigma = \frac{71}{75} \times 100 = 94{\cdot}67$ ....(ii)

(c) $\bar{x} - 3\sigma = 18{\cdot}23 - 3 \times 1{\cdot}86 = 18{\cdot}23 - 5{\cdot}58 = 12{\cdot}65$

$\bar{x} + 3\sigma = 18{\cdot}23 + 3 \times 1{\cdot}86 = 18{\cdot}23 + 5{\cdot}58 = 23{\cdot}81$

Number of terms between $12{\cdot}65$ and $23{\cdot}81$ is

$75 - 1 = 74$

Thus percentage between

$\bar{x} \pm 3\sigma = \frac{74}{75} \times 100 = 98{\cdot}67$ ....(iii)

From (i), (ii) and (iii), we conclude that the given frequency distributed may be considered bell-shaped symmetrical.

Normal distribution, 68-95-99.7 Rule:

https://www.youtube.com/watch?v=mtbJbDwqWLE

Example

Descriptive Statistics

Descriptive Statistics

Example: The Government may want to get some idea about the income of its population to make economic decisions. The first step will be to collect as much data as possible across different classes and age groups. Now, this data will be processed to get meaningful information, e.g., mean, standard deviation, etc. After calculating different quantities, government can make inferences, e.g., the average income of 30–40 years age group is more than 10–20 years age group. Also, the government can use this data to model the income of middle-class population or classify a person as middle-class depending on other factors.

Data, Information, and Description

The national army of a country is composed of 1,500 officers, in different confinements and bases. The commanders of the army have selected a sample of 20 officers for whom four variables were measured: height (in m), **marital status** (0 = single, 1 = married), **education level** (0 = high school, 1 = technician, 2 = graduate, 3 = postgraduate, 4 = Ph.D.), and weight (in kg). The observed data are given in Table (in the next slide) The commanders want to infer certain information from the observed data to plan future supplies for their officers and duties to be assigned based on the skills of their officers. To do this, the provided data should be analyzed

Data, Information, and Description

- In this case, the addressed population size is N = 1500 officers, the sample size is n = 20 officers (observational units), and the domains for all the variables are: sigma1 = [a, b], a < b, a, b ∈ R+, sigma2 = {0, 1}, sigma3 = {0, 1, 2, 3, 4}, and sigma4 = [a, b],

Data, Information, and Description

Observed data

Observed data for

Officer	Height (m)	M.S.	E.L.	Weight (kg)
1	1.76	0	2	83.1
2	1.83	1	2	91.8
3	1.79	1	3	91.3
4	1.71	1	2	85.6
5	1.81	0	0	88.0
6	1.71	0	2	89.2
7	1.96	1	2	92.8
8	1.80	0	2	89.1
9	2.10	0	3	90.8
10	1.89	0	1	87.0
11	2.13	0	3	90.2
12	1.82	0	3	85.9
13	2.07	1	3	93.2
14	1.73	0	2	89.6
15	1.72	1	2	89.1
16	1.86	0	4	90.5
17	1.82	1	2	87.1
18	1.94	1	3	88.5
19	1.74	1	4	89.9
20	1.99	1	2	88.3

Data, Information, and Description

Summary report

Summary report

3.1

Variable	Height (m)	Weight (kg)
Max	2.13	93.2
Min	1.71	83.1
Range R_j	0.42	10.1
Mean $\overline{x}$	1.859	89.05
Variance s^2	0.01736	6.2447
Std. deviation s	0.1317	2.4989
Coeff. of variation c_v	0.071	0.028
Median	1.82	89.15
Mode	(1.715–1.805)	(89.155–91.175)
Geom. mean $\overline{x}_g$	1.855	89.016
Skewenss μ_3	0.7612	−0.4624
Kurtosis β	2.4220	3.0126

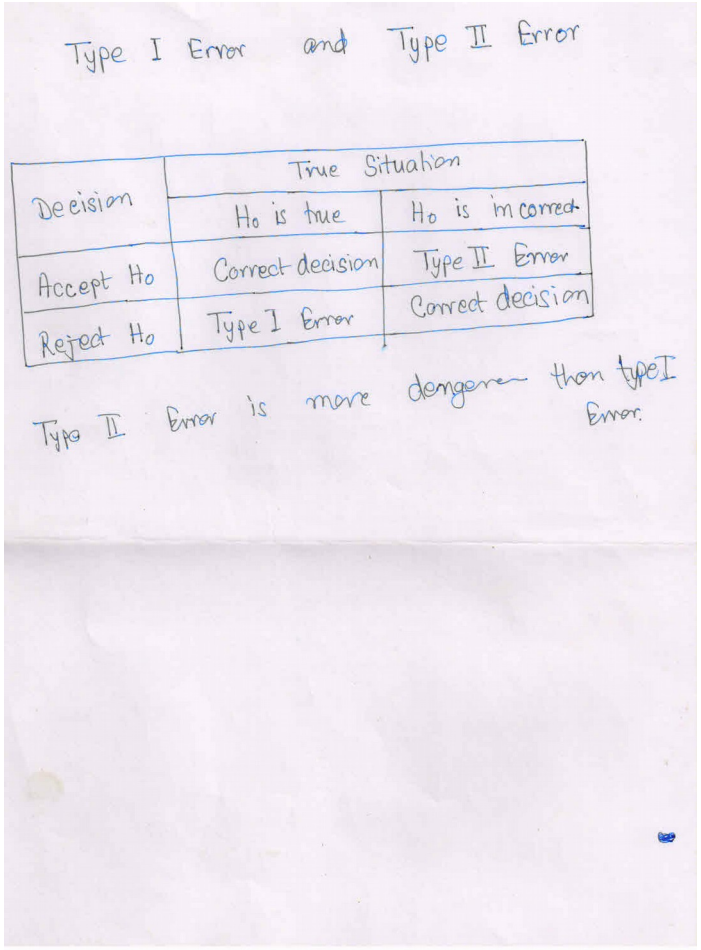
Inferential Statistics

Degrees of Freedom

- For sample populations, often 'N – 1' is used rather than N. This is the simplest calculation of D.O.F., but it can get very complex.

- We assume that the sample mean is the same as the population mean. Therefore, it is related to how many values are free to vary without altering the required mean value. (Rugby example)

Types of Errors in sampling



Type I Error and Type II Error

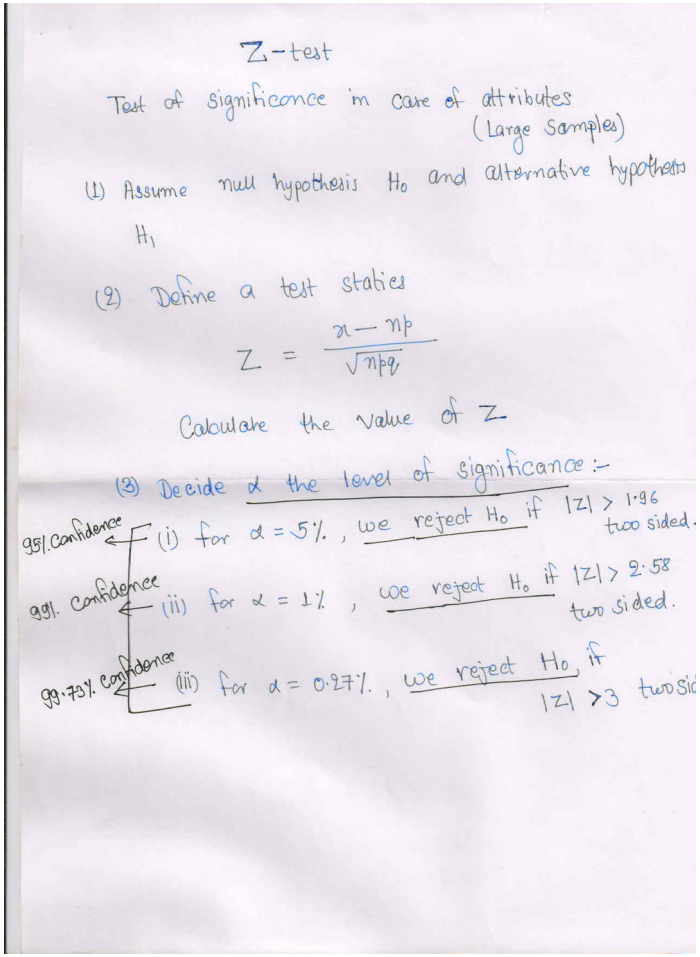
Decision	True Situation	
	H_0 is true	H_0 is incorrect
Accept H_0	Correct decision	Type II Error
Reject H_0	Type I Error	Correct decision

Type II Error is more dangerous than type I Error.

Steps for applying any test

- Assumption
- Calculation
- Comparison
- Decision
- Conclusion

Z- Test



Z-test

Test of significance in case of attributes
(Large Samples)

(1) Assume null hypothesis H_0 and alternative hypothesis H_1

(2) Define a test statics

$$Z = \frac{x - np}{\sqrt{npq}}$$

Calculate the value of Z.

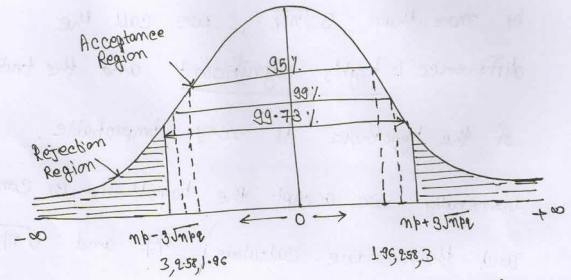
(3) Decide α the level of significance :-

95% Confidence (i) for $\alpha = 5\%$, we reject H_0 if $|Z| > 1.96$ two sided.

99% Confidence (ii) for $\alpha = 1\%$, we reject H_0 if $|Z| > 2.58$ two sided.

99.73% Confidence (iii) for $\alpha = 0.27\%$, we reject H_0 if $|Z| > 3$ two sid

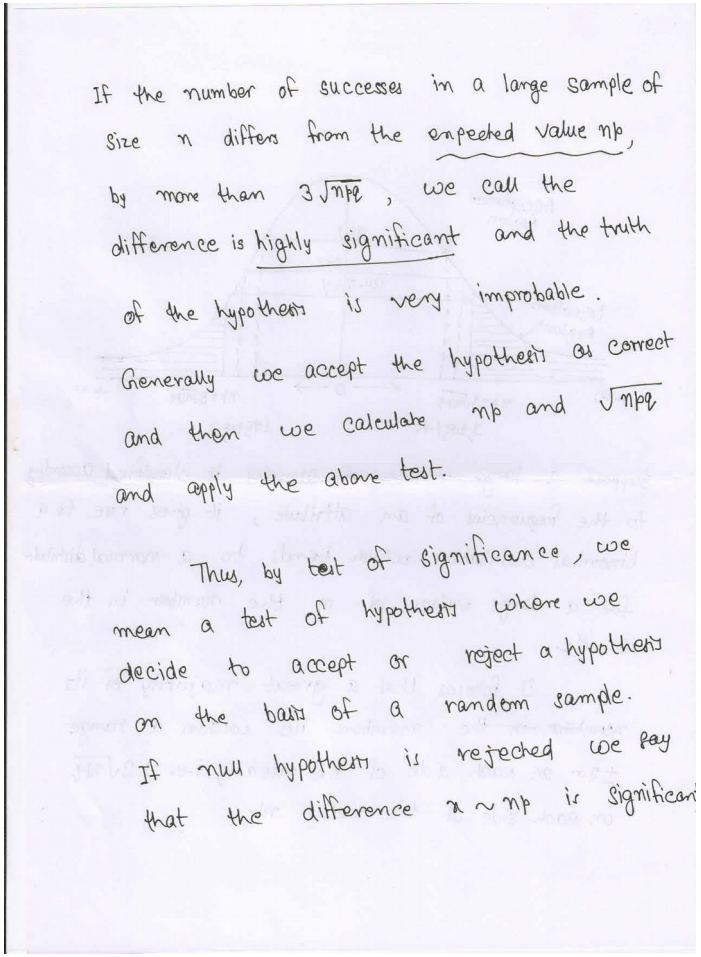
Z- test continue



Suppose a large number of samples is classified according to the frequencies of an attribute, it gives rise to a binomial distribution which tends to a normal distribution for a large values of n, the number in the sample.

It follows that a great majority of its ~~members in the~~ members lie within a range $\pm 3\sigma$ on each side of the mean, i.e. $\pm 3\sqrt{npq}$ on each side of the value np.

Z-test continue

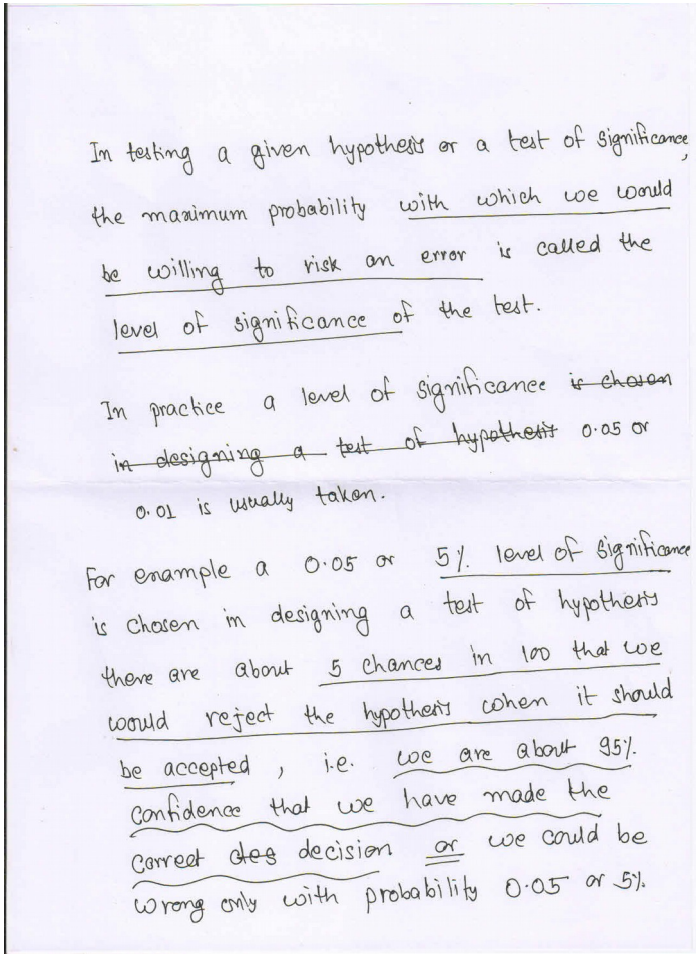


If the number of successes in a large sample of size n differs from the expected value np, by more than $3\sqrt{npq}$, we call the difference is highly significant and the truth of the hypothesis is very improbable.

Generally we accept the hypothesis as correct and then we calculate np and $\sqrt{npq}$ and apply the above test.

Thus, by test of significance, we mean a test of hypothesis where we decide to accept or reject a hypothesis on the basis of a random sample. If null hypothesis is rejected we say that the difference $x \sim np$ is significant

Level of significance



In testing a given hypothesis or a test of significance, the maximum probability with which we would be willing to risk an error is called the level of significance of the test.

In practice a level of significance ~~is chosen in designing a test of hypothesis~~ 0.05 or 0.01 is usually taken.

For example a 0.05 or 5% level of significance is chosen in designing a test of hypothesis there are about 5 chances in 100 that we would reject the hypothesis when it should be accepted, i.e. we are about 95% confidence that we have made the correct ~~dec~~ decision or we could be wrong only with probability 0.05 or 5%.

Standard Error , Probable Error

Standard Error:

The standard deviation of a sampling distribution of a statistics is also called the standard error. This name is given on account of the fact that usually regard the expected value as the true value and divergence from it as error of estimation due to sampling effects.

Therefore frequencies differing from the expected frequency by more than 3 times the standard error are almost certainly not due to fluctuations of sampling. It is some time written as S.E.

Probable Error:

P.E. = 0·67449 . S.E.

Problem on Z-test

Q. A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin may be unbiased one.

Sol Let H_0 : the coin is unbiased.

$$Z = \frac{x - np}{\sqrt{npq}}$$

Here $x = 216$

$n = 400$

$p = \frac{1}{2}$, $q = \frac{1}{2}$

$$Z = \frac{216 - 200}{\sqrt{100}} = 1.6$$

Since the value of $Z = 1.6 < 1.96$

Hence for $\alpha = 5\%$ (level of significance)

H_0 is accepted.

∴ The coin is unbiased one, at 5% level of significance.

Q: In some dice throwing experiment, Sohan throw a ~~dice~~ die 49152 times and of there 25145 yielded 4, 5 or 6. Is this consistent with hypothesis that the die was unbiased ?

Sol.

Problem on Z-test

Let H_0: the die was unbiased one.

Sol: The probability of throwing 4, 5 or 6 with one die

$$p = \frac{3}{6} = \frac{1}{2} \quad \therefore \quad q = 1 - p = \frac{1}{2}$$

$$n = 49152 \quad , \quad x = 25145$$

$$Z = \frac{x - np}{\sqrt{npq}} = \frac{25145 - 24576}{110.9} = 5.1373$$

Since $|Z| > 3$. Hence H_0 is rejected.

$\therefore$ the was biased one. (at 99.73% confidence level)

Comparison of two large sample

Formula $|Z| = \dfrac{p_1 \sim p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$

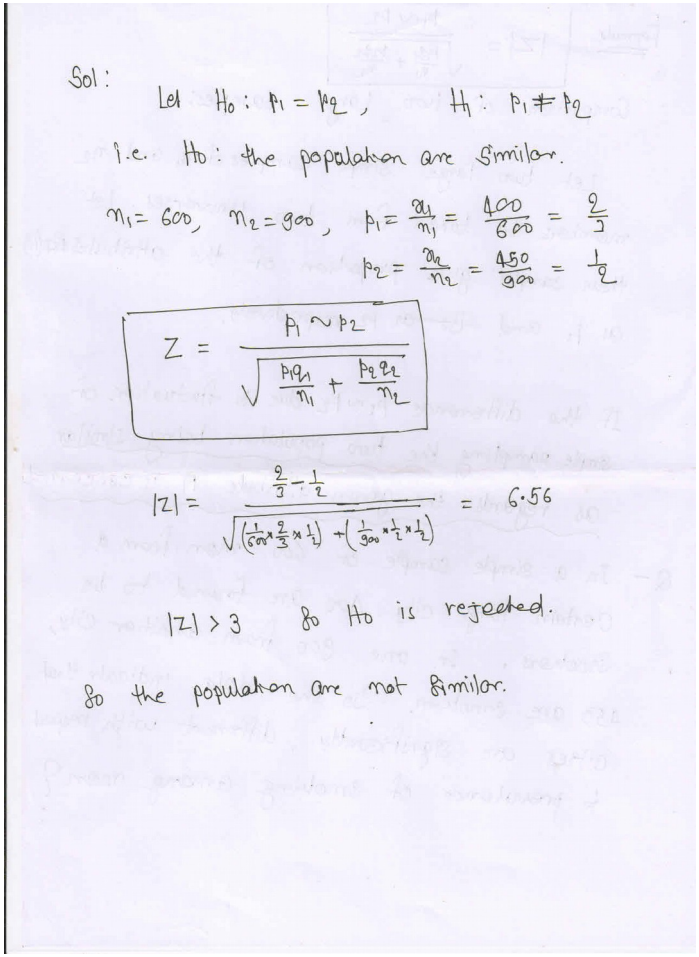
Comparison of two Large Samples:

Let two large simple samples of n_1 and n_2 members be taken from two universes. Let these samples give proportion of the attributes as p_1 and p_2 respectively.

If the difference $p_1 \sim p_2$ due to fluctuation of simple sampling the two population being similar as regards the given attribute A is concerned

Q- In a simple sample of 600 men from a certain large city 400 are found to be smokers. In one 900 from another city, 450 are smokers. Do the data indicate that cities are significantly different with respect to prevalence of smoking among men?

Solution of Problem



Sol:

Let $H_0: p_1 = p_2$, $H_1: p_1 \neq p_2$

i.e. H_0: the population are similar.

$n_1 = 600$, $n_2 = 900$, $p_1 = \frac{x_1}{n_1} = \frac{400}{600} = \frac{2}{3}$

$p_2 = \frac{x_2}{n_2} = \frac{450}{900} = \frac{1}{2}$

$$Z = \frac{p_1 \sim p_2}{\sqrt{\frac{pq_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$$|Z| = \frac{\frac{2}{3} - \frac{1}{2}}{\sqrt{\left(\frac{1}{600} \times \frac{2}{3} \times \frac{1}{2}\right) + \left(\frac{1}{900} \times \frac{1}{2} \times \frac{1}{2}\right)}} = 6.56$$

$|Z| > 3$ so H_0 is rejected.

So the population are not similar.

T - test

19·1. *t*-statistic

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a normal distribution with mean μ (or M) and variance σ^2 (unknown). Let s^2 or $(s')^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \text{ or } \frac{(\bar{x} - \mu)\sqrt{n}}{s}$$

for $n \leq 30$

is called t-statistic with $(n-1)$ degrees of freedom when $n \leq 30$. For large n, t-statistic tends to standard normal variate. However, if we define $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, then

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n-1}}.$$

for $n \geq 30$

19·2. *t*-distribution

The statistic t was introduced by **W. S. Gosset** in 1908 who wrote under the name **'Student'.** That is why it is called Student's t. Later on its distribution was rigorously established by **Prof. R.A. Fisher** in 1926.

T - test

4. Tests of Significance Based on *t*-Distribution

The *t*-distribution is used to test the significance of

1. the mean of the sample.
2. the difference between two means or to compare two samples.
3. sample coefficient of correlation.
4. sample coefficient of regression.

T - test

Question from t-test (Sampling)

The life time of tube lights for a random sample of 10 from a large production gave the following information:

Item	1	2	3	4	5	6	7	8	9	10
Life in '000 hrs	4·0	4·8	3·8	4·2	5·1	3·9	3·7	4·5	4·2	5·8

Can we accept the hypothesis that the average life time of tube lights is 4000 hrs?

(Use t-test)

Given for 9 degree of freedom and

$t_{0.05} = 2.262$

Solution:

Solution of Question from t-test Page (1/2)

Assumption: Null hypothesis H_0: Let the average life of tube light is 4000 hrs.

Calculation: Applying the t-test

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

Calculation of $\bar{X}$ (mean):

$$\bar{X} = \frac{4000 + 4800 + 3800 + 4200 + 5100 + 3900 + 3700 + 4500 + 4200 + 5800}{10}$$

$$\bar{X} = \frac{44000}{10} = 4400 \text{ hrs}$$

Now calculation of S (Standard deviation):

X	$(X-\bar{X})$	$(X-\bar{X})^2$
4000	− 400	160000
4800	400	160000

Solution:

X	$(X-\bar{X})$	$(X-\bar{X})^2$
4000	− 400	160000
4800	400	160000
3800	− 600	360000
4200	− 200	40000
5100	700	490000
3900	−500	250000
3700	−700	490000
4500	100	10000
4200	−200	40000
5800	1400	1960000
$\Sigma X = 44000$		3960000

$$S = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = \sqrt{\frac{3960000}{9}} = 663.3249581$$

Solution:

$$\therefore \quad t = \frac{4400 - 4000}{663.3249581}\sqrt{10}.$$

$$\therefore \quad t = 1.906925 \quad (\text{Calculated})$$

$$\therefore \boxed{t_{\text{Calculated}} = 1.906925} \quad (\text{Approx})$$

Comparision

Since $\nu = n - 1 = 10 - 1 = 9$

For $\nu = 9$ the value $\boxed{t_{0.05} = 2.262}$

$$\therefore \quad t_{cal} (1.906925) < t_{\text{tabulated}} (2.262)$$

$$\therefore \quad t_{cal} < t_{tab}$$

for 9 degree of freedom. at 5% level of Sign.

Decision

We accept the Null hypothesis H_0.

Solution:

Decision

at 5% level ... 0

We accept the Null hypothesis H_0.

that the average life of tube light is 4000 hrs.

Conclusion:

So the average life time of the tube light could be 4000 hours.

—— Ans

Another Problem from t - test

Example 3. *Ten individuals are chosen at random from a population and their heights are found to be in inches* 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. *Discuss the suggestion that the mean height in the universe is* 65 *inches given that for* 9 *degrees of freedom the value of Student's t and* 5 *percent level of significance is* 2·262.

Solution:

The calculation table is :

No. of Individuals	x	$x - \bar{x} = x - 67$	$(x - \bar{x})^2$
1	63	− 4	16
2	63	− 4	16
3	64	− 3	9
4	65	− 2	4
5	66	− 1	1
6	69	2	4
7	69	2	4
8	70	3	9
9	70	3	9
10	71	4	16
$n = 10$	$\Sigma x = 670$	-	$\Sigma (x - \bar{x})^2 = 88$

$\therefore$ Sample mean, $\bar{x} = \frac{\Sigma x}{n} = \frac{670}{10} = 67.$

Sample standard deviation, $s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{88}{9}} = 3 \cdot 13$ inches.

H_0 : the mean of the universe is 65 inches.

Test Statistic :

$\therefore \quad t = \frac{\bar{x} - M}{s}\sqrt{n} = \frac{67 - 65}{3 \cdot 13}\sqrt{10} = 0 \cdot 638 \times 3 \cdot 1622777 = 2 \cdot 02.$

The number of degrees of freedom $= \nu = 10 - 1 = 9.$

Tabulated value for 9 d.f. at 5% level of significance is 2·262.

Since calculated value of t is less than tabulated value for 9 d.f. (2·02 < 2·262).

This error could have arisen due to fluctuations and we may conclude that the data are consistent with the assumption of mean height in the universe of 65 inches.

Next another problem:

Example 5. *Ten boxes are chosen at random from a godown and their weigh*
re found to be in kgs. 15·75, 15·75, 16·0, 16·25, 16·5, 17·25, 17·25, 17·5, 17·
7·75. *Discuss the suggestion that the mean weight in the universe is* 16·25 *kg. Giv*
hat for 9 *degrees of freedom the value of Student's t at* 5 *percent level of significa*
2·262. **(Jiwaji 1993; Sagar 91, 92,**

Solution :

No. of Individuals	x	$x - \bar{x}$	$(x - \bar{x})^2$
1	15·75	− 1·00	1·0000
2	15·75	− 1·00	1·0000
3	16·00	− ·75	0·5625
4	16·25	− ·50	0·2500
5	16·50	− ·25	0·0625
6	17·25	·5	0·2500
7	17·25	·5	0·2500
8	17·50	·75	0·5625
9	17·50	·75	0·5625
10	17·75	1·00	1·0000
$n = 10$	$\Sigma x = 167·50$		$\Sigma (x - \bar{x})^2 = 6·0625$

Sample mean, $\bar{x} = \frac{\Sigma x}{n} = \frac{167·50}{10} = 16·75$ kg.

Sample standard deviation, $s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{6·0625}{9}}$

$= \sqrt{0·6736} = 0·82$ kg.

Null hypothesis H_0 : the mean of the universe, $M = 16·25$ kg.

Test Statistic :

$$\text{Student's } t = \frac{(\bar{x} - M)\sqrt{n}}{s}$$

$$= \frac{(16·75 - 16·25)\sqrt{10}}{0·82} = \frac{·5 \times 3·162}{0·82} = 1·93$$

The number of degrees of freedom $= 10 - 1 = 9$

Solution conti....

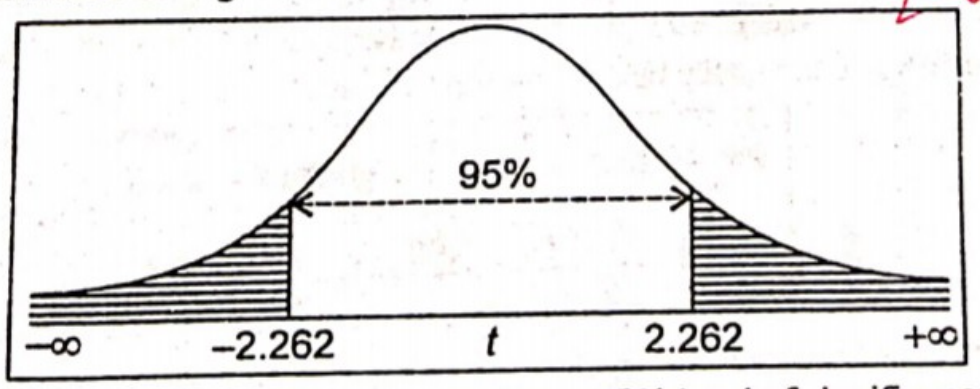
Null hypothesis H_0 : the mean of the universe, $M = 16{\cdot}25$ kg.

Test Statistic :

$$\text{Student's } t = \frac{(\bar{x} - M)\sqrt{n}}{s}$$

$$= \frac{(16{\cdot}75 - 16{\cdot}25)\sqrt{10}}{0{\cdot}82} = \frac{{\cdot}5 \times 3{\cdot}162}{0{\cdot}82} = 1{\cdot}93$$

The number of degrees of freedom $= 10 - 1 = 9$



The value of t for 9 degrees of freedom at 5% level of significance is 2·26

The calculated value of t is 1·93 and is less than the table value at 5% leve
significance. This difference could have arisen due to fluctuations of sampling
it can be said that the mean weight in the universe is 16·25 kg.

t-distribution critical values or t- Table

https://www.stat.tamu.edu/~lzhou/stat302/T-Table.pdf

Table: t value

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

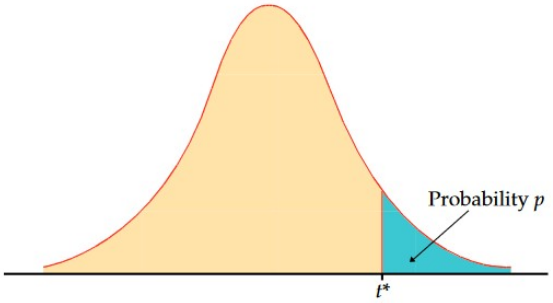


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850

Table: t value

20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

For T F Z Chi tests:

https://www.youtube.com/watch?v=HpWpIY2fhIo

For T, Z F Chi Parametric/ nonparametric test:

https://www.youtube.com/watch?v=gp5xQHdbwwI

Para/ nonpara talvir singh

https://www.youtube.com/watch?v=WxaBoUy5qww

Para/ nonpara sinha

https://www.youtube.com/watch?v=9SEINCx9uM8

Para/nonpara IGNOU

https://www.youtube.com/watch?v=tLcpcqk5I3M

Pearson Vs Spearman:

https://www.youtube.com/watch?v=c5ASFOYd918

Why Data Scientists use R?

https://www.youtube.com/watch?v=u94oFWZCTCU

Some useful links

Skills Needed For Data Scientist and Data Analyst

https://www.youtube.com/watch?v=em8nBc-zRaM

Data Scientist Roles and Responsibilities | Data Scientist Career | Data Science Training | Edureka

https://www.youtube.com/watch?v=nh4RgxaiKgI

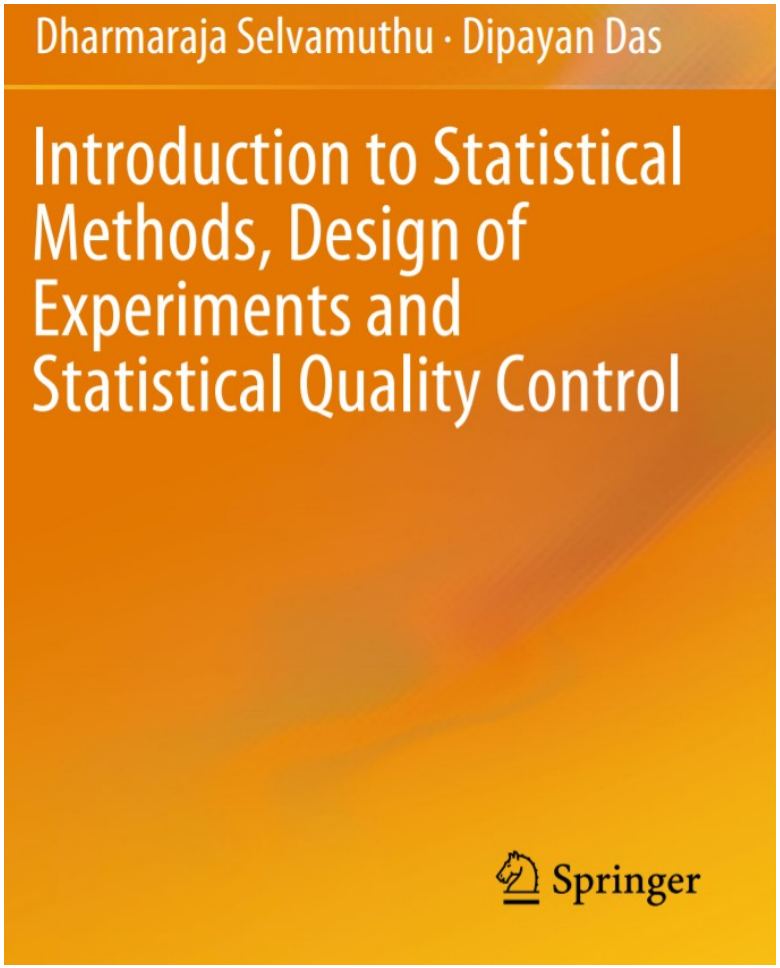
Recommended literature:

- Hanke E. J, Reitsch A. G: Understanding Business Statistics
- Anderson, D.R. - Sweeney, D.J. - Williams, T.A.: Statistics for Business and Economics. South-Western Pub., 2005, 320 p., ISBN 978-032-422-486-3
- Jaisingh, L.R.: Statistics for the Utterly Confused. McGraw Hill, 2005, 352 p., ISBN 978-007-146-193-1
- Everitt, B. S.: The Cambridge (explanatory) dictionary of statistics. Cambridge University Press, 2006, 442 p., ISBN 978-052-169-027-0
- Illowsky, B. - Dean, S. (2009, August 5). Collaborative Statistics. Retrieved from the Connexions Web site: http://cnx.org/content/col10522/1.36
- Valan, J. A. : Statistical Analysis- Descriptive analysis https://www.edureka.co/blog/what-is-data-analytics/
- Tom Methven :Introduction to Statistical Methods www.macs.hw.ac.uk/~mjc/teaching/ResearchMethods

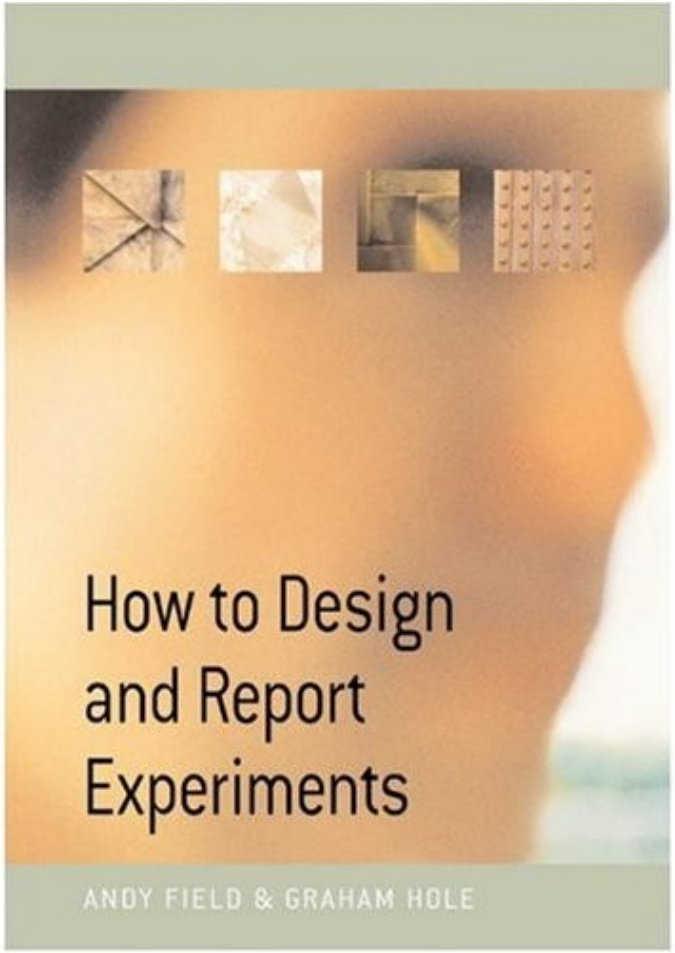
Moodle course by Martina Majorova at:

http://moodle.uniag.sk/fem/course/view.php?id=211

Introduction to Statistical Methods

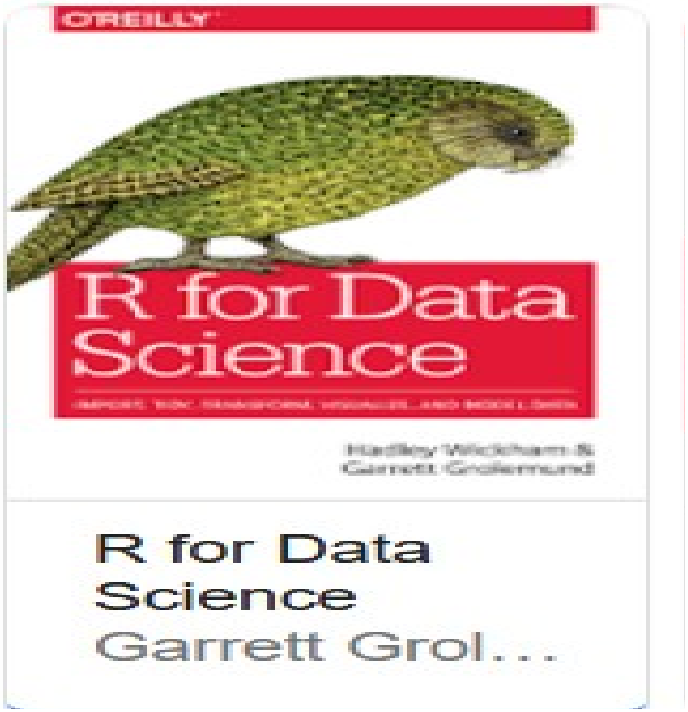


Recommended Reading



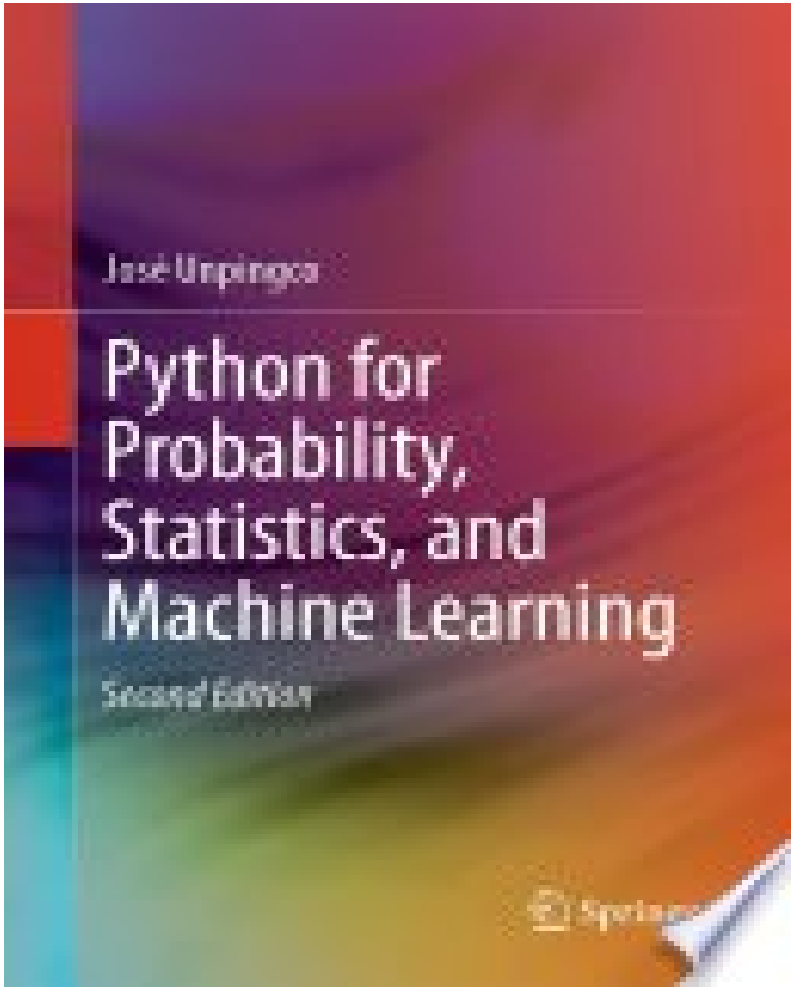
Book by Garrett Grolemund and Hadley Wickham

R Studio book



Python for Probability, Statistics, and Machine Learning

Authors: **Unpingco**, José



Thank You.